SBORNÍK

prací účastníků vědeckého semináře doktorského studia
Fakulta informatiky a statistiky
Vysoké školy ekonomické

Vědecký seminář se uskutečnil dne 14. února 2012
pod záštitou děkana FIS doc. RNDr. Luboše Marka, CSc.

Sestavení sborníku

prof. Ing. Petr Doucek, CSc.

proděkan pro vědu a výzkum

© Vysoká škola ekonomická v Praze
Nakladatelství Oeconomica – Praha 2012
ISBN 978-80-245-1862-6

OBSAH

Předmluva **5**

Stav testování softwaru v České republice **7**
Anna Borovcová (Aplikovaná informatika)

Konceptuální model pro Enteprise Architecture jako služba **16**
Libor Gála (Aplikovaná informatika)

Využití podnikové architektury v závislosti na komplexitě organizace **28**
Martin Hladík (Aplikovaná informatika)

Enterprise Architecture jako podpora transformace **35**
Pavel Hrabě (Aplikovaná informatika)

Praktická ukázka využití standardu COBIT k hodnocení úrovně IT Governance **47**
Dana Malcová (Aplikovaná informatika)

Formalizace nových poznatků ve strukturovaném přístupu k podpoře a údržbě software - kompetenční a sub-kompeteční centra **54**
Petr Marounek (Aplikovaná informatika)

Bezpečnostní aspekty SaaS z pohledu zákazníka **65**
Jan Mészáros (Aplikovaná informatika)

Aplikace pro podporu auditu datové kvality **76**
David Pejčoch (Aplikovaná informatika)

Podpora rozhodování v reálném čase z hlediska datové integrace **86**
Lukáš Pitka (Aplikovaná informatika)

Nástroj pro realizaci dynamického kompozitního programování v Javě **96**
Zbyněk Šlajchrt (Aplikovaná informatika)

Zajištění integrity systému dlouhodobé archivace **105**
Rudolf Vohnout (Aplikovaná informatika)

Koncept stochastické dominance a interaktivní metoda vícekriteriálního programování jako analytické nástroje investičního rozhodování **112**
Adam Borovička (Ekonometrie a operační výzkum)

Hallův model permanentního příjmu **125**
Milan Bouda (Ekonometrie a operační výzkum)

Optimization of Harvesting Natural Resources **135**
Viktor Chrobok (Ekonometrie a operační výzkum)

Lineární regresní model s intervalovými výstupy – vybrané charakteristiky a jejich výpočet **145**
Miroslav Rada (Ekonometrie a operační výzkum)

Dealing with Fiscal Data Uncertainty (DSGE-VAR Approach) **152**
Jakub Ryšánek (Ekonometrie a operační výzkum)

Kvantitativní dopady úrokových sazeb **161**
Michal Řičař (Ekonometrie a operační výzkum)

Modelling generator constraints for the self-scheduling problem **171**
Jiří Šumbera (Ekonometrie a operační výzkum)

Problematika ambientních daní v ekologii z pohledu teorie her **183**
Martina Zouharová (Ekonometrie a operační výzkum)

Předpovědi extrémních říčních průtoků metodou „peaks over threshold“ a jejich kvalita.....190
Adam Čabla (Statistika)

Zobecněné řešení soustavy lineárních algebraických rovnic a zobecněná metoda nejmenších čtverců.....202
Richard Horský (Statistika)

Použití endogenních vah při srovnání vysokých škol pomocí kompozitního indikátoru.....207
Lenka Hudrlíková (Statistika)

Měření efektivnosti v oblasti sociálních služeb.....218
Věra Jeřábková (Statistika)

Porovnanie sekvenčného a dávkového učenia pri metódach samo-organizujúcich sa máp.....227
Lukáš Pastorek (Statistika)

AutoClass – klasifikační metoda založená na bayesovském teorému.....233
Lukáš Sobíšek (Statistika)

Odhady naděje dožití u osob se specifickým onemocněním.....238
Ondřej Šimpach (Statistika)

Analýza CHAID jako nástroj vyhodnocování marketingových kampaní.....250
Jiří Vild (Statistika)

Analýza heterogenních preferencí pomocí hierarchických modelů.....259
Ondřej Vilikus (Statistika)

Neostré shlukování.....269
Tomáš Vintr (Statistika)

Approaches to Outlier Detection.....279
Vanda Vintrová (Statistika)

Předmluva

Dobrou tradici vědecké práce na Fakultě informatiky a statistky VŠE v Praze představuje každoročně pořádaný seminář „Den doktorandů“. Na něm presentují studenti všech oborů doktorského studia Fakulty informatiky a statistiky výsledky své vědecké práce v uplynulém období a seminář pro mnohé z nich představuje první platformu k otevřeným vědeckým diskusím nad problémy jejich témat doktorských prací.

Referáty, přihlášené do letošního Dne doktorandů, byly i v letošním roce rozděleny do tří samostatných sekcí podle jednotlivých studijních oborů. Na devatenáctý Den doktorandů bylo celkem přihlášeno 36 příspěvků. Z toho na obor Informatika 14 příspěvků, na obor Ekonometrie a operační výzkum 9 příspěvků a na obor Statistika 13 příspěvků. Semináře se zúčastnili studenti obou forem studia, jak presenční, tak i kombinované.

Úroveň vystoupení doktorandů, jak po stránce vědecké, tak i prezentační, hodnotily odborné komise pod vedením doc. Doc. Ing. Vojtěch Svátka, PhD. (obor Aplikovaná informatika), prof. Ing. Romana Huška, CSc. (obor Ekonometrie a operační výzkum) a prof. Ing. Hany Řezankové, CSc. (obor Statistika).

Hlavními kritérii pro nelehké rozhodování jednotlivých komisí byly zejména aktuálnost zpracovaného tématu, přístup k řešení vybraného problému, způsob a kvalita použité metodiky, úroveň práce s reálnými daty a v neposlední řadě i schopnost prezentovat a argumentačně obhájit výsledky své vědecké práce v diskusi jak se svými vrstevníky, tak i se staršími kolegy. Ti nejlepší z nich získávají prestižní „Cenu děkana FIS“, s níž je spojena i symbolická finanční odměna.

V letošním roce se dostalo odměny za nejlepší příspěvky v jednotlivých kategoriích následujícím studentům:

Studijní program – Aplikovaná informatika

obor Aplikovaná iInformatika

1. místo: Ing. Pavel Hrabě: Enterprise Architecture jako podpora transformace

2. místo: Mgr. Zbyněk Šlajchrt: Nástroj pro realizaci dynamického kompozitního programování v Javě

3. místo: Ing. Lukáš Pitka: Podpora rozhodování v reálném čase z hlediska datové integrace

Studijní program – Kvantitativní metody v ekonomice

obor Ekonometrie a operační výzkum

1. místo: Ing. Jakub Ryšánek: Dealing with Fiscal Data Uncertainty (DSGE-VAR Approach)

2. místo: neuděleno

3. místo: Ing. Adam Borovička: Koncept stochastické dominance a interaktivní metoda vícekriteriálního programování jako analytické nástroje investičního rozhodování

Ing. Michal Řičař: Kvantitativní dopady úrokových sazeb

Ing. Martina Zouharová: Problematika ambientních daní v ekologii z pohledu teorie her

obor Statistika

1. místo: neuděleno

2. místo: Ing. Tomáš Vintr: Neostré shlukování

Ing. Ondřej Vilikus: Analýza heterogenních preferencí pomocí hierearchických modelů

3. místo: Ing. Věra Jeřábková: Měření efektivnosti v oblasti sociálních služeb

Oceněným studentům doktorského studia upřímně blahopřejeme.

Uznání a poděkování také patří všem vědeckým pracovníkům FIS, kteří svými příspěvky, presentacemi a diskusními příspěvky vytvořili atmosféru semináře. Poděkování a uznání náleží rovněž jejich školitelům, členům hodnotících komisí a dalším kolegům, kteří se na úspěchu devatenáctého „Dne

doktorandů“ na Fakultě informatiky a statistiky Vysoké školy ekonomické v Praze podíleli. Zvláštní poděkování pak patří studijní referentce doktorského studia paní Jitce Krajíčkové, díky níž byl seminář perfektně organizačně zajištěn, a Mgr. Lee Nedomové za obětavou práci při editaci a sestavení tohoto sborníku.

Prof. Ing. Petr Doucek, CSc,
Proděkan pro vědu a výzkum

Stav testování softwaru v České republice

Anna Borovcová
anna.borovcova@gmail.com
Doktorandka oboru informatika
Školitel: doc. Ing. Alena Buchalcevová, Ph. D, (buchalc@vse.cz)

Abstrakt: Cílem tohoto článku je seznámit odbornou veřejnost s částí výsledků průzkumu stavu testování a řízení kvality v softwarových firmách vyvíjejících software na území České republiky, u nichž kvalita, efektivita a spolehlivost vývoje jsou důležitými aspekty konkurenceschopnosti.

Článek se zaměřuje zejména na to, jak vypadá testování v softwarových firmách a na způsob získávání znalostí v oblasti testování, tedy na kvalitu vzdělávání. Přitom zodpovídá následující otázky:

1. Jakým způsobem testují firmy software a koho najímají na testování?
2. Jak velký je rozpočet na testování v poměru k rozpočtu na celý vývoj?
3. Odkud čerpají zaměstnanci své znalosti o řízení kvality a jaké jsou možnosti jejich vzdělávání?

Pro srovnání byl proveden a použit obdobný zahraniční průzkum, kde převažovali respondenti z Indie a USA. Oba průzkumy měly formu internetových anonymních dotazníků a byly určeny pracovníkům, kteří se buď podílí na testování, nebo s ním alespoň každodenně přichází do styku.

Dále článek popisuje dostupné možnosti vzdělávání pracovníků, a jak jsou v České republice využívány, zejména možnosti vysokoškolského vzdělávání v oblasti testování softwaru na základě katalogů předmětů a osnov zveřejněných jednotlivými pražskými veřejnými vysokými školami.

Klíčová slova: testování softwaru, řízení kvality, průzkum kvality v softwarových firmách, projektové řízení, softwarové inženýrství.

Úvod

V dnešním rychlém světě je kvalita používaného softwaru jednou z hlavních konkurenčních výhod a tím i atributem úspěšné existence ekonomických subjektů. Za poslední tři desetiletí vzniklo mnoho různých standardů pro dosažení kvality, pomocných nástrojů pro řízení testování, specializovaných konferencí a byla napsána řada knih o tom, jak lépe testovat a jak zvýšit kvalitu vyvíjeného softwaru. Softwaroví testeři jsou dnes často speciálně školeni, předávají si informace na konferencích, stejně jako na fórech a blozích, a mají k dispozici řadu odborných článků na internetu.

Zdá se, že testování softwaru je rychle se rozvíjející oblastí, které každých pár let vládnou nové myšlenky a ty původní jsou označeny za zastaralé. Nakolik jsou ale nové myšlenky nebo léty prověřené praktiky využívány v běžné praxi? Jak se liší přístup ke kvalitě a testování v České republice od situace popisované v zahraničních publikacích?

V současnosti jsou na českém trhu pouze dvě knihy zabývající se testováním softwaru a i ty jsou pouze českým překladem knih zahraničních odborníků. Původních českých publikací a studií je velmi málo a obecně nejsou známy odborné veřejnosti. To je jeden z hlavních důvodů, proč jsme se rozhodli provést studii zabývající se současným postavením testování softwaru v České republice.

Tato studie předkládá a analyzuje výsledky ze tří pohledů:

- výsledků dotazníkový průzkumu provedeného v českých softwarových firmách,
- průzkumu možností zejména vysokoškolského vzdělávání v oblasti testování softwaru na základě katalogů předmětů a osnov zveřejněných jednotlivými pražskými veřejnými vysokými školami,
- a srovnání se současným postavením testování softwaru v zahraničí na základě obdobných zahraničních studií a průzkumů.

Průběh dotazníkového průzkumu

Pro provedení průzkumu v českých softwarových firmách byl vytvořen internetový dotazník, který byl uveřejněn na webové adrese http://studiekvality.vse.cz, kde probíhal sběr dat po dobu 5 měsíců. Tento průzkum proběhl na konci roku 2010, kdy bylo osloveno přes 300 softwarových firem různých velikostí působících v České republice, které se zabývají zejména vývojem softwaru na zakázku. Bylo preferováno, aby za každou firmu byl vyplněn dotazník pouze jedenkrát a to pracovníkem, který se podílí na testování, nebo s ním alespoň každodenně přichází do styku. Jelikož se jméno zaměstnavatele nikde neuvádělo, tato preference nebyla nijak kontrolována. Dále byla informace o prováděném průzkumu vyvěšena na několika webech a diskusních fórech, které se zabývají různými oblastmi vývoje softwaru. Dotazník byl anonymní, dotazovaní pouze vyplnili obecnou charakteristiku firmy a své pracovní pozice.

Celkem dotazník vyplnilo 84 pracovníků, kteří nějakým způsobem přicházejí do styku s testováním nebo se podílejí na utváření kvality.

Účastníci byli výhradně zaměstnanci softwarových firem, z toho 48% bylo z firem do 100 zaměstnanců a 52% z firem nad 100 zaměstnanců. Alternativní rozdělení firem nabízí následující tabulka:

Tabulka 1: Dělení účastníků dle velikosti zaměstnavatele Zdroj: autor

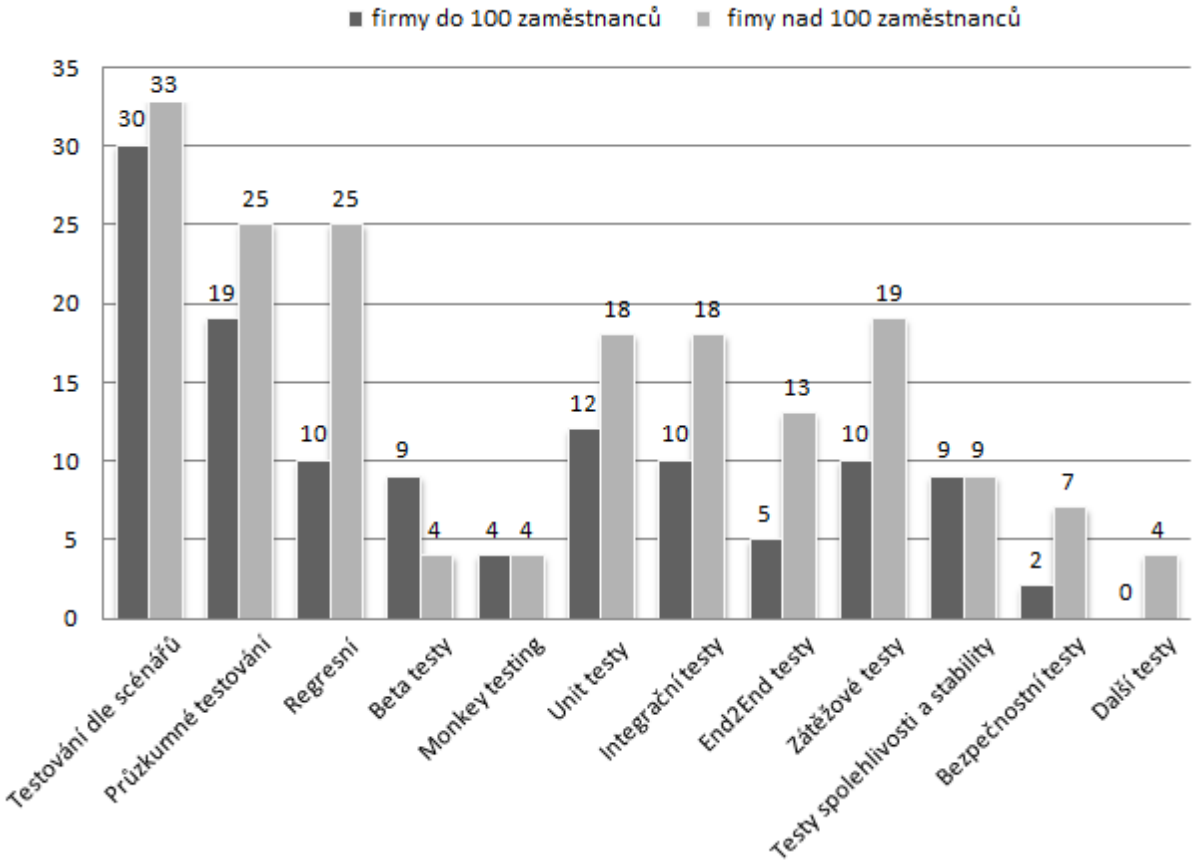
Velikost firmy	Zastoupení v procentech
Do 40 zaměstnanců	34,5 %
Do 300 zaměstnanců	36,9 %
Do 3000 zaměstnanců	16,7 %
Nad 3000 zaměstnanců	11,9 %

Mezi zúčastněnými bylo 41 členů testovacího týmu, 16 projektových manažerů a 27 dalších pracovníků, většinou na pozicích vývojáře nebo analytika. Více jak dvě třetiny účastníků je možné považovat za zkušené, neboť je daném oboru pracují a do styku s řízením kvality přicházejí přes dva roky.

V lednu 2012 byl proveden průzkum s převážně stejnými otázkami a možnostmi zaměřený na testování v zahraničí, zejména v Indii a USA, kterého se zúčastnilo 41 pracovníků softwarových firem, kteří se aktivně účastní procesu testování. Účastníci zahraničního průzkumu pocházeli zejména z větších firem s minimálně 100 zaměstnanci. Výsledky tohoto průzkumu zde byly využity pro srovnání stavu v České republice se zahraničím.

Způsob testování v České republice

Na otázku, jaké techniky jsou obvykle prováděny na projektech, kterých se dotazovaní účastní, jich vybrali z nabídky 12 testovacích technik v průměru 5. Zaměstnanci menších firem uvedli v průměru o jednu testovací techniku méně.



Obrázek 1: Počty dotazovaných, kteří uvedli, že daná technika je u nich na projektech obvykle prováděna
Zdroj: autor

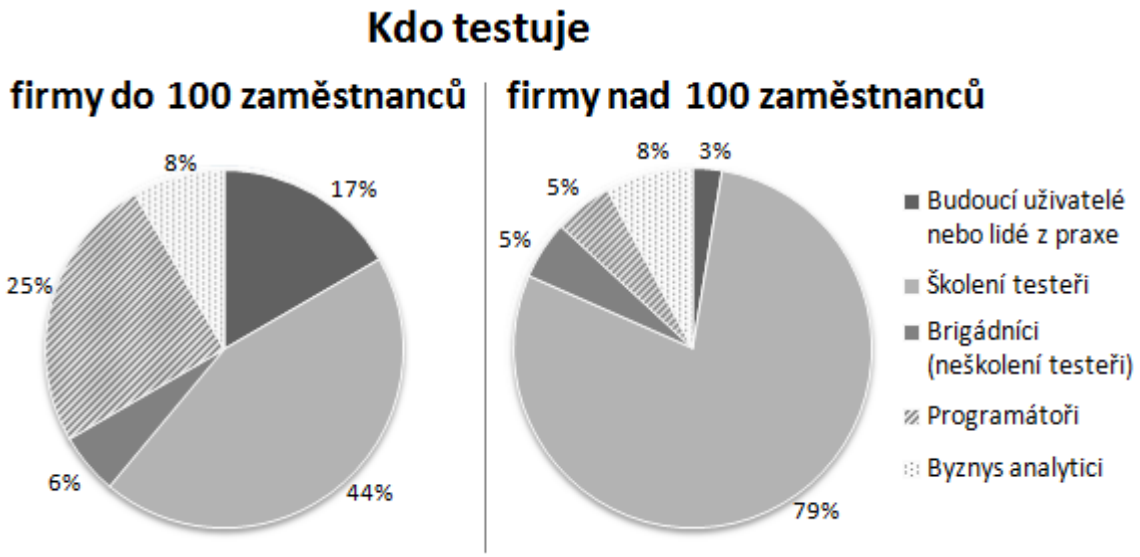
Z obrázku 1 je patrné, že nejčastěji se lze v České republice na projektech setkat s testováním dle scénářů, průzkumným testováním a regresními testy. Tyto testy se soustředí na testování kvality z hlediska funkčnosti softwaru. Zajímavé je, že skupina bezpečnostních technik jako celek není příliš používána. Dle sebraných údajů pouze 5% menších firem a 16% větších firem provádí bezpečnostní testování.

Pro srovnání v zahraničním průzkumu účastníci uvedli v průměru 7,5 technik, které jsou obvyklé na projektech v jejich firmě. To znamená, že jejich přístup k testování je více diverzifikovaný. Také častěji byly uváděny méně známé techniky jako náhodné Monkey testing[1], které se dle výsledků průzkumu používá více jak třetina (36,6%) zahraničních projektů. Nejmenší popularity v zahraničí dosáhlo bezpečnostní testování, které provádí 26,8% firem, což je stále mnohem větší poměr než u větších firem v České republice.

Složení testovacího týmu

Průzkum v oblasti složení testovacího týmu ukázal, že nejčastěji jsou na testování v České republice najímáni speciálně školení testeři. Větší firmy nad sto zaměstnanců preferují tuto variantu v 79 % případů. Přestože se jedná o nejoblíbenější variantu i u menších firem, ty častěji sahají po alternativních možnostech, kdy testují budoucí uživatelé nebo programátoři. Kompletní výsledky srovnání toho, kdo testuje v menších a větších firmách, nalezneme na obrázku 2.

[1] Monkey testing nebo taky náhodné testování je druh automatizovaného testování, kdy nástroj náhodně kliká a zadává náhodná data do testovaného programu.



Obrázek 2: Srovnání toho, kdo testuje v menších a větších firmách v České republice Zdroj: autor

Co do velikosti testovacího týmu zaměstnávají v České republice menší firmy výrazně větší počet testerů ve srovnání s programátory, než je tomu u firem nad sto zaměstnanců. Mezi zaměstnanci menších firem se také našlo 35% respondentů, kteří uvedli, že se u nich počet testerů rovná nebo dokonce převyšuje počet programátorů. Mezi zaměstnanci větších firem se nikdo takový nenašel.

Zajímavé je, že porovnáme-li hodnocení spokojenosti zákazníka celkově u všech respondentů malých firem a těch respondentů, kteří uvedli velký poměr testerů, neshledáme žádný významný rozdíl. Ani porovnání s odpověďmi na další otázky nepřineslo vysvětlení tohoto jevu. Je třeba ale podotknout, že tento vzorek je příliš malý, neboť pouze 13 respondentů uvádí tento poměr.

Oproti tomu v zahraničním průzkumu více jak polovina účastníků uvedla, že počet testerů u nich významně převyšuje počet programátorů a to 4 a více testerů na 1 programátora. Celkově pak na jednoho testera připadá dva a půl programátora, viz tabulka 2. Zároveň v zahraniční průzkum uvádí největší poměr školených testerů (86%), není tedy možné rozdíl vysvětlit využitím nekvalifikovaného personálu. Existují ale různá jiná vysvětlení tohoto poměru, například že dané firmy využívají testery nejen na otestování u nich vyvíjeného softwaru, ale poskytují tyto služby i ostatním firmám. Přesnou příčinu tohoto vysokého poměru testerů není možné z výsledků průzkumu zjistit.

Tabulka 2: Poměr testerů a programátorů ve firmách Zdroj: autor

Charakteristika firmy	Poměr
Česká firma do 100 zaměstnanců	Na 1 testera připadá 1,1 programátora
Česká firma nad 100 zaměstnanců	Na 1 testera připadá 3,8 programátora
Zahraniční firma nad 100 zaměstnanců	Na 1 testera připadá 0,4 programátora

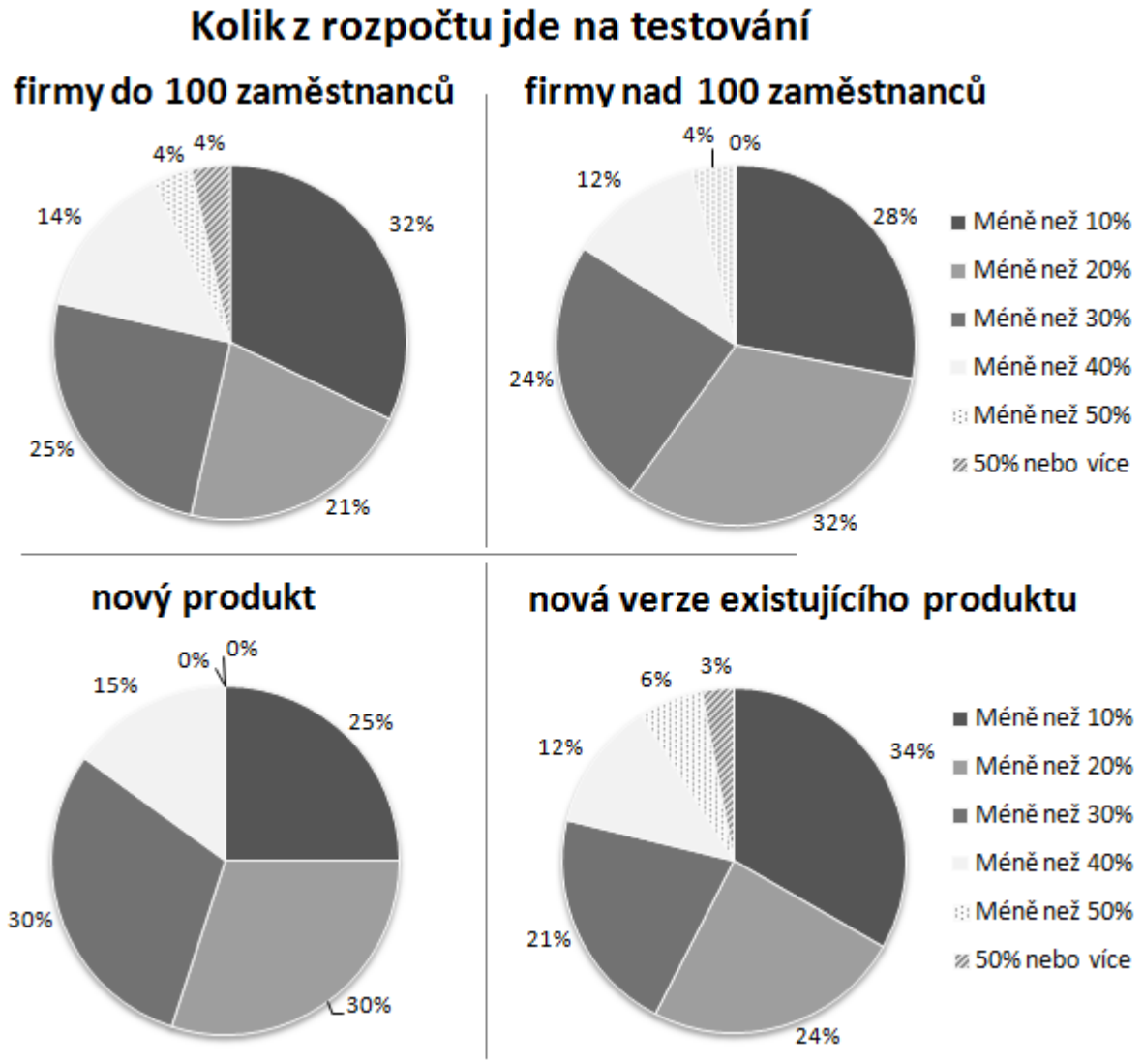
Velikost investice do testování

Testování softwaru je důležitá součást vývoje a utváření kvality, které může dle Beizera spotřebovat více jak padesát procent celkového rozpočtu na vývoj (citováno v Merkel a Kanij, 2010). Výsledky průzkumu v softwarových firmách působících v České republice toto potvrzují, nicméně takový rozpočet je velmi výjimečný.

Otázky spojené s velikostí investic byly nepovinné, v části, které byla označená jako určená pouze manažerům. Výsledky českého průzkumu ukázaly, že jak menší, tak větší firmy investují relativně malou část rozpočtu projektu do testování. Srovnání se zahraničím zde nebylo možné, kvůli příliš malému počtu zahraničních účastníků, kteří odpověděli na tuto otázku.

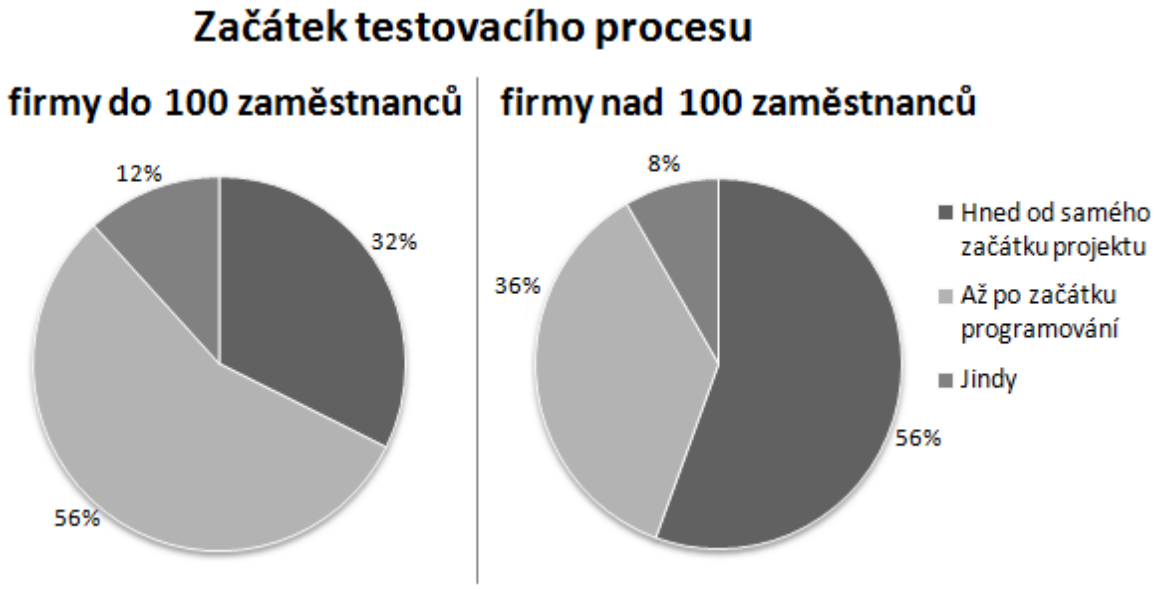
Graf na obrázku 3 ukazuje, jak čeští respondenti odpovídali na otázku, jakou část rozpočtu tvořil rozpočet na testování na jejich posledním projektu. Rozložení odpovědí pro menší i větší firmy vychází podobně, a to i přesto, že se liší jak v nejčastější odpovědi, tak v tom, zda jsou jejich zaměstnavatelé ochotni věnovat na testování více jak polovinu rozpočtu.

V tomto případě byli respondenti rozděleni i podle charakteru projektu, pro který rozpočet uvedli. Dle očekávání, úprava existujícího produktu má obecně menší rozpočet než vytvoření kompletně nového produktu.



Obrázek 3: Kolik z celkového rozpočtu na projekt pojme v České republice testování Zdroj: autor

Představu, jak rychle je tento rozpočet na testování utracen si můžeme udělat z následujícího grafu (obrázek 4), kde je vidět, že více jak polovina firem nad sto zaměstnanců začíná s procesem testování od samotného počátku projektu, zatímco menší firmy dávají přednost začít s testováním, až když je k dispozici spustitelná verze. Jelikož v období před existencí spustitelné verze proces testování zahrnuje plánování a přípravu testů, dá se z následujícího bezpečně usuzovat, že menší firmy musí často minimalizovat čas na přípravu a plánování testů. To je možné výběrem menších projektů a testovacích technik, které nevyžadují tvorbu podrobné dokumentace.



Obrázek 4: Srovnání, kdy firmy začínají s procesem testování Zdroj: autor

Proces testování začíná před vytvořením první nové spustitelné verze programu ve všech významných metodikách, jako je RUP[2], Extrémní programování[3], V-model[4] a další. Včasná a neustálá kontrola kvality je jednou z šesti uznávaných nejlepších praktik[5].

Hlavními důvody, proč je nejefektivnější začít v dostatečném předstihu před samotným programováním, je náročnost přípravy testování a to, že včasným odhalením chyb, dochází k minimalizaci nákladů a omezení problémů s překročeným rozpočtem a nepřipraveností produktu v termínu.

Několik zahraničních studií prokázalo, že významné množství – zhruba 50 % - všech nalezených chyb vzniká v požadavcích (Sheldon udává 41 % (1992), Nakashima udává 50 % (1999), Budde udává 56 % (1984)). Zároveň platí, že čím později jsou chyby odhaleny, tím dražší je jejich oprava. Tento rostoucí trend může být až exponenciální (Vrbka, 2008; Patton, 2002), zvláště když do porovnání zahrneme nejen cenu opravy, ale i další související náklady jako jsou ztráty spojené s existencí chyby.

Z toho vidíme, že i když rozpočet na testování je poměrně malý, bývá firmami často neefektivně alokován na odstraňování chyb a opakované přetestování než na jejich prevenci.

Vzdělávání v oblasti testování a kvality

Při posuzování kvality produktů firmy využívají v čím dál větší míře podrobné testovací scénáře[6] a nástroje na automatizaci testů, čímž delegují provádění i značné části testů na nezkušené pracovníky nebo automaty. Na druhé straně se tím zvětšuje prostor pro specializaci odborníků v oboru testování. Znalosti, schopnosti a zkušenosti v oblastech byznys logiky produktu a testování softwaru tak nadále hrají významnou roli.

Merkel dokládá ve své studii, že jsou pozorovány výrazné rozdíly ve výkonnosti softwarových testerů[7], kdy nejlepší testeři jsou přinejmenším o polovinu výkonnější, než je v daném týmu průměr.

[2] Rational Unified Process (RUP) je komerční tradiční metodika prosazovaná firmou IBM.
[3] Extrémní programování je jednou z významných agilních metodik.
[4] V- model – jeden z modelů životního cyklu projektu, používaný v řadě firemních metodik.
[5] Nejlepší praktiky (Best practices) – soubor šesti praktik adoptovaných řadou metodik. Původně byly identifikovány jako společné vlastnosti řešení, které aplikovaly úspěšné firmy ve snaze předejít problémům při vývoji softwaru.
[6] Testovací scénář je posloupnost jednotlivých kroků testu spolu s popisem zadávaných dat a očekávaných reakcí systému. Testovací scénáře mohou být rozděleny do testovacích případů, kdy každý případ se zaměřuje na otestování jednoho funkčního toku konkrétní funkčnosti.
[7] Výkonnost byla posuzována subjektivně respondenty průzkumu, kteří určili jako důležité pro měření výkonnosti následující faktory: kvalitu hlášení chyb, závažnost a počet objevených chyb, odhalení důsledků chyby, pečlivost při plánování a provádění testů.

Z toho vyplývá, že individuální schopnost a znalosti velkou měrou ovlivňují výsledky jedinců (Merkel a Kanij, 2010).

Iivonan na základě své studie identifikoval společné vlastnosti vysoce výkonných testerů[8] ve třech finských softwarových společnostech.

Následující vlastnosti byly společné u všech zkoumaných testerů s vysokou výkonností (Iivonen, 2009):

- Zkušenost s produktem.
- Zkušenost v dané byznys oblasti.
- Zkušenost s programováním.
- Zkušenost se speciálními testovacími technikami.
- Psaní dobrých hlášení chyb.
- Udržení nadhledu/přehledu .
- Baví ho hledat chyby.
- Pečlivost, svědomitost, trpělivost a výdrž.

V seznamu vlastností není zmíněna práce s nástroji, neboť chyby nalezené automatem se ne vždy dají přiřadit ke konkrétnímu objeviteli.

Při vzdělávání testerů je tedy zapotřebí se zaměřit vedle popisu procesů a technik testování a zajišťování kvality, na znalost dané byznys domény, znalost testování a praktické zkušenosti. Toto vzdělání má tester možnost získat pomocí zdrojů a způsobů vyobrazených v tabulce 3.

Tabulka 3: Možnosti získání znalostí a zkušeností Zdroj: autor

Zdroje získávání znalostí a zkušeností:	Způsoby získávání znalostí a zkušeností:
• Vysokoškolské studium • Zaměstnavatel • Samostudium • Certifikace a rekvalifikace	• Školení a kurzy • Konference a kolegové • Internet (odborné články, blogy, fóra) • Knihy • Praxe

Pomineme-li získávání zkušeností z praxe, převládá jako hlavní zdroj znalostí v České republice internet (46,4 %), následován kolegy (20,2 %) a knihami (17,9 %). Školení a kurzy, ať už odborné nebo univerzitní, mají nejmenší podíl na vzdělávání pracovníků softwarových firem v oblasti testování.

To je výrazný rozdíl oproti zahraničí, kde tito pracovníci softwarových firem čerpají znalosti v oblasti testování zejména ze školení (48,6 %) a dále z internetu a knih.

Zatímco samostudium je neřízená forma vzdělávání, jehož kvalita závisí od zvoleného způsobu získávání znalostí, absolvování speciálně zaměřených kurzů v rámci vysokoškolského studia by mělo poskytovat dostatečnou záruku kvality vzdělání.

Přehled znalostí a dovedností v oblasti testování a řízení kvality, které je možné získat na pražských veřejných vysokých školách a které mají studijní obor zaměřený na informatiku, je v tabulce 4.

[8] Vysoce výkonní testeři byli identifikováni analýzou obsahu databáze systémů pro hlášení chyb na základě počtu, závažnosti a podílu vyřešených chyb.

Tabulka 4: Přehled výuky požadovaných znalostí v pěti významných pražských veřejných vysokých školách obsahujících studijní obor zaměřený na informatiku[9] Zdroj: autor

Výuka	ČVUT[10]	VŠE[11]	CUNI[12]	ČZU[13]	VŠCHT[14]
Testování softwaru	Ano	Ano	Ano	**Ne**	**Ne**
Řízení kvality (prevence, náklady, metriky)	Ano	Ano	**Ne**	Ano	**Ne**
Nástroj(e) na řízení testování	Ano	Ano	**Ne**	**Ne**	**Ne**
Nástroj(e) na funkční testování	Ano	Ano	**Ne**	**Ne**	**Ne**
Nástroj(e) na zátěžové testování	Ano	Ano	Ano	**Ne**	**Ne**
Nástroj(e) na jednotkové testování	Ano	Ano	**Ne**	**Ne**	**Ne**
Nástroj(e) na analýzu kódu	**Ne**	**Ne**	Ano	**Ne**	**Ne**
Programování	Ano	Ano	Ano	Ano	Ano
Databázové systémy	Ano	Ano	Ano	Ano	Ano

Z tabulky jde vidět, že testování a řízení kvality není ještě plně integrovanou součástí výuky informatiky na všech školách, které tak i dnes produkují budoucí IT manažery, softwarové inženýry a analytiky, kterým chybí znalosti o utváření kvality a praktické dovednosti s nástroji pro její zajištění.

Závěr

Výzkum se zaměřoval na úroveň testování a řízení kvality v softwarových firmách vyvíjejících software na území České republiky, u nichž kvalita, efektivita a spolehlivost vývoje jsou důležitými aspekty konkurenceschopnosti. Přičemž byly zodpovězeny následující otázky:

1. Jakým způsobem testují firmy software a koho najímají na testování
2. Jak velký je rozpočet na testování v poměru k rozpočtu na celý vývoj
3. Odkud čerpají zaměstnanci své znalosti o řízení kvality a jaké jsou možnosti jejich vzdělávání

Jako základ pro hodnocení výsledků průzkumu stavu testování v České republice byl použit obdobný průzkum v zahraničí. Počet a charakteristika účastníků tohoto průzkumu odpovídala počtu a charakteristice účastníků českého průzkumu, je však nutné zmínit dva rozdíly, které mohli v menší míře ovlivnit výsledky průzkumů:

1. Zatímco průzkum v České republice probíhal na konci roku 2010, průzkum v zahraničí proběhl o rok a čtvrt později, v lednu 2012. Tento rozdíl by však měl být zanedbatelný, jelikož se stav testování v České republice dle zkušenosti autorky za tuto dobu nijak nezměnil.
2. Zatímco v České republice byla oslovena velká část firem s žádostí o vyplnění dotazníku, zahraniční průzkum žádal o vyplnění dotazníku pouze na zahraničních fórech a diskusních skupinách (náhodní Čeští respondenti nebyli do výsledků zahrnuti). Dá se tedy předpokládat, že větší část zahraničních respondentů má větší zájem o získávání znalostí v oblasti testování, než by tak mohlo být u průměrného českého vzorku.

Tyto rozdíly a jejich dopady byly vzaty při vyhodnocení a analýze výsledků do úvahy.

Z výsledků průzkumů bylo zjištěno, že testování softwaru je v České republice věnována menší pozornost než v zahraničí, což se odráží zejména ve vzdělávání pracovníků a v používání moderních technik a nástrojů. Pro zlepšení tohoto stavu bych navrhla následující kroky:

- Rozšíření nabídky vysokoškolských a odborných kurzů zejména s důrazem na preventivní metody řízení kvality a na seznámení manažerů s ekonomickými aspekty investování do kvality vývoje softwaru.

[9] Údaje z katalogů předmětů 2011/2012 jednotlivých vysokých škol přístupných na internetu k datu 15.11.2011
[10] ČVUT – České vysoké učení technické v Praze
[11] VŠE – Vysoká škola ekonomická v Praze
[12] CUNI – Univerzita Karlova v Praze
[13] ČZU – Česká zemědělská univerzita v Praze
[14] VŠCHT – Vysoká škola chemocko-technologická v Praze

- Rozšíření praktických kurzů zacílených na použití teoretických znalostí o řízení kvality v praxi.
- Rozšířit nabídku školení ve firmách pro zaměstnance aktivně se podílejících na utváření kvality softwaru, případně motivovat zaměstnance k řízenému samostudiu s doporučeným seznamem kvalitních zdrojů.

Literatura

BUDDE, Reinhard. *Approaches to Prototyping*. Berlin: Springer-Verlag, 1984. ISBN 978-0387134901.

MERKEL, Robert a Tanjila KANIJ. Does the Individual Matter in Software Testing? [online]. 2010. 0 [cit. 2012-01-20]. Dostupné z: http://www.swinburne.edu.au/ict/research/sat/technicalReports/TC2010-001.pdf

NAKASHIMA, T., OYAMA, M., HISADA, H. a N. ISHII. Analysis of software bug causes and its prevention. In *Information and Software Technology*. Elsevier Science Publishing Company, Inc., 1999. Volume 41, Number 15, s. 1059-1068. ISSN 0950-5849.

IIVONEN, Joonas. *Identifying and Characterizing Highly Performing Testers – A Case Study in Three Software Product Companies.* Espoo: Helsinki University of Technology, Faculty of Information and Natural Sciences, 2009. Master's Thesis. Dostupné také z: http://www.soberit.hut.fi/espa/publications/MTI.pdf

PATTON, Ron. *Testování softwaru*. Brno: Computer Press, 2002. ISBN: 80-7226-636-5.

SHELDON, Frederick T., et al. *Reliability Measurement from Theory to Practice*. [online]. IEEE Software, July 1992. [cit. 2012-01-20]. Dostupné z: http://www.cs.utexas.edu/~hestness/papers/sheldon-softwarereli.pdf

VRBKA, Zdeněk. *Zpráva o stavu a potřebách v oblasti testování informačních systémů*. Brno: Fakulta informatiky Masarykova univerzita, 2008. Technická zpráva. [cit. 2012-01-20]. Dostupné z: http://kirlab.fi.muni.cz/documents/RR_2008_Zprava_o_stavu_a_potrebach_v_oblasti_testovani_informacnich_systemu.pdf

JEL Classification: A20, C80, Y80.

Summary

State of software testing in the Czech Republic

The aim of this article is to introduce results of the survey of the current state of testing and quality management in software companies developing software in the Czech Republic, as quality, efficiency and reliability of the software development are important competitive aspects. The article focuses on how testing is done in software companies and on methods of acquiring knowledge in the field of testing, thus on quality of education. The article answers following questions:

1. How companies are testing the software and who hire the testing?
2. How big is the budget for testing in relation to the budget for the entire development?
3. Where workers draw on their knowledge of quality management and what are the possibilities of their education?

For a comparison a similar international survey was made and used, where respondents were mainly from India and USA. Both surveys took a form of anonymous online questionnaires and were intended for employees who are either involved in testing, or at least they come in contact with it every day (e.g. developers, project managers,..).

Furthermore, the article describes options available for educating and training employees in software testing and how they are used in the Czech Republic. In particular it provides the overview of the specific knowledge and skills in testing and quality management that can be gained from those public universities in Prague that have study programs focused on information technology.

Keywords: software testing, quality management, research of software quality, project management, software engineering.

Konceptuální model pro Enteprise Architecture jako služba

Libor Gála
gala@vse.cz
Doktorand oboru informatika
Školitel: Prof. Ing. Jaroslav Jandoš, CSc., jandos@vse.cz

Abstrakt: Architektonické principy se stále častěji prosazují v podnicích a aplikace těchto principů se označuje pojmem Enterprise Architecture. Záměrem tohoto příspěvku je využití principů služeb a služební orientace v oblasti Enterprise Architecture (EA) s cílem formulovat návrh hrubého konceptuálního modelu služeb, kterými by EA mohla být v podniku realizována, tzv. Enterprise Architecture as a Service (EAaaS). Práce je rozpracováním případu, kdy služby architektury byly mapovány do modelu řízení informatiky, a v tomto příspěvku jsou nejprve shrnuty klíčové principy služeb a služebně orientovaného přístupu, jež podle mého názoru jsou podstatné pro formulaci „něčeho" co má být v podniku realizováno (nabízeno a konzumováno) jako služba. Dále je analyzována EA, tj. je charakterizován její obsah i rozsah vzhledem k jejímu různému vnímání a jsou představeny klíčové typy aktivit, které jsou v EA realizovány. Výsledky analýzy těchto dvou pilířů vstupují na jedné straně jako omezující faktory modelu (služby) a na straně druhé jako podklad pro formulaci náplně služeb vlastního návrhu konceptuálního modelu EAaaS.

Klíčová slova: Podniková architektura, Enterprise Architecture, EA, Service, Architecture as a Service.

Úvod

Poznání, jak by organizace měla v budoucnu vypadat a jaké realizační kroky by měla udělat, aby byla schopna v budoucnosti přežít, fungovat a rozvíjet se, patří ke klíčovým úkolům strategického řízení. Jednou z metod, či spíše se jedná o celou disciplínu, je architektura, která stále častěji ze své původní domény, tj. stavebnictví, proniká i do oblasti návrhu konstrukce socio-ekonomických systémů. V prostředí podniku, jakožto jednoho z typů socio-ekonomického systému se tato disciplína ujala pod označení Enterprise Architecture (Malveau, 2004). Podnik je chápán v širším pojetí jako „kolekce (soubor) organizací, které mají společný cíl či pracují na stejných principech anebo mají shodný předmět zájmu (Schekkerman, 2004)".

Analýzy Langenbergové a Wegmana (2004) a také Schöenherra (2008) ukazují, že se koncept Enterprise Architeture (EA) nejen na úrovni vládních institucí, ale také v akademické a podnikatelské sféře „ujal" a realizované průzkumy poukazují na to, že „se na podnikové úrovni jedná o koncept již dostatečně zralý" (Aziz a Obitz, 2007), že „se stále zvyšuje role konceptu v oblasti strategických rozhodnutí o transformaci podnikání včetně reakce na výzvy a změny podnikatelského prostředí" (Obitz a Babu, 2009) a také, že „jeho implementace může zlepšit i celkovou výkonnost organizace" (Kamogawa a Okada, 2009).

Na druhou stranu, v okamžiku, kdy se architektonické principy v podnicích přesunuly z aplikací v jeho technických podsystémech a tedy vykročily do „vyšších pater" řízení, řeklo by se, kdy se dostávají do sféry měkkých systémů, se stále častěji objevují kritické hlasy na její adresu. EA je nejčastěji vyčítáno:

- že se nevyrovnala s komplexitou socio-ekonomického systému a není v něm schopna objasnit svůj obsah a rozsah, že její jméno je devalvováno a přetěžováno jejím nadužíváním (van Alst, 2010; Fehskens, 2011).
- že je stále je tažena IT místo toho, aby byla schopna absorbovat i další domény podniku (Gartner, 2010);
- že se utápí v definičních sporech, místo toho, aby se inspirovala disciplínami, kde se pojem architektura používá a inspiraci zužitkovala v přípravě takových podkladů, jejichž implementací by byl „zkonstruován" podnik vyhovující záměrům zainteresovaných i v měnícím se vnějším prostředí (Bloomberg, 2011; Zachman, 2009).

Sama disciplína EA tedy stojí před otázkou, jakým způsobem na tyto výzvy vhodně reagovat. Domnívám se, že inspiraci by si mohla hledat opět v podniku samotném. Podniky nyní velmi často přenáší

jim dobře známý koncept služeb z oblasti služeb pro své zákazníky (Kotler, 1988; Zeithaml, 1996; O'Sullivan, 2002) i do oblasti samotné konstrukce a fungování podniku. Tento přístup se zde objevuje pod označením Service-Oriented Enterprise (Steen, 2005; Khoshafian, 2006). To neznamená vůbec nic jiného než, že podnik své byznys funkce konstruuje jako jasně definované služby (Hrabě, 2010) a přitom využívá přirozené vlastnosti služeb být vhodně organizovány – komponovány do na cíl zaměřených celků. Že koncept služeb je funkční i v této oblasti, lze najít nejen v IT prostředí, kde si již architektura upevnila svoji pozici a kde třeba ITIL definuje řadu služeb (Hanna a další, 2009), ale především v oblasti outsourcingu byznys funkcí (Smejkal a Rais, 2006; Rydvaldová a Rydval, 2007; Dvořáček a Tyll, 2010).

Domnívám se, že pokud podnik je schopen služby akceptovat a absorbovat ve výkonu byznys funkcí i jejich managementu a také architektura sama ve svém výsledku na tento koncept reaguje, neboť je schopna minimálně v IT tuto skutečnost zohlednit (Steen a další, 2005; Velitchkov, 2008), pak by i sama disciplína EA by mohla být předmětem téhož konceptu.

Záměrem této práce je využít principů služeb a služební orientace v oblasti Enterprise Architecture při formulaci návrhu hrubého konceptuálního modelu služeb, kterými by EA mohla být v podniku realizována, tzv. Enterprise Architecture as a Service (EAaaS).

Práce je rozpracováním případu, kdy služby architektury byly mapovány do modelu řízení informatiky (Gála a Jandoš, 2011), a v tomto příspěvku jsou nejprve shrnuty klíčové principy služeb a služebně orientovaného přístupu, jež podle mého názoru jsou podstatné pro formulaci „něčeho" co má být v podniku realizováno (nabízeno a konzumováno) jako služba. Dále je analyzována EA, tj. je charakterizován její obsah i rozsah vzhledem k jejímu různému vnímání a jsou představeny klíčové typy aktivit, které jsou v EA realizovány. Výsledky analýzy těchto dvou pilířů vstupují na jedné straně jako omezující faktory modelu (služby) a na straně druhé jako podklad pro formulaci náplně služeb vlastního návrhu konceptuálního modelu EAaaS.

Služby a služebně orientovaný přístup

Princip služeb je v podnikovém prostředí dostatečně znám (Kotler, 1988; Zeithaml, 1996; O'Sullivan, 2002) a objasněn i z pozice různých vědeckých disciplín (Baida, 2004; Hrabě, 2010). Lze tedy shrnout, že se jedná o funkce, respektive jejich schopnosti, vytvořené propojením různorodých vlastností různých zdrojů, které přinášejí přidanou hodnotu a které jsou vykonané poskytovatelem pro příjemce na základě jeho požadavku a v souladu se vzájemnou dohodou (smlouvou) o parametrech služby.

Z hlediska zaměření příspěvku je nutné sledovat vedle řady atributů a principů služeb především motiv, za jakým je služba konzumována, její rozsah a principy, za který mohou být komponovány do větších celků.

U konzumentů lze identifikovat dva zásadní motivy. První směřuje k tomu, že službou jsou realizovány činnosti, pro které sám konzument nemá schopnosti, tj. pro konzumenta je důležité samotné „konání", které za něho někdo realizuje. Druhý motiv směřuje k tomu, že výsledkem služby je nějaký produkt, který sám konzument není svými schopnostmi vytvořit, tzn., že pro konzumenta je důležitý výsledek a ne samotné konání, které k výsledku směřuje. Podle účelu lze provést členění služeb na služby orientované na úlohu a entitu (Erl, 2008).

Rozsahem služby vyjadřujeme, co vše konzumací služby získáme a často pro něho využíváme pojmu granularita. Za hrubozrnné služby (zjednodušeně) považujeme ty s velkým rozsahem a za jemnozrnné s rozsahem malým. Jde samozřejmě o vyjádření krajních mezí a to za účelem vymezení znuvupoužitelnosti služby v různých situacích. Platí, že služby jemnozrnné mají vyšší znovupoužitelnost (Erl, 2008). Vhodná granularita služeb je jedním z předpokladů pro to, aby bylo možné služby skládat do vyšších celků.

Tabulka 1 ukazuje příklady služeb z různých oblastí.

Tabulka 1: Příklady služeb

Rozsah / Motiv	Jemnozrnná	Hrubozrnná
Úloha	Zaúčtování záznamu; Přeprava mezi dvěma body	Řízení informatiky; Řízení projektů
Entita (produkt)	Reklamní kampaň; Plán materiálových požadavků	Inovace výrobku; Hrubý plán výroby

Služby s různým rozsahem i orientací jsou předmětem kompozice, která má za cíl připravit pro konzumenta takovou složenou službu, která naplní jeho očekávání. V kompozici služby vzájemně spolupracují a spolupráce je koordinována. Podobně jako u spolupráce v pracovních týmech se zde používají metody, které Dillenbourg, a další (1996) označuje jako metody kolaborativní a metody kooperativní. Koordinace kooperace je řízena nějakým prvkem a je označována jako orchestrace, zatímco koordinace kolaborace, tzv. choreografie, řídící prvek nemá.

Rozsah u komponovaných služeb může být ovlivněn úrovní závislosti mezi službami, které vstupují do kompozice. Ačkoliv teorie spojená se službami v oblasti IT, tzv SOA (Earl, 2008) předpokládá trend ke službám se slabou vazbou, kterou Humm (2008) charakterizuje jako princip založený na následujících tvrzeních a) X má minimální znalost o Y; b) X může pracovat, i když Y je dočasně nedostupná a c) Y neočekává, že X splní nějaké vstupní podmínky a X neočekává, že Y naplní nějaké konkrétní podmínky X, musíme počítat s tím, že především z obchodních důvodů, je tento princip často porušován.

Enterprise Architecture

Na rozdíl od konceptu služeb je vymezení Enterprise Architecture obtížné, což dokazuje nepřeberné množství různých definic, které jsou prezentovány jak v příspěvcích, knihách, ale i v diskusích v odborných komunitách. Ve vymezení se totiž skrývá odpověď na otázky co je to architektura, jaký je její rozsah a obsah a jaké metody bude použito (Buchalcevová a Gála, 2008) a to vše je, především v praktických aplikacích disciplíny v podnicích, předmětem rozličných názorů.

Obsah Enteprise Architecture

Při vymezení obsahu architektury se dnes nejčastěji využívá standardu ISO/IEC/IEEE 42010:2011 (2011). Podle tohoto standardu vystižení podstaty nějakého systému, tj. identifikace fundamentálních prvků a vlastností těchto prvků, stejně tak klíčových vazeb mezi nimi, ale také jejich podstatných vazeb k okolí systému, které jsou zásadní pro fungování, chování, charakter i formu systému, jeho vnitřní i vnější harmonii představuje to, co označujeme pojmem architektura systému. Pokud tedy existuje systém samotný, existuje také jeho architektura (Basl a kol., 2011). Je pak pouze na těch, kteří jsou na systému zainteresovaní, zda jsou schopni poznat, co je v systému i pro systém fundamentální, a co by v jeho evolučním či revolučním vývoji mělo či mohlo být modifikováno, ať již vlivem zainteresovaných anebo změnou v okolí systému tak, aby byla zajištěna jeho budoucí existence a fungování.

Poznání architektury zachycujeme jejím popisem, který je schopen zainteresovaným prezentovat fundament systému tak, aby v něm identifikovali svá hlediska zájmu, tj. svoji zainteresovanost na systému, avšak, který také zahrnuje časové hledisko rozvoje systému, tj. popisuje fundament systému v současném stavu, ve stavu budoucím, tzn. vyhovujícím zainteresovaným v budoucnosti včetně toho, jak evolučně či revolučně provést transformaci, aby budoucího stavu bylo dosaženo (Buchalcevová a Gála, 2008).

Popis architektury tvoří komplexní model, respektive vzájemně provázaná soustavou modelů (ISO/IEC/IEEE 42010:2011, 2011). Zda se jedná o modely observační nebo fenomenologické, homomorfní anebo izomorfní, či jaký jazyk modelu je použit, závisí na konkrétním systému, respektive jeho typu a potřebách zainteresovaných (Křemen, 2007). Jiný model je použit, jestliže se jedná o rostlinu a jiný v případě člověka a jiný model je použit třeba u zmíněné rostliny, pokud je zainteresovanou stranou badatel a jiný v případě, že se jedná o pěstitele. Jednotlivým zainteresovaným jsou nad modely poskytovány pohledy, které naplňují jejich hlediska.

Přestože by se zdálo, že je obsah architektury dostatečně zřetelný, upozorňuje Zachman (2007) na to, že velmi často dochází k záměně mezi tím, co je architektura a co je výsledkem její implementace. Zatímco architektura představuje množinu popisných reprezentací, které jsou nutné proto, aby objekt mohl být vytvořen, výsledkem implementace architektury je již nějaký skutečný objekt (včetně jeho chování), který je instancí formulované architektury.

I v podniku obsahem architektury je popsaný model. Teprve implementací podle něho vzniká konkrétní podnik nebo jen jeho část, přičemž jeho forma a chování odpovídá zásadám, které architektura formulovala. Architektura tedy pracuje s reprezentací objektů, které ze kterých je podnik složen. Tyto objekty tvoří a) zdroje podniku, které jsou identifikovány různými metodami (např. Resource-Base View) anebo jsou do podniku převzaty z oborových referenčních modelů (např. skupina modelů APQC), b) schopnosti organizace zdroje vhodně spojovat i c) takto vzniklé nové celky. Množina objektů je v souladu s definicí architektury omezena a architektura se zabývá jen těmi, které jsou dle architektonického rozhodnutí fundamentální pro budoucí funkci podniku. Stanovení prvků patří k významným otázkám EA a pojí se s vymezením rozsahu EA, tj. formulací, kde jsou hranice systému, který je předmětem architektury (Bernus, 2009).

Při modelování Zachman (1987, 2007; 2009) doporučuje zohlednit řadu hledisek, respektive vytvořit si ontologickou představu těchto prvků zohledňující nejen hlediska, tzv. aspekty, ale i to z jakého úhlu pohledu, tzv. perspektivy je vhodné na objekty pohlížet. Zohlednění ontologie se projevilo i v rozšíření názvu „The Zachman Framework™“ o podtitul „The Enterprise Ontology™“ (Zachman, 2008), který představuje ontologický pohled na podnik a který může být v podniku využit jako základ metamodelu architektury. Fundamentální objekt podniku, který je modelovaný v různých aspektech a perspektivách na základě formulované ontologie a jehož popis tvoří soustavu modelů, je základem proto, co následně standard ISO/IEC/IEEE 42010:2011 označuje jako hlediska a pohledy zainteresovaných. Celá EA je pak tvořena soustavou souboru modelů tří typů - současnost, budoucnost, transformace (viz výše).

Rozsah Enteprise Architecture

V úvodu uvedené kritika EA je veskrze dle mého názoru oprávněná. Podnik totiž představuje podle Bouldingovi (2003) hierarchie systémů složitý socio-ekonomický systém a existuje v něm bezpočet vazeb různých typů mezi různými prvky.

Z pohledu architektonického lze podle Douceta a dalších (2009) identifikovat ve vztahu mezi dvěma prvky ten, u kterého je „přání prvku“ (budoucnost, budoucí požadavky) pouze poznáváno a architektonickými modely je toto poznání pouze zachyceno a ten, u něhož navíc dochází k takovému modelování (při respektování zvoleného obsahu, tj. metamodelu), že jeho konstrukce je schopna požadavkům vyhovět a mezi prvky je dosažena harmonie.

V podnikovém prostředí spíše hovoříme o souladu (Chan a Reich, 2007) a nejčastěji se hovoří o tzv. souladu mezi byznysem a IT, jenž jsou založeny na pracích, respektive modelech MacDonalda (1991); Baetse (1992), Hendersona a Venkatramana (1993), a které byly následně rozšířeny Luftmanem (1993) a Maesem (1999).

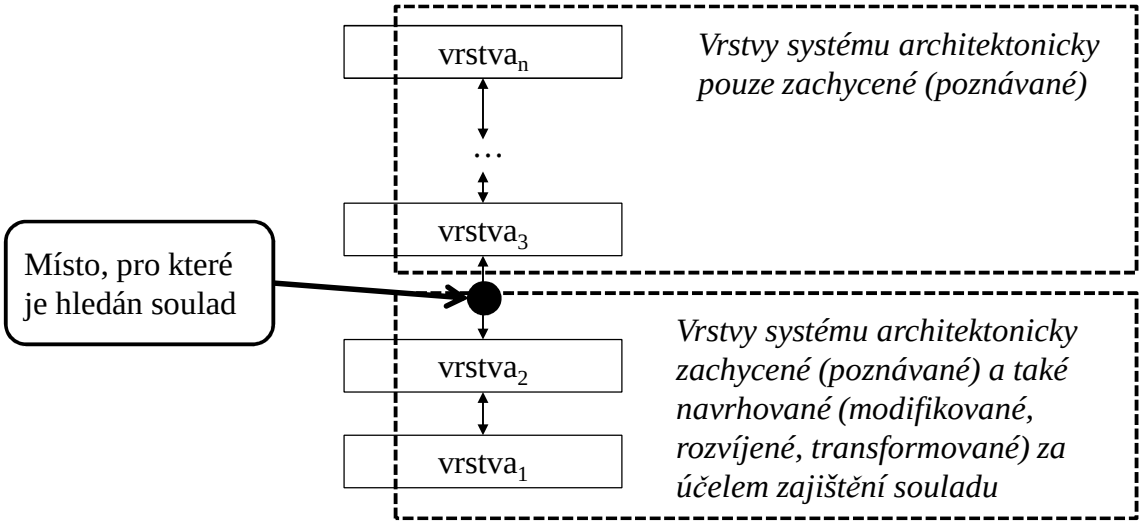
Především praktické postupy dosahování souladu se však ukazují složitější, či spíše granularita prvků byla větší, než předpokládaly teoretické modely. Například Bredemayer a Malanová (2004) identifikovali 3 klíčové případy, které se orientují na dosažení souladu a které jsou předmětem EA. Jedná se o:

- Případ, kdy soulad je hledán mezi vrstvou technologie a vrstvou aplikací informačního systému. EA je vlastně technologickou architekturou.
- Případ, kdy soulad je opět hledán mezi vrstvou technologie a vrstvou aplikací informačního systému avšak napříč celé organizace s cílem formulovat společnou IT infrastrukturu s definovanou množinou služeb za účelem zlepšení spolupráce různých částí podniku např. při zefektivnění řízení portfolia aplikací, eliminaci duplicitních částí ICT apod. (Weill, et al., 2002). EA je tzv. EWITA (Enterpise Wide-IT Architecture).
- Případ, kdy soulad je hledán mezi byznysem a jeho informačním systémem. EA představuje EISA (Enterprise Information System Architecture), tj. architektonicky je koncipován informační systém na základě byznys požadavků.

Doucet a další (2009) definovali také tři případy, s tím, že první z nich je vlastně posledním případem dle Bredemeyera a Malanové. Jedná se o:

- Případ, výše označený jako EISA. Autoři tento případ označují pojmem Foundation Architecture.
- Případ, kdy soulad je hledán mezi strategickými podnikatelskými cíli a byznysem a zároveň mezi byznysem a informačním systémem. Byznys architektura jako součást EA slouží byznys manažerům k plánování transformace byznysu. Autoři tento případ označují jako Extended Architecture. EA v tomto případě je plánovacím nástrojem transformace podniku.
- Případ svým rozsahem stejný předchozímu případu, avšak s tím, že EA je organickou součástí každodenních procesů a je produkována samotnými podnikovými procesy, které jsou v souladu s přijatými architektonickými principy a pravidly. Autoři tento případ označují jako Embeded Architecture.

To znamená, že se praxe orientovala i na hledání souladu mezi jinými prvky než jen mezi byznysem a IT. Systematický přístup k řešení, který napomáhá k rozdělení složitosti systému, je aplikace vrstevnatého modelu (Janssen, 2009), přičemž soulad je formulován mezi jednotlivými vrstvami. ObrázekObrázek 1 znázorňuje tento koncept.



Obrázek 1: Vrstevnatý model a architektonické činnosti

Jsou v něm zároveň uplatněny následující principy (Doucet a další, 2009):

- ne vždy musí být předmětem architektonických činností celá vrstva a
- architektonické činnosti související s budoucností mohou mít charakter:
  - poznávací, tj. v této části jde o zachycení stavu, avšak ne o jeho návrh, tj. ovlivnění anebo
  - i návrhu, včetně návrhu řešení (transformace), které vede k souladu

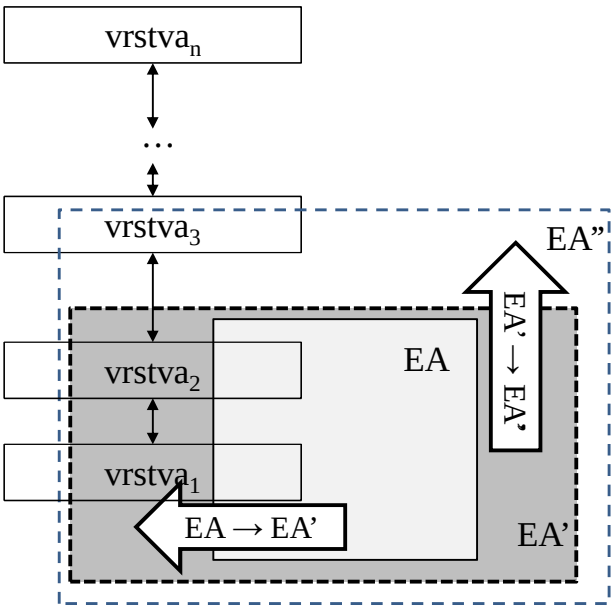
V příkladu na výše uvedeném obrázku jsou architektonicky navrhovány modely vrstvy$_2$, aby byl zajištěn soulad s identifikovanými potřebami vrstvy$_3$. Vrstva$_3$ je zde reprezentována zájmem zainteresovaných, je vnějším prostředím a zároveň tvoří misi vrstvy$_2$.

V návaznosti na výše uvedené se s vrstvami lze setkat i v dalších pracích (viz tabulka 2).

Tabulka 2: Přehled vybraných vrstevnatých modelů

Název	**Počet vrstev**	**Vrstvy (od nejvyšší; v originálním znění)**
NIST (Fong a Goldfine, 1989)	5	Business unit architecture; Information architecture; Information systém architecture; Data architecture; Delivery system architecture
FEAF (The Chief Information Officers Council, 1999)	4	Business architecture; Data architecture; Application architecture; Technology architecture
TOGAF9 (The Open Group, 2009)	3	Business Architecture; Information system architecture; Technology architecture
ST.Gallen (Bucher a další; 2006)	5	Business architecture; Process architecture; Integration architecture; Software architecture; Technology (or infrastructure) architecture
Janssen (2009)	5	Business architecture; Business process architecture; Information architecture; Application architecture; Technical architecture (i.e., shared infrastructure)

Uvedené případy ukazují, že rozsah EA v podniku se mění tak, jak architekturou zasahujeme další a další vrstvy. EA je však i při aplikaci vrstevnatého přístupu vyčítáno, že se stále soustřeďuje na IT, tj. je IT centristická (Gartner, 2010). To třeba lze pozorovat u vrstvy, označené jako technologická nebo technická architektura. Při bližším prozkoumání totiž zjistíme, že této vrstvě jsou pouze prvky, respektive objekty související s informačními technologiemi jako např. počítače, sítě, procesy serverů apod. a zcela zde chybí další podnikové prvky, které tvoří jeho technický, či technologický podsystém, tj. třeba stroje a výrobní linky, budovy, rozvody elektrické energie, vody apod. To znamená, že rozsah může být měněn minimálně ve dvou rozměrech, jak znázorňuje obrázek 2.



Obrázek 1: Změny rozsahu a tedy i obsahu EA

Změna rozměru, pak zákonitě vede i ke změně (rozšíření) ontologického metamodelu architektury a podle mého názoru i k možné změně vrstev a jejich uspořádání. Ačkoli v řešení tohoto problému se v současné době používají služebně orientované principy, které vedou k vytváření doménově orientovaných skupin znovupoužitelných služeb, jež jsou dle potřeby komponovány do vhodných celků, patří stanovení rozsahu spolu s výběrem prvků systému, tj. stanovení obsahu EA k zásadním otázkám EA (viz výše).

Metody Enterprise Architecture

Aktivity spojené s architekturou představují komplexní disciplínu (Reynolds, 2010), která pokrývá „konání“ architektů a architektonického týmu. Tyto aktivity nejsou uspořádány do projektu a projektově řízeny, protože EA v podniku má charakter programu (Gála, 2010). EA totiž zaniká až spolu se zánikem systému, např. po akvizici, při fůzi anebo i opravdu, v okamžiku, kdy podnik zaniká.

První typ aktivit má charakter inženýringu a aktivity do této skupiny zařazené jsou svázány s obsahem architektury, tj. architektonickými modely. Jedná se o aktivity tvorby, upřesňování a změny modelů v rozsahu architektury i zvoleného ontologického metamodelu. Typicky se jedná o „mentální operace s ontologickými objekty, tedy operace s vnitropsychickými znalostními modely těch objektů nebo znalostními modely zapsanými ve vnějším obvykle formálním jazyce, tedy operace se symboly a vždy jsou jen jistou aproximací reality“ (Křemen, 2007). Zároveň, právě z důvodů, že zde existuje jistá míra vágnosti způsobená nedokonalostí formálních jazyků, je významnou součástí inženýringových aktivit komunikace architektury vzhledem k zainteresovaným, těm, kteří budou architekturu implementovat a těm kteří budou implementaci kontrolovat. Typicky se jedná o využití vizualizačních a analytických technik (Jonkers a další, 2003), kterými jsou architektura a v ní obsažené závislosti komunikovány.

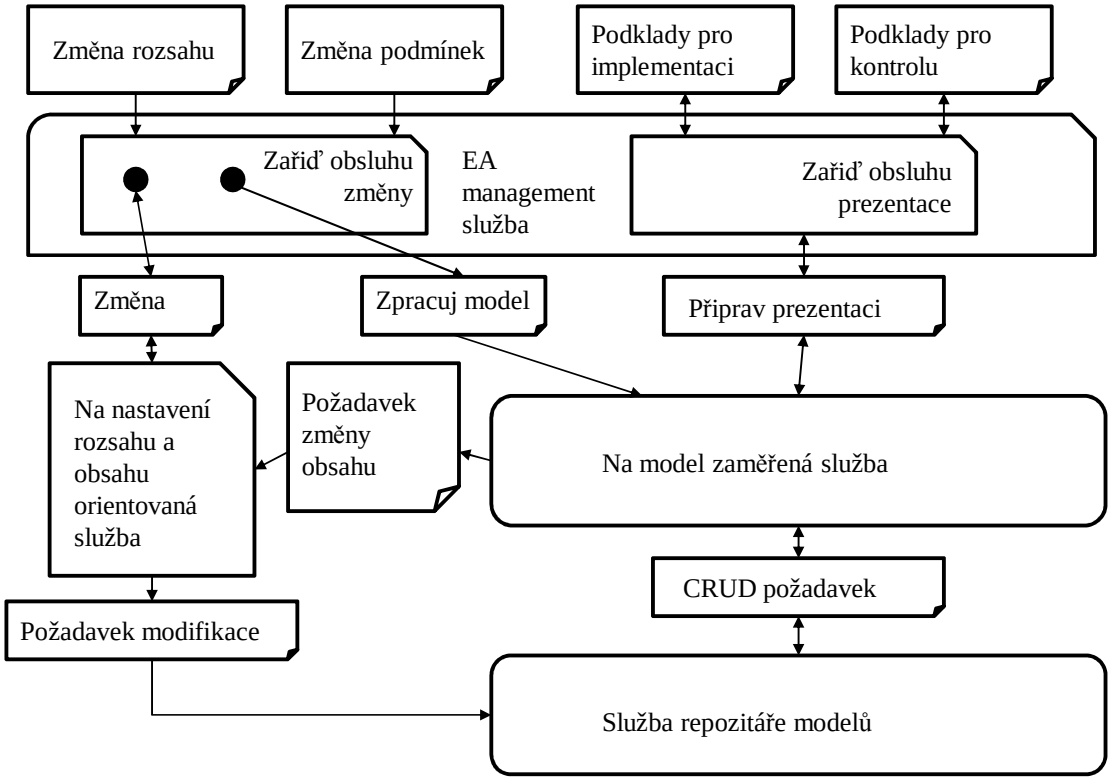
Druhým typem aktivit jsou aktivity spojené s řízením EA během jejího životního cyklu, v rámci kterých jsou aplikovány standardní manažerské postupy a metody (Hladík, 2011), čímž tedy řízení architektury není nijak výjimečné a neliší se od řízení jiných programů, které v podniku probíhají. Možná i to, že podnik může použít pro něho vhodné a mu blízké manažerské metody řízení, vede až k inflaci nejrůznějších procesních modelů, se kterými se můžeme setkat. Základní přehled lze najít třeba v (Schekkerman, 2004) a další se objevují jako příspěvky v časopisech nebo na různých konferencích.

Specifickou oblastí je governance EA, neboť aktivity této oblasti v zásadě z principu zde realizovaných činností nespadají do kompetence architektů a architektonického týmu. V závislosti na rozsahu EA jsou součástí buďto IT Governance anebo tzv. Enterprise governance (Gála, 2010).

Často vznikají otázky ohledně toho, jak jsou zachyceny v modelech EA části, které jsou předmětem outsourcingu, tzn., zda v případě navrhovaného modelu EAaaS konzumovat příslušné služby. Domnívám se, že ano, protože i outsourcované části podniku musí být architektonicky zachyceny, neboť jsou součástí fungování podniku a podnik se bez outsourcovaných funkcí neobejde. Zde máme dispozici v zásadě dvě situace. V první situaci, je požadavek outsourcingu prvků explicitně vyjádřen strategií outsourcingu předem. V tomto případě architektonický inženýring nemodeluje vnitřní strukturu prvků, ale omezuje se na modelování schopností těchto prvků nebo celých oblastí. Architektonicky model, potom představuje to, co v implementaci je formulováno jako outsourcingový kontrakt. Druhá situace znamená, že architektura pracuje s prvky bez ohledu na to, jak budou implementovány a v tomto případě, teprve implementační rozhodnutí realizuje konstrukci jako outsourcingový vztah.

Návrh konceptuálního modelu EAaaS

Koncepce navrhovaného řešení modelu je založena na služebně orientovaných principech a principech samotných služeb a model budou tvořit hrubozrnné služby orientované na úlohu, entitu a orchestraci, přičemž tyto služby realizují aktivity EA inženýringu a managementu a jsou schopny reagovat nejen na změnu podmínek, ale také rozsahu a zároveň jsou schopny poskytovat podklady pro implementaci architektury i pro činnosti kontroly této implementace.



Obrázek 2: Konceptuální model hrubozrnných služeb EAaaS

Obrázek 3 znázorňuje návrh hrubozrnných služeb, které tvoří jako celek EA, která je realizována jako služba. Model zahrnuje dvě služby orientované na entitu. Jedná se o „Na model zaměřenou službu“ a „Služba repozitáře modelů“. Dále zahrnuje jednu samostatnou na úlohu orientovanou službu s názvem „Na nastavení rozsahu a obsahu orientovaná služba“. „EA management služba“ představuje komplexní orchestrační službu, která na této v této verzi konceptuálního modelu obsahuje pouze dvě dílčí na úlohy orientované služby – „Zařiď obsluhu změny“ a „Zařiď obsluhu prezentace“. Jednotlivé další manažerské úlohy nejsou v modelu nyní zachyceny (např. řízení pracovních týmů, kooperace s EA governance apod.).

„Na nastavení rozsahu a obsahu orientovaná služba“ je služba reagující na změnu situace v EA tak, že vyhodnocuje, zda současné nastavení obsahu EA je v souladu s potřebami podniku. Pokud soustava modelů neodpovídá potřebám, zajistí u „služby repositáře modelů“ rozšíření o další modely, tj. navrhne úpravu metamodelu a vyžádá si jeho zaznamenání. Požadavek může být založen:

- Na změně podmínek, které formulují zainteresovaní v důsledku změny cíle podniku nebo změny, která je očekávána v okolí podniku.
- Na změně rozsahu EA, to například znamená, že EA se má orientovat na návrh zajištění nejen souladu mezi technologiemi a aplikacemi IS, ale má být zároveň zajištěn i soulad mezi byznysem a IT.
- Na novém poznání architektů v průběhu samotného řešení architektury, které může vést k nutnosti zařadit třeba další prvky, neboť se pro fungování podniku staly fundamentální.

„Na model zaměřená služba“ realizuje inženýringové aktivity vedoucí k sestavení soustavy modelů v definovaném rozsahu metamodelu, která představuje návrh řešení požadavků zainteresovaných anebo vyplývá z očekávání změny v okolí podniku. Modely jsou předávány „službě repozitáře modelů“ a mohou znamenat vytvoření (C), změnu (U) nebo zrušení (D). Vedle toho služba realizuje pro zainteresované i přípravu vhodných pohledů na soustavu modelů, tzn., připravuje jejich vhodnou prezentaci včetně třeba i vizualizace, a proto komunikuje se „službou repozitáře modelů“ v režimu vyžádání (R). Pohledy prezentuje formou podkladů pro implementaci systému anebo formou informací, které slouží kontrolním mechanismům k prověření, zda implementace je v souladu s architekturou.

Závěr

V příspěvku je představen konceptuální model hrubozrnných služeb, které umožní realizovat program EA jako službu. Účelem rozhodnutí adoptovat služby i v oblasti samotných operací v EA, je odstranění problémů, které vznikají změnou v obsahu a rozsahu EA, tím, že služby by měly přinést do EA interoperabilitu jejich jednotlivých částí, neboť aktivity jsou „ukryty“ za rozhraním a vlastní realizace operací, tj. jaké postupy budou pro sestavení modelů použity u toho kterého prvku systému, již pro EA nebudou podstatné. To umožní i zlepšení flexibility EA při změně rozsahu, který v současné době většinou znamená i změnu dosud použitých rámců.

V další práci se očekává příprava detailního návrh rozhraní služeb a jejich následné rozpracování do obsahu jednotlivých služeb a prověření fungování modelu při změně rozsahu EA. Protože soustava modelů EA tvoří rozsáhlý informační soubor, jsou pro jeho vznik, uložení a správu využívány vhodné IT nástroje pracující nad společnou repository, které mohou anebo nemusí být službám architektury dostupné jako služby. Toto je další část, která bude v následující práci rozpracována.

Literatura

AZIZ, Sohel, Thomas OBITZ. Enterprise Architecture is Maturing : Infosys Enterprise Architecture Survey 2007. [Dokument] s.l. : Infosys Technologies, 2007.

BAETS, W. Aligning Information Systems with Business Strategy. *Journal of Strategic Information Systems*, Vol. 1, No. 4.

BAIDA, Ziv, Jaap GORDIJN, Boris OMELAYENKO. A Shared Service Terminology for Online Service Provisioning. *ICEC '04: Proceedings of the 6th international conference on Electronic commerce*, Delft, The Netherlands, October 25 - 27, 2004, ACM:New York

BASL, Josef, Alena BUCHALCEVOVÁ, Petr DOUCEK, Libor GÁLA, Pavel HRABĚ, Miloš MARYŠKA, David MELICHAR, Jiří TUPA. Inovace podnikových informačních systémů. Praha: PROFESSIONAL PUBLISHING, 2011. ISBN 978-80-7431-045-4.

BERNUS, Peter. The Future of Enterprise Engineering. DOUCET, Gary, John GØTZE, Pallab SAHA a Scott BERNARD. *Coherency Management: Architecting the Enterprise for Alignment, Agility and Assurance*. Blomington: AuthorHouse, 2009, s. 431-449. ISBN 978-1-4389-9606-6.

BLOOMBERG, Jason. Why Nobody is Doing Enterprise Architecture. In: ZapThink, a Dovèl Technologies Company: ZapFlash [online]. April 5, 2011 [cit. 2012-01-22]. Dostupné z: http://www.zapthink.com/2011/04/05/why-nobody-is-doing-enterprise-architecture/

BOULDING, Kenneth E. Obecná teorie systémů jako schéma vědy. MALÝ, Jaroslav V. a Miloš VÍTEK. *Teorie systémů*. Díl II. 3. vyd. Hradec Králové: Gaudeamus, 2003. ISBN 80-7041-833-8.

BREDEMEYER, Dana, Ruth MALAN. What It Takes to Be a Great Enterprise Architect. *Cutter Consortium*. [Online] 2004. [cit. 2008-07-03] Dostupné z http://www.cutter.com/promotions/ear0408/ear0408.pdf.

BUCHALCEVOVÁ, Alena, Libor GÁLA. Architektura v podnikové informatice. *Systémová integrace*. 2008, 15(3), 7–22.

BUCHER, Tobias, Ronny FISCHER, Stephan KURPJUWEIT, Robert WINTER. Analysis and Application Scenarios of Enterprise Architecture – An Exploratory Study. Los Alamitos: IEEE Computer Society. In *IEEE Computer Society, editor, EDOC Workshop on Trends in Enterprise Architecture Research (TEAR 2006)*.

DILLENBOURG, P., M. BAKER, A. BLAYE, C. O'MALLEY. The evolution of research on collaborative learning. *In E. Spada & P. Reiman (Eds) Learning in Humans and Machine: Towards an interdisciplinary learning science*. Oxford: Elsevier. s. 189-211.

DOUCET, Gary, John GØTZE, Pallab SAHA, Scott BERNARD. *Coherency Management. Architecting the Enterprise for Alignment, Agility and Assurance*. Bloomington : authorHouse, 2009. ISBN: 978-1-4389-9606-6.

DVOŘÁČEK, Jiří, Ladislav TYLL. *Outsourcing a offshoring podnikatelských činností*. Praha : C. H. Beck, 2010. xii, 183 s. ISBN 978-80-7400-010-2.

ERL, Thomas. *SOA Principles of Service Design*. Prentice Hall PTR. 2008. ISBN 0132344823

FEHSKENS, Len. Enterprise Architecture's Quest for its Identity. In: *The Open Group: The Open Group Blog* [online]. March 10, 2011 [cit. 2012-01-22]. Dostupné z: http://blog.opengroup.org/2011/03/10/enterprise-architecture%E2%80%99s-quest-for-its-identity/

FONG, Elizabeth N. a Alan H. GOLDFINE. *Information Management Directions: The Integration Challenge*. In: [online]. Washington: U. S. Government Printing Office, 1989 [cit. 2012-01-22]. DOI: NIST Special Publication 500-167. Dostupné z: http://www.itl.nist.gov/lab/specpubs/NIST%20SP%20500-167.pdf

GÁLA, Libor. Prověřování podnikové architektury. *Systémová integrace*. 2010, 17(2), s. 53–69.

GÁLA, Libor, Jaroslav JANDOŠ. Enterprise Architecture Inclusion Into IT Management Model. Milano 14. 11. 2011 – 15. 11. 2011. In: *Creating Global Competitve Economies (A 360 – degree Approach) [CD-ROM]*. Norristown : IBiMA, 2011, s. 1041–1050. ISBN 978-0-9821489-6-9.

Gartner's Enterprise Architecture Hype Cycle Reveals Two Generations of Enterprise Architecture. *GARTNER*. Gartner Newsroom [online]. Egham, UK, August 5, 2010 [cit. 2012-01-22]. Dostupné z: http://www.gartner.com/it/page.jsp?id=1417513

HENDERSON, J. C., N. VENKATRAMAN. Strategic alignment: Leveraging information technology for transforming organization. *IBM Systems Journal*. 1993, 32, 1, s. 4-16. ISSN 0018-8670

HLADÍK, M. Zjednodušení procesu rozvoje podnikové architektury. *Systémová integrace*. 2011, 18(4), (v tisku).

HRABĚ, Pavel. Úloha služeb v rámci podnikové architektury (Enterprise Architecture). *Systémová integrace*. 2010, 17(1), 48–78.

HUMM, Bernhard. Loose Coupling: Service-Oriented Architecture. [Online] 2008. [Cit: 2009-11-25] http://www.fbi.h-da.de/fileadmin/personal/b.humm/SOA_SS08/Lecture/07_Loose_Coupling.pdf.

CHAN, Yolande E., Blaize Horner REICH. IT alignment: what have we learned?. *Journal of Inforation Technology*. 2007, vol. 22(Number 4), 297-315. DOI: 10.1057/palgrave.jit.2000109.

ISO/IEC/IEEE 42010:2011. Systems and software engineering -- Architecture description. Geneva (Switzerland): International Organization for Standardization, 2011.

JANSSEN, Marijn. Framing Enterprise Architecture: A Meta-Framework for Analyzing Architectural Efforts in Organizations. In: DOUCET, G., et al. *Coherency Management : Architecting the Enterprise for Alignment, Agility and Assurance*. Bloomington : AuthorHouse, 2009., p. 99-119. ISBN 978-1-4389-9606-6

JONKERS, Henk, René VAN BUUREN, Farhad ARHAB, Marcello BONSANGUE, Hans BOSMA, Hugo TER DOEST, GROENEWEGEN, Juan Guillen SCHOLTEN, Stijn HOPPENBROUWERS, Maria-Eugenia IACOB, Wil JANSSEN, Marc LANKHORST, Diederik VAN LEEUWEN, Erik PROPER, Andries STAM, Leon VAN DER TORRE a Gert VELDHUIJZEN VAN ZANTEN. Towards a Language for Coherent Enterprise Architecture Descriptions. In: *Proceeding EDOC '03 Proceedings of the 7th International Conference on Enterprise Distributed Object Computing*. Washington: IEEE Computer Society, 2003, s. 28-39.

KAMOGAWA, Takaaki, Hitoshi OKADA. Enterprise Architecture Create Business Value. *Proceedings of the 2009 Ninth Annual International Symposium on Applications and the Internet. Institute of Electrical and Electronics Engineers*. 2009, s. 205-208

KHOSHAFIAN, Setrag. *Service Oriented Enterprises*. New York: Auerbach Publications, 2006. ISBN 978-0849353604.

KOTLER, P. *Marketing Management: Analysis, Planning, Implementation and Control*. 6th ed. Englewodd Cliffs: Prentice Hall. 1988

KŘEMEN, Jaromír. *Modely a systémy*. Praha: Academia, 2007. Knižnice matice technické LANNA. ISBN 978-80-200-1477-1.

LANGENBERG, Kerstin, and Alain WEGMANN. Enterprise Architecture: What Aspects is Current Research Targeting? [Dokument] Lausanne : Ecole Polytechnique Fédérale de Lausanne, 2004. EPFL Technical Report IC/2004/77.

LUFTMAN, J. N., P. R. LEWIS, S. H. OLDACH. Transforming the Enterprise: The Alignment of Business and Informaion Technology Strategies. *IBM System Journal*. 1993, Vol. 32, No. 1, p. 198-221.

MACDONALD, H. The Strategic Alignment Process. In: S. Morton and S Michael (eds.) *The Corporation of the 1990s : Information technology and organizational transformaion*. London: Oxford Press, 1991., s. 310-322.

MAES, R. 1999. A Generic Framework for Information Management. PrimaVera Working Paper 99-03. [online] Universiteit van Amsterdam. 1999. [cit 2010-05-04] Dostupné z <http://imwww.fee.uva.nl/~pv/PDFdocs/99-03.pdf>

MALVEAU, Raphael. Bridging the Gap: Business and Software Architecture, Part 2. *The Cutter Edge*. [Online] 3 February 2004. [cit. 2006-03-13] Dostupné z http://www.cutter.com/research/2004/edge040203.html.

OBITZ, Thomas, Mohan K BABU. Enterprise Architecture Expands its Role in Strategic Business Transformation : Infosys Enterprise Architecture Survey 2008/2009. [Dokument] s.l. : Infosys Technologies, 2009.

O'SULLIVAN, Justin, David EDMOND, Artur H. M. ter HOSTEDE. Service Description: A survey of the general nature of services. *service-description.com*. [online]. 2002. [cit. 2007-11-24] Dostupné z http://www.service-description.com/papers/tr-dapd-osullivan.pdf

REYNOLDS, Chris. Introduction to Business Architecture. Boston: Course Technology PTR, 2010. ISBN 978-1-4254-5422-4.

RYDVALOVÁ, Petra, Jiří RYDVAL. *Outsourcing ve firmě: průvodce pro manažera s tipy pro české prostředí*. Brno: Computer Press, 2007. vii, 102 s. ISBN 978-80-251-1807-8.

SCHEKKERMAN, Jaap. *How to survive in the jungle of Enterprise Architecture Frameworks*. Second Edition. Victoria, Canada: Trafford Publishing, 2004. ISBN 1-4120-1607-X.

SCHÖENHERR, Marten. Towards a Common Terminology in the Discipline of Enterprise Architecture. In: *Service-Oriented Computing – ICSOC 2008 Workshops*. Springer, 2008, s. 400-413. Lecture Notes in Computer Science, Volume 5472/2009. DOI: 10.1007/978-3-642-01247-1_40.

SMEJKAL, Vladimír, Karel RAIS. *Řízení rizik ve firmách a jiných organizacích*. Vyd. 2. Praha: Grada Publishing, 2006, 300 s. ISBN 80-247-1667-4.

STEEN, M. W. A., P. STRATING, M. M. LANKHORST, H. TER DOEST a M.-E. IACOB . Service-Oriented Enterprise Architecture. In: *Doc.telin.nl* [online]. 2005 [cit. 2012-01-26]. Dostupné z: https://doc.telin.nl/dsweb/Get/Document-47205/SOEA_section_v14.pdf

THE CHIEF INFORMATION OFFICERS COUNCIL. 1999. Federal Enterprise Architecture Framework, Version 1.1. Chief Information Officers Council. [Online] September 1999. [Cited: 11 23, 2008.] http://www.cio.gov/Documents/fedarch1.pdf.

THE OPEN GROUP. *The Open Group Architecture Framework (TOGAF)*. [Online] s.l. : The Open Group, 2009. ISBN 978-90-8753-230-7.

VAN ALST, Mike. Why Enterprise Architecture Doesn't Work. In: *Mastering Architecture* [online]. 7 September 2010 [cit. 2012-01-22]. Dostupné z: http://soamastery.blogspot.com/2010/09/why-enterprise-architecture-doesnt-work.html

VELITCHKOV, Ivaylo. Integration of IT Strategy and Enterprise Architecture Models. Gabrovo: CompSysTech'08, 2008. *International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*. ISBN: 978-954-9641-52-3/08/06.

WEILL, Peter, Mani SUBRAMANI, Marianne BROADBENT. Building IT Infrastructure for Strategic Agility. *MIT Sloan Management Review*. 2002, Vol. 44, No. 1.

ZACHMAN, John A. A Framework for Information Systems Architecture. *IBM Systems Journal*. 1987, 26(3), 276-292.

ZACHMAN, John A. Architecture Is Architecture Is Architecture. In: *Zachman International, Inc*. [online]. March 2007 [cit. 2012-01-22]. Dostupné z: http://old.zachmaninternational.com/index.php/ea-articles/68-architecture-is-architecture-is-architecture

ZACHMAN, John A. John Zachman's Concise Definition of The Zachman Framework™. ZACHMAN INTERNATIONAL. *Zachman International, Inc*. [online]. 2008 [cit. 2012-01-24]. Dostupné z: http://www.zachman.com/about-the-zachman-framework

ZACHMAN, John A. Yes, „Enterprise Architecture Is Relative" BUT It Is Not Arbitrary. *Zachman International, Inc*. [online]. 2009 [cit. 2012-01-22]. Dostupné z: http://www.zachmaninternational.com/index.php/ea-articles/117-yes-enterprise-architecture-is-relative-but-it-is-not-arbitrary

ZEITHAML, Valarie A., Mary Jo BITNER. *Services Marketing*. New York: McGraw-Hill. 1996

JEL Classification: L21, M10, M15

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu GAČR P403-10-0303 Enterprise Architecture jako princip v řízení malých a středních organizacích.

Summary

Conceptual model of Enterprise Architecture as a Service

The architectural principles gain the ground in enterprises more frequently. The application of these principles is called Enterprise Architecture. The aim of the paper is to take of the advantage of services and service-orientation and transfer that approach into the Enterprise Architecture area. In the paper there is presented the conceptual model of Enterprise Architecture as a Service (EAaaS). The paper is organized as follows. The background consists of two parts. In the first part the principles of services are analysed and the key attributes of services (the motivation, scope and composability) are presented. The content, scope and methods of Enterprise Architecture are described in the second part. The results of the analysis are applied to the design of the conceptual model. The presented model comprises the task, entity and orchestration services which constitute EAaaS.

Keywords: Enterprise Architecture, EA, Service, Architecture as a Service.

Využití podnikové architektury v závislosti na komplexitě organizace

Martin Hladík
martin_hladik@yahoo.com
Doktorand oboru Informatika
Školitel: prof. Ing. Jaroslav Jandoš, CSc. (jandos@vse.cz)

Abstrakt: Podniková architektura poskytuje nástroje pro podporu řízení transformací, kterými podnik prochází v důsledku vlastního rozvoje nebo změny trhu. Článek zkoumá role podnikové architektury v závislosti na stádiu životního cyklu podniku a jeho komplexitě. Důraz klade na strategickou úroveň řízení v rámci transformace podniku. Všímá si také možného uplatnění principů podnikové architektury v prostředí malých a středních podniků.

Klíčová slova: podniková architektura, strategické řízení, životní cyklus podniku, transformace, komplexita organizace.

Úvod

Podniková architektura (angl. enterprise architecture) je mladá avšak rychle se rozvíjející disciplína, která poskytuje nástroje pro řízení rozvoje podniku (resp. jakékoliv organizace) včetně strategické úrovně řízení. Řada zdrojů řadí podnikovou architekturu mezi manažerské nástroje typu Balanced Scorecard nebo strategické plánování (angl. strategic planning). Setkali jste se však s takovýmto pochopením podnikové architektury ze strany vrcholového managementu?

Z principu věci se podniková architektura nezabývá jen strategickými otázkami, ale také úkoly na nižší úrovni. Přestože řeší architekturu podniku v plném rozsahu (např. uspořádání obchodních procesů), významný podíl svého poslání věnuje informačním technologiím. Některé rámce (angl. framework) považují přínosy na straně informačních technologií za svůj hlavní cíl.

Stejně jako se rozvíjí podnik, také podniková architektura, aby naplnila své poslání, prochází vlastním vývojem. Soulad rozvoje podniku a podnikové architektury je nezbytným předpokladem. Domnívám se, že jedním z kritických faktorů úspěchu využití podnikové architektury je definování jejího účelu, který odpovídá aktuálním potřebám (strategii) podniku.

Většina rámců popisuje podnikovou architekturu velmi rozsáhle. Základní principy podnikové architektury jsou platné pro každý podnik, přesto jsem názoru, že její uplatnění v plném rozsahu má smysl jen u velkých podniků a v konkrétní situaci, obvykle týkající se razatní změny podniku. Na druhé straně lze diskutovat zjednodušenou aplikaci podnikové architektury v prostředí malých a středních podniků.

Správné pochopení situace podniku je, podle mých zkušeností, dalším kritickým faktorem úspěchu, nejen pro využití podnikové architektury, ale především pro zajištění potřebných změn podniku. Význam poznání podniku v takové situaci roste se složitostí organizace (angl. complexity), případně s rozsahem změny. Z tohoto důvodu se zabývám komplexitou podniku, ve které hledám příležitosti pro aplikaci podnikové architektury.

Změny, kterými podnik prochází v důsledku vlastního vývoje a/nebo situace trhu (vnější podmínky), je zajímavé analyzovat v rámci životního cyklu podniku. V každé fázi života podniku lze identifikovat klíčové výzvy, kterými se management podniku musí zabývat. Domnívám se, že tyto výzvy jsou vhodným vstupem pro stanovení cílů podnikové architektury.

Různé využití podnikové architektury v závislosti na životním stádiu a komplexitě podniku (jinými slovy v závislosti na charakteru a rozsahu změny, kterou podnik prochází) vedou k myšlence propojení podnikové architektury a řízení změny podniku (transformace).

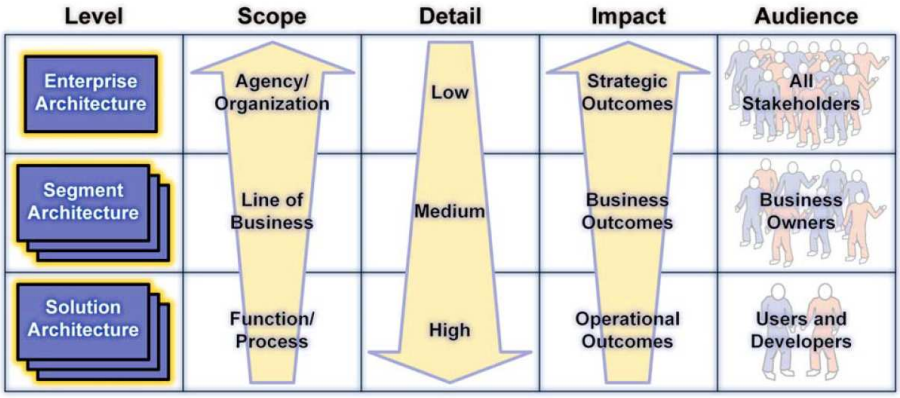
Cílem tohoto textu, vzhledem ke skutečnosti, že zahajuje tvorbu mé doktorské práce, je analyzovat zdroje v oblastech podnikové architektury, životní cyklu podniku, komplexity organizací a řízení změny (transformace). Z uvedených teorií odvozuji hypotézy jejich vzájemného využití pro úspěšnější aplikace podnikové architektury.

Věřím, že výsledná práce nabídne, i s ohledem na ekonomický vývoj posledních let a vizionářské pohledy na budoucí proměny řady oborů, zajímavé návody pro využití podnikové architektury, zejména pak při řešení složitých problémů.

Role podnikové architektury

Podniková architektura poskytuje řadu nástrojů pro podporu řízení rozvoje podniku (příp. v širším významu organizace) a to na strategické i nižší úrovni řízení. Přehledně tuto víceúrovňovou roli podnikové architektury prezentuje rámec Federal Enterprise Architecture (FEA).

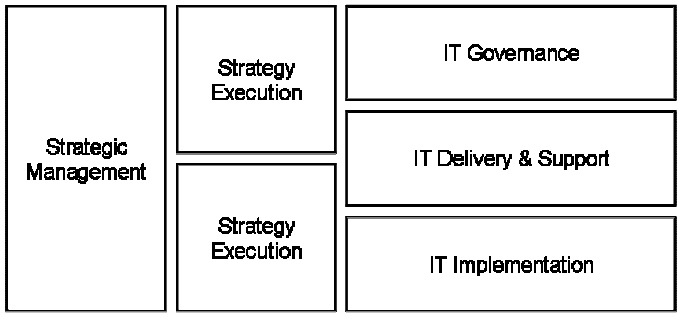
Rámec FEA rozlišuje architekturu na úrovni podniku (angl. enterprise), části podniku (angl. segment – může se jednat např. o divizi podniku) a řešení (angl. solution). Každá úroveň se liší šíří záběru a detailem zpracování, potřeba konkrétní úrovně souvisí se složitostí organizace a řešeným problémem (US Federal Government Office 2007).



Obrázek 1: Hierarchie úrovní architektury Zdroj: US Federal Government Office 2007

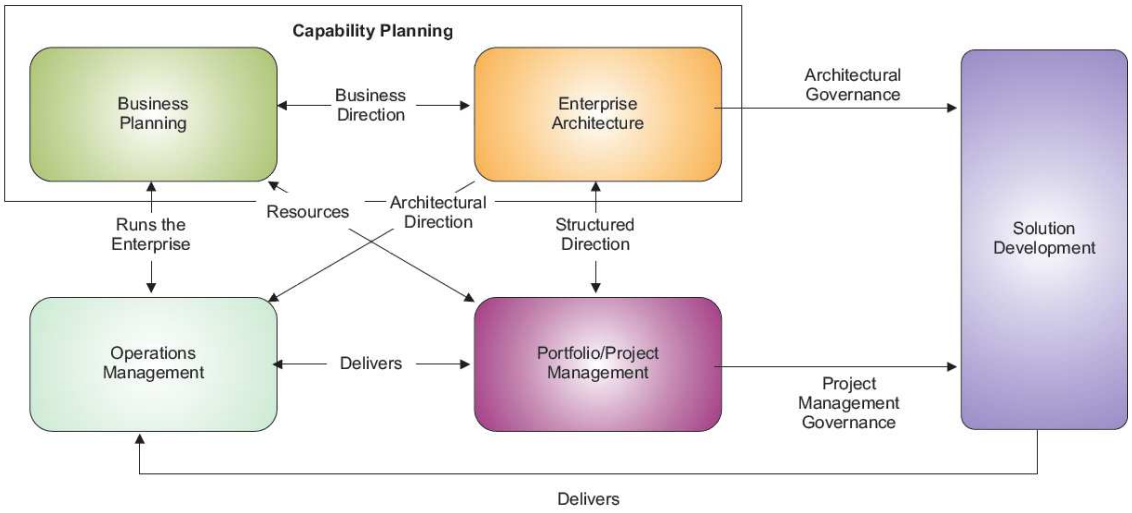
Z vlastních zkušeností se domnívám, že strategické úrovni podnikové architektury je v praxi podniků často věnována nedostatečná pozornost. Přitom strategická úroveň je kritická pro úspěšné využívání podnikové architektury (Ross 2006). Podniková architektura slouží pro orchestraci strategie s podnikovými zdroji (nejen informačních technologií)

Souvislost podnikové architektury s dalšími manažerskými disciplínami vystihuje Lankhorst 2009 (viz obrázek 2). Jako příklady konkrétních nástrojů uvádí Balanced Scorecard (oblast Strategic Management), European Foundation for Quality Management (Strategy Execution), ISO 9001 (Quality Management), COBIT (IT Governance), ITIL (IT Service Delivery and Support) a CMMI (IT Implementation).



Obrázek 2: Oblasti managementu relevantní pro podnikovou architekturu
Zdroj: Lankhorst 2009

Togaf (The Open Group 2009) odhaluje podobný pohled, ve kterém zdůrazňuje roli podnikové architektury coby partnera pro strategické plánování (angl. business planning) a dirigenta pro řízení provozu (angl. operations management) a rozvojových projektů (angl. portfolio/project management).



Obrázek 3: Návaznost podnikové architektury na další manažerské disciplíny Zdroj: The Open Group 2009

Na základě výše citovaných zdrojů a svých zkušeností odvozuji následující role podnikové architektury (podrobněji viz Hladík 2011):

- Podpora při vytváření byznysové strategie podniku, zejména identifikace optimálních řešení v souladu se strategickými cíli;
- Popis stávajícího a budoucího (požadovaného) stavu organizace, což umožňuje zajištění předchozí role; součástí je identifikace standardů, jejichž aplikace ve svém důsledku zvyšuje flexibilitu organizace a efektivitu budoucích změn;
- Definice postupů pro realizaci a využívání podnikové architektury (především návrh a kontrola architektury řešení), včetně organizace samotné podnikové architektury (např. rozvoj architektury, propagace v rámci podniku, sdílení znalostí).

Podobný pohled sdílí Hoogervorst 2004, který ve své publikaci odvozuje role podnikové architektury z funkčního (co podnik dělá) a konstrukčního pohledu (jak) na podnik. Pro detailnější zkoumání je zajímavý návrh struktury architektury v jednotlivých vrstvách (rozlišuje byznysovou, organizační, informační a technologickou).

V tomto příspěvku bych rád vyzdvihl strukturu byznysové a organizační architektury. Byznysová architektura obsahuje Misi a Strategii podniku (zejména z důvodu pochopení rozvojových aktivit podniku). Dalšími elementy jsou Trh a Konkurence, které rozvádějí Strategii z hlediska současné a požadované pozice na trhu. Následují Produkty/Služby (včetně určení míry zákaznických úprav na straně jedné, standardizace na straně druhé atp.), Klíčové zdroje (potřebné pro fungování podniku), Provozní metody (ve smyslu, jaké obchodní kanály a partnery podnik potřebuje), Ekonomický model a Zákazníci. Posledními elementy jsou Prostředí (možné vlivy prostředí na podnik) a Investoři.

Organizační architektura obsahuje obvyklé Procesy, Struktury a Systémy (identifikace klíčových prvků podniků, jejich smyslu a souvislostí). Dále sem patří Podnikové události (na které podnik pravidelně reaguje), Učení (podniku), Řízení lidských zdrojů, Chování zaměstnanců (podstatné pro klíčové oblasti podniku), Kompetence, Kultura, Řízení a Výkonnost.

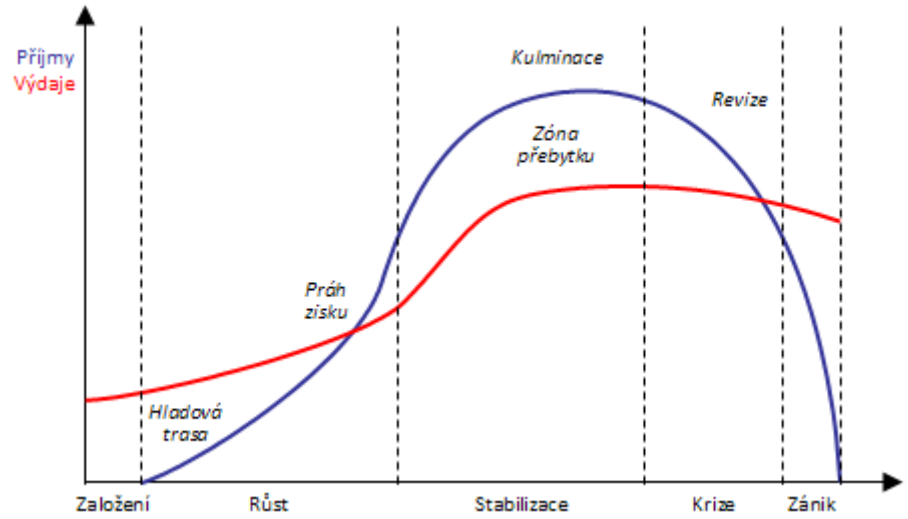
Byznysová a organizační architektura dle Hoogervorsta poskytuje návod pro celkový popis (pochopení) obchodního modelu a fungování podniku. Toto pochopení je důležité pro identifikaci příležitostí pro změnu a tedy i úkolu podnikové architektury.

Správný směr podnikové architektury udává dle většiny zdrojů tzv. vize architektury (např. van den Berg 2006, The Open Group 2009). Dle van den Berga je hlavním cílem vize stanovení a vysvětlení

smyslu a účelu podnikové architektury v návaznosti na strategii podniku, role a služeb, které podniková architektura zajišťuje, a její vnitřní organizace. Tyto zdroje potvrzují mé vnímání kritických faktorů úspěchu využití podnikové architektury.

Životní cyklus podniku a cíle podnikové architektury

Podnik prochází od svého vzniku vývojem, v jehož průběhu čelí různým výzvám. Proto i přínos podnikové architektury se nutně mění. Základní popis tzv. životního cyklu podniku poskytuje např. Synek 1999 (viz obr. 4).



Obrázek 4: Životní cyklus podniku Zdroj: Synek, 1999

Přínosem tohoto poznání je odvození různých úloh podnikové architektury v jednotlivých fázích životního cyklu podniku. Na základě Synkova výkladu a dalších autorů (např. Churchil 1983) odhaduji následující:

- Fáze Založení: Podniková architektura může pomoci navrhnout detaily obchodního modelu (angl. business model), zejména jak realizovat strategii (např. identifikace klíčových zdrojů, rizik, dopadů, doporučení fázového postupu), odhadnout náklady, čas potřebný k zahájení obchodní činnosti.
- Fáze Růst: Přínos podnikové architektury může spočívat v určení řešení pro rozšíření výrobních kapacit, produktového portfolia, zajištění kontroly nad podnikem a zákazníky.
- Fáze Stabilizace: Cílem je stabilizace vnitřního fungování podniku, získání konkurenční výhody na základě stávající pozice (např. specializace výrobků, nízké náklady díky úsporám z rozsahu, pozice značky). Podniková architektura by měla pomoci s nastavením efektivního využití podnikových zdrojů podporujících zvolenou strategii v této fázi životního cyklu („vyladit vnitřní chod podniku“).
- Fáze Krize: Pokud podnik v této fázi zvolí záchranné kroky, pak podniková architektura může pomoci v hledání vhodné, realizovatelné transformační strategie, s využitím disponibilních zdrojů (podniku i externích, například při fúzi). Role podnikové architektury spočívá také v podpoře vytvoření transformačního plánu.
- Fáze Zánik: Zde neuvažuji využití podnikové architektury.

Je zřejmé, že výše uvedený životní cyklus podniku může v praxi nabývat dalších rozměrů. Stejně jako úpadek podniku může nastat v jakékoliv fázi, lze uvažovat i o růstu – například do jiného segmentu trhu za současné stabilizace původního obchodního modelu. Nicméně se domnívám, že pochopení potřeb podniku v průběhu základního životního cyklu usnadní prvotní nasměrování podnikové architektury (resp. vize architektury).

Vliv komplexity podniku

Problematiku komplexity podniku jsem zahrnul do své práce, protože dle mého názoru by podniková architektura měla pomoci řešit především komplexní problémy ve složitých prostředích. Bez správného pochopení prostředí podniku a jeho komplexity to však není možné. Podobně to vystihuje Heylighen: „Zdá se, že porozumění komplexity je jedinou možností, jak se zachránit před evolucí, ve které se vše stává více nejisté, více komplikované a více proměnlivé“ (Schwandt 2009).

Zdroje zabývající se komplexitou podniku rozlišují její příznivé i negativní účinky na výkonnost podniku. Příčiny komplexity leží uvnitř podniku (historický vývoj) i vně (očekávání zákazníků, spolupráce s partnery, regulace). Přehledný výčet teorie komplexity podniku uvádí například Dooley 2002.

Možnosti snižování komplexity uvádějí například Siggelkow 2005 (volba organizace podniku na základě charakteru prostředí) nebo Akella 2009 (zaměřeno na IT architekturu). Hledáním příležitostí v komplexitě se naopak zabývá např. Brown 2010 (vliv komplexity produktů a portfolia zákazníků na výkonnost podniku).

Jsem názoru, že podniková architektura musí poskytnout podporu v obou směrech, tedy identifikovat příležitosti vedoucí ke snížení komplexity i využití příležitostí dané komplexitou pro zvýšení výkonnosti podniku. Klíčem je umět určit správnou a špatnou komplexitu. Studium teorie komplexity přináší inspiraci, kde tyto příležitosti hledat. Předpokládám, toto poznání také podpoří obhajobu využití podnikové architektury (což přiznávají i někteří autoři – např. van den Berg 2011).

Řízení změny

V předchozí části jsem se několikrát zmínil o změnách, kterými podniky ve svém životním cyklu procházejí. Sebelepší plán dosažení růstu však nezajistí úspěšný výsledek. To potvrzuje i lidové rčení: „Cesta do pekla je dlážděná dobrými úmysly“. Kaplan 2001 ve své publikaci uvádí, že schopnost realizovat strategii, tedy schopnost úspěšného provedení změny, je důležitější, než kvalita strategie samotné.

Z tohoto důvodu vnímám klíčovou roli pro podnikovou architekturu právě v podpoře procesu změny. V této oblasti je relevantní hledat inspiraci v projektovém/programovém řízení a řízení změny strategie (např. Kotter 2005, Buckley 1984). Na základě poznání těchto disciplín a rámců podnikové architektury se domnívám, že vzájemné propojení může přinést řadu benefitů.

Například Togaf (The Open Group 2009) poskytuje řadu návodů, počínaje procesu rozvoje architektury (Architecture Development Method, zejm. fáze E, F a G), avšak neakcentuje úplný celopodnikový pohled podobný výše uvedeným zdrojům.

Využití principů podnikové architektury v malých podnicích

Využití podnikové architektury v prostředí malých a středních podniků je další zajímavou úlohou. V kontextu výše uvedeného lze tyto podniky označit za méně komplexní (i když označení malý nebo střední podnik je relativní, neboť existuje více měřítek, navíc moderní obory a obchodní modely – např. outsourcing – mohou tato měřítka zkreslovat).

Zkoumáním charakteristického vývoje malých a středních podniků se zabývá např. Churchil 1983. Mimochodem Jihoafrická republika na tomto základě vyvinula státní program pro podporu malého a středního podnikání. Pohled na možné využití podnikové architektury, v závislosti na Churchilově poznání, včetně návrhu zjednodušené aplikace podnikové architektury v prostředí malých a středních podniků nabízí Jacobs 2011.

Odhaduji, že využití podnikové architektury v prostředí malých a středních firem bude vyžadovat vytvoření specifických, snadno použitelných a opakovatelných vzorů (z důvodu jejich aplikace méně specializovanými pracovníky – viz také Jacobs 2011 a také pro významné snížení souvisejících nákladů). Podobně tomu je u software určeným pro potřeby těchto podniků. Zajímavou úvahou je, v kontextu hledání skutečných přínosů komplexity (viz Brown 2010), zda by takové vzory nenašly uplatnění i u velkých podniků.

Závěr

Schopnost adaptovat se na změnu prostředí je zásadní nejen pro živé organismy, ale také pro podniky. Navíc se podnikatelské prostředí vlivem moderních technologií a globalizace mění stále rychleji a současně se stává složitějším.

Úspěšná adaptace podniku však není možná bez dobré znalosti podniku, jeho částí a vzájemných souvislostí (komplexita podniku). Dalším kritickým bodem je systematický a řízený postup adaptace (transformace). Na základě tohoto poznání se domnívám, že využití podnikové architektury usnadní současná aplikace vybraných pasáží z disciplín zabývajících se řízením strategie, komplexitou organizace a řízením změny.

Cílem tohoto textu je identifikovat relevantní teorie a poznatky ve všech uvedených disciplínách, které mohou pomoci při aplikaci podnikové architektury do praxe podniku. Specifický záběr je soustředěn na návrh a realizaci strategie podniku (transformace). Výsledkem práce je vedle vlastního poznání také stanovení hypotéz pro potřeby dalšího zkoumání.

Hypotéza 1: Poznáním fáze životního cyklu podniku a jeho komplexity lze definovat stereotypy, které určí základní rámec strategie podniku a také vizi podnikové architektury (resp. stereotypy určí specifickou roli podnikové architektury).

Hypotéza 2: Analýza komplexity podniku umožňuje identifikovat příležitosti pro zvýšení výkonnosti podniku. Podniková architektura poskytuje podporu při identifikace těchto příležitostí.

Hypotéza 3: Podniková architektura účinně pomáhá při transformaci podniku a to v průběhu plánování i realizace transformace.

Hypotéza 4: Malé a střední podniky potřebují zjednodušené vzory pro aplikaci podnikové architektury (z důvodu snažšího použití a nižších nákladů).

Hypotéza 5: Tyto zjednodušené vzory mohou být užitečnými standardizačními nástroji i ve velkých podnicích, zejména v těch částech podniku, které nemá smysl diversifikovat (bez přínosu konkurenční výhody).

Literatura

AKELLA, J., BUCKOW, H., REY, S.: IT architecture: Cutting costs and complexity. Časopis McKinsey on Business Technology. New Yourk, McKinsey & Company, Inc., 2009, č. 16.

BROWN, A., ELSER, B., MESSENBOECK, R., MUNNICH, F., KOMIYA, S.: Mastering Complexity – Capture the Hidden Opportunity. Firemní materiál. Boston Consulting Group, 2010.

BUCKLEY, K. W., PERKINS, D. M.: Managing the Complexity of Organizational Transformation. In Adams, J. D., *Transforming work*. Alexandria, Miles Rivers Press, 1984.

DOOLEY, K.: Organizational Complexity. In Warner, M. (ed), *International Encyclopedia of Business and Management*. London, Thomson Learning, 2002.

JACOBS, D., KOTZÉ, P., VAN DER MERWE, A., GERBER, A.: Enterprise Architecture for Small and Medium Enterprise Growth. In. Sborník konference *13th Enterprise Information Systems*. Beijing, ICEIS, 2011. ISBN: 978-3-642-21058.

HLADÍK, M.: Předpoklady úspěšné implementace podnikové architektury. Časopis Systémová integrace. Praha, ČSSI, 2011, č. 2. ISSN: 1214-6242.

HOOGERVORST, J. A. P.: Enterprise Architecture: Enabling Integration, Agility and Change. Časopis International Journal of Cooperative Information Systems. World Scientific, 2004. ISSN: 0218-8430.

CHURCHIL, N. C., LEWIS, V. L.: The Five Stages of Small Business Growth. Časopis Harward Business Review. Boston, Harward Business Review Publishing, 1983. ISSN: 0017-8012.

KAPLAN, R. S., NORTON, D. P.: Alignment – systémové vyladění organizace. Jak využít Balanced Scorecard k vytváření synergií. Praha, Management Press, 2006. ISBN: 80-7261-155-0.

KOTTER, J. P.: Vedení procesu změny. Osm kroků úspěšné transformace podniku v turbulentní ekonomice. Praha, Management Press, 2004. ISBN: 978-80-7261-015-0.

LANKHORST, M.: Enterprise Architecture at Work: Modelling, Communication and Analysis. Dordrecht, Springer, 2009. ISBN: 3642013090.

ROSS, J. W., WEILL, P., ROBERTSON, D. C.: Enterprise Architecture as Strategy. Boston, Harward Business School Press, 2006. ISBN: 1-59139-839-8.

SCHWANDT, A.: Measuring organizational complexity and its impact on organizational performance – A comprehensive conceptual model and empirical study. Doktorská práce, Berlín, 2009.

SIGGELKOW, N., RIVKIN, J. V.: Speed and Search: Designing Organizations for Turbulence and Complexity. Časopis Organization Science. Informs, 2005, č. 16. ISSN 1047-7039.

SYNEK, M.: Podniková ekonomika. Praha, C. H. Beck, 1999. ISBN 80-7179-228-4.

THE OPEN GROUP: TOGAF Version 9. Reading, [online], 2009. ISBN:978-90-8753-230-7. https://www2.opengroup.org/ogsys/jsp/publications/mainPage.jsp

US FEDERAL GOVERNMENT OFFICE: FEA Practice Guide „Value to the Mission“. Washington, [online], 2007. http://www.whitehouse.gov/sites/default/files/omb/assets/fea_docs/FEA_Practice_Guidance_Nov_2007.pdf

VAN DEN BERG, M., VAN STEENBERGEN, M.: Building an enterprise architecture practice. Dordrecht, Springer, 2006. ISBN: 1402056052.

WINTER, R., FISCHER, R.: Essential Layers, Artifacts, and Dependencies of Enterprise Architecture. In. Sborník konference 10th IEEE International Enterprise Distributed Object Computing. Hong Kong, IEEE, 2006. ISBN: 0-7695-2743-4.

JEL Classification: M10, M15, O30.

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu GAČR P 403-10-0303 (Enterprise Architecture jako princip v řízení malých a středních organizací).

Summary

Application of Enterprise Architecture, Depending on Organizational Complexity

Enterprise architecture provides supportive instruments to manage a transformation process, which an enterprise approaches due to its inner development or market changes. This text explores specific roles of the enterprise architecture in dependence on enterprise life cycle and complexity. It emphasizes strategic management and transformation process. It is noticed possible utilization of enterprise architecture principles in the are of small and medium business as well.

Keywords: enterprise architecture, strategic management, life cycle of an enterprise, transformation, organizational complexity.

Enterprise Architecture jako podpora transformace

Pavel Hrabě
pavel.hrabe@iol.cz
Doktorand oboru Aplikovaná informatika
Školitel: prof. Ing. Josef Basl, CSc., (basl@vse.cz)

Abstrakt: Článek se zabývá možnostmi využití celkové podnikové architektury (Enterprise Architecture) jako prostředku celostního poznání organizace, jako podpory pro efektivní rozhodování managementu organizace a jako metody při řízení inovací a transformace organizace.

V článku jsou představeny principy dalšího očekávaného rozvoje a změn celkové architektury, které jsou potřebné, aby mohla EA plnit pozitivní úlohu při podpoře transformace organizací. Tedy zejména aby byla celostní a důvěryhodná.

Celostního přístupu bude dosaženo dalším rozšiřování objektů zájmu EA až do míry poznání získaného podnikovou ontologií. Přitom je třeba celkový koncept všech architektu podniku rozdělit do vrstev s rozdílnou mírou detailu informací tak, aby holistická vrstva zůstala užitečná jako manažerská pomůcka a řídící vrstva pro podrobnější segmentové architektury a architektury řešení.

Článek poukazuje na zkušenosti s využitím EA pro transformaci veřejného sektoru a přináší návrh procesu místní úpravy či vývoje architektonického rámce pro EA v Česku a jeho implementace pro celkovou národní architekturu a vzorové odvětvové architektury pro segment SME. Navržený proces dělí do řady samostatně proveditelných kroků, přinášejících konkrétní výstupy a produkty, umožňující efektivní partnerství státu, výzkumu i soukromého sektoru.

Klíčová slova: Celková (Enterprise) architektura, podniková transformace, veřejný sektor, SME, manažerská metoda.

Úvod

Vymezení předmětné oblasti

Předmětem zkoumání připravované disertační práce je použití metod podnikové architektury (Enterprise Architecture) jako manažerských metod pro podporu inovace a transformace (reformy) organizací, a to jak v podnikatelské sféře, tak ve veřejné správě.

Práce je ve svých výstupech zaměřena jak na celkový pohled na podnikovou architekturu, tak zejména na oblast návrhu dobré logické aplikační architektury na podporu očekávaných business změn v českých organizacích.

Práce vychází ze zjištění, že v českém prostředí je podniková architektura využívána teprve sporadicky, a to zvláště velkými korporacemi a pobočkami zahraničních firem, zejména bank, pojišťoven, telekomunikačních operátorů a energetických společností.

Podniková architektura je zcela pominuta (vůbec se nevyskytuje) ve veřejném sektoru a velmi omezeně v segmentu malých a středních podniků. Práce si klade za cíl porozumět společným potřebám těchto segmentů vůči metodice podnikové architektury a navrhnout taková přizpůsobení, která by mohla přijetí podnikové architektury těmito organizacemi zvýšit.

V detailních doporučeních se práce zabývá zejména celkovou definicí metodiky podnikové architektury, tedy její strukturou, metamodelem a procesy jejího managementu a vývoje. Konkrétní akcelerátory v podobě referenčních modelů již byly v práci vytvořeny pro oblast Aplikační architektury a jsou plánovány pro oblast Business architektury, pro procesy a služby.

Definice cílů disertační práce

Cílem disertační práce je:

- zjistit míru využití EA u českých podniků a organizací na základě šetření prováděných v grantech VŠE KIT zaměřených na EA
- identifikovat důvody akceptace nebo neakceptace metodiky EA v těchto podnicích a organizacích, analyzovat a interpretovat tyto důvody.
- představit metodiku tvorby podnikové architektury v širokém kontextu souvisejících disciplín
- shromáždit a analyzovat mezinárodní zkušenosti s využitím EA na podporu inovací a transformace organizací, zejména v oblasti reformy veřejné správy a formulovat doporučení pro Českou republiku.

V návaznosti na to je cílem disertační práce:

- navrhnout takové úpravy a rozšíření EA rámce TOGAF, které by zvýšily jeho využitelnost jako rámce metodiky pro inovace a business transformace,
- navrhnout takové úpravy EA rámce TOGAF pro použití v ČR, které by usnadnily jeho přijetí, zejména organizacemi veřejné správy a sektorem malých a středních podniků.
- navrhnout celkovou koncepci budování EA v ČR na úrovni státu tak, aby synergicky podpořila přijetí EA v organizacích veřejné správy a sektoru malých a středních podniků.
- navrhnout část obsahu metodiky EA – referenční modely a implementační pomůcky (akcelerátory) pro usnadnění modelování logické aplikační architektury pro inovativní procesy v organizaci, směřující k dosažení strategických cílů organizace.
- ověřit účelnost a použitelnost navržených úprav metodiky TOGAF při praktických projektech u některých klientů SAP.

Současné teoretické a praktické problémy v předmětné oblasti

Problémy v oblasti využití EA pro podporu inovací v podnicích a podporu transformace (reformy) veřejné správy je možné rozčlenit do několika bloků. Jsou to zejména:

- Problémy s vnímáním IT jako zásadního prostředku pro inovace. Jeho přeceňování ve spojení s eGovernmentem a zásadní podceňování při inovaci managementu podniků.
- Problémy s měnící se definicí pojmu Enterprise Architecture a s měnícím se převážným účelem jejího použití – od standardizace IT technologické infrastruktury po business (Enterprise) transformaci.
- Problémy s nalezením správným míst a způsobů transformace, tj. problémy s porozuměním transformujícímu se systému a s řízením transformačních změn v kontextu celé organizace jako systému.
- Problém se zjednodušeným pohledem na návratnost (přínosy) investic a použití tohoto pohledu na přínosy EA, která je spíše prostředkem (enabler) dalších transformačních změn.
- Problém vztahu EA k ostatním disciplínám. Zejména jde o vymezení EA vůči EITA[1], zahrnutí BA do EA, vztah EA k BPM, Information Architecture, Security Architecture a dalším.

Enterprise Architecture jako manažerská metoda řízení transformace

Problémem využití EA jako manažerské metody řízení transformace je kombinace všech výše uvedených faktorů (zaměření pouze na IT, domnělá komplikovanost, nutnost abstrakce, absence rychlých benefitů apod.), které způsobují, že se zatím EA jako manažerská metoda neprosadila. V Česku, zejména u SME a ve veřejné správě, komplikuje situaci navíc to, že neexistuje žádná jazykově lokalizovaná a v daných sektorech již ověřená metodika EA.

Dalším problémem je, že v některých odvětvích není EA pro své silné stránky vůbec žádaná, přestože by byla celospolečensky přínosná. Jde zejména o to, že průvodním jevem zavedení EA je skokové zvýšení transparentnosti celé organizace.

[1] Enterprise IT Architecture, používáno stejně jako EwITA – Enterprise wide IT Architecture

Enterprise Architecture představuje společný jazyk, kterým se mohou business a IT snadněji domluvit. Jako u každého jazyka to ale předpokládá určitou fázi učení slovní zásoby, frází i gramatiky.

V posledních letech (2010 – 2012) se zintenzivňuje diskuse o EA jako prostředku holistického (celkového) pochopení a porozumění podniku v jeho celém kontextu a vazbách na okolí. Autor práce je jedním z propagátorů této myšlenky. Lze předpokládat, že EA se postupně stane základní manažerskou metodou pro řízení převádění strategických cílů do podnikové praxe (operativy).

Enterprise Architecture versus Enterprise IT Architecture

Mezi nejčastěji diskutované problémy patří otázky, co má být výstupem EA, zda má být EA v zodpovědnosti útvaru podnikové informatiky a komu má EA vlastně sloužit.

Výsledkem analýzy je, že záběr a úloha EA se v průběhu historie měnily a podle (BREDEMEYER a MALAN, 2004) je v organizacích za Enterprise Architecture považována architektura s různým rozsahem, od TA[2], přes EWITA až po EA ve smyslu EWITA + Business Architecture (BA).

Poslední názory, například (FEHSKENS, 2011) z Open Group, říkají, že pokud je cílem EA nalezení lepšího, s byznysem sladěného IT řešení, pak se jedná stále o EwITA nebo IT motivovanou architekturu. Přestože její součástí je i BA.

Enterprise Transformation. Nejnověji se vztah EA a podnikové transformace objevuje v sekci Open Group věnované pojmu Enterprise Transformation. Aktuální posun tohoto pojmu z oblasti systémového inženýrství do oblasti řízení podniku demonstruje jak článek (Brown, 2011), tak zejména program výroční konference Open Group 2012 v San Francisku, pro niž se transformace stala vůdčím tématem (The_Open_Group, 2012).

Návrh metod práce

Pro formulování definic, původu, poslání, přínosů a souvislostí EA s ostatními vědami a manažerskými metodami bylo využito rešerší z bohaté literatury.

Pro ověření stavu, porozumění a míry využití EA v České republice bude provedena analýza dotazníkových šetření uskutečněných v průběhu posledních let týmy VŠE v rámci grantů zaměřených na EA.

Formulování návrhů úprav architektonického rámce TOGAF, respektive návrh komponent vlastního rámce vycházejícího z TOGAF 9, je syntézou dosavadních znalostí a vlastní tvorbou autora práce.

Ověření výsledků práce bude provedeno formou případových studií v rozsahu podle toho, jak se podaří zajistit účast doktoranda na projektech, zabývajících se vybudování podnikové architektury.

Dosavadní výsledky výzkumu

Stav rozpracování disertační práce

Doktorand analyzoval problematiku EA z více úhlů pohledu. Na základě těchto analýz formuloval doktorand zobecňující prognózu dalšího vývoje EA jako disciplíny a základní principy tohoto směrování promítnul do návrhů úprav architektonického rámce TOGAF 9.

Těmito základními principy jsou:

- Podíl Business Architecture v EA se bude nadále zvyšovat, ale nedojde ke splynutí EA=BA, nýbrž pod deštníkem holistické EA se BA zařadí po bok dalších disciplín, jako architektura IS (datová a aplikační), technologická, bezpečnostní či znalostní architektura.
- Podniková architektura nebude výlučně využívat jediného architektonického stylu (například SOA), ale bude cíleně heterogenní.
- Metamodel holistické EA najde oporu v Enterprise Ontology a bude se neustále rozšiřovat a zpřesňovat.

[2] Technology Architecture (TA)

- Proces tvorby a údržby artefaktů EA se bude stále zjednodušovat s podporou akcelerátorů, jako jsou referenční modely.

Uvedené principy a všechny získané znalosti uplatnil doktorand ve vlastních návrzích pracovních výstupů disertační práce, kterými již jsou:

- Aktualizace definic a teorie EA.
- Formulace principu cílené heterogenity architektury.
- Formulace podílu EA na inovaci a transformaci.
- Návrh vrstvového modelu struktury EA.
- Návrh nového pohledu na metamodel EA.
- Návrh rozšíření metamodelu v rámci TOGAF.
- Návrh úpravy metody (cyklu) tvorby EA dle TOGAF ADM[3].
- Referenční doménový model logické aplikační architektury.
- Návrh postupu tvorby a implementace Národní celkové architektury ČR (GEA ČR), včetně vytvoření „Katalogu služeb veřejné správy ČR“.
- Návrh produktů a podnikatelských záměrů spojených s implementací a využíváním Národní celkové architektury ČR.

Některé z dosud nepublikovaných výsledků jsou představeny v následujících kapitolách.

Filosofické základy EA

Podniková architektura má usilovat o podchycení všeho, co tvoří podnik, neboť to všechno je alespoň trochu poznatelné a pro porozumění podniku důležité. Architektura naopak nemá usilovat o zachycení (poznání) všeho do posledního detailu, neboť to pro poznávajícího či vysvětlujícího není možné a pro účely architektury ani potřebné.

Často se podnik přirovnává k velice složitému systému. S trochou nadsázky je možné podnik přirovnat i ke světu ve smyslu jeho definice podle Leo Apostela (AERTS a další, 1994). Stejně tak snahu o postižení základní podnikové filosofie je možné přirovnat k hledání a osvojení si světového názoru.

Doktorand navrhuje, aby architektonické metodické rámce pro hledání odpovědí na základní světonázorové otázky spojené s podnikem do svých postupů zahrnuly myšlenky následujících filosofických nauk:

- ontologie
- etiologie
- etika a axiologie
- praxeologie a ekonomika
- noetika, epistemologie

Součástí osobnostní výbavy architektů, alespoň vedoucích osobností mezi nimi, by mělo být široké filosofické myšlení. Jenom tak se mohou stát důvěryhodnými rádci manažerů.

EA jako prostředek podpory inovací

Existence poznané, popsané, udržované a užívané Enterprise Architecture úzce souvisí s procesy inovace byznysu i IT. EA je na jedné straně podmínkou a předpokladem některých inovací, například center sdílených služeb, outsourcingu nebo v IT oblasti využití SOA či Cloud Computingu. Na druhé straně může být nedostatečně zavedená a udržovaná EA blokátorem drobných flexibilních inovací, například v rychlých úpravách prodejních a marketingových procesů a jejich softwarové podpory.

Vlastní zavedení EA pak terminologií inovací představuje inovace procesní a organizační, které významně zlepšují prostředí vedoucí k dosažení optimálního souladu mezi byznysem a ICT.

V souvislosti s inovacemi je klíčovou rolí EA identifikace příležitosti ke změně kteréhokoliv prvku systému (podniku), jenž je architekturou modelován a prověření této změny v celopodnikových sou-

[3] Architecture Development Method

vislostech. Způsob realizace (podstata) této změny je předmětem inovace, inovačního procesu s využitím vhodných známých manažerských metod, např. projektového řízení.

Government EA jako prostředek transformace veřejné správy

Pro rozvoj schopností a dovedností státu (nikoliv jenom jeho formalizovaných či elektronických služeb) je potřebné, aby byly řízeny jako architektura systému státu, s nadhledem a v celkových souvislostech. Architekturou státu se míní tzv. Cross-Government Enterprise Architecture (xGEA, GEA).

Protože většina transformačních kroků státu je stejně jako u podnikových korporací v současné době umožněna jen s pomocí IT, je celková architektura státu současně prostředkem vývoje a řízení transformačních změn a současně IT podpory těchto změn.

Národní celková architektura, angl. Cross-Governmental Enterprise Architecture (xGEA), je EA aplikovaná na stát stejným způsobem, jako je používána ve velkých podnikových korporacích. Je to tedy celková architektura napříč jednotlivými resorty, vytvořená s pomocí některého z osvědčených architektonických rámců jako jsou TOGAF, FEAF[4], MODAF[5] či Zachman, nebo rámce vlastního, vytvořeného z jejich kombinace.

Existuje celá řada přístupů státu k GEA, ale tři z nich jsou podstatné:

- použití GEA jako prostředku transformace veřejné správy, tak se uplatňuje například v Kanadě, Velké Británii, Singapuru či Korei
- použití GEA pro návrh IT podpory eGovernmentu a řízení IT státu, například v Izraeli, Austrálii, Novém Zélandu, Kataru a dalších
- absence GEA – pouze technická architektura: autorovi je z desítek zemí známa v tomto způsobu použití pouze Česká Republika

Zhodnocení trendů působících pro změnu modelu EA

Z požadavku a potenciálu, aby EA a GEA sloužily jako prostředek na podporu inovace podnikání a transformace veřejné správy vyplývá potřeba, aby jejich model a obsah představovaly skutečně úplné, holistické poznání všech typů i výskytů jsoucen v organizaci. To vede na významné rozšiřování metamodelu architektury nad rámec referenčních metamodelů známých například z architektonických rámců TOGAF nebo FEAF.

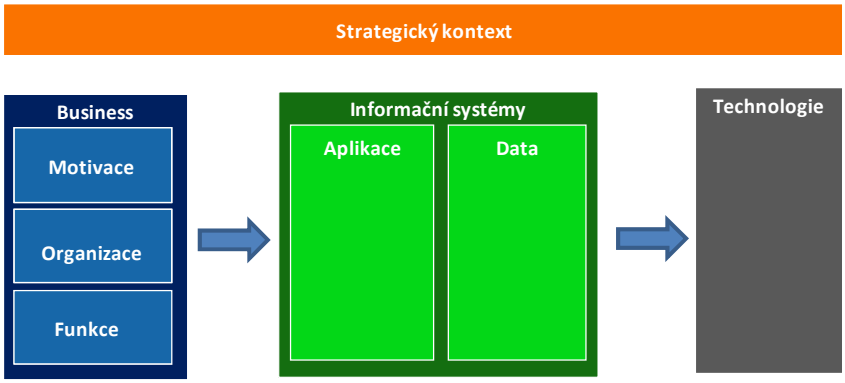
Druhým trendem, se kterým se následující návrh vyrovnává je neustálé soupeření a porovnávání se metod manažerského řízení, průběžného zlepšování a radikální transformace s přístupem EA. Základním axiomem EA (GEA) je úsilí o úplné (holistické) poznání organizace v celé její šíři, na který ostatní disciplíny obvykle rezignují a zaměřují se spíše do hloubky. To vede dále autora k představě architektury členěné na vrstvy s rozdílnou granularitou, kde některé jsou celostní a méně podrobné, další segmentové a detailní.

Třetím trendem je potřeba, zdánlivě zcela protichůdná k první. Pro specifické segmenty použití EA pro malé a střední podniky (SME) a pro agentury veřejné správy je potřebné, aby při zachování možnosti holistického poznání organizace byly všechny postupy, modely, výstupy i řízení architektury, co nejjednodušší. Z toho plyne opět potřeba vytvořit celkovou architekturu s omezenou granularitou informací (a možná i redukovaným ontologickým modelem) a segmentové architektury, které se zacílí na podrobné zachycení informací spojených s dílčími architektonickými angažmá.

Návrh vrstev struktury architektur podniku

Původní členění architektur je například podle TOGAF (The Open Group, 2009) tvořeno business, informační (datovou), aplikační a technologickou architekturu. Významným rysem tohoto rozdělení architektur je v souladu s metodikou vývoje architektury TOGAF ADM jejich postupná kaskádovitá implikace. To znamená, že změny business architektury vyvolávají potřebu změny architektur IS (datová a aplikační) a tyto jsou zadáním pro změny technologické architektury.

[4] FEAF – Federal Enterprise Architecture Framework (USA)
[5] MODAF – Ministry of Defence Architecture Framework (UK)

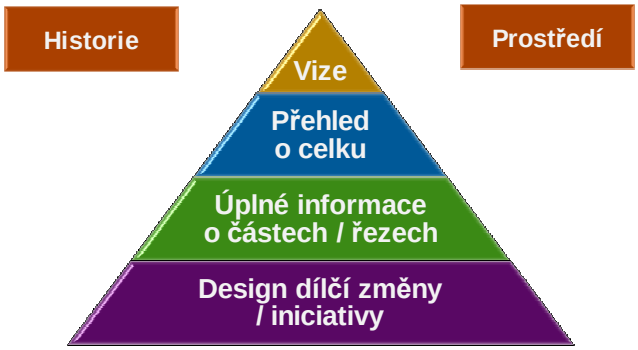


Obrázek 1: Původní implikace architektur dle TOGAF (zleva doprava)

Dokud platí výše uvedená kaskádová implikace mezi architekturami, pak se nejedná o podnikovou architekturu ve smyslu metodiky celostního popisu podniku, ale jedná se stále o účelovou metodiku směřující k co nejlepšímu návrhu informačních technologií organizace (EITA, EwITA).

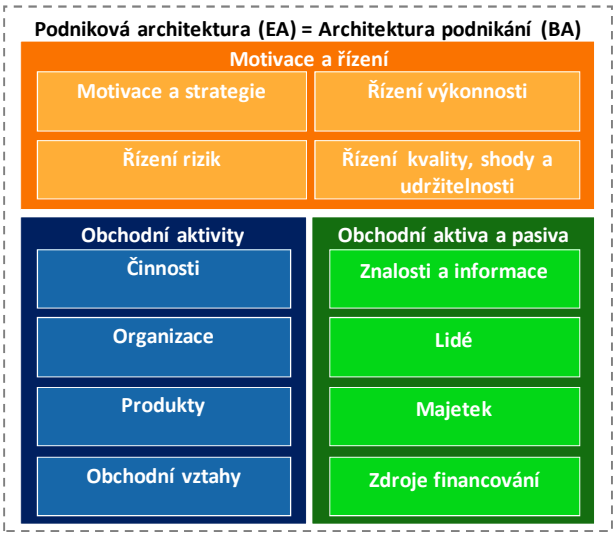
Na základě dlouhodobé znalosti a porozumění struktuře rámce TOGAF, s přihlédnutím zejména k modelům FEAF a Coherent EA byl doktorandem navržen čtyř-vrstvý model architektur podniku a základní dekompozice metamodelu jeho holistické vrstvy, nadále označované jako EA (GEA).

Celkový vrstvový model podnikových architektur, vzájemně se lišící podle míry granularity zachycených informací a šířky záběru reality podniku, se pracovně nazývá „Whole-of-Enterprise Architecture“ – WEA.



Obrázek 2: Model vrstev architektur podniku podle rozdílné granularity obsahu

Základním prvkem vrstevnatého modelu architektury je druhá, holistická vrstva, označovaná jako Enterprise Architecture.



Obrázek 3: Rozdělení druhé holistické vrstvy EA na domény

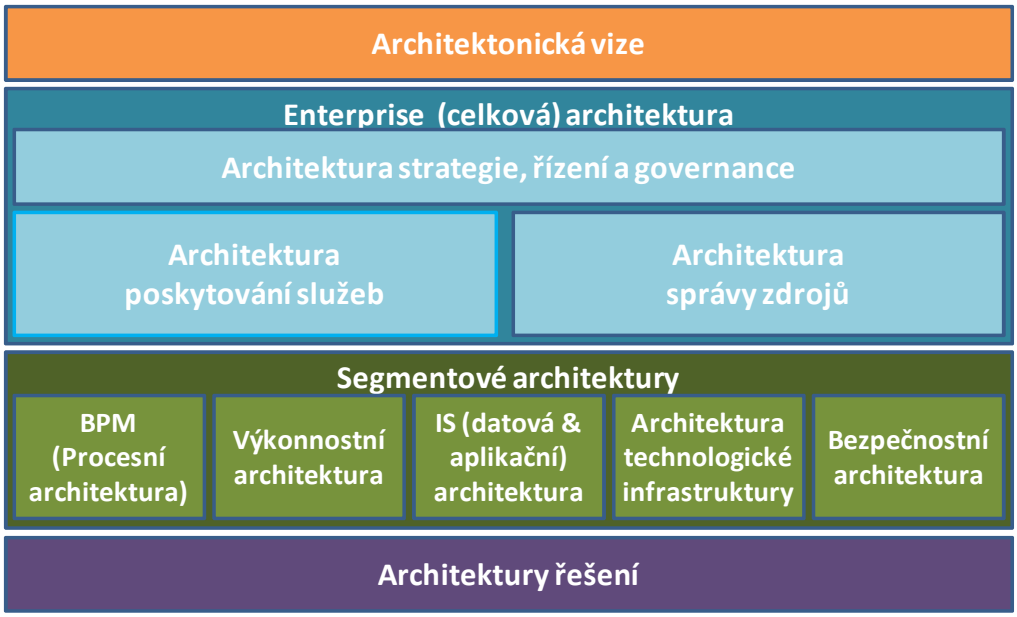
Byla navržena odvozením z Business Architecture dle TOGAF 9, ale s podstatně rozšířeným rozsahem zahrnutých objektů a s odlišnou kompozicí. Základní dekompozice této architektury spočívá v jejím rozdělení na:

- Motivační architekturu (strategie a řízení).
- Architekturu výkonu podnikání (nebo poskytování služeb veřejné správy).
- Architekturu zdrojů organizace (podniku).

Díky tomu, že jednotlivé objekty definované v této úrovni jsou dále rozpracovány v segmentových architekturách, může být koncept postupů, nástrojů, výstupů a governance této vrstvy architektury dále značně zjednodušován, ve shodě s potřebami segmentu SME a veřejné správy.

Do otevřené vrstvy segmentových architektur byly na základě dosavadních zkušeností zařazeny:

- Business Architektura
- Architektura IS (Datová a Aplikační),
- Architektura technologické infrastruktury (všechny ve smyslu definic dle TOGAF 9) a dále
- Bezpečnostní architektura (zatím bez reference)
- a Výkonnostní architektura (dle vzoru FEAF)



Obrázek 4: Návrh vrstev modelu architektury firmy, organizace (a státu)

Součástí segmentové vrstvy může být dále například také Znalostní architektura nebo Stavební architektura či Výrobní (produkční) architektura. Všechny tyto dílčí segmentové architektury jsou také celopodnikové (angl. Enterprise-wide).

Spodní vrstvu architektur podniku tvoří tzv. Architektury řešení (Solution Architecture), které odpovídají jednotlivým dílčím potřebám a příležitostem ke změně.

Architektury řešení musí vyhovovat architektonickým principům architektur z vyšších vrstev modelu. Architektury řešení jsou typicky průřezové přes více segmentů, ale nemusí být a nebývají celopodnikové.

Dopady navrhovaných změn struktury architektur

Uplatnění těchto změn vede k tomu, že se postupně sbližují Enterprise Architecture s Enterprise Ontology, kde EO je teoretickou disciplinou hledající prokazatelně správnou a úplnou formu popisu součástí objektivní existence podniku a EA je manažerská technika sloužící k využití identifikovaných informací o existenci podniku a struktuře jeho výstavby pro podporu koncepčních (transformačních) a každodenních rozhodnutí managementu.

Doména obchodních zdrojů byla navržena v podobě s širším pojetím aktiv, zejména pak znalostí (explicitních i tacitních), lidských zdrojů, majetku a zdrojů financování. Zahrnutí objektů informačních technologií mezi podniková aktiva by mohlo mít zajímavý dvojí, vzájemně protisměrný efekt:

- Zahrnutí objektů IT technologií mezi podnikové zdroje (hmotná i nehmotná aktiva) po bok patentů, licencí na jedné straně a výrobních technologií či budov na druhé straně vyvolá změnu vnímání těchto aktiv mezi IT specialisty.
- Uplatnění architektonických dovedností původně používaných pouze v IT oblasti na ostatní podnikové zdroje umožní řídit a spravovat technologická a informační aktiva jednotně. Tak vznikne například vedle IT strategie a architektury také strategie a architektura výrobních linek nebo strategie a architektura pro ekologii v podniku.

Je navrhováno, aby se posílila role business architektury, která bude na základě ontologií zpřesňovat svůj metamodel a bude vždy postihovat organizaci jako celek. Takto chápaná EA=BA bude ještě natolik přehledná a jednoduchá, aby byly použitelná i u menších podniků a státních organizací jako prostředek celostního poznání. Tato architektura je určena jako nositel **sdílené znalosti** v organizaci.

Návrh postupu vývoje a implementace Government EA pro ČR

Jako výsledek analýzy podoby i procesu implementace GEA v úspěšných zemí byl doktorandem navržen základní postup vývoje a implementace GEA v ČR na různých úrovních veřejné správy a sada očekávaných výsledků práce tzv. výstupů, produktů.

Vzhledem k vysoké míře znovu-využitelnosti prvků a nástrojů národní EA i pro segment SME, jej implementace EA na národní úrovni jedním z předpokladů jejího uplatnění i v segmentu SME.

Postup vývoje a implementace GEA v ČR.

Pro zavedení GEA do reality české veřejné správy a její udržení v chodu je potřeba vykonat celou řadu aktivit, projektů.

- Teoretická (koncepční) přípravná fáze
    - Vyhodnocení stavu GEA ve vyspělých zemích světa a zobecnění doporučení platných pro nejbližší roky pro ČR.
    - Návrh výchozího architektonického rámce ČR jako kombinace nejlepších součástí osvědčených rámců.
    - Analýza zákonných úprav spojených s GEA ve vyspělých zemích a návrh potřebných legislativních úprav v zákonech o IS státu, v kompetenčním zákoně a dalších.
    - Analýza procesu formování GEA, legislativních, organizačních a procesních změn v řízení států v souvislosti se zavedení GEA a návrh principů a postupu implementace GEA v ČR.
- Plánovací fáze
    - Vytvoření Koncepce celkové architektury ČR.
- Referenční fáze
    - Vytvoření referenčních modelů cílového stavu (tam kde bude obsah reforem znám s dostatečnou přesností) anebo alespoň současného stavu pro klíčové objekty (subjekty) architektury státu na celostátní úrovni.
    - Vytvoření výchozího modelu federalizovaných architektur, tj. stanovení oblastí zodpovědnosti za architekturu na úrovni vlády, ministerstev a jejich OSS, krajů, samospráv a jimi zřizovaných organizací.
    - Rozšíření „Slovníku nejčastěji používaných pojmů ve veřejné správě“ (Institut pro VS Praha) jeho uvedení do souladu se strukturami architektury veřejné správy. Všechny pojmy ze slovníku budou mít vyjasněný vztah k objektům architektury nebo jejich vlastnostem.
- Institucionální fáze
    - Vytvoření metodiky řízení životního cyklu GEA v jednotlivých organizacích veřejné správy.
    - Vytvoření modelu posuzování zralosti organizace veřejné správy z pohledu zavedení odpovídající části GEA do praxe řízení této organizace.

    - Vytvoření koncepce uplatnění znalostí o architekturách a míry zralosti organizací při řízení transformačních změn a pořizování odpovídajících IT technologií pro podporu těchto změn, na všech úrovních veřejné správy, vč. ustavení orgánů a institucí (Vrchní architekt státu, Architektonická rada státu, …)
- Pilotní fáze
    - Počátek naplňování obsahu GEA v rámci prvních pilotních architektonických angažmá. Například společně s produktem „Katalog služeb“, neboť katalog služeb je také jedním z klíčových výstupů (produktů, artefaktů - výtvorů).
    - Uplatnění GEA v reformě pilotního resortu nebo agentury (OSS).
    - Zobecnění pilotních zkušeností a formulace aktuálně závazných „Architektonických principů ČR“
- Dobrovolná průkopnická fáze
    - Rozšiřování využití GEA ČR pro transformaci veřejné správy v dalších resortech a OS na dobrovolné bázi.
- Fáze vyhlášení GEA ČR
    - Vyhlášení GEA ČR jako závazné metodiky s náběhem rozsahu povinností podle fází stanovených zákonem

Navrhované produkty (výstupy) výzkumu GEA ČR

Produkty ze skupiny metodiky a pravidel.

- **Architektonický rámec Czech GEA.** Návrh výchozího architektonického rámce ČR jako kombinace nejlepších součástí osvědčených rámců. Včetně výchozího návrhu metamodelu architektury státu a modelu řízení životního cyklu architektury státu. Včetně vzorových architektonických výtvorů a dokumentů.
- **Metodika řízení organizace s pomocí Czech GEA.** Návrh koncepce uplatnění znalostí o architekturách a míry zralosti organizací při řízení transformačních změn a pořizování odpovídajících IT technologií pro podporu těchto změn, na všech úrovních veřejné správy, na základě analýzy využití GEA ve vyspělých zemích.
- **Legislativa Czech GEA.** Návrh potřebných legislativních úprav v zákonech o IS státu, v kompetenčním zákoně a dalších normách na základě analýzy zákonných úprav spojených s GEA ve vyspělých zemích.
- **Národní Roadmapa Czech GEA.** Návrh strategie, principů a postupu implementace GEA v ČR na základě analýzy procesu formování GEA, legislativních, organizačních a procesních změn v řízení vyspělých států v souvislosti se zavedení GEA.
- **Národní referenční modely Czech GEA.** Návrh celostátních (federated) referenčních modelů členění (dekompozice) klíčových objektů architektury státu - funkcí státu (funkcí, služeb, procesů), aktérů (lokality, organizace, role, zodpovědnosti), zdrojů státu (aktiva, pasiva, znalosti, majetek, apod.), IS veřejné správy (IT služby, aplikační komponenty, technologické komponenty).
- **Metodika Governance Czech GEA.** Návrh metodiky řízení životního cyklu GEA v jednotlivých organizacích veřejné správy, včetně modelu posuzování zralosti organizace veřejné správy z pohledu zavedení odpovídající části GEA do praxe řízení této organizace.

Produkty z oblasti vzdělávání, certifikace a publikace

- **Czech GEA Curriculum**. Plán, obsah a dodávka kurzů pro přípravu k certifikaci Czech GEA. Produkt bude připraven ve dvou variantách, jako:
    - licencovaný obsah a školící metodika pro školící partnery Centra
    - kurzy prováděné přímo Centrem
- **Czech GEA certifikát.** Certifikační zkouška k Czech GEA
- **Czech GEA Audit Curriculum.** Plán, obsah a dodávka kurzů pro přípravu k certifikaci a akreditaci k oficiálnímu posuzování zralosti Czech GEA v organizaci.
- **Czech GEA Audit certifikát.** Certifikační zkouška pro auditora jako předpoklad pro provádění oficiálního posudku zralosti organizace z pohledu Czech GEA

- **Czech GEA Audit akreditace.** Osvědčení pro organizace poskytující auditní služby k oprávnění provádět posudek (audit) zralosti organizace podle Czech GEA.
- **Tištěné a elektronické publikace** pro:
  - Architektonický rámec Czech GEA
  - Celostátní referenční modely
  - Referenční modely vybraných částí veřejné správy a samosprávy
  - Metodiku řízení životního cyklu, včetně posuzování zralosti Czech GEA v organizaci a uplatnění znalostí GEA v procesech organizace.

Produkty ze skupiny dodávky architektonického obsahu

- **Pilotní naplnění GEA** ve vybraném resortu, OSS a vybrané části architektury.
  - Například podíl na sestavení katalogu služeb veřejné správy
  - Vytvoření katalogu logických a fyzických aplikačních komponent vybraného resortu
- **Vzory Czech GEA artefaktů.** Vypracování typových modelů GEA a vzorových architektonických výtvorů a dokumentů pro různé dílčí složky veřejné správy (ministerstva, odvětví, kraje, obce apod.)

Produkty ze skupiny SW nástrojů

- **Czech GEA portál.** Nástroj z oblasti znalostního managementu a sociálních sítí, dostupný odkazem jak portálu veřejné správy, tak z portálů jednotlivých podnikatelských komunit. Portál bude zpřístupňovat dokumenty o GEA, vzdělávání o GEA, nástroj pro tvorbu GEA a bude podporovat vzájemnou komunikaci všech zájemců o GEA.
- **Czech GEA Toolset.** Evidenční a modelovací nástroj pro uchovávání entit a artefaktů EA pro menší organizace veřejné správy a střední podniky. Aplikace by byla provozována v cloudu, případně s lokálním editorem modelů. Na vývoji aplikace by se mohly podílet partnerské technické vysoké školy.
- **Referenční databáze Czech GEA.** Dodávka referenční databáze pro hlavní SW nástroje (ARIS, Troux, Enterprise Architect, apod.). Na vývoji těchto referenčních databází by se mohli podílet komerční partneři.

Synergie mezi GEA a EA pro SME

Z působení doktoranda ve třech výzkumných týmech: EA pro inovace, EA pro SME a EA pro konkurenceschopnost veřejné správy vzniklo významné poznání podobnosti potřeb a vlastností EA mezi úřady veřejné správy a segmentem SME.

Je nepochybné, že pro podporu širokého přijetí a snadného užití musí být metodika i obsah plně lokalizována do českého jazyka i českého prostředí. Na rozdíl od velkých podnikatelských korporací, které v Česku již koncepty EA vytvářejí a užívají, obvykle v anglickém jazyce a standardním rozsahu.

Současně bohužel platí, že přinejmenším před reformou veřejné správy musí být tato metodika co nejjednodušší a nejkonkrétnější, aby byla užitečná i těm manažerům a úředníkům veřejné správy, kteří rutinně nevyužívají abstraktní systémové myšlení.

To, co bylo výše řečeno o organizacích veřejné správy, platí plnou měrou i pro většinu středních a malých podniků, i když důvod je jiný. Také tyto podniky potřebují odstranit všechny bariéry (jazyk, abstrakce, rozsah), které by jim mohly bránit dosáhnout na prospěch z užití EA.

Domníváme se tedy, že velká většina uvedených produktů z oblasti GEA bude přímo nebo po malých úpravách (například přizpůsobení obsahu referenčních modelů) přenositelná do prostředí malých a středních podniků na podporu jejich inovací a konkurenceschopnosti.

Také řada navrhovaných produktů procesu budování GEA v ČR přímo použitelná pro podporu využití EA malými a středními podniky. Zejména to bude společný znalostní portál, nástroj pro tvorbu EA, databáze referenčních modelů a vzdělávací produkty.

Závěr

Veškeré dostupné informace potvrzují, že rozhodnutí týmu VŠE KIT považovat EA za manažerskou metodu pro podporu inovací a řízení business transformace organizací bylo správné.

V návaznosti na toto rozhodnutí byly zahájeny práce na výzkumu a vývoji prostředků, které by v Česku usnadnili přijetí EA jako metody řízení v segmentech SME a veřejném sektoru.

Aby mohla EA plnit pozitivní úlohu při podpoře transformace organizací musí být celostní, holistická. To povede na potřebu změny existujících architektonických rámců, například v duchu návrhů této práce.

Lokální úprava či vývoj rámce pro EA v Česku a jeho implementace pro celkovou národní architekturu a vzorové odvětvové architektury pro segment SME bude proces trvající řadu let. V této práci navržený postup dělí celý proces do samostatně proveditelných kroků. Díky řadě produktů a služeb, které v průběhu procesu vzniknou je celý proces vhodný pro partnerství státu, výzkumu a soukromé sféry.

Plán dalšího postupu vedoucí k dokončení disertační práce.

Postup na dokončení disertace bude spočívat v provedení následujících prací:

- Analýza, vyhodnocení a návrhu úpravy vztahů metodiky EA a dosavadních metodik KIT VŠE.
- Prověření ontologických metod BORO a DEMO jako potenciálních akcelerátorů
- Dokončení úprav metodiky tvorby architektury podle TOGAF ADM pro tzv. vrstvový model architektury podniku
- Návrh hrubého referenčního modelu pro funkce (dovednosti, služby, procesy) veřejné správy jako jednoho z pilířů Katalogu služeb veřejné správy.
- Ověření správnosti navržených referenčních modelů (doménového aplikačního modelu a funkční dekompozice)
    - na EA projektech zaměstnavatele doktoranda (SAP)
    - ve spolupráci s odbory Ministerstva vnitra, Ministerstva financí a dalších
    - v další práci grantů VŠE, například při návrhu Katalogu služeb veřejné správy
- Ověření vrstevnaté dekompozice architektury organizace a metamodelu GEA, zejména diskusí v mezinárodní odborné komunitě enterprise architektů
- Ověření navrženého postupu vývoje a implementace GEA v ČR podrobnějším studiem postupů v úspěšných zemích (Singapur, Kanada, …) a výměnou zkušeností.

Literatura

AERTS, D., APOSTEL, L., DE MOOR, B., HELLEMANS, S., MAEX, E., VAN BELLE, H., a další, 1994. *World Views. From fragmentation to integration.* Načteno z CLEA Books: http://www.vub.ac.be/CLEA/pub/books/worldviews.pdf

BREDEMEYER, Dana, & MALAN, Ruth, 2004. *What It Takes to Be a Great Enterprise Architect.* Načteno z http://www.cutter.com/content/architecture/fulltext/summaries/2004/08/index.html

BROWN, Allen, 2011. *Today's projects, tomorrow's bigger picture: an examination of Enterprise Transformation.* Získáno 29. 1 2012, z The Open Group Blog: http://blog.opengroup.org/2011/10/14/todays-projects-tomorrows-bigger-picture-an-examination-of-enterprise-transformation/

FEHSKENS, Len, 2011. *Enterprise Architecture's Quest for its Identity.* Získáno 28. 10 2011, z The Open Group Blog: http://blog.opengroup.org/2011/03/10/enterprise-architecture%e2%80%99s-quest-for-its-identity/

THE OPEN GROUP, 2009. *TOGAF Version 9. The Open Group Architecture Framework (TOGAF).* ISBN:978-90-8753-230-7

THE OPEN GROUP, 2012. *Gain the tools and knowledge you need to introduce meaningful Enterprise Transformation....* Získáno 29. 1 2012, z The Open Group Conference, San Francisco, January 30 - February 3, 2012: http://www3.opengroup.org/sanfrancisco2012

JEL Classification: H11, L20, M15, O31.

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu GAČR P403-10-0303 (Enterprise Architecture jako princip v řízení malých a středních organizací).

Summary

Enterprise Architecture as a Transformation Support

The article deals with the possibilities of using the overall Enterprise Architecture as a means of holistic understanding of organizations, as support for effective decision making of organization's management and as a method of driving innovation and transformation of the organization.

The article presents the principles of further development and expected changes in the Enterprise Architecture needed to allow the EA to play a positive role in promoting and management of the transformation of organizations. Thus it is in particular the holistic approach and trustworthiness.

Holistic approach will be achieved by further expanding the range of objects which are subject of interest of the EA to the extent of knowledge acquired by enterprise ontology. By that journey overall concept of enterprise architectures should be divided into layers with different levels of detail information (granularity) so that a holistic layer with limited granularity remains useful as a managerial tool and control layer for further segmented architectures and solution architectures.

Article refers to the experience of using EA to transform the public sector and provides process design changes or local development of architectural frameworks for EA in the Czech Republic and its implementation for the total sample of national architecture and industrial architecture for SME segment. The proposed process is divided into a number of separately executable steps, bringing concrete outputs and products, enabling effective partnership law, research and private sector.

Keywords: Enterprise Architecture, Business Transformation, Public Sector, SME, Management Method.

Praktická ukázka využití standardu COBIT k hodnocení úrovně IT Governance

Dana Malcová
mald01@vse.cz
Doktorandka oboru informatika
Školitel: Prof. Ing Jiří Voříšek, CSc., (vorisek@vse.cz)

Abstrakt: Článek se věnuje tématu IT governance tak, jak je definováno organizací ISACA v rámci COBIT 5. V úvodu je nastíněn přínos přístupu governance obecně a následně je popsán přístup k IT governance dle zatím nevydané verze rámce COBIT 5. Následuje popis 5 principů, ze kterých tento rámec vychází, a popis jeho struktury. Důležitou součástí definice COBIT 5 je jeho procesní referenční model, jehož draft verze je již k dispozici. Na základě tohoto procesního referenčního modelu bylo provedeno zhodnocení pokrytí oblastí řešených rámcem COBIT 5 konkrétní směrnicí reálné firmy. Porovnání odhalilo větší zaměření na oblast procesů Managementu než na oblast vlastní Governance, tak jak jsou definovány procesním referenčním modelem COBIT 5. V oblasti managementu se směrnice dotýkala všech procesů procesní domény Sladění, plánování a organizace, následované procesní doménou Tvorba, nákup a implementace, kde se směrnice dotýkala asi poloviny procesů. Procesní domény Dodávka, servis a podpora a Monitoring, zhodnocení a vyhodnocení však nebyly příliš dotčeny a směrnice neobsahovala ani odkazy na související normy řešící tuto problematiku. Doplnění těchto chybějících vazeb tedy do budoucna pomůže zajistit větší compliance zkoumané směrnice s požadavky rámce COBIT 5.

Klíčová slova: IT Governance, COBIT, 5 principů COBIT 5, sladění IT s obchodem, celopodnikové cíle, IT cíle, enablery governance

Úvod

Přínos governance tkví v zajištění sladění cílů jednotlivých složek podniku od úrovně jednotlivých společností v případě skupiny, přes jednotlivé úrovně organizačních jednotek až po jednotlivce. Každá z těchto složek se snaží optimalizovat své chování vzhledem k pro ni definovaným ukazatelům výkonnosti, na jejichž základě je hodnocena její úspěšnost za dané období. Úkolem governance je zajistit, aby aktivita vyvíjená jednotlivými organizačními složkami podniku v každém jednotlivém případě přispívala k tvorbě celkové hodnoty podniku a k realizaci celopodnikových obchodních cílů.

Úvod do rámce COBIT 5

Oblast řešená rámcem COBIT 5 je definována jako governance pro oblast "IT". Cílem COBIT 5 tak je definovat způsoby zajištění souladu oblasti IT a jejich aktivit s celopodnikovými cíly firmy. A pomoci zajistit, že hodnota vytvořená IT bude přínosem pro tvorbu celkové přidané hodnoty tvořené firmou pro všechny zainteresované subjekty.

Plánovaná verze COBIT 5 vychází z verze COBIT 4.1, jejíž struktura byla upravena tak, aby nový rámec lépe podporoval sladění oblasti IT s obchodními cíli firmy. Toto je v reakci na reálnou poptávku trhu, kde se celosvětově stále zvyšuje požadavek top managementu na doložení přímých přínosů IT projektů a aktivit pro obchodní aktivity a cíle firem. K tomuto trendu přispívá samozřejmě i aktuální ekonomická situace, kdy je většina firem nucena v důsledku probíhající ekonomické krize optimalizovat své fungování. Tyto snahy se tedy stále více promítají i do oblasti firemního IT. COBIT 5 také zahrnuje související obsah z původně samostatně existujících frameworků jako jsou Val IT 2, Risk IT, BMIS a další). Vydání první oficiální verze COBIT 5 je plánováno na počátek roku 2012, zatím je k dispozici draft verze vydaná v červnu 2011.

K užšímu provázání IT aktivit a cílů na obchodní aktivity a cíle firmy má přispět aplikace níže jmenovaných 5 principů (v originále Principles) rámce COBIT 5:

1. Integrující rámec (Integrator Framework)

- COBIT 5 pokrývá definovanou problematiku napříč celou firmou a umožňuje integraci se souvisejícími rámci, standardy a praktikami pokrývajícími související či dílčí problematiky. Výsledkem je pak konzistentní a integrovaný celopodnikový rámec, který poskytuje netechnický, technologicky nezávislý jazyk ke komunikaci napříč pokrytými tématy.

2. Řízený hodnotou pro zainteresované osoby (Stakeholder Value Driven)

- Každá firma (ať již založená za cílem realizovat zisk či nikoliv) je realizována s cílem tvorby nějaké přidané hodnoty. Protože každý podnik má mnoho zainteresovaných osob uvnitř i vně firmy, které mohou mít i protichůdné potřeby, je třeba sladit hodnototvorný proces napříč firmou. A pro každou hodnototvornou složku podniku je třeba zhodnotit, jaké přínosy a rizika s sebou její aktivita přináší a jaké zdroje jsou zapotřebí pro její realizaci. Následně lze optimalizovat celkové přínosy, celková rizika i celkovou potřebu zdrojů na úrovni celého podniku.

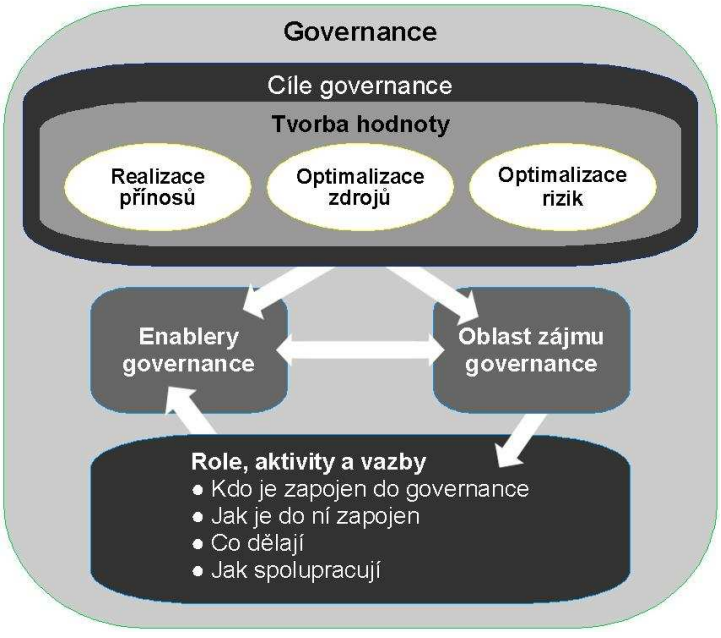
3. Zaměřený na obchod a kontext (Business and Context Focussed)

- Tento princip akcentuje potřebu navázanosti IT cílů a aktivit na obchodní cíle a aktivity firmy. IT neexistuje v podniku samo o sobě, ale slouží k dosažení celopodnikových cílů. Tento důležitý princip je třeba mít na zřeteli v každém okamžiku při návrhu a realizaci IT governance v konkrétní firmě. Zároveň je třeba při její implementaci vždy brát v úvahu jedinečný kontext každé firmy a jejího IT, kterému je třeba vlastní implementaci IT governance přizpůsobit.

4. Přístup ke governance založený na enablerech (Enabler Based Governance Approach)

- **Enablery governance** jsou představovány organizačními zdroji, vůči a díky kterým jsou namířeny jednotlivé governance aktivity směřující k dosažení cílů. Mezi enablery počítáme např. rámce, principy, struktury, procesy a praktiky, schopnosti aplikací, technologií či lidí a informace.
- **Řešenou oblast governance** je třeba vhodně vymezit před započetím návrhu a realizace governance v takto definovaném rozsahu.
- Role, aktivity a vazby jsou čtvrtým elementem governance. Takto definujeme:
  - Kdo je do governance zapojen,
  - Jakým způsobem,
  - Co dělají
  - A jak spolupracují a navzájem se ovlivňují

v rámci definovaného rozsahu systému governance.



Obrázek 1: Struktura rámce governance dle COBIT 5 Zdroj: ISACA, 2011a

5. Strukturovaný dle governance a managementu (Governance and Management Structured)

- COBIT 5 jednoznačně rozlišuje mezi aktivitami v rámci governance a aktivitami v rámci managementu. Tyto dvě disciplíny vyžadují různé typy aktivit, odlišné organizační struktury a slouží různým účelům.

(dle ISACA, 2011a, s. 9-14)

Struktura obsahu rámce COBIT 5

Obsah rámce COBIT 5 vychází z identifikace zainteresovaných osob (uvnitř i vně podniku) a identifikuje jejich základní potřeby vycházející z popisu jejich rolí.

Na základě takto identifikovaných potřeb je definována kaskáda cílů.

Nejprve jsou potřeby zainteresovaných osob provázány s jedním či více cíly governance (těmi jsou: realizace přínosů, optimalizace rizika a optimalizace zdrojů).

Cíle governance jsou následně „přeloženy do“ a mapovány na celopodnikové cíle, kteréžto jsou členěny dle přístupu BSC (= Balanced Scorecard) do čtyř dimenzí (finanční, zákaznické, interní a učení a růstu).

A následně jsou na celopodnikové cíle mapovány relevantní cíle IT, kteréžto podporují dosažení cílů celopodnikových. Cíle IT jsou opět rozděleny do čtyř dimenzí BSC.

Pro každý cíl IT jsou následně popsány potřebné enablery v jedné ze sedmi definovaných forem:

- Procesy,
- Principy a politiky,
- Organizační struktury,
- Schopnosti a kompetence,
- Kultura a chování,
- Schopnosti ve formě služeb,
- Informace.

Pro každý enabler je možno definovat související sadu cílů, která podporuje dosažení relevantního IT cíle.

Enablery spolu interagují a chovají se ve výsledku jako systém, což znamená, že je třeba k nim přistupovat jako k celku a analyzovat jejich hlavní vazby, abychom na ně mohli uplatnit governance přístupy a skutečně je řídit.

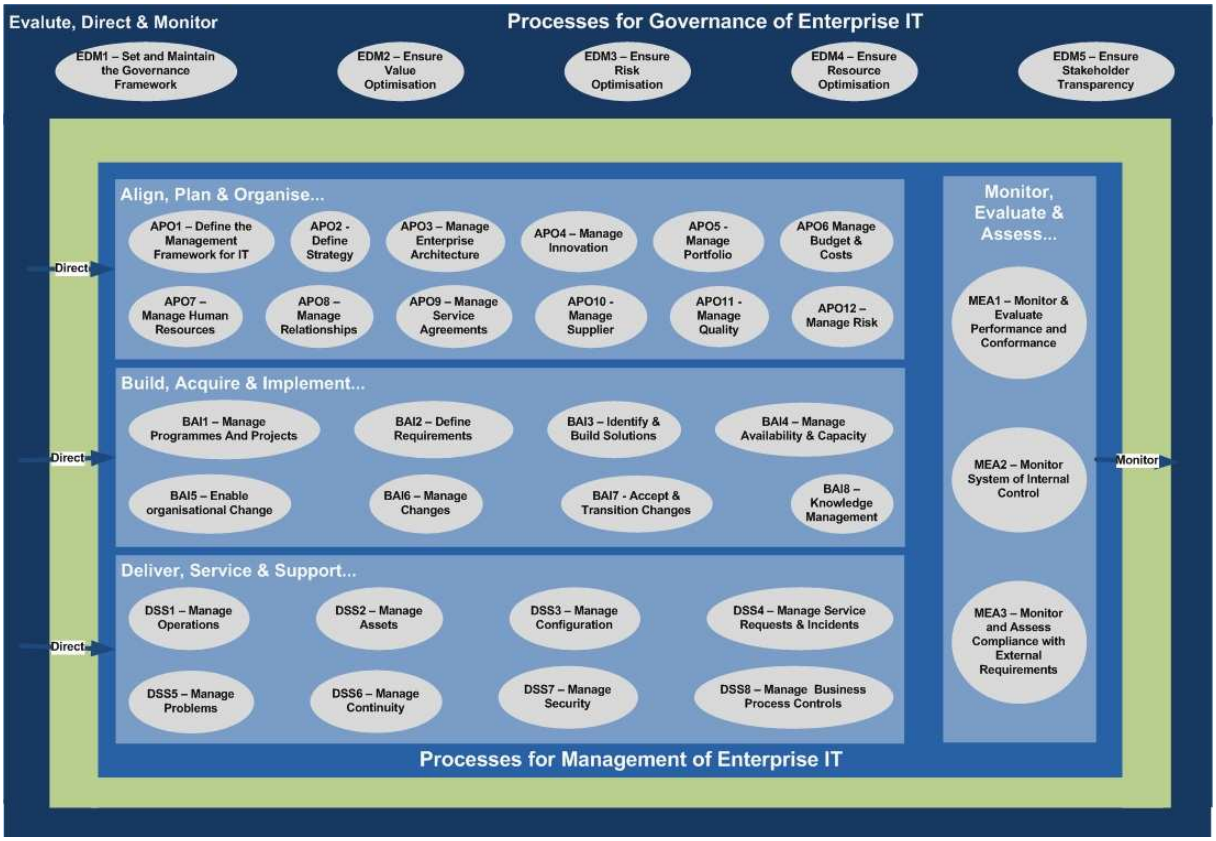
Rámec COBIT 5 zahrnuje obecný model všech enablerů, poskytující tak návod, jak s nimi standardně zacházet. (volně dle ISACA, 2011a, s. 9-14)

Využití procesního referenčního modelu COBIT 5 k rychlému zhodnocení souladu interní směrnice s tímto rámcem

V dalším textu se zaměříme na rámcem COBIT 5 definované procesy a jejich realizaci v hodnocené firmě. Pro účely tohoto zhodnocení máme k dispozici interní směrnici upravující IT governance ve zmiňované anonymní firmě. Jedná se o velkou firmu, která se skládá z několika regionálních společností a je součástí nadnárodního koncernu. Jako taková podléhá definovanému předpisu řešícímu IT governance na úrovni koncernu. Avšak je třeba implementovat tyto přístupy na národní urovni, kde je nutné reflektovat národní specifika.

Pro účely rychlého zhodnocení tohoto vnitropodnikového předpisu podle COBIT 5 se zaměříme na zhodnocení realizace jednotlivých dle COBIT 5 definovaných procesů tak, jak jsou pokryty dle popisu v uvedené interní směrnici.

Detailní popis procesů dle COBIT 5 je k dispozici v draftu dokumentu (ISACA, 2011b). Níže naleznete přehled procesů (převzato z ISACA, 2011a, s. 69).



Obrázek 2: Přehled procesů definovaných v rámci COBIT 5 a jejich členění Zdroj: ISACA, 2011b

Následující bylo zjištěno porovnáním definice procesů dle COBIT 5 a interní směrnice IT Governance (názvy procesů uvedené tučnou kurzívou jsou určitým způsobem dotčeny ve sledované směrnici):

Oblast: Governance

Procesní doména: Hodnocení, řízení a monitoring

EDM1 – Nastavení a správa rámce governance

EDM2 – Zajištění optimalizace hodnoty

EDM3 – Zajištění optimalizace rizika

EDM4 – Zajištění optimalizace zdrojů

EDM5 – Zajištění transparentnosti zainteresovaných osob

Oblast: Management

Procesní doména: Sladění, plánování a organizace

APO1 – Definice rámce řízení pro IT

APO2 – Definice strategie

APO3 – Správa podnikové architektury

APO4 – Správa inovací

APO5 – Správa portfolia

APO6 – Správa rozpočtu a nákladů

APO7 – Řízení lidských zdrojů

APO8 – Řízení vztahů

APO9 – Správa dohod o službách

APO10 – Správa dodavatelů

APO11 – Řízení kvality

APO12 – Řízení rizik

Procesní doména: Tvorba, nákup a implementace
BAI1 – Správa programů a projektů
BAI2 – Definice požadavků
BAI3 – Identifikace a tvorba řešení
BAI4 – Správa dostupnosti a kapacit
BAI5 – Umožnění organizačních změn
BAI6 – Řízení změn
BAI7 – Přijetí a přechod změn
BAI8 – Řízení znalostí

Procesní doména: Dodávka, servis a podpora
DSS1 – Řízení provozu
DSS2 – Správa aktiv
DSS3 – Správa konfigurací
DSS4 – Správa servisních požadavků a incidentů
DSS5 – Správa problémů
DSS6 – Správa kontinuity
DSS7 – Správa bezpečnosti
DSS8 – Správa kontrol podnikových procesů

Procesní doména: Monitoring, zhodnocení a vyhodnocení
MEA1 – Monitoring a vyhodnocování výkonu a vhodnosti
MEA2 – Monitoring systému interních kontrol
MEA3 – Monitoring a zhodnocení souladu s externími požadavky

Z výše uvedeného je patrné, že směrnice pokrývá spíše oblast managementu než oblast governance, kde se blíže dotýká pouze procesu EDM1 – Nastavení a správa rámce governance, díky definici zodpovědnosti za definici řídicích procesů.

V rámci oblasti managementu pokrývá směrnice, či se alespoň dotýká, prakticky všech procesů procesní domény Sladění, plánování a organizace. Zatímco v procesní doméně Tvorba, nákup a implementace směrnice neřeší ani polovinu procesů. V procesní doméně Dodávka, servis a podpora se směrnice pouze lehce dotýká procesu DSS7 – Správa bezpečnosti díky definici odpovědnosti za aktivitu definice požadavků na bezpečnost. V rámci procesní domény Monitoring, zhodnocení a vyhodnocení je směrnicí pouze částečně řešen proces MEA1 – Monitoring a vyhodnocování výkonu a vhodnosti.

Pozitivní informací je zmínka o navázání IT governance na Corporate Governance. To značí, že IT governance neexistuje ve firmě odděleně od ostatních disciplín, ale je začleněna do celopodnikového rámce řešícího problematiku Governance.

Doporučení k úpravám interní směrnice

Přestože směrnice nezmiňuje aktivity, které by přispívali k realizaci některých dle COBIT 5 definovaných procesů, je pravděpodobné, že mnohé z těchto aktivit jsou ošetřeny samostatně v rámci procesního modelu firmy. Do budoucna by však bylo vhodné, aby směrnice obsahovala alespoň odkazy na tyto související definice, snaží-li se firma o budoucí dosažení compliance s tímto rámcem. V dalším plánu je samozřejmě pro zajištění funkčnosti definovaných opatření IT governance třeba zajistit konzistenci a vhodnou návaznost všech takovýchto procesů a jejich dílčích aktivit. Směrnice IT Governance by se tak do budoucna stala jednotným místem umožňujícím přístup ke komplexním informacím z oblasti IT governance. Toto by též usnadnilo zajištění a udržování souladu s rámcem COBIT 5, případně s dalšími souvisejícími rámci, normami a předpisy. Samozřejmě, zajištění souladu firemních praktik s nějakým standardem nemůže být cílem samo o sobě. COBIT 5 však představuje koncentrovanou podobu nejlepších praktik v oblasti IT governance a jeho aplikace díky tomu představuje nástroj realizace přínosů IT governance.

Výsledky hodnocení interní směrnice v kontextu globálního stavu IT Governance

Nejnovější zpráva společnosti ISACA o globální úrovni zavedení Governance podnikového IT z roku 2011 (ISACA, 2011c), která vychází z informací poskytnutých 384 výkonnýcmi manažery z oblasti businessu a 450 vedoucími oddělení IT z 21 zemí, reprezentujících malé i velké firmy z 10 odvětví, uvádí, že dvě třetiny společností mají zavedenu nějakou formu IT Governance a jen 5 procent firem nepovažuje toto téma za důležité (ISACA, 2011c, s.24). Je tedy zřejmé, že hodnocená firma je na správné cestě k zefektivnění svého IT a tím i ke kvalitnější podpoře svého businessu. Pozitivní vliv zavedení IT governance na finanční výsledky firmy dokládá mj. studie brazilských firem z roku 2009 (Lunardi, 2009). Ziskovost sledovaných firem, které zavedli disciplínu IT governance se dle výsledků této studie pozorovatelně zvýšila v porovnání s kontrolní skupinou – v druhém roce po zavedení IT governance byl nárůst ziskovosti ještě markantnější než v roce jejího zavedení (toto je logický výsledek, uvážíme-li, že náklady vynaložené na zavedení disciplíny v prvním roce ziskovost snižují). Avšak aktuálně jsou hlavními důvody zavádění IT governance jmenovanými respondenty výzkumu pro (ISACA, 2011c) spíše způsoby zefektivnění vnitřního fungování firmy jako způsob proaktivních opatření na obranu proti dopadům ekonomické krize. Jako hlavní důvody byly jmenovány zajištění sladění aktuální funkcionality IT s aktuálními potřebami businessu, což v důsledku minimalizuje náklady na IT aktivity, které přímo nepodporují obchodní cíle firmy. V souladu s tím bylo na druhém místě jmenováno řízení nákladů. K dosažení těchto cílů má potenciál přispět i implementace IT Governance ve sledované firmě.

Literatura

ISACA, 2011a, *COBIT 5: The Framework Exposure Draft* [online]. ISACA, ©2011 [cit. 2011-01-30]. Dostupné z: http://www.isaca.org/Knowledge-Center/Research/Documents/COBIT5-Framework-ED-27June2011.pdf

ISACA, 2011b, *COBIT 5: Process Reference Guide Exposure Draft* [online]. ISACA, ©2011 [cit. 2011-01-30]. Dostupné z: http://www.isaca.org/Knowledge-Center/Research/Documents/COBIT5-Framework-ED-27June2011.pdf

ISACA, 2011c, *Global Status Report on the Governance of Enterprise IT (GEIT)—2011*. ISACA, ©2011 [cit. 2011-01-30]. Dostupné z: http://www.isaca.org/Knowledge-Center/Research/Documents/Global-Status-Report-GEIT-10Jan2011-Research.pdf

Lunardi, Guilherme Lerch, João Luiz Becker a Antonio Carlos G. Maçada, 2009, *The Financial Impact of IT Governance Mechanisms' Adoption: an Empirical Analysis with Brazilian Firms* In: Proceedings of the 42nd Hawaii International Conference on System Sciences – 2009 [cit. 2011-01-30]. Dostupné z: http://www.ea.ufrgs.br/professores/acgmacada/pubs/Artigo%20Hawaii%20final%202009.pdf

JEL Classification: M15

Summary

Practical example of using the COBIT framework for the assessment of level of IT Governance

This paper examines the topic of IT governance, as defined by the ISACA organization in the COBIT 5 framework. Benefits of the generic governance approach mentioned in the introduction are followed by a description of the IT governance as defined in currently available draft version of the COBIT 5 framework. The 5 Principles which represent the basis for approach defined in this framework and the framework's structure are also described. An integral part of the framework is represented by its Process Reference Model, which is also available in a draft version. An evaluation of the level of COBIT 5 processes covered in one anonymous real company's internal guideline was performed based on this Process Reference Framework. Comparison of those materials showed that the internal guideline was mainly focused on the Management area compared to the Governance area, as they are defined in the COBIT 5 Process Reference Framework. The guideline touched in some way each process in the Align, plan and Organise process domain, followed by almost one half of processes being to certain level solved for the Build, Acquire and Implement process domain. Whereas the

process domains Deliver, Service and Support and Monitor, Evaluate and Assess where not much covered and the guideline did also not contain references to other internal guidelines covering these topics. Adding these references to the IT governance guideline will help to ensure higher level of compliance with the COBIT 5 framework.

Keywords: IT Governance, COBIT, 5 Principles of COBIT 5, Business-IT alignment, business goals, IT goals, governance enablers.

Formalizace nových poznatků ve strukturovaném přístupu k podpoře a údržbě software - kompetenční a sub-kompeteční centra

Petr Marounek
Petr_marounek@centrum.cz
Doktorand oboru informatika
Školitel: doc. Ing. Prokop Toman, CSc., (toman@vse.cz)

Abstrakt: Efektivní realizace (z pohledu času, nákladu, využití lidských zdrojů atd.) provozu informačních systémů je strategickou záležitostí. Náklady spjaté s údržbou, provozem software a řízením jeho rozvoje nyní reprezentují více jak 90% jeho celkových nákladů.

Software maintenance je množina aktivit potřebných pro nákladově efektivní podporu IT řešení. Kompetenční centra IS/IT nepokrývají oblast podpory a údržby software, čili neexistuje formalizovaná metodika nebo procesní rámec pro kompetenční centrum podpory a údržby – tj. chybí předpis, jak takové centrum vytvořit, řídit, vyhodnocovat, jakou má mít organizační strukturu, jaké má poskytovat služby a výstupy, a jak kontinuálně toto centrum zlepšovat. Autor navrhuje vlastní řešení v podobě definice a implementace kompetečního centra podpory a údržby, a sub-kompetenčních center podpory a údržby dílčích aplikací.

Dle autora aktuální ontologie podpory a údržby nezachycuje potřebné komponenty a vazby - zejména zde chybí pohled řídící, pohled plánovací a odhadování úloh. Proto autor navrhuje redefinici a rozšíření ontologie organizační struktury o prvky kompetenčního a sub-kompetenčního centra, typologii úloh (management, maintenance) a jejich řízení tj. odhadování, plánování a relizaci. Na závěr autor diskutuje přínosy implementace kompetenčního a sub-kompetenčího centra do podpory a údržby software.

Klíčová slova: podpora a údržba software, metodika, model, ontologie, organizační struktura, kompetenční centrum, sub-kompetenční centrum, pracnost, plánování.

Úvod

Vzrůstající komplexita požadavků uživatelů a penetrace ICT technologií do života každého jedince, vyžadují stále rozsáhlejší, komplexnější řešení, která jsou dodávána pomocí projektů. Tyto projekty již nejsou doménou několika osamělých jedinců, kteří se rozhodují ad hoc nebo pouze na základě vlastní zkušenosti. Pro zaručení úspěchu jsou projekty realizovány pomocí ucelené množiny pravidel, metodiky.

Metodiky, které jsou dnes používány ve světě vývoje software, zahrnují pravidla, doporučení a návody, jak dodat jeden izolovaný projekt – jeden informační systém od formulace a analýzy požadavků zákazníka až po jejich realizaci a nasazení do provozu.

Z pohledu globální ekonomiky, ekonomiky podniku a ceny produktu, tvoří provoz, podpora a údržba zásadní podíl na nákladech spjatých s pořízením a provozem řešení. „Přestože jsou uváděna různá čísla, mnoho studií [1, 2, 7, 10, 34, 46, 54, 56, 58] indikuje, že software maintenance spotřebovává 60% až 80% z celkových nákladů během životního cyklu software, výzkumy dále ukazují, že náklady na podporu jsou vysoké zejména díky požadavkům na úpravy/vylepšení (často 75–80%), než na opravu chyb.“ (Canfora, 2000). Náklady spjaté s údržbou, provozem software a řízením jeho rozvoje nyní reprezentují více jak 90% jeho celkových nákladů.“ (Koskinen, 2010).

Autor se dlouhodobě zaměřuje na problematiku podpory a údržby ICT řešení a to z pohledu organizačního, metodického a řídícího, přičemž identifikoval následující témata k rozvoji:

- **Metodiky, modely a procesní rámce** - nepokrývají řídící a plánovací pohled, organizační strukturu, odhadování pracnosti, plánování, alokace zdrojů.
- **Modely pro odhadování pracnosti podpory a údržby** - jsou velmi komplexní a nevhodné pro odhad a nacenění několikadenních úloh.

- **Ontologie podpory a údržby** - nepokrývá řídící a plánovací pohled jednotlivých útvarů ani úloh. Subontologie nezahrnuje možnost správy více aplikací/řešení atd. Dále nepokrývá odhadování pracnosti, ukládání reality do knowledge base a následnou úpravu odhadů o historickou zkušenost.

V tomto článku se autor detailněji zabývá implementací konceptu kompetenčních center jako prostředku pro zavedení organizační struktury a organizace práce do podpory a údržby software.

Podpora a údržba IS

Posláním „podpory, údržby a provozu“ je udržovat a rozvíjet aplikaci v produkčním prostředí a také podporovat koncové uživatele, kteří ji používají. Zpravidla se jedná zejména o opravy chyb, rozvoj a modifikace funkcionalit, dat, optimalizace výkonu atd.

Potřebou těchto úprav se zabýval Lehman, který definoval tzv. Lehmanovy zákony. „Pokud řešení, které je používáno ve skutečném světě, není aktualizováno dle potřeb, poměrně rychle zastará. Významné změny se také objevují díky integraci řešení s externími entitami, jako jsou lidé, organizace a systémy. Ve skutečnosti, software je možno nekonečně upravovat, a proto je často vnímán jako nejsnažší část změny systému.“ (Brooks, 1987).

„Po dodání systému nastává podpora a údržba software, což je proces modifikace systému a jeho komponent, s cílem opravit chyby, zlepšit výkon nebo jiné atributy, a nebo jej pozměnit tak, aby i nadále vyhoval změněnému prostředí.” (IEEE, 1990).

„Mnoho autorů (např. Schneidewind, Osborne and Chikofsky, Pigoski atd.) nesouhlasí s tímto pohledem, tj. že provoz, údržba a podpora software jako pouze post delivery aktivita a tvrdí, že počátky software maintenance mají začínat před tím, než je řešení nasazeno do produkce. Pigoski [62] zachytil potřebu zahájit provoz, údržbu a podporu software již na začátku developmentu v nové definici „Software maintenance je souhrn aktivit potřebných pro nákladově efektivní podporu IT řešení. Tyto aktivity jsou realizovány v rámci pre-delivery stejně tak i post-delivery fázích. Pre-delivery aktivity zahrnují plánování postdelivery provozu, podporovatelnost a logistická rozhodnutí. Post-delivery aktivity zahrnují úpravy software, školení a provozování help desku“.

Tato definice je konzistentní s přístupem k provozu, údržbě a podpoře dle ISO (International Organization for Standardization, Mezinárodní organizace pro normalizaci) a jejích standardů ohledně procesů životního cyklu software [44]. To definitivně bortí obrázek, že provoz, údržba a podpora software je jen a pouze o opravách různých chyb. (Canfora, 2000).

Autor se plně ztotožňuje s pohledem, kdy provozu, údržby a podpory software začínají již hluboko v dodávce řešení. Jednou z best practises IT odvětví je podílení se budoucích pracovníkú podpory a provozu na projektu dodávky, neboť to usnadňuje následný knowledge transfer a hladší průběh náběhu do produkčního prostředí. Metodiky dodávky SW specifikují, jaké všechny výstupy pro potřeby provozu mají vzniknout. Například řešení má být navrženo s ohledem na nefunkční požadavky, přičemž jeden z nich je i Supportability neboli podporovatelnost (v produkčním prostředí).

Modely, metodiky a procesní rámce podpory a údržby software

- Organizace IEEE, Institute of Electrical and Electronics Engineers, (česky „Institut pro elektrotechnické a elektronické inženýrství“).
    - IEEE-1219 Proces podpory a údržby.
    - SWEBOOK, tzv. Body of knowledge, které pokrývá softwarové inženýrství. Kapitola 6 je věnována podpoře a údržbě.
- ISO (International Organization for Standardization, „Mezinárodní organizace pro normalizaci“).
    - ISO/IEC 12207 (2008) Systems and software engineering – Software life cycle processes.
    - ISO/IEC 14764 Software Engineering – Software Maintenance.
- SEI, Carnegie Mellon University of Pittsburgh.
    - Software Maintenance Model (SMmm).

- ISACA – dříve známá jako Information Systems Audit and Control Association.
  - COBIT.
- Stockholmská univerzita, Švédsko.
  - Cm3 – Corrective Maintenance Model. Autorkou modelu je Mira Kajko-Mattsson a její kolegové ze stockholmské univerzity ve Švédsku.
- Spolupráce pracovníků z holandského Software Engineering Research Centre a Division of Mathematics and Computer Science, Faculty of Sciences na univerzitě v Amsterdamu.
  - IT service Information Technology Service Capability Maturity Model, dále jen IT Service CMM.
- Hlavní best practises.
  - ITIL Service Suppport.
  - ITIL Service Delivery.
  - Malcolm Baldrige.
  - Rational Unified Process rozšířený o část podpory a provozu.

Implementace kompetenčních center do podpory a údržby

Kompetenční centra v IT

Posláním kompetenčních center je sdružit a rozvíjet unikátní znalosti, dovednosti vybraných pracovníků ve vybrané oblasti a pomocí definovaných procesů a organizační struktury poskytovat služby s vysokou přidanou hodnotou.

„CoE je centralizovaný soubor znalostí a experní zkušenosti v určité oblasti, funkci nebo technologii, který využívá strukturovaný soubor procesů, postupů a činností, které podporují vysokou úroveň výkonnosti. Na podporu tohoto přístupu je COE obvykle obsazeno odborníky ve zvoleném oboru, kteří podporují spolupráci a aplikaci znalostí, technik, nástrojů a procesů.“ (Williams, 2011).

„Kompetenční centrum reprezentuje vybranou organizační entitu, která se skládá z vybraných procesů, úloh a „mozkového trustu”. Kompetenční centrum představuje vyšší úroveň synergie ve srovnání s tradiční organizační strukturou.“ (Allen, 2004).

„COE umožňuje fyzickou nebo virtuální reorganizaci lidí, procesů, které podporují, a dokonce i technologie, kterou používají, v jeden propojený celek.“ (Meta Group, 2004).

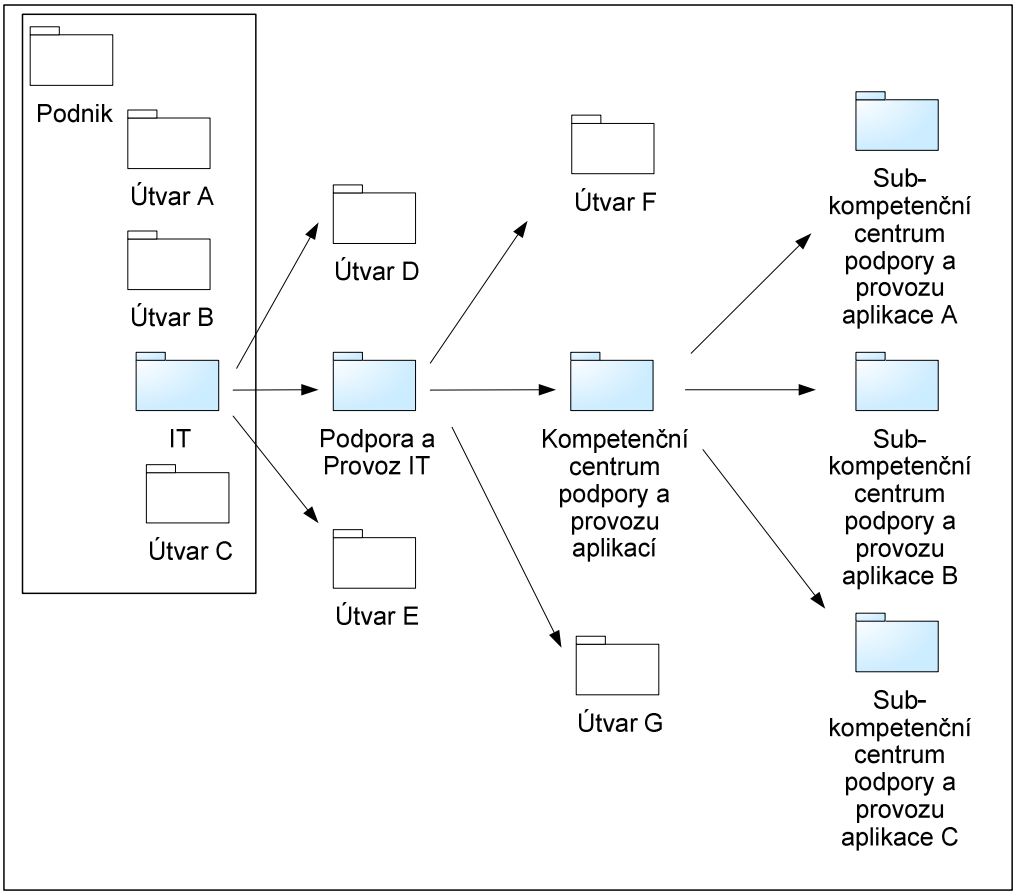
Přes různé pojeti kompetenčních center v oblasti IS/IT, tato kompetenční centra nepokrývají oblast podpory a údržby software – tj. chybí předpis, jak takové centrum vytvořit, řídit, vyhodnocovat, jakou má mít organizační strukturu, jaké má poskytovat služby a výstupy a jak kontinuálně toto centrum zlepšovat.

Definice kompetenčního centra podpory a údržby IS/IT

Kompetenční centrum podpory a údržby je organizační útvar, který využívá strukturovaný soubor procesů, postupů a činností s cílem co nejefektivnější dodávky služeb v oblasti podpory a údržby software. Na podporu tohoto přístupu je COE obvykle obsazeno odborníky ve zvoleném oboru, kteří podporují spolupráci a aplikaci znalostí, technik, nástrojů a procesů.

Ani jedna z libovolných definic kompetenčního centra, ani nově autorem definované kompetenční centrum podpory a údržby nepokrývají problematiku podpory, údržby a provozu vybrané aplikace nebo systému. Pro tuto problematiku se chybně v laické veřejnosti používá také pojem kompetenční centrum. Chybnost je v definici kruhem, neboť dle této logiky je kompetenční centrum složeno z dalších kompetenčních center – čemuž tak ani v teorii ani v praxi není a nemůže být.

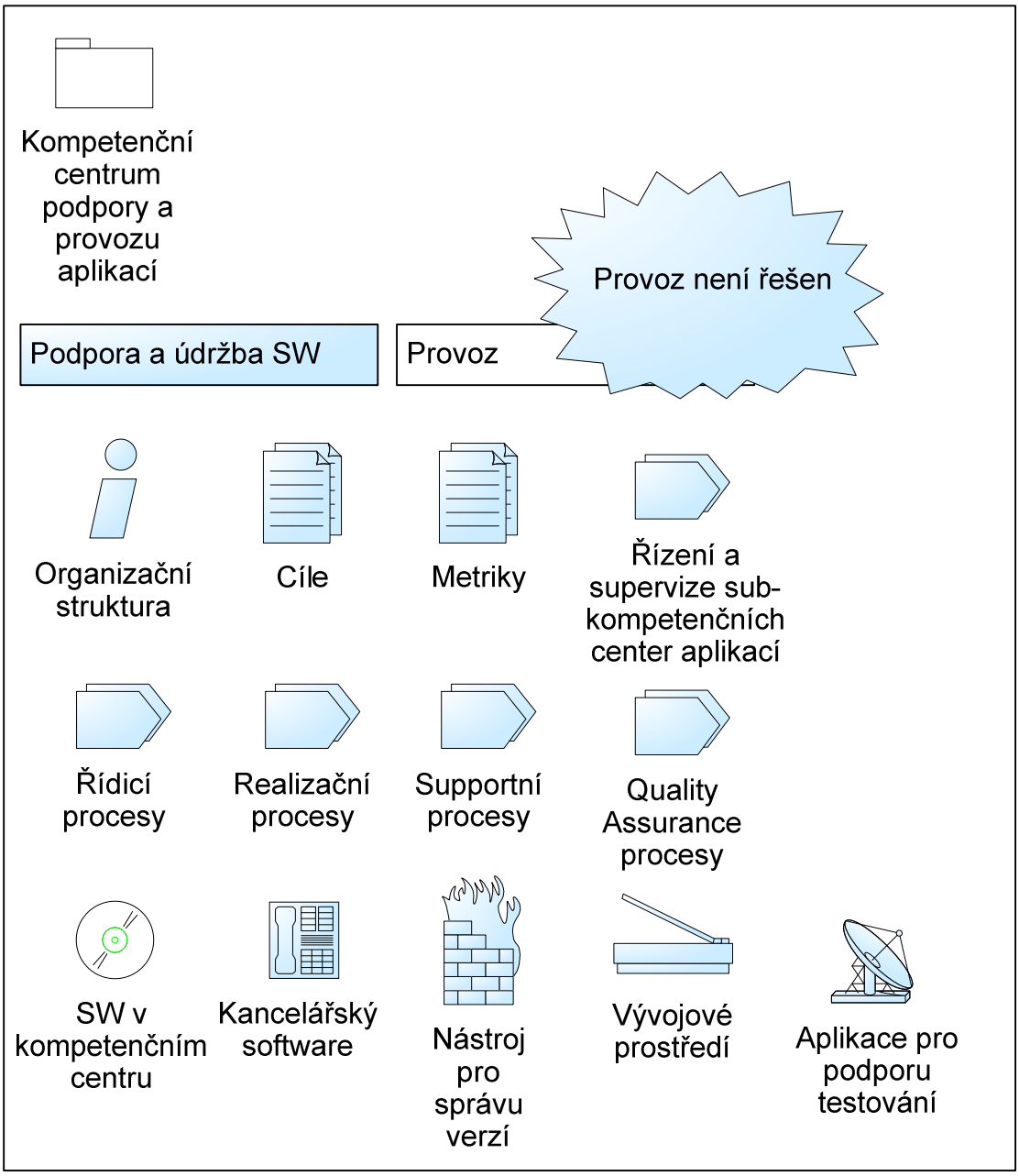
Z tohoto důvodu autor zavádí pojem sub-kompetenční centrum jako podřízené organizační jednotky, která se specializuje na dodávku právě jednoho řešení.



Obrázek 1: Vztah mezi KC a SUB-KC v rámci organizační struktury podniku Zdroj: autor

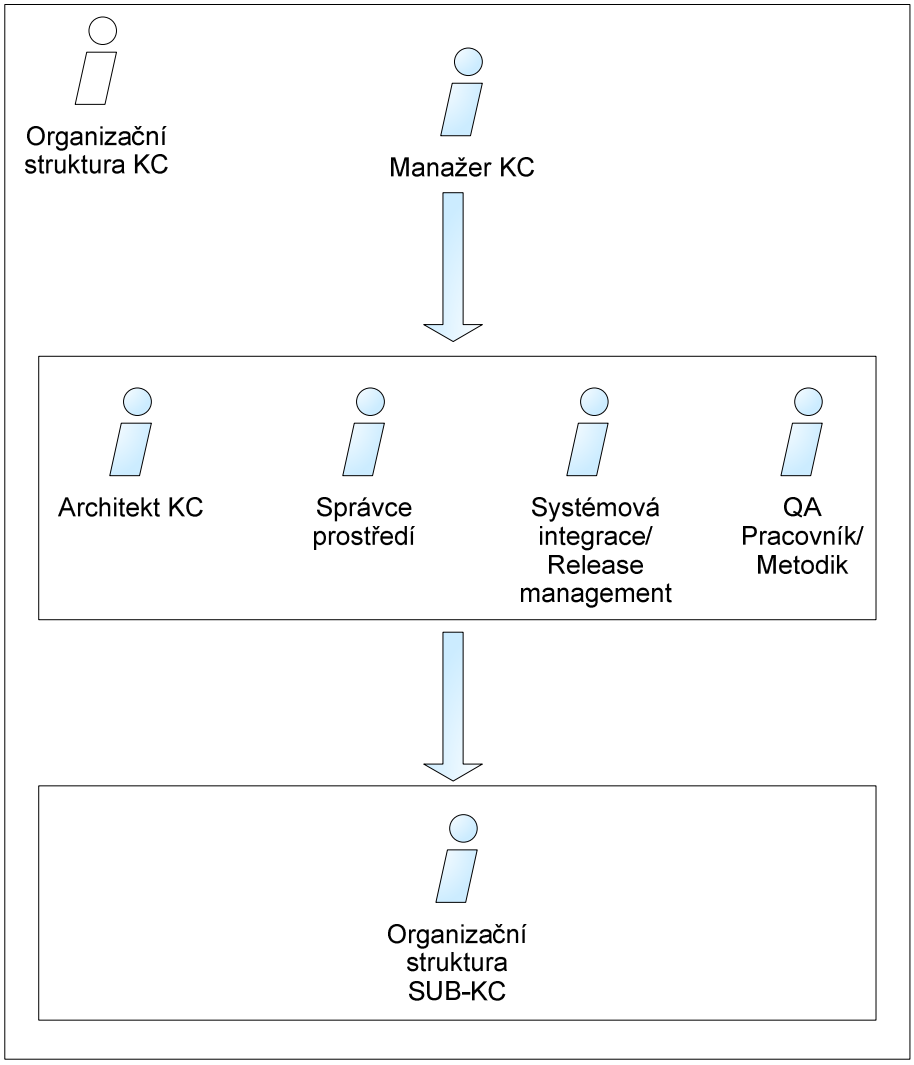
Kompetenční centrum podpory a údržby tedy zabezpečuje podporu a údržbu vybraných řešení v organizaci formou sub-kompetenčních center. Kompetenční centrum podpory a údržby lze dekomponovat následujícím způsobem:

- Organizační struktura (viz dále).
- Cíle KC – nákladově efektivní podpora a údržba vybraných aplikací a jejich koncových uživatelů.
- Metriky – technicko-ekonomické ukazatele nákladovosti a výnosovosti.
- Řízení a supervize sub-kompetenčních center – stanovení dílčích cílů, řízení a podpora dosahování těchto cílů.
- Metodika procesů – tj. řídící, realizační, podpůrné a QA.
- Vybraný software, který je používán napříč sub-kompetenčními centry – kancelářský software, plánovací, různá vývojová, testovací, akceptační prostředí, nástroje pro správu verzí atd.



Obrázek 2: Dekompozice KC pro podporu a provoz aplikací Zdroj: autor

Organizační struktura kompetenčního centra podpory a údržby zahrnuje vedení, architektonickou autoritu, správce sdílených prostředí, systémovou intergaci, release management a dohled kvality.



Obrázek 3: Organizační struktura KC pro podporu a provoz aplikací Zdroj: autor

Definice sub-kompetenčního centra podpory a údržby IS/IT

Sub-kompetenční centrum podpory a údržby autor definuje následujícím způsobem:

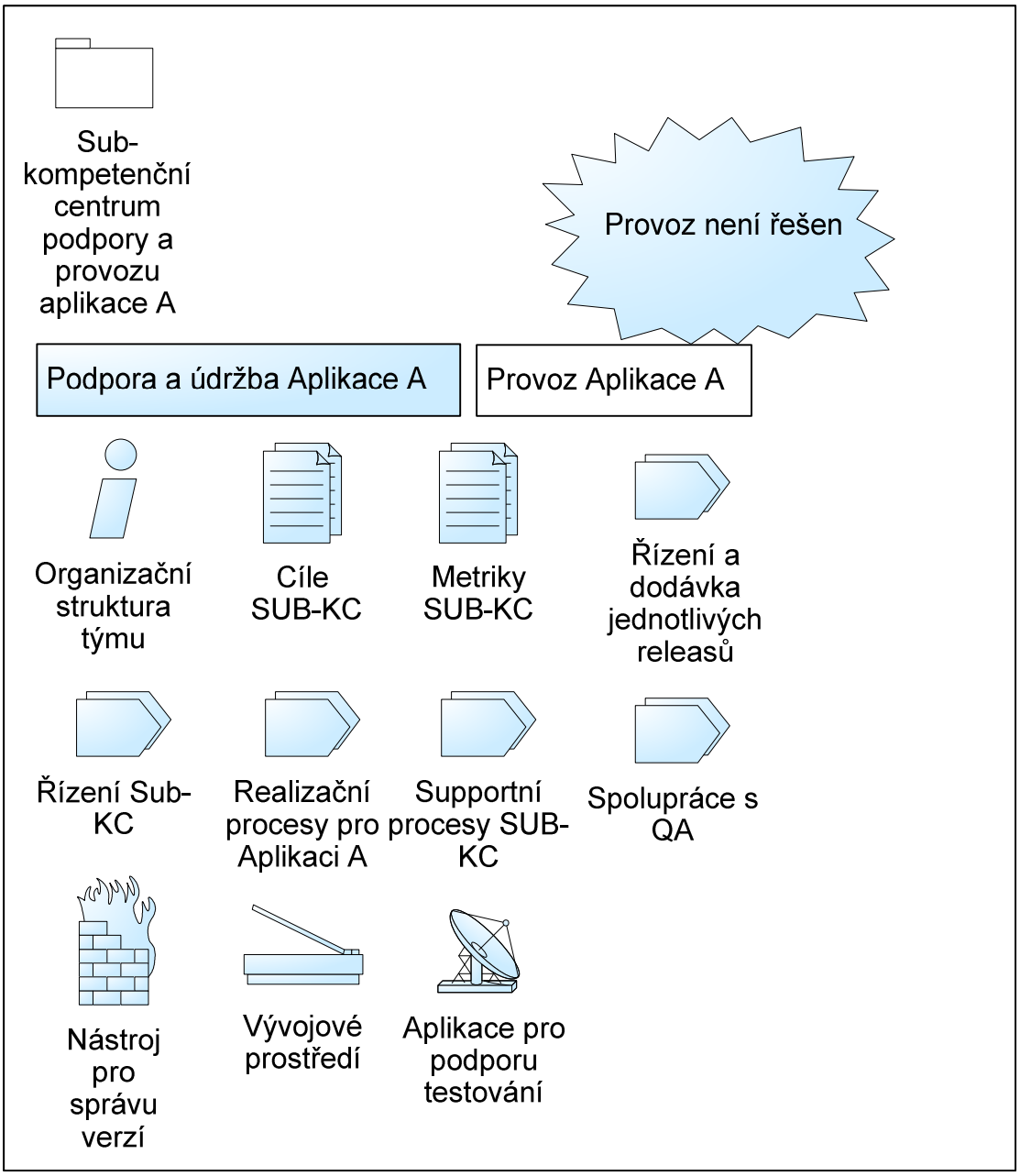
SUB-Kompetenční centrum podpory a údržby je organizační útvar, který využívá strukturovaný soubor procesů, postupů a činností s cílem co nejefektivnější dodávky služeb v rámci vybrané aplikace nebo systému, nebo logicky souvisejících řešení (např. Šablony, tisky atd.). Na podporu tohoto přístupu je SUB-COE obvykle obsazeno odborníky ve zvoleném oboru, kteří podporují spolupráci a aplikaci znalostí, technik, nástrojů a procesů.

Z kompetenčního centra přepoužívá následující oblasti:

- Převzaté cíle.
- Metodiku řízení a vytváření výstupů.
- Metriky a Reporting.
- Systémovou integraci a plán Releasů.
- QA.

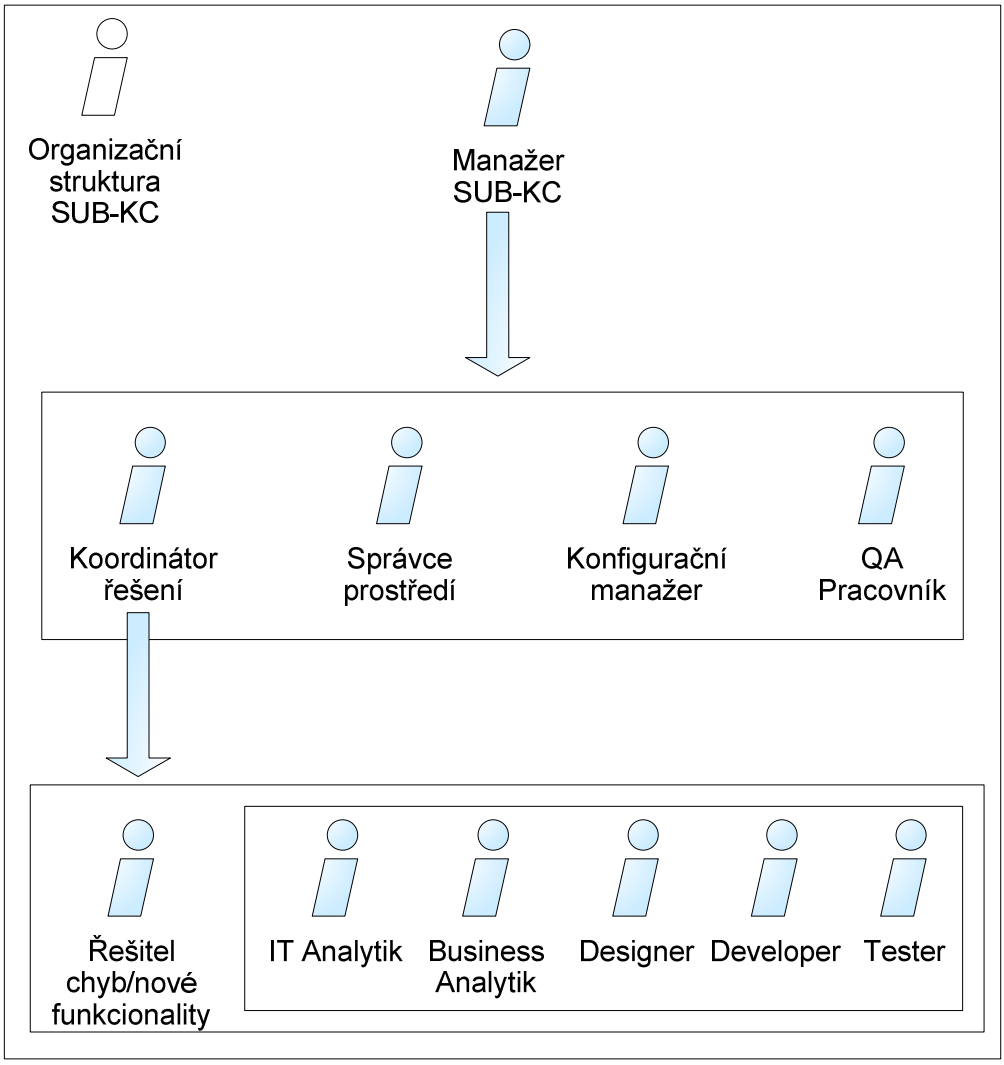
Pro své potřeby subkompetenční centrum definuje zejména tyto oblasti:

- Plán a řízení dodávek jednotlivých releasů (nebo oprav).
- Vybrané specifické procesy pro řízení, dodávku, podpory a QA dle podporovaného řešení a technologie (různé pro SAP vs. J2EE atd.).
- Specifické nástroje a prostředí opět dle řešení a technologie (různé pro SAP vs. J2EE atd.).



Obrázek 4: Dekompozice SUB-KC pro podporu a provoz dílčí aplikace Zdroj: autor

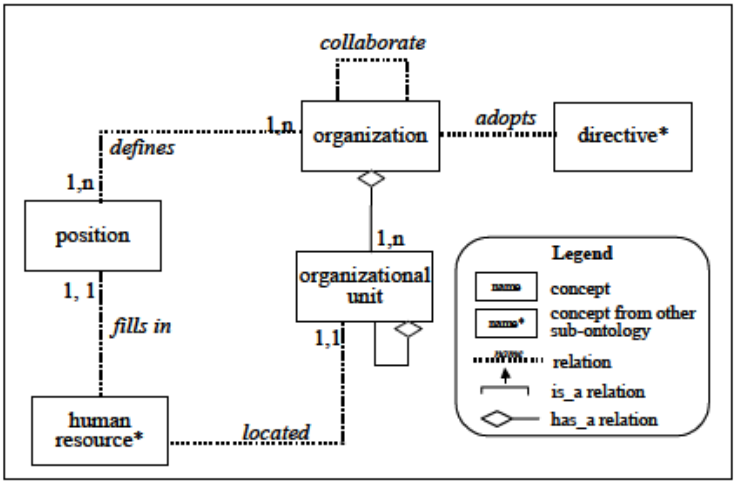
Organizační struktura kompetenčního centra podpory a údržby vybrané aplikace zahrnuje vedení sub-KC, koordinátora vlastního řešení, správce prostředí, konfiguračního manažera a dohled kvality. V oblasti vlastních dodávek oprav chyb nebo nových funckionalit je definován řešitel a poté typické role převzaté z vývoje software počínaje analytikem a konče testerem. Navržená organizační struktura může být libovolně měněna, tj. rozšiřována o další role, nebo naopak vybrané role mohou být sloučeny do jedné, pokud je podporované řešení malé (např. Manažer KC a koordinátor řešení).



Obrázek 5: Dekompozice SUB-KC pro podporu a provoz dílčí aplikace Zdroj: autor

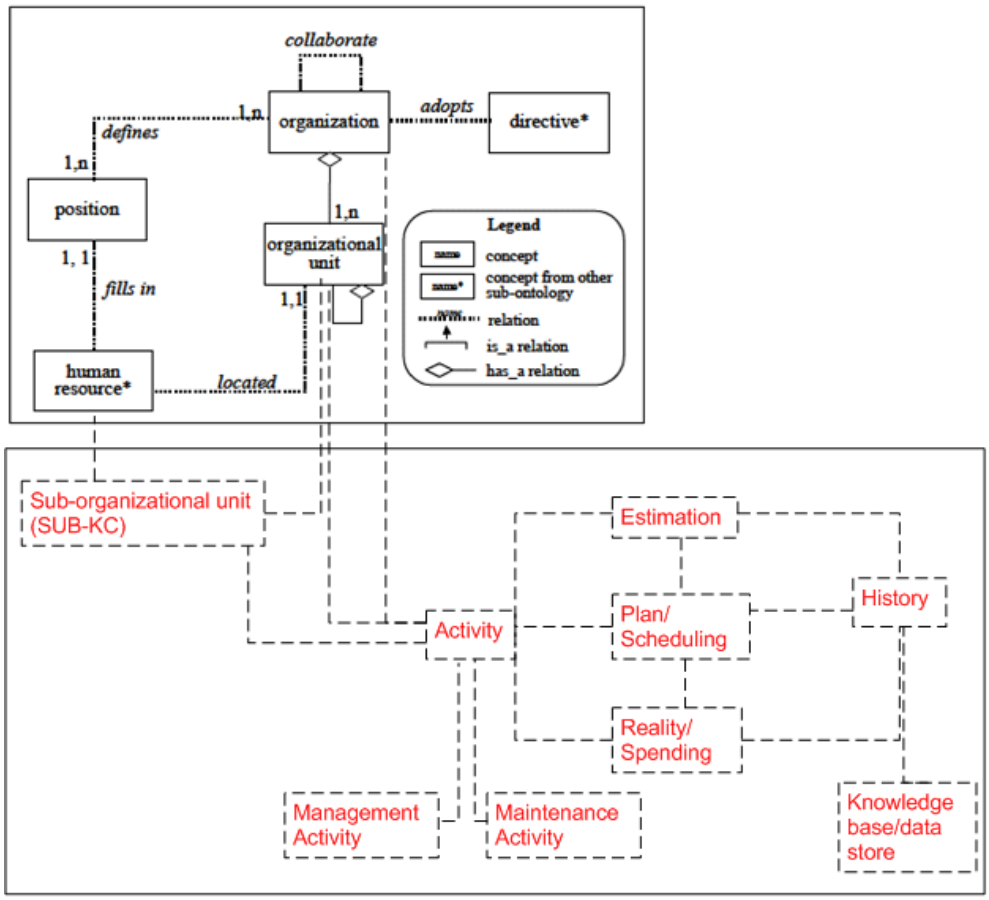
Rozšíření ontologie podpory a údržby

Následující schéma zobrazuje sub-ontologii z pohledu organizační struktury. Dle autora sub-ontologie nezachycuje potřebné komponenty a vazby. Zejména zde chybí pohled řídící, pohled plánovací, odhadování úloh. Dále subontologie nezahrnuje možnost správy více aplikací/řešení atd.



Obrázek 6: Sub-ontologie organizační struktury Zdroj: Dias, 2003

Proto autor této práce navrhuje následující redefinici a rozšíření ontologie organizační struktury o prvky kompetenčního a sub-kompetenčního centra. Dále navrhuje rozšíření o typologii úloh (management, maintenance), jejich řízení tj. odhadování, plánování a relizaci. Navíc uvádí historické vstupy – odhadování pracnosti a následnou skutečnost při realizaci a jejich ukládání.

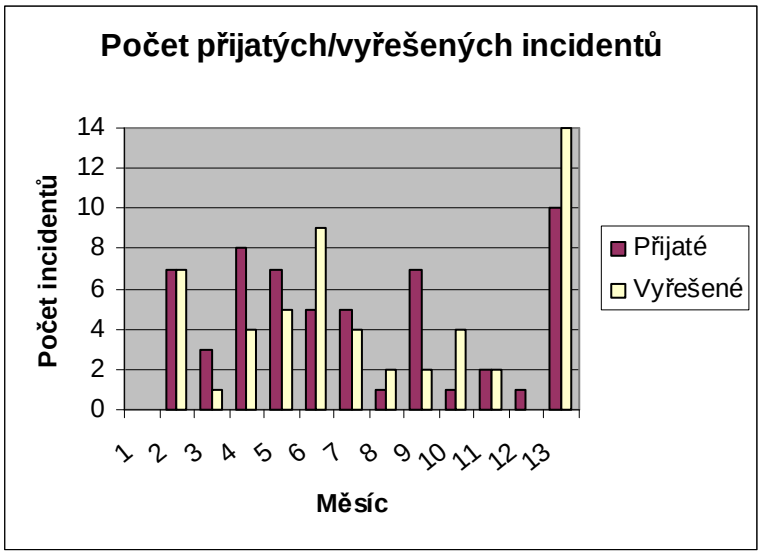


Obrázek 7: Rozšíření Sub-ontologie organizační struktury Zdroj: autor

Závěr - přínosy řešení

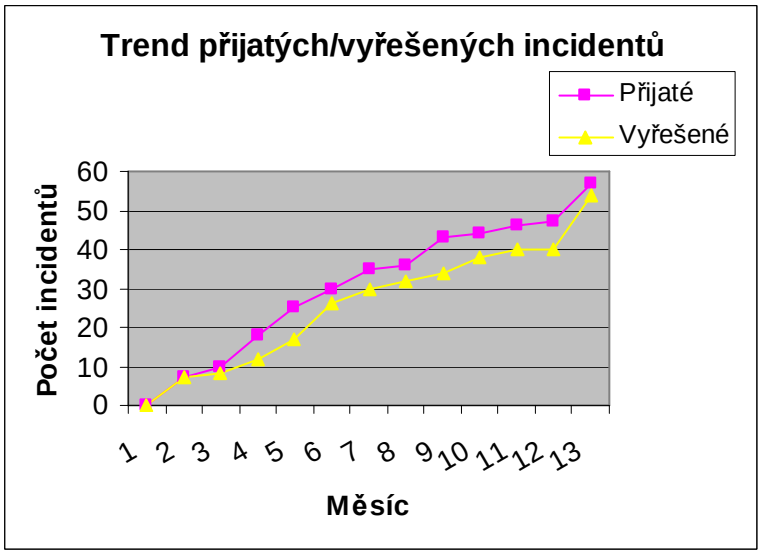
Implementace konceptu kompetenčního centra podpory a údržby a sub-kompetenčního centra podpory a údržby bylo ověřeno na oblasti bankovních IS. Implementace, kromě vyřešní specifických požadavků, přinesla zejména následující přínosy:

- Zřetelná formalizace a oddělení společných úloh pro všechny aplikace v kompetenčním centru v rámci podpory a údržby (např. Systémová integrace nebo release management).
- Zřetelná formalizace a oddělení vlastních úloh pro dílčí aplikaci v sub-kompetenčním centru podpory a údržby aplikace (vlastní opravy chyb, testy atd.).
- Vytvoření funkční formalizované organizační struktury.
- Kodifikaci, formou interních směrnic, pracovních postupů a jejich výstupů.
- Průběžně ověřování kvality na všech úrovních (vstup, výstup, audit, indikátor).
- Přesnější formální odhadování pracnosti podporuje plánování a alokaci zdrojů a vlastní řízení SUB KC (SUB KC disponuje vlastním mechanismem pro odhadování pracnosti a zlepšování odhadů na základě zkušenosti, které není diskutováno v tomto článku).
- Probíhá měření kvality a spokojenosti pomocí následujících indikátorů:
    - Počet přijatých, vyřešených chyb po jednotlivých měsících:
        - Indikátor zachycuje poměr mezi počtem nahlášených a počtem vyřešených chyb za měsíc. Cílem (z pohledu KC) je vyřešit za měsíc co nejvíce hlášených chyb.



Obrázek 8: Počet přijatých, vyřešených incidentů měsíčně na jeden produkt
Zdroj: SUB KC podpory bankovních IS

- Trend přijatých, vyřešených incidentů po jednotlivých měsících:
  - Indikátor zachycuje trend počtu přijatých a řešených incidentů za nějakou periodu. Cílem je, aby trend byl v čase klesající – neboť tak deklaruje zvýšení kvality dodané aplikace i jejích záplat (v čase je identifikováno stále méně chyb).



Obrázek 9: Trend přijatých, řešených chyb Zdroj: SUB KC podpory bankovních IS

Literatura

ALLEN, Bruce, 2004. *Building Operational Excellence: IT People and Process Best Practices.* Intel Press. ISBN-13: 978-0974364971.

ANON,1990. *IEEE Standard Glossary of Software Engineering Terminology.* IEEE Std 610. E-ISBN 0-7381-0391-8.
URL:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=159342&isnumber=4148

ANON, 2004. *Maturing to Centers of Excellence: The Next Step in IT Organization Evolution.* META Group.
URL:http://www.msqaa.org/Best_Practices/Miscellaneous/Maturing_to_Centers_of_Excellence.pdf

BROOKS, F. P. Jr, 1987. *No Silver Bullet Essence and Accidents of Software Engineering*. IEEE Computer vol.20. ISSN 0018-9162.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1663532&isnumber=34828

CANFORA, Gerardo a CIMITILE, Aniello, 2000. *Software Maintenance*. Roma: University of Sannio. URL: ftp://cs.pitt.edu/chang/handbook/02.pdf

DIAS, Márcio, et al., 2003. *Organizing the Knowledge Used in Software Maintenance*. Journal of Universal Computer Science, vol. 9, no. 7.
URL:http://www.jucs.org/jucs_9_7/organizing_the_knowledge_used

KOSKINEN, Jussi, 2010. *Software Maintenance Costs*. Jyväskylä: University of Jyväskylä.
URL:http://users.jyu.fi/~koskinen/smcosts.htm

WILLIAMS, Paul, 2011. *Innovation Centers of Excellence: Next Practices in Innovation Management.* Wisconsin: An American Institute for Innovation Excellence. URL: http://thinkforachange.com/wp-content/uploads/2011/05/2011-02-CoE-Research-Report-Formatted-Draft.pdf

JEL Classification: D23, L23, L25, M15.

Summary

Formalization of new knowledge in structured approach to software support and maintenance – centers and sub-centers of excellence

Effective implementation (in terms of time, cost, utilization of human resources, etc.) of operation of information systems is a strategic issue in today's time when business processes are integrally aligned with the business informatics. Currently, costs associated with software support and maintenance represent more than 90% of total costs.

Software maintenance is a set of activities needed for cost-effective support of IT solution. IS / IT Centers of excellence (COE) do not cover the area of software support and maintenance, so there is no formalized methodology or procedural framework for COE for support and maintenance. Therefore author proposes his own solution by definition and implementation of center of excellence for support and maintenance and sub-centers of excellence for support and maintenance of particular applications.

Current ontology of support and maintenance does not capture the necessary components and links. Therefore author proposes his redefinition and enrichment of ontology of organizational structure about elements of competency and sub-competency center, typology of tasks (management, maintenance), and management - estimating, planning and realization.

In conclusion author discusses the benefits of the implementation of center of excellence for support and maintenance and sub-centers of excellence for support and maintenance of particular applications in banking industry.

Keywords: software support and maintenance, methodology, model, ontology, organizational structure, center of excellence, sub-center of excellence, effort, planning.

Bezpečnostní aspekty SaaS z pohledu zákazníka

Jan Mészáros
jan.meszaros@vse.cz
Doktorand oboru Informatika
Školitel: doc. Ing. Alena Buchalcevová, Ph.D. (alena.buchalcevova@vse.cz)

Abstrakt: Článek se zabývá problematikou bezpečnosti SaaS, uvádí 25 faktorů, jejichž vyhodnocení poskytne celistvý pohled na míru zabezpečení SaaS poskytovatelem. Zjištěná míra zabezpečení slouží potenciálním zákazníkům SaaS k porovnání služeb nabízených různými poskytovateli.

Hodnotící faktory jsou rozděleny do pěti skupin dle jejich zaměření. První skupina hodnotí základní předpoklady pro užití SaaS v organizaci zejména z pohledu legislativních předpisů, druhá skupina pojednává o nejpoužívanějších typech certifikací SaaS. Třetí skupina je zaměřena na smluvní hledisko a hodnotí, do jaké míry zveřejňuje poskytovatel SaaS texty smluv, podmínek užití a záruk formou SLA. Poslední dvě skupiny jsou věnovány technologickým bezpečnostním aspektům, které lze snadno měřit bez smluvního vztahu s poskytovatelem, a dále těm technologickým aspektům, které sice poskytovatel deklaruje, ale nelze je jednoduchým způsobem prověřit.

Článek přináší také podrobnou analýzu úrovně bezpečnosti nejvýznamnějších SaaS dle prezentovaných hodnotících kritérií.

V závěru článku je diskutována provedená analýza a dále je prezentována závěrečná sada doporučení pro potenciální zákazníky SaaS.

Klíčová slova: SaaS, bezpečnost, cloud computing, Software-as-a-Service.

Úvod

Cloud computing jako obchodní a technický model poskytování informatických služeb představuje pro zákazníka i dodavatele těchto služeb nesporně řadu výhod, zejména v oblasti finančních nákladů a technické pružnosti. Kromě řady výhod prezentovaných v marketingových materiálech existuje také množství nevýhod a problémů, které se mohou stát překážkou pro odebírání či poskytování služeb formou cloud computing.

Tento příspěvek je věnován problematice bezpečnosti software jako služby (SaaS, Software-as-a-Service) v rámci podnikového užití pomocí modelu cloud computing.

Hlavním cílem práce je prezentovat soubor doporučení a kontrol pro podporu rozhodování managementu podniku o využití software jako služby. Dílčím cílem je provést hodnocení vybraných poskytovatelů SaaS dle prezentovaných doporučení a kontrol.

Definice pojmů

Vzhledem k různorodosti výkladů pojmu *cloud computing* je pro účely následujícího textu tento termín určen na základě definice NIST (Mell, 2011, s. 2) jako model, který vykazuje pět základních charakteristik:

- zákazník si opatřuje výpočetní kapacitu samoobslužně, automatizovaně, bez nutnosti osobního kontaktu s personálem poskytovatele služby;
- výpočetní kapacita je dostupná přes počítačovou síť a je k ní přistupováno skrze standardní mechanismy, které umožňují užití služby pomocí různorodých klientských platforem (např. tenký / tlustý klient, mobilní telefony, tablety, notebooky, pracovní stanice);
- výpočetní prostředky poskytovatele jsou sdílené mezi množství zákazníků;
- výpočetní kapacita je zákazníkům pružně přidělována i odebírána tak, aby odpovídala aktuálním potřebám zákazníků;

- automatická kontrola a optimalizace zdrojů poskytovatele na základě měření; používání zdrojů může být monitorováno, kontrolováno a vykazováno jak poskytovateli, tak zákazníkům.

Pojem *cloud computing* je dále definován výčtem třech modelů poskytovaných služeb (Mell, 2011, s. 2–3):

- *Software as a Service (SaaS, software jako služba)* – poskytovaná výpočetní kapacita představuje aplikaci, provozovanou poskytovatelem na *infrastruktuře cloudu*. Aplikace je dostupná pomocí různých technických prostředků skrze rozhraní tenkého klienta (např. webový prohlížeč) nebo skrze programové rozhraní. Zákazník má možnost přizpůsobovat aplikaci svým potřebám pomocí poskytnuté sady konfiguračních parametrů.
- *Platform as a Service (PaaS, platforma jako služba)* – poskytovaná výpočetní kapacita pro zákazníka představuje možnost nasazení vlastní aplikace vytvořené pomocí programovacích jazyků, knihoven, služeb a nástrojů podporovaných poskytovatelem. Zákazník má plnou kontrolu nad nasazenou aplikací a má k dispozici sadu konfiguračních parametrů pro nastavení poskytnutého prostředí pro běh aplikace.
- *Infrastructure as a Service (IaaS, infrastruktura jako služba)* – poskytovaná výpočetní kapacita představuje výpočetní výkon, úložiště, sítě a další podstatné výpočetní zdroje. Zákazník může nasazovat a provozovat libovolný software včetně operačních systémů a aplikací. Zákazník má plnou kontrolu nad operačním systémem, úložišti, aplikacemi, případně nad vybranými síťovými komponentami (např. firewally).

Posledním definičním kritériem je přehled čtyř modelů nasazení (Mell, 2011, s. 3):

- *private cloud (soukromý cloud)* – infrastruktura cloudu je poskytována výhradně pro soukromé užití v rámci jedné organizace,
- *community cloud (komunitní cloud)* – infrastruktura cloudu je poskytována výhradně pro užití v rámci určité komunity,
- *public cloud (veřejný cloud)* – infrastruktura cloudu je poskytována k užití veřejnosti,
- *hybrid cloud (hybridní cloud)* – infrastruktura cloudu je poskytována užitím kombinace alespoň dvou výše uvedených modelů.

Výše uvedené definice užívají pojem *infrastruktura cloudu (cloud infrastructure)*, která je definovaná (Mell, 2011, s. 2) jako soubor hardwaru a softwaru, které zpřístupňují pět základních charakteristik modelu cloud computing. Infrastruktura cloudu je členěna na dvě vrstvy, základní fyzickou vrstvu, nad kterou je vrstva abstrakce.

V následujícím textu jsou použity termíny, charakterizující různé role, nazývané *cloud computing actors* (Badger, 2011, s. 19–22):

- *cloud consumer (zákazník)* – osoba nebo organizace, která má obchodní vztah s poskytovatelem služby cloudu a tuto službu používá,
- *cloud provider (poskytovatel)* – osoba nebo organizace zodpovědná za zpřístupnění služby zainteresovaným stranám,
- *cloud auditor (auditor)* – vykonává nezávislou kontrolu služeb cloudu za účelem vyslovení posudku na zkoumanou službu,
- *cloud broker (broker)* – řídí užití, výkon a dodávky služeb cloudu a vyjednává smluvní podmínky mezi poskytovateli a zákazníky,
- *cloud carrier (přepravce)* – prostředník, který poskytuje konektivitu a transport služeb cloudu mezi zákazníky a poskytovateli. Přepravci poskytují zákazníkům přístup ke službám skrze síť, telekomunikaci a další zařízení.

Vymezení rozsahu

Na základě uvedených definic je rozsah tohoto příspěvku vymezen modelem poskytovaných služeb typu *SaaS* a modelem nasazení *veřejný cloud* z hlediska zákazníka.

Z hlediska migrace softwaru do modelu cloud computing definuje Mohan (2011, s. 49) pět úrovní: aplikace, kód, návrh, architektura a užití. Tento příspěvek je věnován výhradně úrovni užití.

Příspěvek poskytuje úvodní hodnocení produktů dle informací veřejně dostupných na webových stránkách provozovatelů, nikoli hodnocení vlastních zakoupených služeb. Kritéria hodnocení jsou navržena pro organizace se sídlem v České republice. Na základě vyhodnocení uvedených kritérií může organizace realizovat předběžný výběr vhodných potenciálních poskytovatelů.

Východiska a motivace

Z pohledu řízení bezpečnosti je na základě vymezení rozsahu práce zaměřena pozornost na řízení bezpečnosti služeb třetích stran. Důvodem je povaha veřejného cloudu, kde je veškerá infrastruktura cloudu řízena i vlastněna třetí stranou. Fyzické umístění této infrastruktury je rovněž pod výhradní kontrolou třetí strany (CSA, 2011, s. 19) v prostorách ovládaných dodavatelem, což je v anglické literatuře označováno jako *off-premise*[1].

Řízení bezpečnosti služeb třetích stran jsou věnovány dvě sekce normy ISO/IEC 27002. Dle sekce A6.2 nesmí být informační bezpečnost ohrožena zavedením produktů nebo služeb třetích stran. Dle sekce A10.2 je při dodávkách služeb třetích stran nutné brát v úvahu bezpečnostní požadavky počínaje smluvními podmínkami (SLA[2]) po následné monitorování a management změn (Doucek, 2008, s. 142). Této problematice je také věnován proces „DS2 Řízení služeb třetích stran" kontrolního rámce COBIT 4.1, který zdůrazňuje význam SLA, řízení rizik a monitorování kvality dodávaných služeb (ITGI, 2008, s. 106).

Z hlediska modelů poskytovaných služeb je SaaS modelem, ve kterém je nejvyšší míra zodpovědnosti za implementaci bezpečnosti a za její řízení na straně poskytovatele služby (CSA, 2011. s. 22), srovnání s ostatními modely poskytovaných služeb je znázorněno na obrázku 1. Zákazník se musí ujistit, že míra jím požadované úrovně bezpečnosti odpovídá úrovni bezpečnosti deklarované poskytovatelem. Zákazník by se měl dále ujistit, že deklarovaná úroveň bezpečnosti odpovídá reálné úrovni bezpečnosti.

Obrázek 1: Zodpovědnost za implementaci bezpečnosti a její řízení

Služby, poskytované pomocí modelu veřejného cloudu, jsou dostupné a využívané veřejností bez omezení. Z tohoto důvodu je nutné považovat všechny uživatele služeb za nedůvěryhodné (CSA, 2011, s. 19). Důsledkem tohoto faktu pro zákazníka je riziko úniků informací k jinému zákazníkovi stejného poskytovatele, který může být například konkurencí (CSA, 2011, s. 109).

Služby SaaS jsou většinou v modelu veřejného cloudu přístupné přes rozhraní webového prohlížeče formou webové aplikace. Technické zabezpečení SaaS je tedy ve většině případů ekvivalentní s technickým zabezpečením webových aplikací. Organizace OWASP[3] sdružuje řadu projektů, které se

[1] Opakem je pojem „*on-premise*", který označuje fyzické umístění infrastruktury cloudu v prostorách zákazníka.
[2] Smluvní podmínky jsou v kontextu poskytování IT služeb obsaženy v dokumentu SLA (Service Level Agreement – dohoda o úrovni poskytovaných služeb).
[3] The Open Web Application Security Project, http://www.owasp.org/

věnují bezpečnosti webových aplikací formou nástrojů a metodických publikací věnovaných obraně před útoky a jejich detekci z hlediska řízení i z technologického hlediska. Mezi relevantní publikace patří Testing Guide (OWASP, 2008) a Top Ten Project (OWASP, 2010).

Definice kritérií pro hodnocení úrovně bezpečnosti SaaS

Výběr hodnotících kritérií a jejich členění do skupin vychází zejména ze základních požadavků na řízení bezpečnosti informací (Doucek, 2008), z požadavků na bezpečnost služeb cloudu definovaných CSA[4] (2011) a z příručky pro provádění penetračních testů webových aplikací vydané organizací OWASP[5] (2008).

Celkem 25 hodnotících kritérií je členěno do pěti skupin:

Z. základní předpoklady užití SaaS (3 kritéria),
A. certifikace, shoda a audit (7 kritérií),
B. smluvní podmínky a SLA (5 kritérií),
C. testovatelné technologické faktory (6 kritérií) a
D. ostatní technologické faktory (4 kritéria).

Konkrétní kritéria, uvedená v následujících kapitolách, jsou vybrána a podrobně vysvětlena autorem příspěvku na základě jeho zkušeností v oblasti zabezpečování webových aplikací.

Z. Základní předpoklady užití SaaS

Jedná se o hodnocení organizace, která plánuje využít pro svoji činnost SaaS, tedy o hodnocení potenciálního zákazníka. Základem pro hodnocení je stanovení charakteru informací, které organizace zamýšlí zpracovávat užitím SaaS. Kritéria vychází zejména z legislativních faktorů, které mohou organizaci zabránit použití SaaS ve veřejném cloudu. Výstupem základního hodnocení jsou omezující podmínky, které ovlivní hodnocení poskytovatelů SaaS dle dalších faktorů v následujících kapitolách.

Tabulka 1: Přehled kritérií základního hodnocení zákazníka

Z.1	**Organizace je správcem osobních údajů dle zákona č. 101/2000 Sb., o ochraně osobních údajů.**
	Poskytovatel SaaS, který zpracovává osobní údaje, se stává dle § 4 *zpracovatelem* a musí dle § 6 uzavřít se správcem smlouvu o zpracování osobních údajů. Kvalita této smlouvy je zákonem dále konkretizována: „Smlouva musí mít *písemnou formu*. Musí v ní být zejména výslovně uvedeno, v jakém *rozsahu*, za jakým *účelem* a na jakou *dobu* se uzavírá a musí obsahovat *záruky zpracovatele o technickém a organizačním zabezpečení* ochrany osobních údajů.“ Pokud s poskytovatelem SaaS není možné uzavřít smlouvu o zpracování osobních údajů, pak není možné používat SaaS pro práci s osobními údaji. Dle § 27 odst. 1 musí být zajištěno, že jsou osobní údaje zpracovávané poskytovatelem SaaS na území ČR nebo na území některého z členských států EU. Pokud výše uvedenou podmínku není možné splnit, pak je nutné postupovat dle § 27 odst. 2, 3 a 4 zákona č. 101/2000 Sb., o ochraně osobních údajů, viz kritérium A.1 níže.
Z.2	**Organizace nakládá s utajovanými informacemi dle zákona č. 412/2005 Sb., o ochraně utajovaných informací a o bezpečnostní způsobilosti.**
	Poskytovatel SaaS musí disponovat osvědčením NBÚ[6] pro stejný nebo vyšší stupeň utajení oproti osvědčení organizace.
Z.3	**Organizace je certifikována dle některé z norem.**
	Je nutné zajistit, aby použití SaaS ve veřejném cloudu nevedlo k pozastavení nebo k odebrání certifikace.

[4] Cloud Security Alliance, https://cloudsecurityalliance.org/
[5] Open Web Application Security Project, https://www.owasp.org/
[6] Národní bezpečnostní úřad, http://www.nbu.cz/

A. Certifikace, shoda, audit

V této sekci je uveden přehled nejpoužívanějších typů certifikací, kterými může být certifikován poskytovatel služeb SaaS.

Tabulka 2: Certifikace, shoda, audit

A.1	**U.S.-EU Safe Harbor Framework**
	Z důvodů přísné legislativní ochrany osobních údajů (viz kritérium Z.1) v zemích EU zavedlo ministerstvo obchodu USA[7] rámec pro dosažení shody v ochraně osobních údajů mezi EU a USA nazvaný *U.S.-EU Safe Harbor Framework* (Export.gov, 2011). Pokud poskytovatel SaaS na území USA prokáže shodu s tímto rámcem, organizace sídlící v EU mohou užít SaaS pro zpracování osobních údajů (Europa - Press Releases, 2010, s. 1–2). V těchto případech musí být mezi zákazníkem a poskytovatelem SaaS uzavřena smluvní doložka o zpracování osobních údajů dle uvedeného rámce, vzory smluvních doložek je možné získat z publikací Evropské Komise (Rozhodnutí Komise ze dne 5.2.2010, s. 10–21).
A.2	**Statement of Auditing Standards No. 70 (SAS 70)**
	SAS 70 je v USA široce respektovaným standardem pro provádění auditu. Audit dle tohoto standardu zahrnuje podrobné prověření kontrolních cílů a kontrolních aktivit poskytovatele SaaS, které zahrnují kontroly nad informačními technologiemi a souvisejícími procesy. Provedený audit dokládá, že poskytovatel SaaS aplikuje odpovídající kontroly a zabezpečení při hostování a zpracovávání dat svých zákazníků. Tento standard je v USA používán pro splnění požadavků sekce 404 zákona SOX[8] (SAS70.com, 2012). Dle SAS 70 existují dva typy reportů (SAS70.com, 2012). První typ je označován jako *SAS 70 Type I Report*, obsahuje hodnocení auditora pouze k určitému okamžiku. Druhý typ nese označení *SAS 70 Type II Report*, obsahuje hodnocení v rámci časového rozmezí obvykle stanoveného na šest a více měsíců. Tento druhý typ je tedy vhodnější pro prezentaci třetí straně.
A.3	**TRUSTe**
	Organizace TRUSTe provádí certifikaci služeb cloudu a SaaS nazvanou *TRUSTed Cloud*. Tato certifikace poskytuje ujištění o následujících oblastech (TRUSTe, 2012): • Sběr dat, toky dat, bezpečnostní metriky související s daty a politiky pro užití dat jsou v souladu s definovanými kritérii[9]. • Technologie poskytovatele je v souladu s definovanými kritérii. • Deklarované zásady pro ochranu osobních údajů odpovídají skutečným praktikám zacházení s osobními údaji. Organizace TRUSTe poskytuje rovněž služby pro zavedení a ověření shody poskytovatele SaaS s rámcem U.S.-EU Safe Harbor (viz kritérium A.1) pomocí certifikace *TRUSTe EU Safe Harbor Seal Program*.
A.4	**Řada norem ISO/IEC 27000**
	Řada norem ISO/IEC 27000 je věnována problematice řízení bezpečnosti informací. Klíčovým prvkem této řady je norma ISO/IEC 27001 s názvem „Systém řízení bezpečnosti informací – požadavky“, která definuje jádro normalizace řízení bezpečnosti informací formou definice systému (Doucek, 2008, s. 90). Druhou nejdůležitější normou je ISO/IEC 27002 „Soubor postupů pro řízení bezpečnosti informací“, obsahující podrobný výklad vhodných bezpečnostních opatření (Doucek, 2008, s. 90).

[7] U. S. Department of Commerce, http://www.commerce.gov/
[8] Sarbanes-Oxley Act, 2002, http://www.soxlaw.com/
[9] Souhrn kritérií pro certifikaci je dostupný na http://www.truste.com/privacy-program-requirements/trusted-cloud/.

A.5	**PCI Data Security Standard (PCI DSS)**
	Standard byl vytvořen pro podporu a zvýšení bezpečnosti držitelů platebních karet a pro podporu širokého přijetí konzistentních metrik pro bezpečnost dat na globální úrovni. PCI DSS se týká všech stran, které se podílejí na zpracování platebních karet včetně obchodníků, zpracovatelů, nabyvatelů, vydavatelů, poskytovatelů služeb a dalších entit, které ukládají, zpracovávají nebo přenášejí data o držitelích platebních karet. Standard obsahuje množinu minimálních požadavků pro ochranu dat držitelů platebních karet (PCI Security Standards Council, 2010, s. 5). Požadavky jsou členěny do šesti základních oblastí: vybudování a údržba bezpečné počítačové sítě, ochrana data držitelů platebních karet, údržba programu řízení zranitelností, implementace silných metrik pro řízení přístupů, pravidelný monitoring a pravidelné testování počítačové sítě, údržba politik informační bezpečnosti (PCI Security Standards Council, 2010, s. 5).
A.6	**Health Insurance Portability and Accountability Act (HIPAA)**
	Zákon USA, který mimo jiné stanovuje pravidla pro ochranu soukromí údajů o zdravotním stavu osob a pravidla pro bezpečné nakládání s těmito údaji na úrovni administrativní, fyzické a technické s cílem zajistit důvěrnost, integritu a dostupnost elektronicky chráněných zdravotních informací (U.S. Department of Health & Human Services, 2012).
A.7	**SysTrust**
	SysTrust je služba zajišťující zvýšení úrovně spokojenosti se systémy pro podporu podnikání z hlediska managementu, zákazníků a obchodních partnerů. Závazek dodržování principů SysTrust zahrnuje oblasti dostupnosti systémů, logické a fyzické bezpečnosti systémů a oblast systémové integrity (SAS70.com, 2012).

B. Smluvní podmínky a SLA

Tato kapitola podává přehled různých druhů smluvních podmínek, které může poskytovatel SaaS prezentovat na svém webu jako podklady pro rozhodování potenciálního zákazníka.

Tabulka 3: Smluvní podmínky a SLA

B.1	**Garance dostupnosti poskytovaných služeb**
	Dostupnost poskytovaných služeb je obvykle uváděna v procentech za určené měřené období, obvykle za kalendářní rok (Voříšek, 2004, s. 155). Někteří poskytovatelé SaaS umožňují veřejný přístup ke statistikám dostupnosti svých služeb.
B.2	**Garance doby odezvy poskytovaných služeb**
	Souvisí úzce s dostupností, je obvykle definována jako doba odezvy jedné transakce v milisekundách (Voříšek, 2004, s. 157).
B.3	**Dostupnost smluvních podmínek**
	Někteří poskytovatelé na své webové prezentaci nezveřejňují žádné smluvní podmínky a ponechávají tuto agendu až do fáze vyzkoušení služby potenciálním zákazníkem. V některých případech je obtížné dohledat odkaz na patřičný dokument obsahující smluvní podmínky. Toto kritérium vyjadřuje míru dostupnosti textu smluvních podmínek poskytovatele SaaS.
B.4	**Dostupnost textu bezpečnostních záruk**
	Kritérium hodnotí kvalitu formulace bezpečnostních záruk bez ohledu na certifikace uvedené v sekci A.
B.5	**Dostupnost textu záruk ochrany osobních údajů**
	Kritérium hodnotí kvalitu formulace záruk ochrany osobních údajů bez ohledu na certifikace uvedené v sekci A.

C. Testovatelné technologické faktory

Faktory hodnotí kvalitu technického zabezpečení poskytovaných služeb, kterou lze změřit ještě před vlastním použitím aplikace jak ve zkušební době, tak na základě smlouvy. Postupy pro měření kvality zabezpečení jsou striktně stanoveny tak, aby neohrozily kvalitu služeb ani provoz testovaných SaaS.

Technické faktory se věnují zejména bezpečnosti webových aplikací, které poskytují rozhraní pro užití poskytovaných SaaS. Většinu faktorů lze jednoduše testovat na přihlašovací obrazovce pro zákazníky SaaS formou tzv. *black box testing*[10].

Tabulka 4: Testovatelné technologické faktory

C.1	**Zabezpečení přenášených dat pomocí protokolu HTTPS**
	Certifikát serveru poskytovatele SaaS musí být platný, musí být podepsán důvěryhodnou certifikační autoritou, název domény se musí shodovat s názvem uvedeným v certifikátu (OWASP, 2010, s. 15). Šifrování mezi aplikací klienta a mezi serverem poskytovatele musí splňovat aktuální požadavky na druh šifrovacího algoritmu, velikost klíče pro šifrování (OWASP, 2008, s. 82) a typ hash algoritmu použitého pro tvorbu otisku klíče. K datu vzniku této práce jsou za dostatečně bezpečné považovány algoritmy AES o délce klíče 256 bitů pro symetrické šifrování, SHA-1 o délce alespoň 2048 bitů pro tvorbu otisku (hash) klíče.
C.2	**Zabezpečení domény poskytovatele pomocí DNSSEC**
	Poskytovatel SaaS může zajistit bezpečnost svého doménového jména pomocí DNSSEC. Tento druh zabezpečení doménového jména poskytuje uživatelům jistotu korektnosti zdroje informací získaných z DNS[11] a jistotu úplnosti a integrity těchto informací.
C.3	**Verze komponent platformy poskytovatele**
	U některých SaaS je možné jednoduše zjistit verze platforem, které tyto služby používají. Platformami jsou obvykle operační systém a webový server. Zjištěné informace můžou odkrýt známé bezpečnostní zranitelnosti (OWASP, 2008, s. 64) zastaralých komponent platforem SaaS.
C.4	**Nebezpečné metody protokolu HTTP**
	Webový server poskytovatele SaaS nesmí podporovat některé nebezpečné metody protokolu HTTP, které mohou být použity ke kompromitaci webové aplikace (OWASP, 2008, s. 113). Webový server by měl podporovat pouze takové metody, které jsou nezbytně nutné pro poskytování služeb.
C.5	**Zabezpečení a obsah cookies**
	Cookies obsahují informace pro autentizaci a autorizaci uživatele oproti webové aplikaci poskytovatele SaaS. Tyto informace musí být dostatečně zabezpečeny pomocí konfiguračních atributů cookies, data cookies nesmí obsahovat jednoduše rozluštitelné informace (OWASP, 2008, s. 168–169).
C.6	**Úroveň ochrany infrastruktury poskytovatele pomocí firewallu**
	Firewall musí co nejlépe chránit infrastrukturu poskytovatele a zároveň umožnit plnohodnotné poskytování služeb. Toto kritérium hodnotí počet a druh otevřených portů TCP/IP (OWASP, 2010, s. 12). Správně nastavený firewall povoluje veřejnou komunikaci na portech 80 a 443.

D. Ostatní technologické faktory

Přehled obsahuje takové faktory, které poskytovatel sice deklaruje ve svých dokumentech, ale zákazník je nemůže legálním, snadným a nenákladným způsobem otestovat.

Těmito faktory jsou: D.1 deklarované použití systémů IDS/IPS[12], D.2 deklarované šifrování záloh, D.3 deklarovaná redundance fyzické infrastruktury, D.4 deklarované provádění penetračních testů.

[10] Testování probíhá za podmínek, kdy nejsou k dispozici žádné informace o vnitřní struktuře testovaného systému.
[11] Podrobné informací o DNSSEC jsou dostupné na http://www.dnssec.cz/.
[12] Intrusion Detection System – systém pro detekci průniků, Intrusion Prevention System – systém pro prevenci průniků

Hodnocení úrovně bezpečnosti vybraných SaaS

Cílem hodnocení je poskytnout podrobné hodnocení vybraných SaaS z hlediska bezpečnosti s využitím výše uvedených hodnotících kritérií.

Metodika

Hodnocení je provedeno na základě faktorů, prezentovaných v předchozích kapitolách s výjimkou faktorů uvedených v kapitole „Z. Základní hodnocení organizace“. Tato skupina faktorů je vynechána, protože se hodnocení nevztahuje k žádné konkrétní organizaci. Celkem je tedy hodnoceno 22 faktorů ve čtyřech skupinách.

Pro každou skupinu je stanovena váha skupiny, pro každý faktor ve skupině je stanovena váha faktoru jako celé číslo v intervalu <0, 5>. Hodnocení faktorů, které lze ohodnotit dosaženou úrovní, je vyjádřeno celým číslem v intervalu <0, 5>, kde číslo 0 znamená nevyhovující a číslo 5 znamená nejlepší hodnocení. Pro faktory, které lze hodnotit dichotomicky, je hodnocení „ano“ vyjádřeno hodnotou 1, hodnocení „ne“ je vyjádřeno hodnotou 0.

Celkové hodnocení relativní úrovně bezpečnosti v procentech H_x jednoho SaaS ve vztahu k m hodnoceným SaaS je vyjádřeno vzorcem

$$H_x = \frac{v_A \sum_{n=1}^{7} v_{An} h_{An} + v_B \sum_{n=1}^{5} v_{Bn} h_{Bn} + v_C \sum_{n=1}^{6} v_{Cn} h_{Cn} + v_D \sum_{n=1}^{4} v_{Dn} h_{Dn}}{\max\{H_i, 1 \le i \le m\}} .100,$$

kde
v_A – váha skupiny faktorů „A. Certifikace, shoda, audit“,
v_B – váha skupiny faktorů „B. Smluvní podmínky a SLA“,
v_C – váha skupiny faktorů „C. Testovatelné technologické faktory“,
v_D – váha skupiny faktorů „D. Další technologické faktory“,
v_n – váha n-tého faktoru v konkrétní skupině faktorů,
h_{Yn} – hodnocení n-tého faktoru v Y-té skupině faktorů.

Váhy skupin byly pro účely hodnocení stanoveny tak, aby byly zvýhodněny skupiny A. a C. oproti zbývajícím skupinám: $v_A = 2$, $v_B = 1$, $v_C = 2$, $v_D = 1$.

Váhy jednotlivých faktorů byly stanoveny na hodnotu 1 s výjimkami zvýhodnění certifikace SaaS dle některé z norem řady ISO/IEC 27000 (A.4), deklarovanou dostupnost (B.1), úroveň zabezpečení přenášených dat (C.1) a úroveň zabezpečení cookies (C.5) takto: $v_{A4} = 2$, $v_{B1} = 2$, $v_{C1} = 2$, $v_{C5} = 2$.

Výběr produktů

Poskytovatelé byli vybráni na základě taxonomie vytvořené společností OpenCrowd[13]. Pro každou kategorii SaaS jsou dle tabulky 5 zvoleny dva nejvýznamnější produkty.

Tabulka 5: Přehled vybraných poskytovatelů SaaS

Kategorie	Vybrané produkty SaaS
1. Bezpečnost	PingConnect On-Demand SSO for SaaS, MessageLabs Hosted Email AntiSpam
2. CRM	Salesforce.com, Oracle CRM On Demand
3. ERP	NetSuite, OrderHarmony
4. Finance	RA.Pid Self Service Spend Analytics, Marketing Budget Management
5. Lidské zdroje	Taleo, staffRostering
6. Osobní produktivita	Microsoft Office Web Apps, Google Docs
7. Prodej	Kadient, Success Metrics

[13] http://cloudtaxonomy.opencrowd.com/taxonomy/software-as-a-service/

Kategorie	**Vybrané produkty SaaS**
8. Řízení IT služeb	Remedy on Demand, Tivoli Live
9. Řízení projektů	Basecamp, Innotas
10. Sociální sítě	Ning, Jive SBS
11. Spolupráce	Egnyte, MindQuilt
12. Správa dokumentů	KnowledgeTree Document Management Software, NetDocuments
13. Správa obsahu	Mosaic, Clickability
14. Účtování	Aria Systems, Zuora
15. Zdraví a wellness	SNO, CECity

Výsledky hodnocení

Na základě provedeného hodnocení byly vybrány produkty, jejichž hodnocení přesahuje 75 %, viz tabulka číslo 6. Uvedené výsledky jsou platné k datu 30. 1. 2012.

Tabulka 6: Produkty, jejichž celkové hodnocení přesahuje 75%

Celkové hodnocení v [%]	**Název produktu**	**Kategorie produktu**	**Hodnocení dílčích oblastí[14] v [%]**			
			A	**B**	**C**	**D**
100,00	Salesforce.com	CRM	100,00	100,00	78,95	0,00
95,65	Basecamp	Řízení projektů	0,00	60,00	100,00	0,00
93,48	Microsoft Office Web Apps	Osobní produktivita	50,00	100,00	78,95	0,00
91,30	MessageLabs Hosted Email AntiSpam	Bezpečnost	0,00	100,00	84,21	0,00
86,96	Egnyte	Spolupráce	50,00	60,00	78,95	50,00
86,96	MindQuilt	Spolupráce	0,00	20,00	100,00	0,00
82,61	RA.Pid Self Service Spend Analytics	Finance	33,33	20,00	84,21	100,00
82,61	CECity	Zdraví a wellness	0,00	40,00	89,47	0,00
80,43	Clickability	Správa obsahu	16,67	0,00	94,74	0,00
78,26	Google Apps	Osobní produktivita	0,00	20,00	89,47	0,00
78,26	Tivoli Live	Řízení IT služeb	0,00	20,00	89,47	0,00
76,09	Kadient	Prodej	16,67	20,00	84,21	0,00
76,09	Ning	Sociální sítě	16,67	60,00	73,68	0,00

Nejvyšší celkové hodnocení získal produkt Salesforce.com, který se tak stal referenčním produktem pro hodnocení ostatních produktů. Z hlediska hodnocení dílčích oblastí jsou z produktů v tabulce 6 nejlépe hodnoceny produkty:

- oblast A – Salesforce.com,
- oblast B – Salesforce.com, Microsoft Office Web Apps a MessageLabs Hosted Email AntiSpam,
- oblast C – Basecamp, MindQuilt,
- oblast D – RA.Pid Self Service Spend Analytics.

[14] Dílčí hodnocení *DH* skupiny faktorů *Y* produktu *X* bylo vypočteno dle vzorce $DH_{X_Y} = \frac{\sum_{n=1}^{k} v_{Yn} h_{Yn}}{\max\{DH_i, 1 \le i \le m\}}$, kde *k* je počet hodnotících kritérií ve skupině *Y*, v_{YN} je váha kritéria *n* a h_{YN} je hodnocení kritéria *n*.

U produktů Basecamp, MindQuilt, CECity, Google Apps a Tivoli Live existují značné rozdíly mezi deklarovanou úrovní zabezpečení (skupiny A, B a D) a reálnou testovanou úrovní zabezpečení vyjádřenou skupinou C.

Produkt RA.Pid Self Service Spend Analytics jako jediný z třiceti hodnocených produktů disponuje zabezpečením DNSSEC dle kritéria C.2.

U třech produktů z třiceti hodnocených nebylo možné zjistit žádné údaje o bezpečnosti.

Závěr

Příspěvek přináší postupy pro hodnocení míry zabezpečení SaaS z pohledu organizace, která zvažuje užití SaaS. Organizace na základě tohoto hodnocení provádí předběžný výběr vhodných poskytovatelů, se kterými následně může zahájit další fázi výběrového řízení.

První krok hodnocení zahrnuje určení druhu informací, které budou zpracovávány pomocí SaaS, a nalezení souvisejících legislativních předpisů. Na základě prvního kroku jsou stanoveny omezující podmínky, které jsou vyjádřeny jako váhy jednotlivých kritérií následného hodnocení poskytovatele SaaS.

Podle navržených hodnotících postupů byla provedena ukázková analýza třiceti nejvýznamnějších poskytovatelů SaaS. Z provedené analýzy vyplývá, že nejlépe zabezpečenými produkty jsou Salesforce.com, Basecamp a Microsoft Office Web Apps. U některých produktů byl zjištěn rozpor mezi prezentovanou úrovní zabezpečení a změřenou reálnou úrovní zabezpečení. Proto je nutné, aby potenciální zákazník přihlédl při výběru konkrétního poskytovatele ke všem skupinám hodnotících kritérií.

Provedení hodnocení dle uvedených hodnotících postupů vyžaduje širokou škálu odborných znalostí, jednotlivými skupinami hodnotících faktorů by se proto měly zabývat různé organizační útvary.

Poskytovatelé SaaS mohou využít hodnotící postupy ke stanovení úrovně zabezpečení svých vlastních produktů a k nalezení problémových oblastí, ve kterých by měli zvýšit míru bezpečnosti.

Mezi cíle dalšího výzkumu patří podrobnější vyhodnocení zjištěných výsledků analýzy, formulace podrobných kritérií pro hodnocení úrovně bezpečnosti z pohledu poskytovatele SaaS a rozšíření oblasti zkoumání o služby typu IaaS a PaaS.

Literatura

BADGER, Lee, Robert BOHN, Shilong CHU, Mike HOGAN, Fang LIU, Viktor KAUFMANN, Jian MAO, John MESSINA, Kevin MILLS, Annie SOKOL, Jin TONG, Fred WHITESIDE a Dawn LEAF. NIST. Special Publication 500-293 (Draft). *US Government Cloud Computing Technology Roadmap Volume II Release 1.0 (Draft): Useful Information for Cloud Adopters*. Gaithersburg, MD: NIST, 2011. Dostupné z: http://www.nist.gov/itl/cloud/upload/SP_500_293_volumeII.pdf

CSA. *Security Guidance for Critical Areas of Focus in Cloud Computing v3.0*. Cloud Security Alliance, 2011, 177 s. Dostupné z: http://www.cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf

DOUCEK, Petr, Luděk NOVÁK a Vlasta SVATÁ. *Řízení bezpečnosti informací*. 1. vyd. Praha: Professional Publishing, 2008, 239 s. ISBN 978-808-6946-887.

ITGI: *Cobit 4.1*. IT Governance Institute, 2007, ISBN 1-933284-72-2, 213 s.

JANSEN, Wayne a Timothy GRANCE. NIST. Special Publication 800-144. *Guidelines on Security and Privacy in Public Cloud Computing*. Gaithersburg: NIST, 2011. Dostupné z: http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf

MELL, Peter a Timothy GRANCE. NIST. Special Publication 800-145. *The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology*. Gaithersburg: NIST, 2011. Dostupné z: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

MOHAN, T. S. Migrating into a Cloud. In: BUYYA, Rajkumar, James BROBERG a Andrzej GOSCINSKI. *Cloud computing: principles and paradigms*. Hoboken N. J.: John Wiley & Sons, Inc. 2011.

OWASP. *OWASP Testing Guide v3.0*. The OWASP Foundation, 2008, 374 s. Dostupné z: http://www.lulu.com/items/volume_66/5691000/5691953/9/print/5691953.pdf

OWASP. *OWASP Top 10 - 2010: The Ten Most Critical Web Application Security Risks*. The OWASP Foundation, 2010, 22 s. Dostupné z: http://www.lulu.com/items/volume_71/11443000/11443706/1/print/1400974_1400974.pdf

PCI DSS. *Payment Card Industry Data Security Standard, Version 2.0*. Wakefield, MA: PCI Security Standards Council, 2010, 75 s. Dostupné z: https://www.pcisecuritystandards.org/documents/pci_dss_v2.pdf

Rozhodnutí Komise ze dne 5. 2. 2010: o standardních smluvních doložkách pro předávání osobních údajů zpracovatelům usazeným ve třetích zemích podle směrnice Evropského parlamentu a Rady 95/46/ES. In: *Official Journal of the European Union*. Brusel, 2010. Dostupné z: http://ec.europa.eu/justice/policies/privacy/docs/modelcontracts/c_2010_593/c_2010_0593_cs.doc

Safe Harbor Overview. *Export.gov: Helping U. S. Companies Export* [online]. 4.11.2011 [cit. 2012-01-27]. Dostupné z: http://export.gov/safeharbor/eu/eg_main_018476.asp

Safer standards for European citizens' data transfers to processors in third countries. *EUROPA - Press Releases* [online]. 2010[cit. 2012-01-27]. IP 10/130 Dostupné z: http://ec.europa.eu/justice/policies/privacy/docs/modelcontracts/ip_10_130_en.pdf

SAS70.COM. *SAS 70 Service Organisation Auditing Standards, Public Accounting Information* [online]. 2012 [cit. 2012-01-27]. Dostupné z: http://sas70.com/

TRUSTE. *Online Trust & Safety from TRUSTe* [online]. 2012 [cit. 2012-01-27]. Dostupné z: http://www.truste.com/

Understanding Health Information Privacy. U. S. DEPARTMENT OF HEALTH & HUMAN SERVICES. HHS.gov [online]. 2012 [cit. 2012-01-27]. Dostupné z: http://www.hhs.gov/ocr/privacy/hipaa/understanding/index.html

VOŘÍŠEK, Jiří. *Aplikační služby IS/ICT formou ASP: proč a jak pronajímat informatické služby*. 1. vyd. Praha: Grada, 2004, 213 s. ISBN 80-247-0620-2.

ČESKO. Zákon č. 101/2000 Sb., o ochraně osobních údajů a o změně některých zákonů, navazující vyhlášky. In: Sbírka zákonů České republiky. 2000.

JEL Classification: L86

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu GAČR č. P403/11/0574.

Summary

SaaS security facets in terms of cloud service consumer

The article is dedicated to the SaaS security issues presenting 25 criteria of evaluation. The criteria bring holistic view of SaaS provider's services security level. Security rating findings are useful for consumer to compare various SaaS providers.

Evaluation criteria are divided to five different areas. The first area evaluates the basic SaaS usage assumptions regarding consumer's organization legal constraints. The second area covers the most adopted SaaS certification types. The third area focuses on agreements and SLA guarantees. The other areas pursue technological facets which are easily measurable without a contractual relationship and supplier declared immeasurable facets.

The article yields the results of detailed SaaS analysis accomplished in according to present evaluation criteria and methods.

The analysis is discussed and the final recommendations for potential SaaS consumers are expressed in the article conclusion chapter.

Keywords: SaaS, security, cloud computing, Software-as-a-Service.

Aplikace pro podporu auditu datové kvality

David Pejčoch
david@pejcoch.com
Doktorand oboru informatika
Školitel: prof. RNDr. Jan Rauch, CSc., (rauch@vse.cz)

Abstrakt: Článek shrnuje současné přístupy k auditu datové kvality, přičemž navazuje na postupy mapování vlastností dat a jejich důsledků pomocí kauzálních map publikované v (Pejčoch, 2011). Po stručném seznámení s typickou funkcionalitou dostupnou v rámci obecných nástrojů pro počítačovou podporu auditu (CAAT) a relevantní funkcionalitou obsaženou v nástrojích pro řízení datové kvality představuje aplikaci CADAQUES vyvíjenou autorem této práce.

Klíčová slova: Audit datové kvality, kanonický datový model, CAAT, CADAQUES.

Úvod

V příspěvku (Pejčoch, 2011) prezentovaném v rámci minulého ročníku Dne doktorandů byl naznačen přístup dávající do souvislosti úroveň vlastností jednotlivých atributů s jejich užitím a náklady plynoucími pro toto užití z naměřené úrovně vlastností. Výsledkem byla čtyřrozměrná OLAP kostka, která se následně stala zdrojem pro generování kauzalit, které demonstrovaly vztah vlastností dat co by metriky výkonnosti řízení dat do metrik výkonnosti řízení ostatních oblastí informatiky. Kauzality bylo možné vizualizovat pomocí kauzálních map a charakterizovat pomocí pravidel. V závěru citované práce bylo poněkud ambiciózně konstatováno, že uvedený přístup představuje návod pro realizaci auditu datové kvality. S odstupem času bych toto tvrzení přeformuloval tak, že uvedený přístup představuje základ pro postup realizace auditu datové kvality. V rámci této práce bych rád na dosud publikované závěry navázal a poskytl popis konkrétního návrhu aplikace pro podporu auditu datové kvality vycházející z těchto závěrů. Tato aplikace postupně vzniká na stránkách dqassessment.com. Její vybrané komponenty by mimo jiné měly později posloužit pro účely výuky v rámci mnou přednášeného předmětu 4IZ562 – Řízení datové kvality.

Připomeňme, že ve zmiňovaném příspěvku (Pejčoch, 2011) jsem vymezil na základě kompilace přístupů z více zdrojů celkem čtyři dimenze vlastností dat: (1) Endogenní vlastnosti, zahrnující důvěryhodnost, unikátnost, sémantickou a syntaktickou správnost, (2) Časovou dimenzi zahrnující aktuálnost, včasnost, volatilitu a časovou synchronizaci, (3) Kontextuální dimenzi, zahrnující konzistentnost vůči ostatním zdrojům a atributům, úplnost, pokrytí všech potenciálních hodnot a konečně (4) Dimenzi užití zahrnující vlastnosti jako je dostupnost, srozumitelnost, interoperabilita a bezpečnost přístupu. Dodatečně jsem tyto čtyři dimenze rozšířil ještě o dimenzi pátou, Ekonomickou dimenzi, reflektující (1) náklady na pořízení a aktualizaci datových zdrojů, (2) náklady na uložení, sdílení, distribuci, zálohování a archivaci a v neposlední řadě (3) náklady na ochranu dat.

Již uvedená práce (Pejčoch, 2011) naznačovala určité závislosti mezi jednotlivými vlastnostmi dat. Tyto kauzality měly však spíše povahu konstatování, že ke stanovení některých vlastností (zejména z časové dimenze) je nutné stanovit vlastnosti jiné. Na základě dalšího zkoumání jsem objevil i jiné, zajímavější kauzality, např. určité trade-off mezi vlastností „dostupnost“ a „bezpečnost přístupu“, či postupné promítání vlastností „spolehlivost“ a „úplnost“ do úrovně vlastnosti „důvěryhodnost“.

V uvedené práci (Pejčoch, 2011) jsem pouze konstatoval, že v rámci kauzalit figurují nějakým způsobem náklady plynoucí z úrovně vlastností daného atributu pro konkrétní užití. Na tomto místě bych rád na toto konstatování navázal a na základě přístupů publikovaných Loshinem (2011) a Englishem (1999) doplněných o své vlastní kategorie vymezil kompletní klasifikaci nákladů na nekvalitní data.

Audit datové kvality (Data Quality Assessment) představuje zpravidla první fázi implementace opatření pro zvýšení kvality dat / informací. Dosavadní přístupy se vesměs shodují, že kromě zjištění aktuálního stavu vlastností dat by měl obsahovat též analýzu užití (stávajícího i potenciálního) jednotlivých datových atributů a zjišťování příčin nekvalitních dat pomocí Root-Cause analýzy.

Podle Lee a kol. (2006) jej lze realizovat pomocí následujících technik: (1) kvalitativní šetření o datové kvalitě, (2) aplikace kvantitativních metrik, (3) analýza integrity dat podle TDQM (Total Data Quality Management). V praxi se podle Lee a kol. (2006) uplatňují tyto přístupy:

- Komparativní přístup, tj. srovnání výsledků subjektivního a objektivního assessmentu, zjištění příčin rozdílného vnímání a vzniku nekvalitních dat, návrh nápravných opatření,
- Subjektivní šetření hodnocení kvality dat, důležitosti vlastností, procesů, programů a nástrojů, jehož výsledky jsou často porovnávány v rámci benchmarkingu na globální úrovni nebo na úrovni hodnocení různými rolemi respondentů,
- Data Integrity Assessment, v rámci něhož je nejprve definováno, co znamená datová kvalita z pohledu dat a jejich kontextu (výstupem jsou integritní pravidla pro data) a následně je provedeno porovnání skutečného stavu s nadefinovanými pravidly, na které navazuje analýza kauzálních příčin narušení pravidel a plán nápravy nebo redefinice pravidel (v případě, že se ukážou jako chybná).

V rámci metodiky TQdM (Total Quality data Management) celosvětově uznávaného guru v oblasti datové a informační kvality Larry Englishe (1999) lze identifikovat kroky, které odpovídají auditu datové kvality. Jedná se o první tři fáze uvedené metodiky: (1) assessment definice dat a kvality informační architektury, (2) assessment informační kvality a (3) měření nákladů na nekvalitu.

V rámci prvního kroku uvažuje English tyto konkrétní kroky: (1) identifikaci metrik kvality definice dat, (2) identifikaci podstatných skupin informací k hodnocení, (3) identifikaci kategorií zainteresovaných subjektů pro vybrané skupiny informací (producenti, znalostní pracovníci, externí zákazníci, ...), (4) technický assessment kvality definice pro zjištění shody s datovými standardy a směrnicemi, (5) assessment kvality informační architektury a návrhu databází, (6) měření zákaznické spokojenosti s definicí.

V rámci druhého kroku uvažuje tyto činnosti: (1) potvrzení nebo definice skupin informací k hodnocení, (2) definice charakteristik, které budou měřeny, (3) definice vztahu informační hodnoty a nákladů pro jednotlivé skupiny informací, (4) identifikace zdrojů a procesů, které budou měřeny, (5) identifikaci zdrojů dat, proti nimž je měřena správnost dat, (6) náhodný výběr dat pro assessment, (7) měření provedené na vzorku dat, (8) interpretaci zjištění a prezentaci / reporting úrovně informační kvality.

V rámci třetího kroku uvažuje: (1) nalezení metrik výkonnosti byznysu, (2) analýzu nákladů na informace, (3) odvození skupin nákladů pocházejících z chybějících a nekvalitních dat a kalkulaci měřitelných nákladů v důsledku selhání procesů, (4) identifikaci segmentů klientů, (5) výpočet životní hodnoty klienta a (6) kalkulace hodnoty informací jako dopadu do hodnoty klienta prostřednictvím nákladů na ztracené a propásnuté příležitosti.

McGilvray (2008) publikovala alternativní metodiku zvyšování kvality dat pod názvem „The Ten Steps Process“. Auditu datové kvality odpovídají kroky (1) definice byznys potřeb a přístupu, (2) analýza informačního prostředí, (3) data quality assessment založený na oskórování jednotlivých vlastností dat a (4) hodnocení byznys dopadu naměřených úrovní vlastností dat.

Přizpůsobením uvedených metodik jsem vytvořil postup, jehož kroky aktuálně převádím do konkrétní podoby aplikace. Metodiku auditu, stejně jako vznikající aplikaci, jsem pracovně nazval CADAQUES (Complex Audit of Data Quality in Enterprise Systems). Kombinuje v sobě kvalitativní a kvantitativní assessment uvažovaný Lee a kol. (2006) a modelování dopadu do skupin nákladů uvažovaných Englishem (1999) a McGilvray (2008).

Aplikace pro podporu auditu

Jak uvádí Svatá (2011), aplikace pro podporu auditu (CAAT – Computer-Assisted Audit Techniques) mohou výrazně zefektivnit práci auditora. Vzhledem k vysokému objemu dat, který je nutné zanalyzovat v rámci auditu informačních systémů, stává se dle mého názoru využití automatizovaných aplikací dokonce věcí nezbytnou. ISACA (2008) člení CAAT do pěti skupin: (1) obecný auditní software, (2) customizované dotazy nebo skripty, (3) utility, (4) software pro monitoring a mapping, a konečně (5) auditní expertní systémy.

Hamřík (2008) uvádí alternativní členění technik auditu s využitím CAAT: (1) GAS (Global Audit Software; obecné auditní systémy), (2) CAS (Custom Audit Software) vytvořené speciálně pro použití u konkrétního klienta, (3) paralelní simulace, tj. počítačová simulace imitující běžné fungování testovaných klientských programů a (4) integrovaná testovací zařízení, kdy auditor vkládá testovací data vedle aktuálních dat v běžně používaných aplikacích. Podobné členění technik auditování pomocí počítače uvažuje též Svatá (2011) a rozlišuje: (1) techniku testovacích dat, kdy se data zpracují na stejném systému jako živá data, (2) paralelní simulaci, kdy jsou stejná data zpracována systémem se stejnou funkcionalitou, (3) tzv. Integrated Test Facilities (pozn. autora: zřejmě lze přeložit jako integrovaná testovací zařízení), kdy se testovací data zpracovávají souběžně s „živými" daty a (4) zabudovaný auditní modul, kdy jsou nastavovány různé reporty a logy monitorující průběh zpracování živých dat.

Funkcionalita dostupná v rámci nástrojů pro podporu auditu

Computer Assisted Audit Group (2009) uvádí následující specifické funkce díky níž se CAAT ukazují jako užitečné: (1) vytváření vzorků dat, (2) řízení souborů (slučování, porovnávání, řízení, oddělování) a (3) generování reportů. Na jiném místě dokumentu CAAG (2009) upřesňuje, že pod vytvářením vzorků dat rozumí stratifikovaný výběr. Tato forma výběru, uvažuje nejprve výběr části pozorování (tzv. strata), o níž se domníváme, že je více homogenní než celý základní soubor a teprve poté realizován náhodný výběr v rámci strat.

Sayana (2003) uvádí jako typickou funkcionalitu obecných auditních systémů extrakci dat z obecně používaných datových formátů a většiny databázových systémů, přičemž uvažuje přímé napojení CAAT na produkční data. Aplikace by podle něj rovněž měla umožňovat vytváření dotazů nad daty, statistickou analýzu dat, vytváření vzorků dat, horizontální spojování dat z různých zdrojů a identifikovat chybějící data v rámci řady. Svatá (2011) k této skupině CAAT ještě uvádí další příklady funkcionality: (1) identifikace zdvojených dat a (2) testování správného výpočtu časové struktury. Podobné funkce má podle Svaté (2011) poskytovat i druhá skupina aplikací, tj. customizované výběry a skripty. Třetí skupinu aplikací charakterizuje tentýž zdroj jako, SW utility „pro prověřování provozu systémů, testování, analýzu systémů, analýzu čerpání zdrojů, atd." Další skupina nástrojů, tzv. „Mapující a monitorující aplikační software", podle téhož zdroje slouží k dokumentaci logiky, cest, podmínek kontrol a posloupnosti operací.

Poslední skupina nástrojů uvažovaná Svatou (2011), auditní expertní systémy, je z pohledu této práce nejzajímavější, neboť konkrétní řešení popsané dále spadá právě do této kategorie. Podle Feigenbauma (1979) je expertní systém: „*inteligentní počítačový program, který užívá znalosti a inferenční procedury k řešení problémů, které jsou natolik obtížné, že pro své řešení vyžadují významnou lidskou expertízu*". Alternativní definice poskytuje Gosman (1990), když považuje expertní systém za: „*počítačový program simulující rozhodovací činnost lidského experta při řešení složitých úloh a využívající vhodně zakódovaných speciálních znalostí převzatých od experta s cílem dosáhnout ve zvolené problémové oblasti kvality rozhodování na úrovni experta*".

Cannon (2008) uvádí následující členění SW nástrojů a technik obsažených v CAAT: (1) nástroje pro hodnocení slabin nastavení systému, (2) analýzu síťového provozu a protokolů s využitím analyzátorů packetů, (3) mapovací a monitorovací nástroje, (4) software pro testování konfigurace specifických aplikací, (5) software pro kalkulaci počtu licencí napříč sítí a (6) testování shody hesel (pozn. autora: s interními standardy) u uživatelských účtů.

V případě kontinuálního online auditu rozlišuje Cannon (2008) tyto techniky: (1) online monitory událostí procházející logy a upozornění, (2) vestavné „pojistky" pro audit programů označující transakce, které je vhodné přezkoumat, (3) kontinuální a namátková simulace (CIS) založené na auditu transakcí splňujících určitá předdefinovaná kritéria, (4) snímkování dat a sestavení sekvence zachycených snímků do souvislého souboru kroků v rámci transakce, (5) vestavné auditní moduly (EAM) zpracovávající fiktivní transakce společně s těmi pravými a (6) SCARF/EAM, tj. auditní programy na úrovni systému monitorující vestavné auditní moduly uvnitř aplikačního software.

Vhodná funkcionalita pro nástroje auditu datové kvality

Funkcionalita pro posouzení úrovně aktuální kvality dat je zpravidla v určité míře součástí každého nástroje pro řízení kvality dat. Jedná se hlavně o komponenty, které souvisejí s profilací dat a monitoringem datové kvality. Není snad jediné schéma procesu řízení datové kvality, ve kterém by tyto dvě komponenty chyběly. V prvním případě se jedná o soubor funkcí pro základní popisné statistiky (medián, modus, maximum, minimum, frekvenční analýza), analýza syntaktických vzorů, porovnání dat s metadaty, ověření referenční integrity, zatímco v druhém případě se jedná převáženě o sledování výstupů nadefinovaných byznys pravidel. Výstupy monitoringu mimo jiné ukáží, jak účinné jsou současně platné kontroly a slouží jako podklad pro kalkulaci ROI programu řízení datové kvality. Pro účely permanentního sledování je často řešení pro řízení datové kvality vybaveno komponentou reagující na konkrétní události spuštěním komponenty jiné, záznamem do logu, anebo inicializací tzv. upozornění. V rámci komponenty pro monitoring je též často k dispozici funkcionalita pro zobrazení přehledů a sledování trendu.

Pro audit datové kvality lze úspěšně použít i další komponenty sloužící pro verifikaci proti externím datovým zdrojům jako je např. ověřování adres proti územně identifikačnímu registru UIR-ADR, číselníkům MV ČR, komerčním kompilacím více registrů jako je např. Credit Info a Registru ekonomických subjektů (v tomto případě je ale nutné postupovat obezřetně, neboť se nejedná o datově „čistý“ zdroj). Funkcionalita z této oblasti by měla umožňovat jak přesný join na základě jedinečného klíče, tak přibližný join, realizovaný optimálně (zejména z výkonnostních důvodů) pomocí porovnávacích kódů.

Pro účely posouzení aktuálního stavu duplicit s cílem stanovit kvantitativní metriky unikátnosti je vhodné též začlenit funkcionalitu pro shlukování záznamů se shodnými porovnávacími kódy vygenerovanými na základě apriorních znalostí o sémantickém typu daného atributu a zadané míry požadované shody. Podrobně jsem se problematice generování porovnávacích kódů věnoval např. v jednom ze svých starších příspěvků na Den doktorandů (Pejčoch, 2009).

Zejména pro účely generování porovnávacích kódů je naprosto nezbytnou součástí řešení pro datovou kvalitu tzv.Quality Knowledge Base (QKB). Obsahuje zpravidla součásti specifické pro jednotlivá jazyková / národní prostředí, jako jsou na jednotlivé sémantické datové typy navěšené gramatiky (syntaktická pravidla s odhadem pravděpodobnosti jejich výskytu jako váhou), fonetické knihovny, obsahující pravidla pro záměnu různých hlásek (např. zdvojených samohlásek) a snažící se eliminovat překlepy u shodně znějících slov, slovníky, standardizační schémata, knihovny regulárních výrazů a tabulky přípustných znaků. Všechny zmiňované komponenty slouží při auditu datové kvality jednak pro účinné generování porovnávacích kódů a jednak pro validační pravidla sloužící pro stanovení míry syntaktické a sémantické správnosti. Pro účely monitoringu datové kvality znalostní báze též obsahuje nadefinovaná byznys pravidla.

V úvodu této práce jsem vyjádřil svůj záměr poskytnout úplnou klasifikaci nákladů plynoucích z nekvalitních dat. Jako základ mi posloužilo členění uvedené Loshinem (2011) a Englishem (1999). David Loshin člení náklady do následujících kategorií: (1) Finanční dopad (přímé operační náklady, režijní náklady, dodatečné poplatky, změny v cash-flow, dopad do odpisů a úniku peněž z firmy např. v důsledku fraudů), (2) dopad do spokojenosti spotřebitele a jeho očekávaného tržního chování, (3) dopad do rizika a compliance a konečně (4) dopad do produktivity firmy. Larry English uvažuje následující alternativní členění: (1) náklady plynoucí přímo z nekvalitních informací, (2) náklady na assessment nebo kontrolu, (3) náklady spojené se zlepšováním procesů a předcházením defektům. Pokud tyto dvě klasifikace podrobíme hlubšímu zkoumání, zjistíme, že se v podstatě doplňují a představují dvě dimenze pohledů na náklady, které je vhodné uvažovat současně. Larry English v podstatě pouze dále člení náklady, které D. Loshin označuje za „finanční dopad“. Výsledný přehled doplněný o další dle mého názoru oprávněné podskupiny nákladů znázorňuje Tabulka č. 1.

Tabulka 1: Klasifikace nákladů

Kategorie	Podkategorie
FI: Finanční dopad	FI1: Přímé náklady FI2: Náklady na kontrolu FI3: Náklady na zlepšení datové kvality FI4: Náklady ušlé příležitosti
CS: Dopad do spokojenosti klienta	CS1: Negativní word-of-mouth CS2: Odchod klienta
RC: Dopad do rizika a compliance	RC1: Postihy regulátorem RC2: Soudní žaloby
PM: Dopad do produktivity firmy	PM1: Snížení výkonnosti IT PM2: Snížení výkonnosti celé firmy PM3: Dopad do strategického řízení PM4: Analytické důsledky

Zkratky uvedené před názvy přestavují jazykově nezávislé kódy příslušných číselníků použitých v aplikaci CADAQUES. Stejně tak jako v případě vlastností dat, je zde zřejmé, že některé kategorie nákladů spolu budou kauzálně souviset. Např. RC2: Soudní žaloby budou mít zřejmě též zprostředkovaně dopad do FI1: Přímé náklady, stejně tak jako do CS2: Odchod klienta.

V rámci shromažďování požadavků na funkcionalitu nástroje pro audit datové kvality jsem identifikoval potřebu existence tzv. Kanonického datového modelu jako součásti znalostí báze. Pojem Kanonický datový model (též Společný datový model) pochází z oblasti datové integrace. Význam slova kanonický lze v českém jazyce chápat jako „odvozený, vztahující se ke kánonu (tedy měřítku, pravidlu, souboru zásad, představě o ideálních proporcích). V matematice představuje kanonický tvar formu, ve které může být objekt jednoznačně prezentován. V oblasti datové integrace jej Štumpf a Džmuráň (2008) zmiňují jako model na konkrétní aplikaci nezávislý. Pro účely znalostní báze řešení pro audit datové kvality uvažuji Kanonický datový model jako referenční sadu atributů nazvanou jednotnými syntaktickými pravidly, na které se odkazují další části znalostní báze, např. byznys pravidla, syntaktická pravidla, knihovny regulárních výrazů, ale i znalosti naakumulované v rámci předchozích auditů (např. formy užití atributu pro různé role, související typy a výše nákladů, sada vlastností dat, které je vhodné pro daný sémantický datový typ sledovat, apod.). Na jednotlivé atributy Kanonického datového modelu jsou v průběhu mého pojetí auditu namapovány skutečné atributy, zařazené do assessmentu.

Štumpf a Drmuráň (2008) odkazují na řadu existujících Kanonických modelů pro různé vertikály. Pro oblast pojišťovnictví je to ACORD (Association for Cooperative Operations Research and Development), pro telekomunikace SID, pro oblast veřejných služeb CIM (Common Information Model), pro energetiku PPDM, resp. MMDM, pro oblast výroby a dodavatelsko-odběratelské řetězce OAGIS (Open Application Group Integration Specification), pro zdravotnictví HL7 (Health Level Seven International) a HIPAA, pro oblast prodeje ARTS (The Association for Retail Technology Standards) a konečně pro kapitálové trhy FPML či SWIFT.

CADAQUES, aplikace pro podporu auditu datové kvality

Aplikace CADAQUES představuje konkrétní realizaci principů popsaných výše. Aplikace je programována v jazycích PHP, JavaScript a Perl. Jako databázovou platformu používá primárně MySQL. Postupně jsou do ní přidávány též konektory na jiné RDBMS. Aktuálně je hotový konektor na Informix a MSSQL, připravovány jsou konektory na Firebird, DB2, PostgreSQL a Oracle 11g. Data lze do aplikace importovat též prostřednictvím txt souborů s oddělovačem. Standardním úložištěm pro data auditu, vstupní data a znalostní bázi (QKB) je MySQL. V současné době není možná online integrace voláním webových služeb aplikací třetích stran.

QKB obsahuje kromě výše zmíněných gramatik, knihoven a pravidel též následující komponenty specifické pro audit datové kvality:

- Kanonický datový model,

- Výběr typických atributů Kanonického modelu pro vybranou vertikálu,
- Přiřazení typických rolí uživatelů pro danou vertikálu,
- Přiřazení typického užití dat k rolím,
- Přiřazení vlastností dat typicky sledovaných pro jednotlivé atributy Kanonického modelu v rámci příslušného užití,
- Přiřazení typických druhů nákladů a jejich obvyklé „fuzzy“ úrovně na jednotlivé úrovně vlastností dílčích atributů,
- Pravidla pro fuzzifikaci pro jednotlivé druhy nákladů v rámci užití,
- Pravidla pro fuzzifikaci pro jednotlivé vlastnosti dat podle atributů Kanonického modelu,
- Sady pravidel vyjadřující apriorní znalosti o kauzalitách mezi atributy Kanonického modelu, úrovní vlastností dat a souvisejícími náklady pro jednotlivé vertikály,
- Sadu otázek pro kvalitativní šetření subjektivních vlastností dat.

Jak již vyplynulo z použitých technologií, interakce s uživatelem probíhá pomocí webové aplikace. V rámci ní je možné administrovat jednotlivé části znalostní báze, na „dashboardu“ sledovat aktuální stav jednotlivých auditů, v rámci jednotlivých auditů sledovat jejich výsledky, generovat zprávu auditora, zakládat nové audity, řídit běh již vytvořených auditů. Postup přípravy auditu zahrnuje podle metodiky CADAQUES následující kroky:

- Založení metadat auditu (výběr partnera, výběr vertikály, vložení popisných údajů nutných pro vygenerování zprávy auditora na závěr auditu),
- Definici zdrojových systémů,
- Přiřazení rolí jednotlivým respondentům na základě QKB a jejich následnou úpravu,
- Přiřazení užití jednotlivým rolím na základě QKB a jejich následnou úpravu,
- Výběr relevantních domén z Kanonického datového modelu,
- Přiřazení nákladů z QKB a jejich následnou úpravu,
- Prioritizaci,
- Přiřazení reálných dat ke kanonickému modelu,
- Vygenerování dotazníku pro kvalitativní šetření.

Uživatel je provázen jednotlivými kroky zakládání auditu formou wizardu. V rámci jednotlivých kroků je možné nahrát znalosti z QKB, ručně vkládat / editovat údaje nebo je nahrát z externího textového souboru s oddělovačem. K jednotlivým fázím je možný dodatečný návrat kliknutím na záložku příslušného kroku.

Vlastní spuštění je možné provést z přehledu vytvořených auditů. Po spuštění jsou odemknuty pohledy na dotazníky kvalitativního šetření a jednotlivým respondentům jsou na e-mail zaslány pozvánky k jejich vyplnění. Audit kvantitativních vlastností dat extrahovatelných z dat je vhodné spustit vzhledem k zátěži separátně. Načasování běhu jednotlivých komponent lze při instalaci na operačním systému Linux administrovat přímo z rozhraní aplikace. Při spuštění auditu je též dogenerován zbytek tabulek představujících OLAP kostku výsledků auditu v tzv. star schématu. Prvky jejích dimenzí byly nadefinovány během průchodu wizardem při zakládání auditu.

Porovnávací kódy jsou v CADAQUES vytvářeny podle následujícího postupu: (1) převedení všech znaků na kapitálky, (2) odstranění diakritiky, (3) aplikace fonetických pravidel, (4) tokenizace, (5) rozpoznání významu jednotlivých tokenů, (6) aplikace standardizačních schémat, (7) vytvoření vlastního porovnávacího kódu.

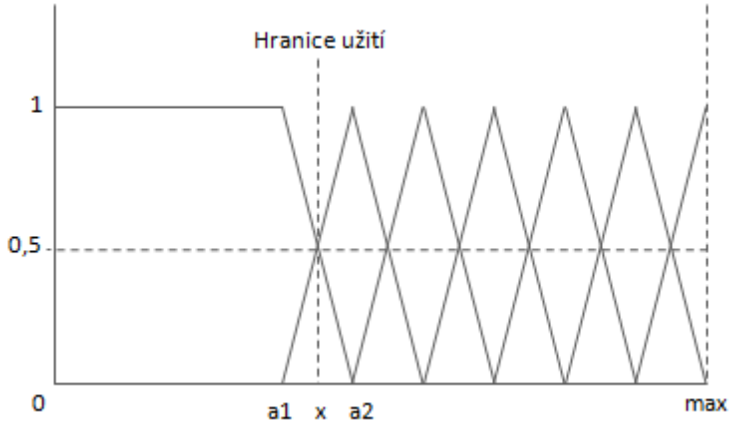
Největší slabinou CADAQUES v současné verzi je nízká podpora pro statistickou analýzu v rámci průběhu auditu. Je sice možné zvolit způsob náhodného výběru dat, součástí automatického vyhodnocování kvantitativních vlastností je výpočet základních popisných statistik, ale pokud nepovažujeme za statistický nástroj integrovaný formulář pro vkládání SQL (Structure Query Language) příkazů, je v této oblasti funkcionalita zcela nesrovnatelná s jinými nástroji.

Lee a kol. (2006) představuje značně inspirativní materiál pro definici konkrétních otázek kladených v rámci kvalitativního šetření. Publikace obsahuje řadu otisků obrazovek z konkrétního realizovaného assessmentu. Ze zkušenosti doporučuje zařazení i takových otázek, které směřují k objektivním vlastnostem dat, jejichž úroveň je zjišťována přímo analýzou dat. Současným zařazením do kvalitativního

šetření je možné porovnat vnímání těchto vlastností uživateli s realitou. I v tomto směru jsem se inspiroval a v rámci zobrazení výsledků auditu je možné u vybraných vlastností jejich subjektivní i objektivní hodnocení porovnat.

Nutno poznamenat, že formulář metadat auditu svou strukturou odpovídá formálním požadavkům §20 zák. 93/2009 Sb. o auditorech (Česko, 2009), standardu S7-Reporting a návodu G20-Reporing publikovanými ISACA (Information Systems Audit and Control Association) v (ISACA, 2003) a (ISACA, 2004). Důvodem je zajištění možnosti snadného automatického vygenerování zprávy v požadované struktuře v závěru auditu.

Jak bylo již naznačeno v (Pejčoch, 2011), pro další práci s informacemi zjištěnými v rámci auditu je vhodné spojité proměnné fuzzifikovat. Experimentální fuzzifikace nad vlastnostmi z reálného auditu potvrdila domněnku, že ne ke všem vlastnostem je možné v rámci fuzzifikace přistupovat stejně napříč všemi atributy. V některých případech, např. při srovnání úplnosti u rodného čísla a čísla mobilního telefonu, nejen že prahová hodnota pro fuzzy množinu leží jinde z toho důvodu, že zarážející v případě rodného čísla je již výskyt chybějících hodnot nižší než o10 %, než je potenciální pokrytí (reálný výskyt chybějících hodnot je v případě rezidentů cca 0,5 %), zatímco v případě mobilního telefonu 10% deficit vlastnosti „úplnost“ není nic překvapivého. Uvedené atributy jsou rovněž dobrým příkladem pro demonstraci možné míry potenciálního pokrytí. V případě rodného čísla je mezi fyzickými osobami s trvalým bydlištěm na území ČR maximální pokrytí 100%. V případě čísla mobilního telefonu se pohybuje v průměru někde na hranici 90 % a se zvyšující věkovou skupinou se též zvyšuje četnost jedinců, kteří mobilní telefon nevlastní. Ještě nižší potenciální pokrytí bychom byli nuceni uvažovat v případě e-mailové adresy. Uvedené skutečnosti mě vedly k tomu, abych do QKB též ke každému sémantickému typu přiřadil apriorní znalost o použitém způsobu fuzzifikace. Obecné schéma použité pro vlastnosti dat je znázorněno na Obrázku č. 2. Jeho aplikace na různé atributy se liší pouze reálnou výší potenciálního pokrytí vyjádřenou hodnotou max a výší označenou jako „Hranice užití“ odpovídající úrovni vlastnosti, při níž již není myslitelné daný atribut využívat pro žádné zamýšlené užití.



Obrázek 1: Fuzzy množiny použité pro vlastnosti dat

Váhy příslušnosti k jednotlivým fuzzy množinám jsou definovány podle vzorce

$v_1 = (x - a_1)/(a_2 - a_1)$, resp. $v_2 = 1 - (x - a_1)/(a_2 - a_1)$,

kde
vi – míra příslušnosti k fuzzy množině *Si*,
x – přesná hodnota atributu,
ai – hranice fuzzy množin na ose x.

Jednotlivým množinám odpovídá následující popis pojmy přirozeného jazyka: Vynikající, Velmi dobrá, Dobrá, Spíše horší, Špatná, Velmi špatná, Kritická.

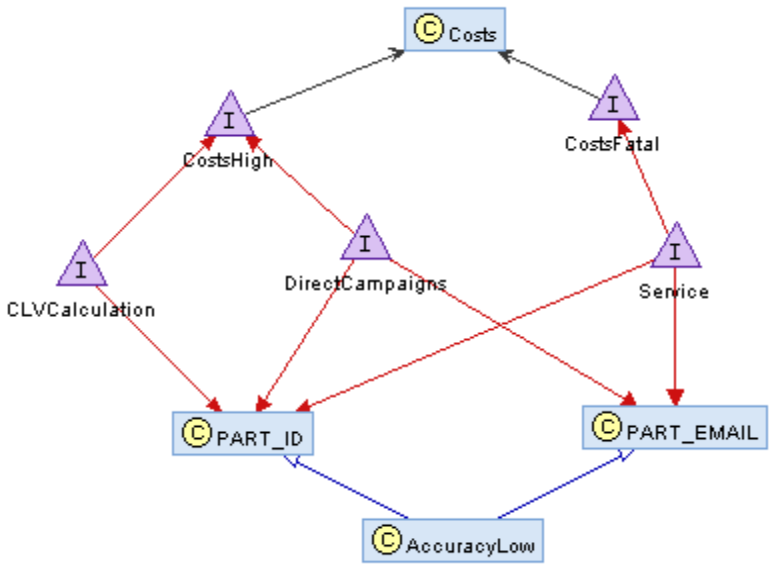
V případě aplikace na fuzzifikaci finančních nákladů se tvar použitých fuzzy množin liší pouze v tom směru, že schéma je symetrické podle mediánu. Platí tedy, že stejně tak jako když při dosažení určité dolní hranice je příslušnost ke kvalitativně „nejhorší“ množině rovna jedné pro všechna *x* nižší, než je prahová hodnota (viz Obrázek č. 2), i pro taková *x*, která jsou vyšší než prahová hodnota na opačném

konci spektra, je již rovna jedné příslušnost k fuzzy množině kvalitativně „nejlepší“. Jednotlivým množinám odpovídá následující popis pojmy přirozeného jazyka: Velmi vysoké, Vysoké, Spíše vyšší, Střední, Spíše nižší, Nízké, Velmi nízké.

Apriorní náklady v QKB jsou chápány jako náklady jednotkové, které lze při kalkulaci celkových nákladů vážit mírou příslušné vlastnosti jako odhadem pravděpodobnosti výskytu a počtem jednotek, u nichž k daným jednotkovým nákladům dojde. V řadě případů je však nutné uvažovat jiný přístup, který vychází z principu existence KO kritérií pro výskyt nákladů v určité výši. V některých případech, např. u nedodržení compliance s určitou normou, nastávají náklady v konkrétní výši bez ohledu na skutečnost, jaká konkrétní úroveň dané vlastnosti dat byla reálně naměřená. Ve znalostní bázi je tato skupina nákladů označena příznakem.

Apriorní pravidla jsou v QKB uložena jako asociační pravidla ve struktuře definované čtyřmi na sebe navázanými tabulkami. Základ této struktury tvoří tabulka hypotéz. Obsahuje základní metadata hypotéz a jejich vazbu na příslušnou vertikálu. Ve vztahu N:1 je k tabulce hypotéz tabulka jednotlivých cedentů. Obsahuje označení typu cedentu (Antecedent, Succedent, Condition) a vazbu na jednotlivé dílčí cedenty uložené v další tabulce (k tabulce cedentů ve vztahu N: 1). Poslední prvek báze uložených pravidel tvoří tabulka jednotlivých literálů, která je ve vztahu N:1 k tabulce dílčích cedentů. Tabulky cedentů a dílčích cedentů obsahují informaci, zda jsou v rámci nich prvky nižší hierarchie uvažovány jako spojené konjunkcí či disjunkcí. Jelikož dílčí cedenty uvažuji jako jednotlivé atributy Kanonického modelu, odkazuje se tabulka dílčích cedentů přímo na primární klíč příslušné tabulky modelu.

Výsledky auditu jsou v současné době pro vizualizaci vybírány na základě manuálního označení jednotlivých pravidel seřazených podle míry dopadu. Do budoucna plánuji jejich automatickou extrakci. Jelikož možnosti vizualizace kauzalit jsou v případě jazyka PHP poněkud omezené, rozhodl jsem se pro zobrazení výsledných pravidel použít externí aplikaci. Abych zajistil co největší míru abstrakce od konkrétní aplikace, CADAQUES umožňuje export výsledků auditu do formátů RDF (Resource Description Framework) a OWL (Web Ontology Language) V obou případech se jedná o aplikace jazyka XML (eXtensible Markup Language). OWL je založený na základních elementech RDF, ale poskytuje více pojmů pro popis vlastností a tříd. Pro uvedené formáty jsem též našel vhodnou aplikaci pro vizualizaci; v případě RDF např. RDF Gravity a v případě OWL nástroj Protégé (resp. jeho zásuvné moduly NavigOwl a OntoGraf). V obou případech se jedná o volně dostupnou aplikaci. Příklad vizualizace výsledků auditu ve formátu RDF pomocí nástroje RDF Gravity znázorňuje Obrázek č. 2.



Obrázek 2: Výsledky vizualizace pomocí nástroje RDF Gravity

Závěr

V rámci své výzkumné činnosti jsem se pokusil zkombinovat hlavní rysy stávajících přístupů pro audit datové kvality a na výsledném konceptu založit vlastní aplikaci spadající do kategorie (CAAT – Computer-Assisted Audit Techniques). V této snaze se cítím být stále ještě na začátku. Řada modulů aplikace je sice již v tuto chvíli hotova, ale mnoho funkcionality dostupné v rámci standardních nástrojů pro podporu auditu zbývá ještě vyvinout.

Jedním ze zamýšlených směrů dalšího rozvoje je přímá integrace nástroje pro vizualizaci výsledků do aplikace. S tím souvisí i experimentování s dalšími alternativními aplikacemi jazyka XML. Jako možná cesta se nabízí jazyk SWRL (Semantic Web Rule Language) nebo přímo jazyk RuleML (Rule Markup Language), ze kterého je SWRL odvozen. Další zamýšlenou oblastí rozvoje je již zmiňovaná automatická extrakce pravidel z výsledků auditu.

V současné době má začleňování apriorních znalostí z QKB převážně podobu možnosti importu pravidel z báze. Do budoucna uvažuji rozšíření těchto možností o funkcionalitu inteligentního agenta, který by v průběhu vytváření auditu „našeptával“ jak při daném kroku postupovat, přičemž by následná doporučení upravoval podle toho, zda se auditor rozhodl jeho předchozí rady poslechnout či vybral alternativní variantu.

Obecně zvažuji převedení aplikace do jiného programovacího jazyka. Vzhledem k množství volně dostupných potenciálně využitelných knihoven vytvořených v jazyku Java (např. knihovna Simmetrics pro míry podobnosti řetězců) vážně uvažuji o J2EE (Java 2 Enterprise Edition).

Literatura

COMPUTER ASSISTED AUDIT GROUP. A Guide to Computer Assisted Audit Techniques, In: Commonwealth of Massachusetts [online]. @2009[cit. 2012-01-20] Dostupné z: http://www.mass.gov/dor/docs/dor/publ/pdfs/caat.pdf

ENGLISH, Larry P. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. New York: Wiley & Sons, 1999. xxvi, 518 s. ISBN-10 0-471-25383-9.

FEIGENBAUM, E. A. Themes and Case Studies of Knowledge Engineering. In Expert Systems in the Micro-Electronic Age. *Edinburgh: University Press, 1979.*

CANNON, David, L. *CISA Certified Information Systems Auditor Study Guide*, Second Edition. Indianapolis: Wiley, 2008. ISBN 978-0-470-23152-4.

GOSMAN, S., ET AL. Umělá inteligence a expertní systémy. Výběr informací z organizační a výpočetní techniky. *Praha, 1990.*

HAMŘÍK, A., STRÁNSKÝ, M. Auditorská rizika a postupy vyplývajícíc z využití ICT v účetnictví. In: PWC [online]. 2008-06-04[cit. 2012-01-29]. Dostupné z: http://www.pwc.com/cz/cs/clanky-2008/auditorska-rizika-a-postupy-vyplyvajici-z-vyuziti-ict-v-ucetnictvi.jhtml

ISACA. IT Audit and Assurance Guideline G20 Reporting. In: ISACA [online]. @2003, 2010[cit. 2012-01-22]. Dostupné z: http://www.isaca.org/Knowledge-Center/Standards/Documents/G20-Reporting-13Aug2010.pdf

ISACA. IS Auditing Standard Reporting Document #S7. In: ISACA [online]. @2004[cit. 2012-01-22]. Dostupné z: http://www.isaca.org/Knowledge-Center/Standards/Documents/Audit20Standards/S7ReportingStandardRevised2Mar05.pdf

ISACA. IS Auditing Guideline G3 Use of Computer-Assisted Audit Techniques (CAATs). In: ISACA [online]. ©1998, 2008[cit. 2012-01-29]. Dostupné z: http://www.isaca.org/Knowledge-Center/Standards/Documents/G3-Use-of-CAATs-16Jan08.pdf

LEE, Yang W., PIPINO, Leo L., FUNK, James D., WANG, Richard Y. *Journey to Data Quality*. The MIT Press, 2006. 240 s. ISBN-10 02-621-2287-1.

LOSHIN, D. *The Practitioner's Guide to Data Quality Improvement*. Burlington: Morgan Kaufmann as inprint of Elsevier, 2011. ISBN 978-0-12-373717-5.

McGILVRAY, D. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann, 2008. xviii, 325 s. ISBN 978-0-12-374369-5.

PEJČOCH, D. Vztah řízení dat k ostatním oblastem řízení informatiky. In: *Sborník prací účastníků vědeckého semináře doktorského studia*. Praha: Fakulta informatiky a statistiky VŠE, 2011. ISBN 978-80-245-1761-2.

PEJČOCH, D. Benchmark přístupů k Fuzzy Match / Merge. In: *Sborník prací účastníků vědeckého semináře doktorského studia*. Praha: Fakulta informatiky a statistiky VŠE, 2009. ISBN 978-80-245-1524-3.

SAYANA, S. Anantha. Using CAATs to Support IS Audit. In: ISACA Journal [online]. ©2003[cit. 2012-01-28]. Dostupné na: http://www.isaca.org/Journal/Past-Issues/2003/Volume-1/Pages/Using-CAATS-to-Support-IS-Audit.aspx

SVATÁ, V. *Audit informačního systému*. Příbram: Professional Publishing, 2011. ISBN 978-80-7431-034-8.

ŠTUMPF, J., DŽMURÁŇ, M. Datová integrace prostřednictvím společného datového modelu. In: *Proceedings of the 16th International Conference on Systems Integration 2008, Prague, Czech Republic, June 10 – 11*. Praha: CSSI, 2008, ISBN 978-80-245-1373-7

ČESKO. Zákon č. 93 ze dne 26. března 2009 o auditorech a o změně některých zákonů (zákon o auditorech). In: *Sbírka zákonů České republiky*. 2009, částka 30, s. 1166-1204. Dostupný také z: www.mvcr.cz/soubor/sb030-09-pdf. ISSN 1211-1244.

JEL Classification: M15, M42

Summary

Application supporting audit of data quality

This article summarizes the current approaches to data quality audit. It follows the concept of mapping levels of data characteristics to their impacts using causal maps which has been published in (Pejčoch, 2011). At first it provides a brief introduction to the typical functionality available within the general tools for Computer-Assisted Audit Techniques (CAAT) and relevnat functionality contained in the tools for data quality management. Later there is presented CADAQUES (Complex Audit of Data Quality in Enterprise Systems), the application being developed by the author of this work. The application is currently implemented using the programming languages PHP, Perl and Java Script. As the primary repository for data it use MySQL database platform. The base of CADAQUES is represented by Knowledge Base (QKB) containing a priori information about the relationships between properties of data, methods of usage of individual data attributes and associated costs calculated for different levels of characteristics. It also includes the semantic data type dependent grammars, phonetic libraries, dictionaries, standardization schemes and libraries of regular expressions allowing the creation of matching-codes used for comparative measurement of the uniqueness of data and verification against external data sources.

Keywords: Data Quality Assessment, Data Quality Audit, Common Data Model, CAAT, CADAQUES.

Podpora rozhodování v reálném čase z hlediska datové integrace

Lukáš Pitka
xpitl05@vse.cz
Doktorand oboru Informatika
Školitel: prof. Ing. Josef Basl, CSc., (basl@vse.cz)

Abstrakt: Článek se zabývá problematikou datové integrace v oblasti podpory rozhodování v reálném čase. Cílem příspěvku je identifikovat aktuální technologické překážky datové integrace v reálném čase. Druhým cílem je představení konceptuálního modelu, který tyto problémy a překážky minimalizuje.

Real-time podpora rozhodování vnáší do oblasti integrace dat z různých datových zdrojů nové překážky. V úvodu článku je objasněna problematika podpory rozhodování v reálném čase.

Druhá část článku se věnuje otázce potřebnosti rozhodování se v reálném čase. V kapitole jsou uvedeny rozdíly mezi rozhodováním v reálném čase a rozhodováním ve správný čas. Kapitola také obsahuje praktické příklady potřeby rozhodování se v reálném čase.

Dále se příspěvek detailněji zabývá aktuálním stavem v problémové oblasti a nabízí grafické porovnání určení jednotlivých dostupných technologií a přístupů.

Ve čtvrté kapitole jsou určeny slabiny aktuálních přístupů k datové integraci z hlediska potřeby integrovat data v (téměř) reálném čase.

Článek dále obsahuje návrh pro řešení překážek identifikovaných v předchozí kapitole. V této části je popsán konceptuální model pro datovou integraci v reálném čase, který vychází z konceptu Data Warehouse 2.0. Koncept datového skladu druhé generace, resp. primárně jeho část Interactive Sector, je rozšířena tak, aby byly co nejvhodněji podpořeny možnosti datové integrace v reálném čase.

Potřeba návazného výzkumu a detailizace některých oblastí je uvedena v závěru článku.

Klíčová slova: podpora rozhodování v reálném čase, real-time BI, datová integrace, federace, EII, CDC, CTF, Data Warehouse 2.0.

Úvod

Zrychlující se tržní prostředí nevyhnutelně vede k potřebě rychlejšího rozhodování na všech úrovních organizace. Těžko předvídatelné, rychle se měnící, silně konkurenční a zákaznicky orientované tržní prostředí přináší zcela nové nároky na osoby v rozhodovacích pozicích v organizacích. Na měnící tržní prostředí musí reagovat také oblast IS/ICT, kde v současné době nalézáme řadu inovačních trendů (Voříšek, 2008).

Výše uvedené změny a inovační trendy vyvolávají značný tlak na osoby v rozhodovacích pozicích. Na všech úrovních organizace je zapotřebí provádět rozhodnutí rychleji, ale zároveň kvalitněji. Pokud organizace neučiní rozhodnutí o zavedení či inovaci výrobku nebo služby dostatečně rychle a na základě správných a aktuálních podkladů, může to znamenat pro společnost značné ztráty, případně i existenční problémy. V současné době se stále více hovoří o tzv. *Real-Time Enterprise* (Fingar a Bellini, 2004). Tento koncept zavedený analytickou společností Gartner (Soejarto, 2002) přináší organizaci teoretické přínosy na všech úrovních řízení:

- zlepšení zákaznické zkušenosti, snížení stavu zásob, snížení rizik a nižší procesní náklady na operativní úrovni,
- na taktické úrovni se jedná o rychlejší využití vznikajících obchodních příležitostí, snížení dopadu v případě vzniku kritických situací a zvýšení rychlosti reakce při vzniku ohrožení,
- strategická úroveň je podpořena zejména rychlejším stanovením a zavedením strategií, které se pružně přizpůsobují měnícímu se tržnímu prostředí.

Koncept real-time enterprise vyžaduje velmi rychlou a kvalitní podporu v podobě nástrojů pro podporu rozhodování, ať už se jedná o rozhodování automatizované či rozhodování prováděné fyzickou

osobou. V souvislosti s real-time enterprise a související podporou rozhodování v reálném čase se často zmiňují pojmy BPM (Business Process Management), BI (Business Intelligence), CI (Customer Intelligence), CPM (Corporate Performance Management) a řada dalších.

V souvislosti s real-time enterprise ale také vyvstávají problémy a otázky, které je třeba řešit. Detailně jsem se jednotlivým otázkám a problémům v této oblasti věnoval v článku (Pitka, 2011). Identifikoval jsem dvě základní skupiny problémů, které souvisejí s potřebou rozhodování se v reálném čase jakožto nedílnou součástí konceptu real-time enterprise.

První skupinou jsou překážky organizační související s organizační strukturou, definicí procesů a obecně se strategií řízení organizace. Druhou skupinu překážek představují překážky technologické. Všechny operace s daty musí probíhat v téměř reálném čase tak, aby bylo možné se v reálném čase také rozhodovat.

Tento příspěvek je koncepčně postaven na rozsáhlejší studii (Pitka, 2011) a obsahuje nejpodstatnější části z něj. Hlavní důraz je kladen na objasnění koncepce systému pro real-time datovou integraci.

Cílem příspěvku je:

- identifikovat a jasně pojmenovat současné technologické problémy a překážky v oblasti analytických systémů pro podporu rozhodování,
- navrhnout a objasnit koncepční řešení, které zmírní jednotlivé překážky v oblasti datové integrace pro podporu real-time analytických a rozhodovacích systémů.

Pro dosažení výše uvedených cílů je příspěvek rozčleněn následovně: v následující části se stručně zaměřím na potřebu podpory rozhodování v reálném čase, tj. na otázku, zda se organizace potřebují rozhodovat v reálném čase. Další kapitola se zabývá aktuálním stavem v oblasti datové integrace a možnostmi využití těchto technologií pro podporu rozhodování v reálném čase. Následující část příspěvku obsahuje analýzu aktuálních problémů a překážek souvisejících s datovou integrací v prostředí real-time podpory rozhodování. Další kapitola obsahuje návrh směrů řešení a dalšího výzkumu, který směřuje k překonání těchto překážek, či alespoň ke zmírnění problémů. V závěru bude provedena rekapitulace příspěvku včetně specifikace rozšíření představeného koncepčního modelu.

Potřeba analytické podpory rozhodování v reálném čase

Článků zabývajících se real-time analýzou, real-time podporou rozhodování, real-time organizací či real-time BI je celá řada. Tyto články se však velmi zřídka zabývají otázkou, zda je podpora rozhodování v reálném čase pro organizaci skutečně potřebná.

Real-time nebo right-time?

Spojení *real-time* nahrazuje Eckerson (2006) termínem *right-time*, tj. potřebou dodání podkladů pro rozhodnutí ve správný okamžik. Jaký je ovšem ten správný okamžik pro dodání této informace? Jedná se o co nejrychlejší dodání? Nebo o co nejvýhodnější? Nejvýhodnější pro zákazníka? Nebo jde o dosažení ekonomicky nejvýhodnějšího času pro organizaci (např. dle Hackathornova (2004) modelu *Time Value Curve*)? Termín real-time naopak přináší polemiku o tom, co znamená „dodání podkladů v reálném čase“ – vždy dochází ke vzniku určitého zpoždění. V dalším textu se budeme držet rozšířenějšího spojení *real-time podpora rozhodování*.

Potřeba podpory rozhodování v reálném čase začíná být v současné době vnímána jako důležitý směr rozvoje analytických nástrojů. Většina výzkumných úvah se zabývá speciálními aplikacemi v této oblasti, tj. např. využitím pro monitorování, analýzu a automatickou reakci v rámci rozvodné sítě elektrické rozvodné soustavy. Obecnější implementace myšlenky real-time podpory rozhodování v jakémkoli podnikovém procesu prozatím není realizována.

Dle výzkumu analytické společnosti Aberdeen Group (Hatch, 2008) je možné oblast real-time podpory rozhodování dále rozdělit do několika skupin (detailněji v (Pitka, 2011; Hatch, 2008), v následujícím seznamu jsem uvedl pouze oblast a časový interval, pro který je oblast určena):

- *Transactional Business Intelligence*: ad-hoc či denní časový interval,
- *Real-Time Analysis*: sekundy, desetiny sekund,

- *Near Real-Time Analysis*: sekundy až hodiny,
- *Operational Reporting*: jakýkoli časový interval,
- *BAM, BPM*: minuty, sekundy, desetiny sekund,
- *Decision Management*: jakýkoli časový interval.

Praktické příklady využití real-time podpory rozhodování

Následující přehled ukazuje několik konceptů, které využívají podporu rozhodování v reálném čase:

- plánování výrobního programu při zpracování kovů a nerostných surovin,
- automatický systém monitoringu a rozhodování v oblasti stavebnictví a stavebních konstrukcí,
- rozhodování v oblasti kritické péče o pacienta,
- detekce podvodného chování (Fraud Detection),
- call centra,
- automatické monitorování rozvodných sítí,
- monitorování a rozhodování v případě vzniku živelných katastrof – evakuační plány.

Detailnější informace o použití v jednotlivých oblastech jsou uvedeny v (Pitka, 2011).

Aktuální stav datové integrace v reálném čase

Nástroje označované výrobci jako *Data Integration Suite* zpravidla obsahují několik konceptů datové integrace. Běžně je zahrnuta datová konsolidace (Data Consolidation), propagace dat (Data Propagation) a datová federace, neboli virtuální konsolidace dat (Data Federation). Další možný přístup za pomocí vyhledávacích indexů je nazýván Data Access.

Datová konsolidace zahrnuje zachycení dat z různých datových zdrojů a jejich integraci do jednoho trvalého datového úložiště. Časová prodleva závisí na frekvenci spouštění tzv. datové pumpy, zpravidla nazývané ETL (Extraction, Transformation, Loading), nebo ELT (Extraction, Loading, Transformation). Standardní technika datové konsolidace není vhodná pro datovou integraci v reálném čase – dávkové procesy je nutné nahradit pružnějším mechanismem aktualizace dat v datovém skladu.

Datová propagace zahrnuje replikaci dat z různých datových zdrojů do různých cílových databází. Používané technologie jsou replikace či CDC (Change Data Capture). Technologie CDC je do určité míry vhodná jako základ pro tzv. mikrodávková ETL (Gathibandhe a kol., 2010; Levy, 2009; White, 2005). Tuto technologii podrobněji rozeberu v následující kapitole.

Virtuální konsolidace dat poskytuje unifikovaný pohled na data z různých datových zdrojů. Data ovšem nejsou nikam přenášena a zůstávají v databázích primárních informačních systémů, popř. v jiných typech zdrojových souborů. Technologickým představitelem datové federace je technologie EII (Enterprise Information Integration), kterou jsem podrobně popsal v (Pitka, 2007).

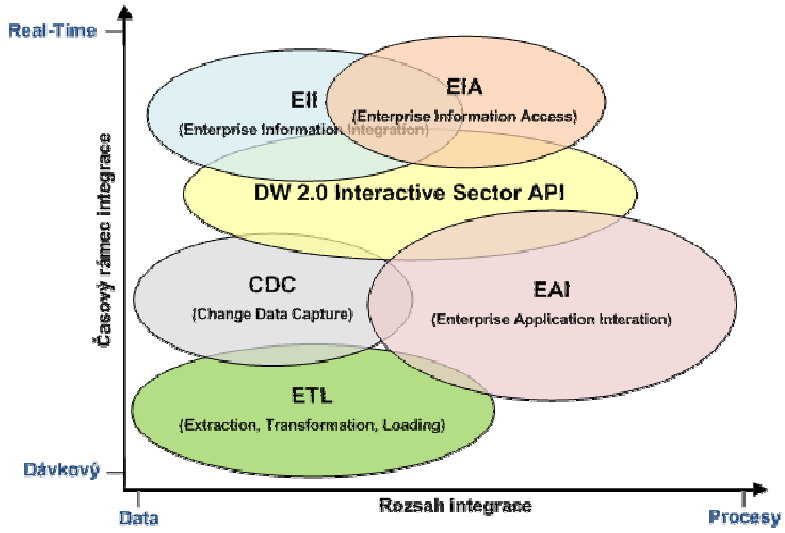
Metoda datové integrace nazvaná *Data Access* je založena na principu zpřístupnění dat, který je podobný prohledávání obsahu Internetu (indexy, agregace, cache). Komerční technologie je nazývána EIA (Enterprise Information Access).

Tyto stávající technologie do určité míry podporují datovou integraci v reálném čase. EII je primárně určeno pro práci s aktuálními daty primárních podnikových systémů, což ovšem přináší na druhé straně několik problémů – výkonnost, zachycení změnových dat, náročnost na správu metadat. Technologie CDC založená na zachycení změnových dat představuje vhodnou komponentu do celkové architektury systému pro podporu rozhodování v reálném čase. Aktuální pojetí technologie CDC je ale založeno na dávkovém zpracování ETL při přenosu dat do centrálního datového skladu.

Potřebu analýzy dat v reálném čase si uvědomili také tvůrci původního konceptu datových skladů a vznikla myšlenka tzv. DW 2.0 (Inmon, Strauss a Neushloss, 2008). Architektura datového skladu nové generace přináší mj. rozdělení datového skladu na sektory – jedním z nich je *Interactive Sector*, který slouží pro zpracování dat a jejich analýzu pro potřebu podpory rozhodování v reálném čase.

Obrázek obsahuje přehled aktuálních integračních technologií, které jsou uvedeny výše. Do schématu jsem pro úplnost zařadil také technologii aplikační integrace EAI (Enterprise Application Integration). Schéma jsem vytvořil rozšířením a úpravou grafu z článku (Imhoff, 2005).

Schéma zobrazuje jednotlivé aktuální technologie a konceptuální návrhy nových technologií ve dvourozměrném grafu. První rozměr (vodorovná osa) představuje *Rozsah integrace*, tj. zda se jedná o technologii datové integrace, aplikační integrace či procesní integrace. Vzhledem k tématu této práce je většina technologií v levé polovině schématu. Druhý rozměr ukazuje *Časový rámec integrace*, tj. zda se jedná o technologii dávkové integrace či integrace v reálném čase.



Obrázek 1: Přehled aktuálních/konceptuálních integračních technologií Zdroj: Pitka, 2011

Analýza přístupů k datové integraci z real-time hlediska

V předchozí části jsem uvedl aktuální možnosti v oblasti datové integrace. Potřeba rozhodování v reálném čase ovšem vnáší do oblasti datové integrace nové požadavky, na které není žádná ze stávajících technologií plně připravena. Požadavky na tzv. **Real-Time Data Warehouse** jsou poměrně jasně specifikovány např. v (Bruckner, List a Schiefer, 2002) nebo v (Inmon, Strauss a Neushloss, 2008). *Požadavky na komponentu plnící datový sklad ovšem již tak jasně specifikovány nejsou*.

Technologické požadavky datové integrace v reálném čase

Následující výčet obsahuje specifikaci nejdůležitějších požadavků datové integrace v reálném čase:

- získání, transformace a uložení dat pro potřeby následné analýzy v téměř reálném čase,
- trvalý a nepřetržitý tok relevantních dat z primárních systémů do analytického úložiště dat,
- trvalý a nepřetržitý běh centrálního datového úložiště (na rozdíl od dávkového nahrávání dat není možný výpadek funkčnosti analytického řešení),
- synchronizace datových entit napříč všemi systémy zákazníka, tj. možnost získat např. komplexní objekt zákazníka z několika primárních systémů s minimální latencí,
- zajištění kvality dat tak, aby do centrálního úložiště k analytickému zpracování vstupovala v maximální možné kvalitě,
- minimální nároky na primární systémy při extrakci dat z jejich databází,
- škálovatelnost řešení musí zajistit schopnost zpracovat tzv. mixed workloads, tedy zpracování velmi rozdílných objemů dat,
- architektura datového skladu musí být připravena na potřebu real-time analýz, historických analýz či analýz nad historickými a aktuálními daty zároveň, tj. na integraci dat různé povahy v rámci datového skladu.

Výše uvedené požadavky představují dle mého názoru v současné době základní rozdíly mezi tradičním řešením pro podporu rozhodování a řešením pro podporu rozhodování v reálném čase. Technologie uvedené v předchozí kapitole některé z těchto požadavků řeší, ovšem žádná z nich je komplexně

nepokrývá. V následujících částech se budu podrobněji (detailněji jsem jednotlivé oblasti rozebral v (Pitka, 2011)) zabývat těmito dílčími technologiemi a jejich problémy z hlediska datové integrace v reálném čase.

Aktuální metody datové integrace z pohledu real-time zpracování

V následujících částech jsou uvedeny překážky jednotlivých technologií uvedených v předchozí kapitole z hlediska jejich připravenosti ke zpracovávání dat v reálném čase. Obrázek obsahuje navíc metody EAI a DW 2.0. Metoda EAI není dále rozpracována - nejedná o metodu datové integrace. Interaktivním sektorem DW 2.0 se budu zabývat v následující kapitole, v současné době se jedná pouze o koncept, který není použit v reálných aplikacích.

Detailní specifikace jednotlivých přístupů a technik datové integrace včetně schémat řešení, základních komponent a odkazů na literaturu je uvedena v (Pitka, 2011).

ETL (Extraction, Transformation, Loading)

Aktuální překážky bránící v použití tradičního ETL konceptu při datové integraci v reálném čase jsou dle mého názoru zejména následující:

- ETL nástroje jsou založené na dávkovém zpracování dat, nikoli na průběžném zpracování změnových dat, na kterém je založena datová integrace v reálném čase.
- Spouštění ETL úloh je naplánováno mimo špičku. Z tohoto důvodu nejsou ETL úlohy příliš optimalizovány pro minimální zátěž zdrojových databází.
- Při provádění ETL úloh je centrální datový sklad pravidla nedostupný pro provádění analytických operací. Real-time datová integrace probíhá prakticky neustále, není možné jakkoli omezit uživatele analytického řešení znepřístupněním centrálního úložiště dat. Potřeba max. přenosu dat do centrálního úložiště zpravidla koliduje se špičkou zatížení primárního IS.
- Datová čistota (transformace a čištění dat představuje časově nejnáročnější operaci v ETL) je v tradičních ETL nástrojích řešena během přenosu dat do centrálního úložiště. V případě přenosu dat v reálném čase vzniká problém s vyčištěním dat dostatečně rychle.

EII (Enterprise Information Integration)

Z hlediska datové integrace pro real-time podporu rozhodování brání v nasazení EII následující:

- EII značným způsobem zatěžuje primární IS. Veškeré dotazy jsou prováděny přímo nad produkčními databázemi, jejichž uživatelé by měli být od datové integrace pro analytické účely odstíněni. Problém by byl zvládnutelný při optimalizovaných standardních dotazech, ovšem v případě ad-hoc analýz mohou dotazy způsobit značné výkonnostní problémy.
- EII neřeší historii dat a jejich změny. Data primárních informačních systémů zpravidla neumožňují časové analýzy. EII data nikam nepřenáší a pracuje tedy s daty v nevhodné formě pro analytické zpracování.
- EII pro potřeby podpory rozhodování v reálném čase není pružné ke změně struktury dat primárních systémů. V případě tradičních řešení s meziúložištěm ve formě datového skladu je tento problém řešitelný.
- Datová kvalita představuje v případě EII stejný problém jako v předchozím případě real-time ETL – data primárních systémů je nutné v reálném čase transformovat a vyčistit. V případě EII je situace ještě složitější tím, že pracujeme pouze s virtuálními daty – data primárních systémů je možné měnit pouze ve velmi omezené míře.

EIA (Enterprise Information Access)

Technologie EIA není pro analýzu dat v reálném čase vhodná. Slouží k naprosto jiným účelům, jedná se pouze o technologický způsob prohledávání podnikového obsahu, nikoli o jeho přesun, agregace a přípravu pro analytické zpracování. V rámci výčtu jsem tuto metodu uvedl pouze pro úplnost.

CDC (Change Data Capture)

Jako překážky využití CDC pro datovou integraci v reálném čase vnímám zejména:

- Log záznamy primárních systémů nejsou určeny pro zpracování jiným systémem. I když jsou tyto záznamy zpravidla uložené v relační databázi, jejich struktura není vhodná pro zjišťování změn v datech/zjišťování nových dat.
- Aplikační log je velmi rozsáhlý a operace nad ním mohou způsobit výkonnostní potíže IS.
- CDC je určeno pouze pro strukturovaná data. V případě nestrukturovaných dat neexistuje transakční log změn. Potřeba real-time analýzy nestrukturovaných dat ale není v současné době primárním požadavkem v oblasti podpory rozhodování v reálném čase.
- CDC řeší pouze nalezení změn v databázi transakčního systému, ale již se nezabývá přenosem těchto změn do centrálního datového skladu. Pro tyto účely je nutné využít jiné technologie, např. ETL, která přenese data z CDC databáze do centrálního datového skladu.

Konceptuální model komplexního řešení datové integrace

Všechny tři relevantní způsoby (ETL, EII, CDC) datové integrace uvedené v předchozí kapitole mají specifický způsob použití. Ani jedna z nich ale není sama o sobě vhodná pro real-time integraci.

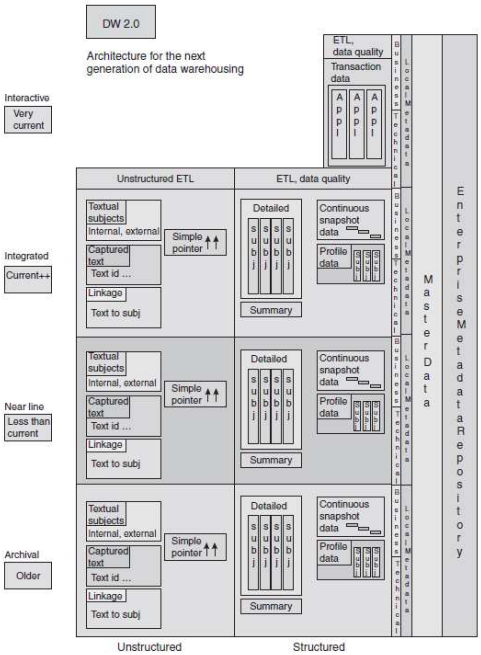
V této kapitole představím koncept architektury (detailněji v (Pitka, 2011)) datové integrace v reálném čase, přičemž tento bude v souladu s architekturou tzv. DW 2.0 (Inmon, Strauss a Neushloss, 2008).

Data Warehouse 2.0

Tradiční datové sklady přestávají dostačovat aktuálním potřebám organizací. Byla vytvořena koncepce datového skladu nové generace – *Data Warehouse 2.0* (Inmon, Strauss a Neushloss, 2008), která reaguje na následující změny v prostředí, ve kterém organizace působí:

- požadavek na intenzivnější a širší využití technologií zpracování informací,
- *online zpracování dat, tj. zpracování, analýza dat a reakce na události v reálném čase,*
- požadavek po integrovaných korporátních datech, tj. odstranění jednotlivých datových sil,
- požadavek na zpracování a analýzu semistrukturovaných a nestrukturovaných dat,
- zvyšování výkonu IT – je požadováno sofistikovanější zpracování a analýza dat rychleji,
- na rozdíl od zvyšování výkonu technologií klesá jejich cena – jednotka výkonu je neustále levnější, což tvoří příznivé prostředí pro inovace.

Z hlediska tématu článku je nejdůležitější vyznačený požadavek na online zpracování dat. Architektura datového skladu dle koncepce DW 2.0 je uvedena na schématu Obrázek .



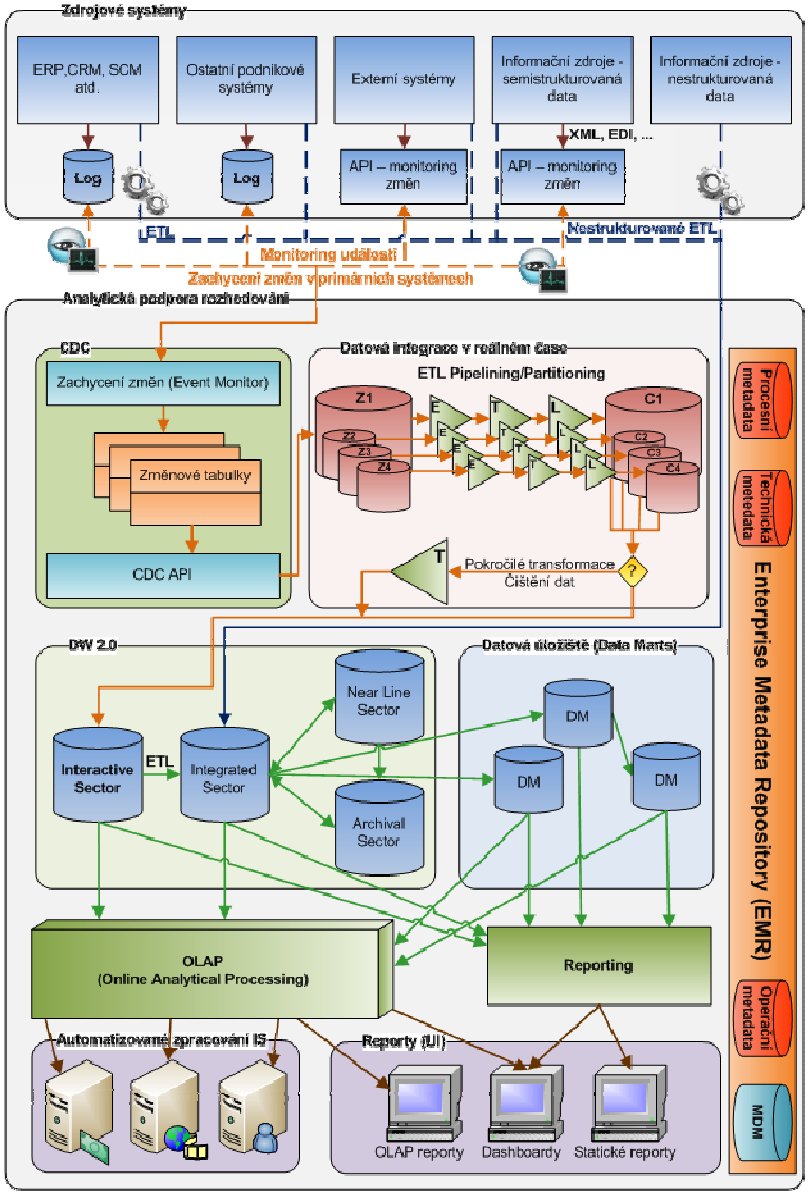
Obrázek 2: Architektura Data Warehouse 2.0 Zdroj: Inmon, Strauss a Neushloss, 2008, s. 25

Pro splnění požadavku real-time datových analýz přináší DW 2.0 tzv. *Interactive Sector*, který obsahuje aktuální data primárních informačních systémů. Data mohou do interaktivního sektoru DW 2.0 vstupovat dvěma způsoby. Prvním je mikrodávková ETL pumpa s metodou CDC. Druhý způsob představují aplikace implementované přímo v prostředí interaktivního sektoru DW 2.0.

Koncept řešení datové integrace v reálném čase

Koncepce DW 2.0 uvedená v předchozí části neobsahuje detailnější architekturu datové integrace v reálném čase. ETL datová pumpa přenášející data do Integrated Sector je standardní dávková ETL úloha.

Rozdělení datového skladu do několika sektorů dle aktuálnosti a rozdílných požadavků na zpracování dat je vhodné. Do celkové architektury je ovšem nutné doplnit prvotní vstup dat do interaktivního sektoru, tj. koncepční způsob datové integrace mezi transakčními systémy a interaktivním sektorem. Komplexní koncepci celkového řešení analytické podpory rozhodování se zapojením DW 2.0 jsem znázornil na Obrázek . Koncepce představená tímto schématem přináší výrazné zmírnění problémů datové integrace pro podporu rozhodování v reálném čase uvedených v předchozí kapitole.



Obrázek 3: Koncepce analytických systémů pro podporu rozhodování v reálném čase Zdroj: Pitka, 2011

Zdrojové systémy obsahují data různých formátů – jedná se o data strukturovaná, nestrukturovaná a semistrukturovaná. Pomocí ETL a nestrukturovaných ETL procesů je možné data přímo přenášet do prostředí DW 2.0, konkrétně do integrovaného sektoru datového skladu. Standardní, dávkový přenos informací pomocí procesu ETL není v koncepci detailněji uveden. Historická data uložená v integrovaném sektoru je možné pro následnou analýzu zpřístupňovat přímo datovým skladem nebo prostřednictvím datových tržišť (Data Mart). Analýza a reporting historických dat je následně určen zejména koncovým uživatelům, kteří dostávají pravidelně aktualizované statické reporty, mají aktualizované dashboardy či vznášejí ad-hoc analytické dotazy. V určitých případech mohou být na základě těchto historických dat prováděna automatizovaná rozhodnutí v procesech IS.

Analytická podpora rozhodování v reálném čase, resp. datová integrace v reálném čase je ovšem velmi odlišná. Ve stěžejní části schématu jsem navrhnul koncepci přenosu dat z primárních informačních systémů do interaktivního sektoru DW 2.0.

Přenos dat v reálném čase kombinuje a rozšiřuje přístupy uvedené v předchozích kapitolách. Základní bloky této koncepce jsou CDC, ETL Pipelining/Partitioning a Metadata - detailněji v (Pitka, 2011).

Uvedený koncept zmírňuje problémy a překážky uvedené v předchozí kapitole. Jeho funkčnost v reálném prostředí je ovšem nutné ověřit detailní analýzou a implementací.

CDC

Úlohou části CDC je extrakce změnových dat z primárních systémů na základě monitorování a analýzy logů těchto systémů. Přístup k získání těchto nových/upravených dat je různý. Nejčastější možnost představuje analýza databázových logů, jinou možností je monitorování aplikačního rozhraní (API) systémů.

Komponenta *Event Monitor* provádí monitoring jednotlivých primárních systémů (za pomoci vrstvy metadat, která odstiňují zejména technologickou náročnost mapování na jednotlivé systémy). Změny zachycené v log tabulkách/log souborech/předané pomocí programového rozhraní API jsou následně uloženy prostřednictvím změnových tabulek v definované podobě a zpřístupněny pomocí CDC API k dalšímu zpracování v rámci řešení datové integrace v reálném čase.

ETL Pipelining/Partitioning

CDC API předává ke zpracování minimální množství dat, která je nutné v téměř reálném čase zpracovat. K tomu je v představené koncepci využito především možností ETL Pipelining a ETL Partitioning (Kozielski a Wrembel, 2009), resp. jejich vhodné kombinace.

Metoda *Pipelining* rozděluje úlohu ETL do několika paralelních subprocesů, kdy je možné simultánně data načítat (Extraction), zpracovávat (Transformation) a nahrávat do cílových systémů (Loading). Už v současné době navíc řada úloh ETL procesů probíhá v paralelních větvích, kdy jsou nezávislé oblasti dat zpracovávány simultánně. Metodu *Partitioning* je vhodné použít jako doplněk k ETL Pipelining. Jedná se o rozdělení datového souboru do dávek s menším počtem záznamů a jejich paralelní zpracování. Současné prostředky výpočetní techniky nejsou překážkou ani pro jednu z těchto technik.

V koncepci datové integrace v reálném čase uvedené na Obrázek je navíc přidána možnost tzv. *ETLT* (Extraction, Transformation, Loading, Transformation), tedy možnost pokročilých transformací a čištění dat poté, co byla tato data vložena do cílového úložiště. V případě potřeby je možné aplikovat složitější business pravidla až následně po uložení do *Interactive Sector*.

Enterprise Metadata Repository (EMR)

Vrstva metadat se vztahuje ke všem komponentám řešení. Jednotná a centralizovaná vrstva metadat je pro uvedenou koncepci kritická, řešení počítá s celopodnikovou správou metadat, nikoli s oddělenými úložišti metadat. Přínosy jednotného EMR řešení v rámci DW 2.0 jsou uvedené v (Inmon, Strauss a Neushloss, 2008). Myšlenka je uvedena již v koncepci DW 2.0, v rámci své práce jsem ji rozšířil na veškeré komponenty analytické podpory rozhodování – detailněji opět v (Pitka, 2011).

Metadata je možné rozdělit do několika skupin:

- *procesní metadata* – definice datových souborů a jednotlivých prvků řešení pomocí jazyka podnikového prostředí. Definice business pravidel, pravidel datové kvality či jednotlivých

datových prvků pomocí jazyka koncového uživatele umožňuje snadnou správu řešení a tvorbu návazných sestav a analýz.
- *technická metadata* – zajišťují připojení jednotlivých částí řešení do databází či na rozhraní primárních informačních systémů, nastavení datových pump, definice dočasných tabulek atd.
- *operační metadata* – data o samotném analytickém systému pro podporu rozhodování, resp. o běhu (zejm. identifikace skriptů pro tvorbu dočasných datových úložišť a historie běhů).
- *MDM (Master Data Management)* – MDM představuje širší oblast než pouze správu metadat. V rámci představené koncepce řešení má tato oblast metadat naznačit nutnost aplikace principů MDM na celou architekturu. Datová kvalita, úplnost a čistota je pro oblast real-time analýz kritická – časy na čištění a transformaci dat jsou minimální, proto je nutné data administrovat co nejlépe přímo na místě jejich vzniku, tedy v primárních systémech.

Závěr

V rámci příspěvku jsem představil technologický pohled na oblast datové integrace v reálném čase. Potřebu podpory rozhodování v reálném čase jsem dokumentoval v kapitole 2 za použití oblastí, ve kterých by zavedení těchto systémů znamenalo nezanedbatelné přínosy. Současné technologie (uvedené v kapitole 3) ovšem nejsou schopny plně zajistit datovou integraci potřebnou pro rozhodování v reálném čase. Kapitola 4 přináší analýzu překážek a problémů, které v současné době brání dosažení potřebné rychlosti datové integrace. Jasné pojmenování a analýza těchto slabých míst současných přístupů k datové integraci je prvním přínosem této práce a prací předcházejících (Pitka, 2011). Jasná definice současných slabých míst mně umožnila rozpracovat koncepční návrh řešení překážek v následující kapitole. V rámci této práce neuvádím detailní technologické řešení, ale spíše architektonický koncept, který je nutné v návazném výzkumu dále rozpracovat.

Příspěvek si nekladl za cíl detailně prozkoumat oblast datové integrace pro potřeby rozhodování v reálném čase. Proto také vnímám určité nedostatky, které je nutné odstranit návazným výzkumem v této oblasti. Jedná se především o:
- rozpracování jednotlivých oblastí navrženého řešení na úroveň detailních návrhů,
- praktické ověření konceptu řešení jednotlivých překážek a zpracování případové studie,
- analýzu Data Integration Suite z hlediska jejich připravenosti na real-time datovou integraci,
- analýzu přínosů a nákladů řešení pro podporu rozhodování v reálném čase – příspěvek se zabývá otázkou potřeby rozhodování v reálném čase, nezabývá se ale ekonomickou otázkou.

Literatura

BRUCKNER, R. M., B. LIST a J. SCHIEFER. Striving towards Near Real-Time Data Integration for Data Warehouses. In: *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*. London, 2002, s. 10. ISBN: 3-540-44123-9.

ECKERSON, Wayne W. A Business Approach to Right-Time Decision Making. In: *TDWI Monograph Series* [online]. June 1, 2006, [cit. 2011-12-29]. Dostupné z: http://tdwi.org/research/2006/06/mr-a-business-approach-to-right-time-decision-making.aspx

FINGAR, Peter a Joe BELLINI. *The Real-Time Enterprise : Competing on Time with the Revolutionary Business S-Ex Machine.* Meghan-Kiffer Press, 2004. ISBN: 978-0929652306.

GATHIBANDHE, H., S. DEOGIRIKAR a A. K. GUPTA. How Smart is Real-Time BI?. In: *Information Management Direct* [online]. 2010-01-28 [cit. 2011-12-29]. Dostupné z: http://www.information-management.com/infodirect/2009_152/real_time_business_intelligence-10017057-1.html?pg=1

HACKATHORN, Richard. The BI Watch: Real-Time to Real-Value. *DM Review*. 2004, January. ISSN: 1521-2912.

HATCH, David., Operational BI: Getting 'Real Time' About Performance. In: *Information Week* [online]. 2008-01-28 [cit. 2011-12-29]. Dostupné z: http://www.informationweek.com/news/software/bi/205920233

IMHOFF, Claudia. Understanding the Three E's of Integration EAI, EII and ETL. In: *Information Management Magazine* [online]. April, 2005, [cit. 2011-12-29]. Dostupné z: http://www.information-management.com/issues/20050401/1023893-1.html

INMON, W. H., D. STRAUSS a G. NEUSHLOSS. *DW 2.0: The Architecture for the Next Generation of Data Warehousing.* Morgan Kaufmann, 2008. ISBN: 978-0123743190.

KOZIELSKI, Stanislaw a Robert Wrembel. *New Trends in Data Warehousing and Data Analysis.* Dordrecht: Springer Science, 2009. ISBN: 978-0-387-87430-2.

LEVY, Evan. Operational BI From the Trenches. In: *Information Management Blogs* [online]. 2009-02-17 [cit. 2011-12-29]. Dostupné z: http://www.information-management.com/blogs/10014977-1.html

PITKA, Lukáš. EII (Enterprise Information Integration). *Systémová integrace*. 2007, č. 4, s. 21. ISSN: 1210-9479.

PITKA, Lukáš. Real-time podpora rozhodování v podnikových procesech primárních informačních systémů. *Systémová integrace*. 2011, č. 2, s. 25. ISSN: 1210-9479.

SOEJARTO, Alex. Now is the Time for Real-Time Enterprise. In: *Gartner* [online]. 2002-09-12 [cit. 2011-12-29]. Dostupné z: http://www.gartner.com/pages/story.php.id.2646.s.8.jsp

VOŘÍŠEK, Jiří. *Principy a modely řízení podnikové informatiky.* Praha: Oeconomica, 2008. ISBN: 978-80-245-1440-6.

WHITE, Colin. The Next Generation of Business Intelligence: Operational BI. In: *Information Management Magazine* [online]. May 2005 [cit. 2011-12-29]. Dostupné z: http://www.information-management.com/issues/20050501/1026064-1.html?pg=1

JEL Classification: L86, M15

Summary

Data Integration for Real-Time Decision Support

Today's market environment is getting more and more competitive and customer oriented. Corporations and state organizations need to make better decisions faster. Nowadays there is a trend to deliver the outcome of the analytical process faster, ideally in (near) real-time. This phenomenon creates quite a few issues which need to be overcome first. One of these issues is real-time data integration.

Paper deals with up-to-date development in the area of data integration in the context of real-time decision support systems. First goal of this paper is to identify and clearly state current data integration technological issues and obstacles in real-time decision support. Conceptual analysis and design of a solution to these issues is another goal of this paper.

Real-Time decision support brings quite a lot of new obstacles into the area of data integration from disparate data sources. An introduction into real-time decision support is the topic of first chapter.

Second chapter contains justification of the need for real-time data integration. The difference between real-time and right-time decision making is explained. This chapter also contains several real life examples.

Actual state of development of data integration techniques and methods is the topic of third chapter. A comparison of techniques is done as well.

Following chapter contains the analysis of weaknesses of individual integration techniques. Weaknesses are analyzed with the real-time element in mind.

A solution concept is introduced in fifth chapter – this conceptual model of real-time data integration is based on the concept of Data Warehouse 2.0, especially the Interactive Sector of the Data Warehouse 2.0 is redesigned.

The need for a further research and detailed specification of some areas is the main subject of the last chapter of the presented paper.

Keywords: real-time decision support, real-time BI, data integration, data federation, EII, CDC, CTF, Data Warehouse 2.0.

Nástroj pro realizaci dynamického kompozitního programování v Javě

Zbyněk Šlajchrt
xslaz900@vse.cz
Doktorand oboru informatika
Školitel: Ing. Rudolf Pecinovský, CSc, (rudolf.pecinovsky@i.cz)

Abstrakt: Článek představuje nástroj Chaplin ACT pro realizaci dynamického kompozitově orientovaného programování (DCOP). Tento nástroj vyplňuje mezeru v oboru, kdy současné programovací jazyky přímo neposkytují plnohodnotné prostředky pro tento styl programování. Nástroj Chaplin ACT je implementován pomocí standardních prostředků jazyka a platformy Java s využitím knihoven pro transformaci Java byte kódu. Jinými slovy, nástroj nemění ani nerozšiřuje syntaxi jazyka Java. Článek se v úvodu zabývá motivacemi a základními principy DCOP a na jednoduchých příkladech srovnává tradiční objektový přístup s DCOP. Čtenář má tak možnost seznámit se se základními stavebními kameny nástroje Chaplin ACT a získat představu na jeho využití v praktických aplikacích. Součástí příspěvku je také katalog návrhových vzorů, které podobně jako v objektově orientované programování nabízí sadu ověřených postupů a konstrukcí při praktikování DCOP. Katalog byl sestaven na základě zkušeností a experimentování během vývoje nástroje Chaplin ACT. Některé vzory byly pouze převzaty z praxe a upraveny pro potřeby DCOP, zatímco jiné byly identifikovány během experimentů a aplikací DCOP.

Klíčová slova: dynamická kompozice, DCOP, COP, objektově orientované programováním, OOP, mixins, traits, Chaplin ACT.

Úvod

Kompozitově orientované programování se snaží modelovat realitu pomocí skládání jednodušších entit do složitějších celků, tzv. kompozitů. Kompozity lze dále skládat s dalšími entitami či kompozity do kompozitů vyššího řádu. Kompozice obvykle vede k vytvoření nové úrovně abstrakce, ve které se operuje s novými pojmy a vztahy. Jedná se vlastně o paralelu k rozvrstvení do úrovní abstrakce tak, jak jej pozorujeme v našem světě: kvarky, subatomární částice, atomy, molekuly, polymery, organely, buňky, orgány, organismus a společnost. V každé úrovni platí specifická terminologie a vztahy, přičemž je zřejmé, že pojmy z jedné vrstvy musí být jaksi propojeny s pojmy ze sousední nižší vrstvy. (Öberg, 2009).

Na rozdíl od klasického OOP, které se při modelování odpoutává od modelovaných objektů a více se zaměřuje na jejich třídy, DCOP pracuje především s objekty a rozhraními. Zařazení objektů do tříd je z hlediska DCOP podružné a může být i nejednoznačné. Jedním ze zajímavých důsledků je, že použitím DCOP se lze vyhnout dědičnosti.

Návrh aplikací pomocí DCOP

Základní myšlenky DCOP bych rád ilustroval na příkladu jednoduchého grafického editoru. Tento editor umožňuje uživateli kreslit jednoduché geometrické tvary na obrazovce a provádět s nimi základní operace jako jsou např. přesun, rotace, změna velikosti atp.

Kompozice

Uvažujme operaci *Move* na geometrickém obrazci, která jej posune na novou pozici podle zadaných souřadnic vektoru posunutí. Uživatel vyvolá tuto operaci výběrem obrazce v editoru. Kliknutím pravým tlačítkem nad obrazcem se vyvolá kontextové menu, ze kterého si uživatel vybere položku *Move*. Otevře se okno, do kterého uživatel zadá souřadnice vektoru posunutí. Po kliknutí na tlačítko OK se provede operace. Kód, který zpracovává uvedenou interakci s uživatelem, může vypadat následovně:

```
...
Shape selected = getSelected();
selected.move(x, y);
...
```

Kód 1 – posunutí jednoho obrazce

Stejným postupem by měl uživatel dokázat posunout i více obrazců naráz. Výběr provede klikáním na obrazce za současného držení klávesy Control. Nad tímto výběrem vyvolá kontextové menu a zvolí operaci Move. Pro zpracování této události je třeba upravit Kód 1 následovně:

```
...
Shape[] selected = getSelected();
for (sel : selected) {
  sel.move(x, y);
}
...
```

Kód 2 – posunutí vybraných obrazců

Metoda `getSelected` nyní vrací místo jediného objektu pole vybraných objektů. Pro provedení hromadné operace je nutno zavést cyklus a operaci `move` zavolat na každém objektu zvlášť.

Uvažujme dále, že program zobrazuje ve spodní části obrazovky rozměr vybraného objektu, případně rozměr sjednocení rozměrů vybraných objektů. Kód, který zachycuje událost výběru objektu na obrazovce a vypisuje rozměr na stavové řádce, může vypadat následovně:

```
...
Shape[] selected = getSelected();
List<Rectangle> bounds = new List<Rectangle>();
for (sel : selected) {
  bounds.add(sel.getBounds());
}
Rectangle union = computeUnion(bounds);
statusBar.showSize(union);
...
```

Kód 3 – zobrazení sjednocení rozměrů vybraných obrazců

V tomto okamžiku bychom si však měli říci dost a zamyslet se nad celkovým návrhem programu. Nebylo by jednodušší, kdybychom považovali výběr geometrických objektů také za geometrický objekt, na kterém lze zavolat metody `move` a `getBounds`? Kód zpracování odpovídajících událostí by se nám značně zjednodušil:

```
…
Shape selected = getSelected();
selected.move(x, y);
…
```

Kód 4 – posunutí po zavedení kompozitu

```
…
Shape selected = getSelected();
statusBar.showSize(selected.getBounds());
…
```

Kód 5 – zobrazení rozměru po zavedení kompozitu

Právě jsme použili návrhový vzor Složenina (Composite) (Gamma, 1994). Jeho zavedením se kód zjednodušuje a zlepšuje se zapouzdření, neboť výběr objektů je nyní chápán také jako objekt a kód manipulující nad výběrem je jeho součástí.

Soustřeďme se nyní na samotný kompozit. V klasickém OOP bychom vytvořili třídu např. `CompositeShape`, která by držela referenci na pole vybraných objektů. Tato třída by implementovala rozhraní `Shape` a její metody `move` a `getBounds` by vypadaly jako Kód 2 a 3. V následujících odstavcích představím alternativu k uvedenému postupu, která nevyžaduje vytváření třídy kompozitu a umožňuje skládat kompozity dynamicky v čase běhu programu.

Chaplin ACT

Nástroje, které jsou zapotřebí pro dynamickou kompozici, nejsou standardně ve výbavě masově používaných programovacích jazyků, jako je Java, C#, C/C++ atd. Jeden ze způsobů, jak je získat, je využití jazyku Java spolu s nástrojem pro manipulace s kódem zkompilovaných tříd a doplnit potřebou funkcionalitu transformací byte-kódu. Tento přístup jsem zvolil a vyvinul nástroj Chaplin ACT (http://www.iquality.org/chaplin). Vývoj s použitím tohoto nástroje probíhá standardně s tím, že zkompilovaný kód je třeba transformovat pomocí nástroje Chaplin ACT. V následujícím textu se budu na tento nástroj odvolávat.

Dynamická kompozice

Vraťme se nyní ke Kódu 4, který má za úkol reagovat na příkaz uživatele k posunutí vybraných objektů. Metoda `getSelected` vrací buď vybraný objekt, nebo kompozit složený z vybraných objektů. Bez použití dynamické kompozice by tato metoda vytvářela instanci třídy `CompositeShape` a předávala jí seznam vybraných objektů. Jedna z pozitivních vlastností dynamické kompozice v Chaplin ACT je, že nevyžaduje vyvíjet speciální třídu pro kompozit. Namísto toho se k vytváření kompozitu použije operátor `$`, kterému se předají komponenty. Výsledkem volání tohoto operátoru je objekt-kompozit. Následující kód ukazuje, jak by vypadala metoda `getSelected` vyvinutá s použitím Chaplin ACT:

```
public Shape getSelected() {
  Shape[] selection = editor.getSelection();
  Shape composite = $(selection);
  return composite;
}
```

Kód 6 – dynamické vytvoření kompozitu

Voláním operátoru `$` a předáním pole vybraných objektů získáme kýžený kompozit. Vysvětlit podrobně fungování tohoto operátoru je nad rámec tohoto článku (viz (Šlajchrt, 2009a)). V podstatě jde o to, že operátor `$` ‚vidí' typ proměnné, do které je přiřazen výsledek jeho volání. To mu umožní použít tento cílový typ (`Shape`) a vytvořit kompozit takovým způsobem, aby vyhovoval přiřazení.

Zavoláním metody `move` na takto vytvořeném kompozitu dojde k distribuci tohoto volání mezi všechny komponenty. Jelikož tato metoda nemá žádnou návratovou hodnotu, není třeba řešit problém slučování více návratových hodnot pocházejících z jednotlivých volání. Tento problém je však třeba řešit v případě volání metody `getBounds` vracející rozměr geometrického tvaru.

Zavoláním metody `getBounds` dojde také k distribuci volání mezi všechny komponenty. Nicméně toto volání skončí výjimkou, která si bude stěžovat na vícenásobnou návratovou hodnotu, protože není možné vrátit více hodnot tam, kde se vyžaduje právě jedna. V tomto případě dynamická kompozice nabízí zařazení speciální komponenty do kompozitu, která bude pouze zachytávat volání metody `getBounds` a slučovat rozměry jednotlivých vybraných geometrických tvarů do jedné hodnoty. Taková komponenta se nazývá *obal* (*wrapper*), což vystihuje její chování, kdy po odchycení se volání předává dál ke zpracování, přičemž obal může provádět nějaké akce před nebo po předání. V našem případě se po předání slučují rozměry jednotlivých objektů.

Obalovou komponentu navrhneme jako abstraktní třídu `GetBoundsAggregator`, která implementuje rozhraní `Shape`, přičemž jediná metoda v této třídě je `getBounds`.

```
public abstract class GetBoundsAggregator
 implements Shape {
        public Rectangle getBounds() {
         MessageReplies replies =
                                 new MessageRepliesImpl();

// předá volání a naplní replies všemi
 // návratovými hodnotami
          $next(replies);
          return computeUnion(replies.getReplies());
        }
```

```
        private Rectangle computeUnion(List rectangles) {
                ...
        }
}
```

Kód 7 – agregační obalující komponenta

Metodu `getSelected` je třeba upravit tak, aby vracela kompozit obohacený o obalující komponentu `GetBoundsAggregator`:

```
public Shape getSelected() {
 // řekneme si editoru o pole vybraných objektů uživatelem
 Shape[] selection = editor.getSelection();
 // vytvoříme novou instanci třídy GetBoundsAggregator
 GetBoundsAggregator getBoundsAgg = $();

 // sloučíme agregační komponenty a pole selection do
 // jednoho pole privátní pomocnou metodou
 Shape[] components = concat(getBoundsAgg, selection);

 // vytvoříme kompozit
 Shape composite = $(components);
 return composite;
}
```

Kód 8 – dynamické vytvoření kompozitu s obalující komponentou

Pokud nyní zavoláme na kompozit vytvořený upravenou metodou `getSelected` metodu `getBounds` dojde k zachycení tohoto volání agregační komponentou, která volání obratem předá dál zavoláním operátoru `$next`. Po distribuci volání mezi zbývající komponenty agregační komponenta sloučí výsledky do jednoho objektu a ten je vrácen volajícímu. Z pozice volajícího se jedná o transparentní chování, kdy se nerozlišuje, zda volání proběhlo nad jedinou či více komponentami.

Mixiny

Dosud kompozit sloužil pouze k distribuci volání metod rozhraní `Shape` mezi komponenty a nenabízel žádnou dodatečnou funkcionalitu specifickou pouze pro kompozit. Jako příklad takové funkcionality, která má smysl pouze u kompozitu a nikoliv u jeho komponent, lze vzít operaci, která dokáže vytvořit nový geometrický objekt složený z komponent. Dejme tomu, že toto chování zachytíme v rozhraní `ShapeComposite`, které bude mít prozatím jedinou metodu `merge` vytvářející nový geometrický útvar z komponent kompozitu.

Naším cílem nyní je obohatit kompozit o komponentu, která dodá funkcionalitu definovanou rozhraním `ShapeComposite`. Takováto komponenta funguje jako *mixin*. K tomuto účelu vytvoříme třídu, která toto rozhraní implementuje:

```
public class ShapeCompositeMixin implements
        ShapeComposite {

        public Shape merge() {
                ... vytvoří a vrátí nový geom. objekt z komponent ...
        }
}
```

Kód 9 – třída komponenty rozšiřující funkcionalitu kompozitu (mixin)

Nyní stačí metodu `getSelected`, která vyrábí kompozit, doplnit o obohacení kompozitu objektem třídy `ShapeCompositeMixin`.

```
public Shape getSelected() {
 Shape[] selection = editor.getSelection();
 // vytvoří novou instanci třídy GetBoundsAggregator
 GetBoundsAggregator getBoundsAgg = $();

 // vytvoří novou instanci třídy ShapeCompositeMixin
 ShapeCompositeMixin shapeMixin = $();
```

```
 // sloučení agregační komponenty, mixinu a pole
 // selection do jednoho pole
 Shape[] components = concat(getBoundsAgg, shapeMixin, selection);

 Shape composite = $(components);
 return composite;
}
```

Kód 10 – začlenění nové funkcionality do kompozitu

Kód, který volá metodu `getSelection`, může návratovou hodnotu (kompozit), přetypovat na typ libovolné komponenty uvnitř kompozitu, tedy také na typ `ShapeComposite`:

```
…
Shape selected = getSelected();
ShapeComposite composite = (ShapeComposite)selected;
Shape merged = composite.merge();
…
```

Kód 11 – přetypování na typ `ShapeComposite`

Zavoláním metody `merge` se deleguje volání na mixin v kompozitu.

Interakce mezi komponentami

Doposud jsem ukázal, jak lze původně nezávislé objekty sloučit do jednoho celku a zajistit, aby se tento celek-kompozit choval obdobně jako jeho komponenty (*fusion*). Dále jsem ukázal, jak lze tuto novou entitu obohatit o dodatečnou funkcionalitu, která má smysl pouze na úrovni kompozitu a nikoliv u komponent (*enhancement*). Takto v podstatě došlo k přechodu na vyšší úroveň abstrakce. Komponenty v takovém kompozitu však stále existují jako samostatné objekty, které nikterak nezajímá fakt, že jsou součástí nějakého celku a že v jejich okolí se nacházejí další komponenty. Jejich chování je nezávislé na *kontextu*. Nicméně dalším důležitým rysem dynamické kompozice je, že umožňuje, aby přítomnost okolních komponent měla vliv na chování dané komponenty. Pro ilustraci interakce mezi komponentami přejdu k dalšímu příkladu.

Dejme tomu, že ve vyvíjené aplikaci potřebujeme objekt, který bude vypisovat výjimky ve specifickém formátu. Výpis bude obsahovat zprávu výjimky, výpis zásobníku a název vlákna, ve kterém došlo k výjimce. Dále požadujeme, aby tento logovací objekt byl **nezávislý** na výstupním zařízení, do kterého se výpis zapisuje. Třídu tohoto objektu lze s využitím anotací Chaplin ACT navrhnout takto:

```
public abstract class SimpleLogger {
@FromContext
 abstract void println(String line);
public void log(Exception e, Thread thread) {
        println(e.getMessage());
        println("Thread:" + thread.getName());

        StackTraceElement[] stackTrace = e.getStackTrace();
        for (int i = 0; i < stackTrace.length; i++) {
         println(i + ": " + stackTrace[i].toString());
        }
 }

}
```

Kód 12 - třída logovacího objektu

Během výpisu výjimky volá metoda `log` abstraktní metodu `println`, která slouží k výpisu prostého textu na výstupním zařízení. Abstraktnost této metody vychází z požadavku, aby byl logovací objekt nezávislý na cílovém výstupním zařízení.

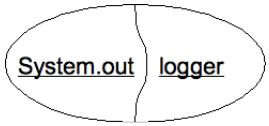
Tuto třídu lze velmi snadno testovat; postačí vyrobit testovacího potomka této třídy `TestSimpleLogger` a v jeho metodě `println` testovat, zda se během testování výpisu připravené výjimky předávají správné texty.

Použití logovacího objektu v programu je také jednoduché. Využijeme faktu, že objekt standardního výstupu obsahuje metodu `println` se stejnou signaturou jako ve třídě `SimpleLogger` (což byl samozřejmě záměr) a vytvoříme kompozit složený z logovacího objektu a objektu standardního výstupu:

```
SimpleLogger logger = $(System.out);
logger.log(exception, Thread.currentThread());
```

Kód 13 – použití logovacího objektu v programu

Schematicky lze výsledný kompozit znázornit takto:



Obrázek 1: Kompozit logovacího objektu a objektu standardního výstupu

V okamžiku, kdy se z metody `log` zavolá abstraktní metoda `println`, se vyšle zpráva do kontextu kompozitu a hledá se vhodný příjemce. V tomto případě se nalezne objekt `System.out`, který nabízí metodu s odpovídající signaturou.

Na tomto příkladě jsem ilustroval další důležitou vlastnost dynamické kompozice, kterou je *interakce* mezi komponentami.

Návrhové vzory

Používání kompozitního programování v praxi ukázalo, že podobně jako v případě objektového programování tak i zde se opakovaně vyskytují jisté postupy, které lze zobecnit do podoby znovupoužitelných návodů či praxí ověřených postupů – tedy návrhových vzorů. Během vývoje a aplikování DCOP jsem identifikoval řadu návrhových vzorů. Některé z nich se již léta používají v praxi, zatímco jiné se vynořily v rámci mého experimentování s možnostmi dynamické kompozice. V následujících odstavcích stručně představím několik zástupců zmiňovaných vzorů. Z důvodu omezeného rozsahu tohoto příspěvku se bohužel nelze věnovat podrobněji jednotlivým případů. To je koneckonců cílem mé dizertační práce.

Dědičnost (Inheritance)

Jako první vzor záměrně uvádím dědičnost, abych poukázal na společné rysy DCOP a OOP. V objektovém programování se dědičnost neuvádí jako samostatný návrhový vzor, neboť se přirozeně chápe jako základní vztah mezi třídami, na kterém staví komplexnější návrhové vzory. V DCOP se však dědičnost nechápe takto fundamentálně, místo toho se jeví pouze jako speciální uspořádání fragmentů uvnitř kompozitu, které může být kombinací tří postupů: převažování metod (Overriding), definice šablonové metody (Template Method, Thin/Rich interface) a rozšiřování (Extending). Blíže o této problematice viz (Šlajchrt, 2009b).

Thin/Rich interface

Pomocí tohoto vzoru se řeší obvyklý problém volby mezi „chudým“ (thin) a „bohatým“ (rich) rozhraním. Chudé rozhraní nabízí pouze malou sadu elementárních (ortogonálních) metod pro práci s daným objektem, zatímco bohaté rozhraní nabízí širokou škálu metod. Oba přístupy mají svá pro a proti: chudé rozhraní je snadné implementovat, avšak jeho použití je komplikovanější, neboť vyžaduje „doprogramování“ chybějících metod klientem rozhraní. Na druhou stranu lze snadno vyměnit stávající implementaci chudého rozhraní za jiné při zachování kódu klientských metod. K implementaci bohatého rozhraní je zapotřebí více úsilí, jeho použití klientem je ale snadnější – vše potřebné má již k dispozici na rozhraní. Problém nastává v situaci, že je zapotřebí změnit chování kanonických metod rozhraní, na kterých jsou postaveny obohacující metody rozhraní.

V kompozitním programování lze tento problém elegantně vyřešit zavedením dvou typů místo jednoho: chudé rozhraní obsahující pouze kanonické metody a abstraktní třídou, která dědí, ale neimplementuje chudé rozhraní a navíc definuje obohacující metody, které jsou sestaveny nad (šablonovými) metodami chudého rozhraní. Klient si pak vybere vhodnou implementaci chudého rozhraní a složí

kompozit složený z vybrané instance chudého rozhraní a abstraktní třídy poskytující bohaté rozhraní. Výhodou dynamického kompozitního programování oproti statickému je možnost definovat složení kompozitu za běhu, případně jej za běhu měnit.

Interceptory

Interceptory umožňují zachytit volání metody na objektu a provést nějaké akce před či po samotném provedení metody. Tento přístup inspirovaný aspektovým programováním (Dessi, 2009) se ukazuje vhodným v situacích, kdy potřebujeme oddělit business logiku od doprovodných aktivit, jako například logování, správa transakcí, kontrola vstupních a výstupních podmínek, zabezpečení atp. Tyto aktivity lze rozdělit do tří skupin:

- **Side effects –** Aktivity, které nemají přímou vazbu na business logiku „obalené“ metody. Např. logování, bezpečností filtr, transakce.
- **Concerns –** Aktivity, které jsou přímo svázány s business logikou „obalené“ metody a které ji nějakám způsobem doplňují.
- **Constraints –** Omezení na vstupní a výstupní hodnoty metody.

V dynamickém kompozitním programování lze interceptory vrstvit a za běhu přidávat či odebírat. Každý interceptor je fragment, který implementuje „obalenou“ metodu a předává ji řízení pomocí speciálního operátoru.

Chameleon

Chameleon je vzor úzce spjatý s koncepcí DCOP. Pomocí tohoto vzoru lze dosáhnout, aby se chování kompozitního objektu přizpůsobilo kontextu, ve kterém je k němu přistupováno. Dochází k tomu tak, že volající objekt, tedy objekt z jehož metody je volána nějaká metoda kompozitu, se po dobu volání stává součástí (fragmentem) volaného kompozitu. Takto lze dočasně ovlivnit vnitřní strukturu a chování kompozitu daty v atributech a metodami, které patří volajícímu objektu. Metody i atributy volajícího (kontextového) objektu mohou převážit stejnojmenné interní metody a atributy v kompozitu.

Vzor Chameleon lze použít i jako variantu vzoru Thin/Rich interface nebo šablonové metody. Volající objekt, který současně implementuje „chudé“ rozhraní, volá metodu „bohatého“ rozhraní. Ta ke svému fungování potřebuje metody z „chudého“ rozhraní, které je však v tomto okamžiku k dispozici díky dočasné přítomnosti volajícího objektu v kompozitu.

Tento vzor lze použít např. v situacích, kdy různorodí klienti volají stejnou službu, avšak její vykonání musí být přizpůsobeno konkrétnímu typu klienta. Místo aby se požadavek z klientů směroval přímo na služební objekt, tak klienty na serveru reprezentujeme různými objekty, které přijímají od požadavky od „svých“ reálných klientů a obratem je předávají na objekt implementující vlastní službu (služební objekt). Tyto klientské reprezentace implementují „chudé“ rozhraní, které je v okamžiku volání služby využito služebním objektem, který naopak implementuje „bohaté“ rozhraní. Služební objekt je defacto „abstraktní“ objekt (viz vzor Zombie), který musí být volán pouze v nějakém kontextu, ve kterém je k dispozici nějaká implementace „chudého“ rozhraní.

Zombie

O tomto vzoru jsem se zmínil v souvislosti se vzorem Chameleon. Jelikož fragmenty kompozitů jsou často reprezentovány jako abstraktní třídy a současně je třeba, aby fragment existoval ještě před vložením do kompozitu, musí existovat možnost, jak vytvořit instanci abstraktní třídy. Ačkoliv v OOP nelze z definice vytvářet instance abstraktních tříd, v DCOP je to zcela běžné. Takovéto instance fragmentů, které se nazývají *zombie* či *pseudoinstance*, nelze použít, dokud se nestanou součástí nějakého kompozitu.

Kompozitní továrna (Composite Factory)

Kompozitní továrna je vzor, pomocí kterého lze složit více továrních objektů do jednoho – kompozitní továrny. Produktem této továrny je pak složenina (kompozit) produktů jednotlivých továrních objektů

uvnitř kompozitu. Tento vzor lze použít pro výrobu kompozitů, které obsahují předem definovanou strukturu, např. interceptory starající se o logování či bezpečnost.

Aggregator

DCOP umožňuje, aby kompozitní objekt obsahoval více fragmentů, které implementují stejné rozhraní. Jelikož kompozit toto rozhraní vystavuje klientům, je třeba definovat chování kompozitu při volání metody, jež se vyskytuje ve více fragmentech. Pokud v takovém případě klient zavolá na kompozitu metodu, která nic nevrací a pouze přebírá vstupní parametry, zavolá se jednoduše odpovídající metoda na všech fragmentech. Problém nastává, pokud metoda vrací nějakou hodnotu či referenci na objekt. V takovém případě DCOP umožňuje specifikovat tzv. agregátor, speciální fragment, který se vloží do kompozitu a je zodpovědný za agregaci výsledků volání metody na jednotlivých fragmentech.

DCOP definuje jeden implicitní agregátor, který se uplatňuje ve vzoru Kompozitní továrna. Tento agregátor slučuje jednotlivé produkty továrních metod na různých fragmentech do jednoho kompozitu.

Recept (Recipe)

Tento vzor představuje návod na sestrojení kompozitu. Klíčovou roli v tomto vzoru hraje tzv. sestavovatel (assembler). Idea je taková, že sestavovatel obdrží návod, podle které sestrojí požadovaný kompozit. Návod může být specifikován pomocí různých programovacích jazyků, vhodných pro danou doménu (zde se nabízí využití DSL – doménově specifických jazyků ušitých na míru řešené problematice). Návod může obsahovat např. instrukce typu: „*Vezmi první parametr z dotazu a použij jej do daného SQL SELECT dotazu. Jediný výsledek dotazu použij k inicializaci instance třídy User. Dále vytvoř zombie instanci fragmentů UserPresentation a UserValidator. Všechny tři instance slož společně do jednoho kompozitu, který je návratovým objektem tohoto receptu.*“

Páteř (Backbone)

Tento návrhový vzor umožňuje vnést řád do struktury a chování kompozitu. Lze jej považovat za jakýsi protokol a schéma pro kompozit. Páteř definuje, jaké metody lze na kompozitu volat (tzv. *context services*) a na jaké události lze reagovat (tzv. *context events*). Páteř se definuje pomocí rozhraní. Toto rozhraní je však určeno pouze pro vnitřní použití v kompozitu fragmenty a nikoliv pro klienty kompozitu. Páteř je vhodné použít tehdy, když je předem znám řád, jakým mají mezi sebou fragmenty komunikovat. Tento vzor lze uplatnit např. při sestavování prvků uživatelského rozhraní např. podle vzorů Model-View-Controller/Presenter.

Další vzory

Z důvodu omezeného rozsahu tohoto článku se v tomto odstavci omezím na stručný výčet dalších návrhových vzorů, které jsem během vývoje a experimentování s DCOP identifikoval:

- **Nesting** – Vzor, podle kterého lze vytvářet hierarchie kompozitních objektů
- **Component proxy** – Každý fragment může být součástí nejvíce jednoho kompozitu. Pokud je zapotřebí jeden fragment použít ve více kompozitech, lze ho „zabalit“ do proxy fragmentu a ten vložit do kompozitu namísto originálu.
- **Default (singleton) fragment** – V DCOP lze definovat globální implicitní (default) fragment, který slouží k doplňování chybějících fragmentů v kompozitu.
- **Enricher** – Slouží k dynamickému obohacení perzistentních entit např. o validační nebo autorizační fragment.
- **State Machine** – Kompozit se může v různých okamžicích skládat z různých fragmentů. Strukturu a složení kompozitu lze popsat a řídit pomocí stavového automatu.

Závěr

Dynamické kompozitově orientované programování (DCOP) přináší další rozměr do objektově orientovaného návrhu aplikací. V současnosti programovací jazyky přímo neposkytují plnohodnotné prostředky pro tento styl programování, nicméně lze použít nástroj Chaplin ACT, který umožňuje dynamickou kompozici v jazyku Java a který vyvíjím v rámci své dizertační práce. Nedílnou součástí práce

je také katalog návrhových vzorů, přičemž některé se již používají v praxi, jiné se podařilo identifikovat během vývoje a experimentování s DCOP.

Mezi nejdůležitější cíle dalšího výzkumu patří podrobná analýza jednotlivých programovacích vzorů a vyvíjení doprovodných ukázek kódu pomocí Chaplin ACT. Dalším cílem je vypracování architektury a návrhu vzorové aplikace za použití DCOP a Chaplin ACT.

Literatura

DESSI, Massimiliano, 2009. *Spring 2.5 Aspect Oriented Programming.* Birmingham: Packt Publishing. ISBN 978-1-847194-02-2.

GAMMA, Erich, HELM, Richard, JOHNSON, Ralph a VLISSIDES, John, 1994. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison-Wesley Professional, 1994, ISBN 0-201-63361-2.

ÖBERG, Rickard. *Introduction to Qi4j, 2009.* http://www.qi4j.org/160.html.

ŠLAJCHRT Zbyněk, 2009a. *Chaplin - The Agile Java Class Transformer.* http://www.iquality.org/chaplin/generated/doc/book.html.

ŠLAJCHRT Zbyněk, 2009b. *Kompozitově orientované programování, sborník Objekty 2009.* Hradec Králové: Gaudeamus. ISBN 978-80-7435-009-2.

JEL Classification: *L86*

Summary

A tool for dynamic composite oriented programming in Java

This article presents tool Chaplin ACT, whose purpose is to allow dynamic composite programming (DCOP). This tool fills a gap in the field of contemporary object oriented tools and languages that do not provide a fully-fledged support for dynamic composition of objects. Chaplin ACT is implemented on the top of the Java language and and the Java platform with the help of libraries for byte code transformation. In other words, the tool does not changes nor extends the Java syntax. In the introduction the article deals with motivations and basic principles of DCOP and presents simple examples on which it compares the traditional object-oriented approach with DCOP. Thus the reader has an opportunity to get acquainted with the basic building blocks of Chaplin ACT and acquire a hint for using the presented approach practically. The article also presents a catalog of DCOP design patterns that, similarly to OOP, provides a set of guidelines and constructions that have proved useful in applications. The catalog has been compiled on experience and experimental basis during development of Chaplin ACT. Some patterns were just taken from practice and adapted to the needs of DCOP, while the others have been recognized during experimenting with DCOP.

Keywords: dynamic composition of objects, DCOP, COP, object oriented programming, OOP, mixins, traits, Chaplin ACT.

Zajištění integrity systému dlouhodobé archivace

Rudolf Vohnout
xvohr900@vse.cz
Doktorand oboru informatika
Školitel: doc. Ing. Zora Říhová, CSc. (zora.rihova@vse.cz)

Abstrakt: Tento příspěvek se zaměřuje na mechanismy zajišťující integritu objektů v dlouhodobém archivu. Ukazuje, jak vytvořením integrační informace a metodami schémat provázaných otisků lze zajistit nepopiratelnost a autenticitu archivovaných dokumentů.

Druhá kapitola se nejprve soustředí na shromaždování informací o archivovaném objektu a popisu jejich významu pro vytvoření souhrnné informace. Dále pak detailně, ve dvou krocích, rozebírá, jakým způsobem jsou otisky archivovaných dokumentů hierarchicky, pomocí stromového schématu, provázány. Vrchol stromu, respektive jeho kořen představuje tzv. souhrnný důkaz. Ten je vytvořen na základě provázání otisků dokumentů registrovaných do archivu během předem stanoveného časového úseku, resp. cyklu. Souhrnné důkazy jsou pak na základě binárního schématu dále provázány a to vždy za pomoci hashovací funkce. Ta je aplikována na posloupnost znaků tvořených z výsledného otisku ze všech předchozích kroků a souhrnného důkazu z kroku aktuálního.

Článek se v následující kapitole věnuje provádění auditu a konstantnímu monitoringu integrity archivovaných objektů. Jsou zde popsána systémová metadata a druhy zpráv, které při auditu hrají klíčovou roli. V závěru jsou pak zmíněna základní systémová oprávnění, která jsou důležitá pro přidělování rolí v systému.

Klíčová slova: dlouhodobá archivace, dokument, otisk, bezpečnost, integrita, metadata.

Úvod

Jedním z problémů archivace je zajištění dlouhodobé integrity archivovaných dokumentů a k nim se vážících objektů. Postupem času může být integrita dokumentů v archivu narušena z několika různých důvodů (např. hardwarová selhání nebo migrační techniky). Dále je třeba vyřešit postupné „stárnutí“ kryptografických technik, na základě kterých jsou vytvářeny otisky archivovaných dokumentů. Ty se budou postupem času stávat více zranitelnými a budou vyžadovat jejich nahrazení novějšími a bezpečnějšími. K zajištění integrity všech dokumentů je v tomto příspěvku prezentován univerzální model, kombinující techniky schémat provázaných otisků (hashů) a důkazních informací vytvořených z každého archivovaného objektu.

Zajištění integrity

Základním principem zajištění integrity je konstantní monitoring archivovaných dokumentů. Tato funkce může být následně použita k provedení auditu integrity archivu a k použití nezávislých auditorských nástrojů třetích stran. Dalším předpokladem je, že nástrojům třetích (pokud budou použity) nebude dovoleno modifikovat obsah archivu a to ani v tom případě, když bude detekován problém. Jeho případné odstranění je ponecháno na správci archivu, či vlastníkovi objektu (záleží na politice). Tento přístup může být následně implementován do libovolné archivační architektury a projektů zaměřených na dlouhodobou archivaci.

Objekty

Po přijetí dokumentu do archivu (první krok procesu archivace) jsou vytvořeny dva objekty. Konkrétně se jedná o:

- Otisk integrity (Integrity Token - IT).
- Souhrnný otisk (Agregated Hash Value - AHV).

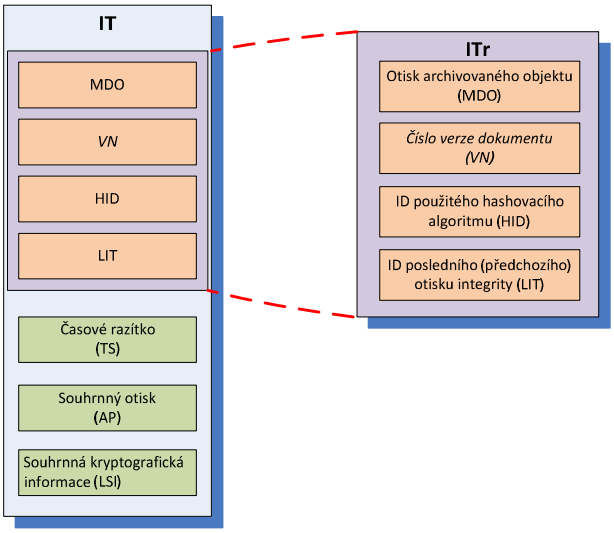
První objekt se z hlediska časové posloupnosti vytváří jako první a to ihned po předání dokumentu do archivu (spolu s příslušnými metadaty dle modifikovaného doporučení Dublic Core). Detailní popis bude následovat v další podkapitole. Druhý objekt se vytváří na základě nastavené politiky, ale vždy až po prvním objektu.

Otisk Integrity

Otisk integrity může být součástí metadat archivovaného objektu nebo být zapsán do separátně spravované databáze. Tento otisk sestává jak z dokumentu samotného, tak z poskytnutých metadat. Dále obsahuje kritické informace systémového charakteru, které jsou nezbytné pro zajištění integrity a monitoring (blíže v předposlední kapitole). Skládá se z:

- Otisku archivovaného objektu (Message Digest Object - MDO).
  V současné době je zajištěn pomocí hashovací funkce SHA-2 s délkou klíče 512-bitů.
- Čísla verze dokumentu (Version Number - VN).
  Slouží především k vzájemnému linkování jednotlivých verzí dokumentu. Pokud v průběhu archivačního procesu nedochází k verzování, lze uvádět stále stejnou hodnotu (například 1).
- ID použitého hashovacího algoritmu (Hash Algorithm ID - HID).
  Slouží k identifikaci použitého hashovacího algoritmu. Je předpoklad, že ten se bude v průběhu času měnit, a tím by mohlo dojít ke zpochybnění integrity archivovaných dokumentů. Tato část neřeší změnu již existujících otisků archivovaných dokumentů v minulosti.
- ID posledního (předchozího) otisku integrity (Last Integrity Token - LIT).
  ID je ukládáno separátně, mimo vlastní IT. Následně je použito při archivaci n+1 dokumentu jako součást IT.
- Časového razítka (Time Stamp - TS).
  Nepovinný atribut. Potvrzuje existenci archivovaného dokumentu v čase, tj. čas, ve kterém je dokument svěřen archivační autoritě. Samotný dokument může již sám o sobě obsahovat časové razítko. Pokud je dokument opatřen také elektronickým podpisem, je před vlastní archivací nejdříve ověřena jeho platnost.
- Souhrnný důkaz (Agregated Proof - AP).
  Nepovinný atribut. Jedná se agregační důkaz složený z jednotlivých otisků všech dokumentů ve struktuře agregačního stromu.
- Souhrnné kryptografické informace (Last Security Information - LSI); (CSI (n-1)).
  Nepovinný atribut. Souhrnné kryptografické informace vygenerované v předchozím kroku buď po dosažení určitého počtu vytvořených IT (tj. archivovaných dokumentů) nebo po uplynutí dynamicky nastavené časové periody.

Podrobněji se nepovinným atributům věnuje následující kapitola. Detailní struktura IT, viz Obrázek.



Obrázek 1: Otisk integrity (IT) Zdroj: (SMORUL,2009), autor

Nutné atributy pro vytvoření IT jsou shrnuty pod pojmem ITr. Atributy napsané na zeleném pozadí jsou obecně tzv. povinně-volitelné atributy, tj. atributy které nejsou nezbytně nutné pro sestavení IT, ovšem pro požadovanou funkcionalitu (tj. zajištění integrity) je nutné vybrané z nich použít.

Souhrnný otisk

Jsou informace o všech objektech archivovaných během dynamicky nastavitelné časové periody nebo po dosažení stanoveného počtu archivovaných objektů. Tyto informace jsou uchovávány mimo samotný archiv. Jak IT, tak souhrnný otisk (AHV) slouží k periodickému ověřování integrit archivovaných objektů. U IT se jedná o ověření jednotlivých objektů, u AHV pak o ověření pravosti části archivu. Platí implikace - pokud je v pořádku AHV pak je v pořádku IT a naopak.

Další možností jak zajistit nepopiratelnost archivovaných dokumentů, je po ověření jednotlivých IT a vygenerování AHV (který je v době vytvoření důkazem pravosti jednotlivých IT v nižších úrovních stromu) jej opatřit souhrnným časovým razítkem, které bude demonstrovat autenticitu celého stromu.

Dva kroky integrity

Celkový záznam o integritě digitálního objektu sestavuje ve dvou krocích, resp. časových úsecích:

- V prvním, který byl již částečně popsán v předchozí podkapitole, se vytvoří IT pro každý archivovaný dokument a jeho otisk se stává součástí agregačního cyklu.
- Ve druhém se pak vytvoří tzv. souhrnná informace (Computed Summary Information - CSI), která může být součástí IT vytvořeného v předchozím kroku.

Využívá se v nich technik kombinace stromového - např. Merkle Tree 0 a lineárního 0 schématu. Tím lze provést ověření autenticity jakéhokoliv verze archivovaného objektu (jeli požadováno).

První krok

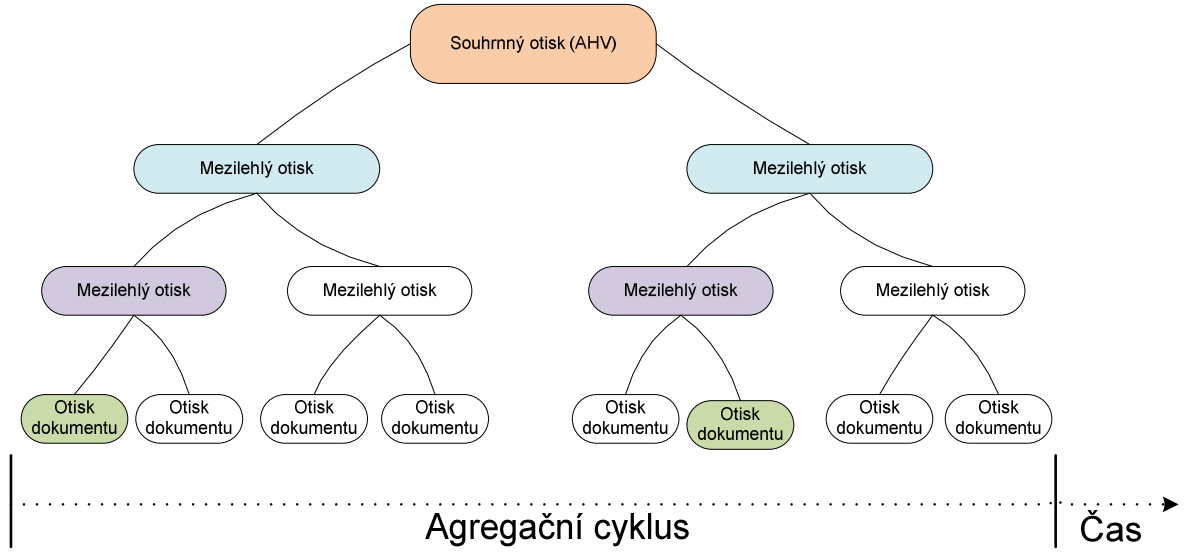
Hierarchické sestavení integritních informací, což představuje v této fázi vytvoření stromu integrity. Strom se skládá z množiny otisků dokumentů poskytnutých archivu během dynamicky nastaveného časového intervalu. Časový interval je proto také nazýván agregačním cyklem. Během agregačního cyklu jsou kromě otisků poskytnutých během předarchivační fáze také vkládány náhodně vygenerované otisky a to v případě kdy není v během cyklu dosaženo jejich požadovaného počtu. To je například v případě, kdy není během nastaveného časového intervalu archivován žádný dokument. Autentizační strom lze také popsat následujícím vztahem:

$$AHVn = \sum_{x}^{1} IHVx$$

$$platí\ IHVx = h\left(h(Doc_x) + h(Doc_y)\right),$$

kde
AHV = souhrnný otisk,
IHV = mezilehlý otisk,
$h(Doc_x)$ = otisk x-tého dokumentu.

Schéma autentizačního stromu viz Obrázek.

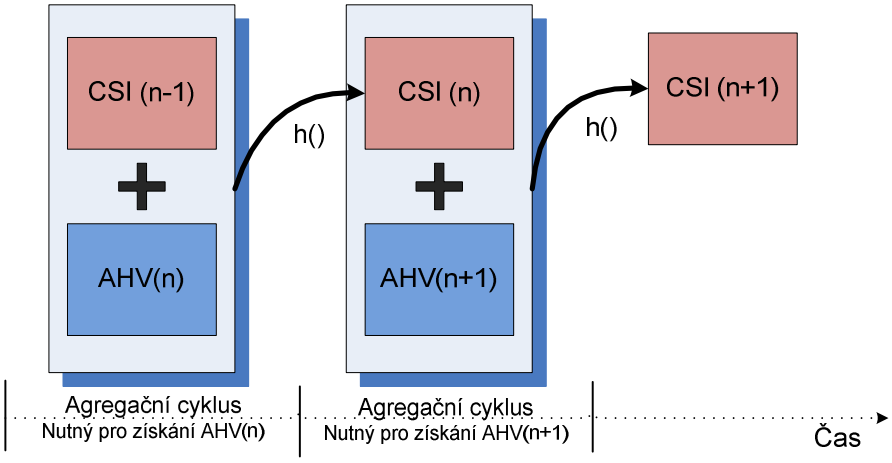


Obrázek 2: Autentizační strom Zdroj: autor

Na konci agregačního cyklu je vytvořen tzv. souhrnný otisk (AHV), Obrázek. Ten představuje otisk vygenerovaný z posloupnosti otisků všech archivovaných objektů během agregačního cyklu. Posledním krokem při sestavování stromu je orazítkování všech objektů archivovaných během agregačního cyklu časovým razítkem. To značí, že přesný čas kdy byl dokument svěřen archivační autoritě (pokud nebyl využit atribut TS při sestavování IT) není potvrzen časovým razítkem, ale důkazem o existenci dokumentu v čase je pouze relativní čas konce agregačního cyklu. Ten je tedy společný pro všechny archivované dokumenty během agregačního cyklu.

Druhý krok

Jeho podstata spočívá ve vytvoření tzv. souhrnné kryptografické informace (CSI). Ta vznikne provázáním kořenového otisku vygenerovaného na konci agregačního cyklu (AHV) s otisky vytvořenými v předchozích cyklech jako souhrnné informace (důkazy) o předchozích cyklech ($CSI_{(n-1)}$). Struktura závisí na použitém schématu provázaných otisků. V tomto případě bylo použito klasické binární schéma provázaných otisků 0. Schéma lze vidět znázorněné na Obrázek. Z něho je vidět, že k tomu abychom dostali příslušnou hodnotu CSI, potřebujeme mít pouze CSI z předchozího kroku a kořenovou hodnotu otisku (AHV) z kroku současného. Hodnota předchozího CSI navíc může být obsažena jako nepovinný atribut v IT (hodnota LSI) a lze ji tedy znovu spočítat a ověřit.



Obrázek 3: Lineární schéma Zdroj: autor

Audit

Důležitou součástí každé systémové implementace dlouhodobé archivace je audit. Auditem se rozumí kontrola integrity archivu a je uskutečňovaná zpravidla externími subjekty. Díky funkci konstantního ověřování, mají externí auditoři tuto činnost usnadněnou. Pro každý ověřovaný objekt je vypočítán otisk (za použití stejné hashovací funkce dle HID) a porovnán s otiskem uloženým v databázi (nebo v systémových metadatech). Jsou podporovány dva způsoby auditování a to:

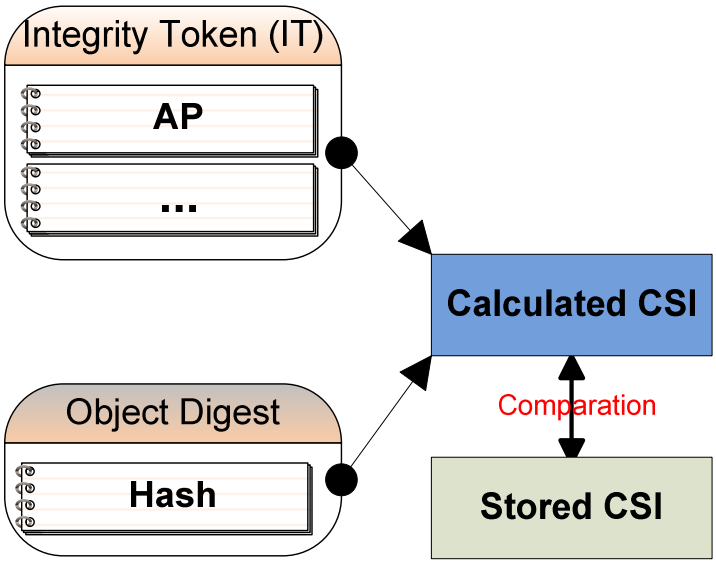
- Pravidelný (periodický).
  Externí zabezpečený server ověřující v pravidelných intervalech integritu archivu.
- Náhodný (na vyžádání).
  Při požadavku kontroly externím auditorem nebo uživatelem nad archivním objektem, na který má oprávnění.

Ověření integrity

Obecně lze ověřit integritu v rámci auditu ve dvou krocích:

- Vygenerování otisku určitého dokumentu a jeho kontrola s údajem uloženým v příslušném IT. Pokud následuje shoda lze pokračovat dalším krokem, pokud ne je třeba iniciovat kroky spojené s porušením integrity archivu. Tato akce však již přísluší pouze pověřeným osobám (např. správce archivu), nikoliv auditorovi samotnému.
- Použít vygenerovaný otisk spolu s AHV pro ověření CSI vytvořené v příslušném kole, ve kterém byl přidán do archivu. Shoda znamená, že objekt je nedotčený.

Princip ověřování integrity spočívá v provázání IT (především je důležitý AP atribut) a uloženého otisku auditovaného objektu k vypočítání hodnoty CSI. Schéma viz Obrázek.



Obrázek 4: Ověřování integrity objektu Zdroj: autor

Při samotném auditu je hlavním podkladem ověřování objektů, které archiv zajišťuje průběžně. Výsledkem je několik druhů zpráv, které charakterizují stav objektu:

- Obecná zpráva – např. že byl soubor přidán do archivu.
- Chybová hlášení – např. soubor je poškozen.

Zprávy dále obsahují následující atributy:

- Typ události – např. soubor nenalezen.
- Popis události – detailní popis události.
- Datum – datum, kdy se daná událost stala.

Systémová Metadata

Během auditu jsou dále prověřována uložená systémová metadata, vážící se ke každému archivovanému objektu. Atributy kolekcí (kolekce metadat se váže ke každému archivovanému dokumentu) systémových metadat jsou následující:

- File name.
  Jméno souboru včetně koncovky, case sensitive.
- Item path.
  Kompletní relativní cesta.
- First Seen.
  Datum kdy byl dokument poprvé zaregistrován do archivu.
- Last Seen.
  Datum a čas kdy byl objekt naposledy prověřen. V případě přidání do archivu má stejnou hodnotu jako „First Seen“.
- State.
  Reprezentuje současný stav souboru.
  - Valid.
    Objekt je archivován, čitelný a otisk souhlasí.
  - Invadlid.
    Objekt je archivován, ale příslušný otisk nesouhlasí.
  - Missing.
    Objekt chybí nebo není čitelný.
  - Registered.
    Objekt je archivován, ale chybí příslušné IT.
- Change date.
  Datum a čas kdy byl naposledy změněn stav objektu.
- Owner.
  Vlastník objektu.

Přístupová oprávnění

Důležitým prvkem v každém archivu jsou správně nastavená přístupová oprávnění. Od těch se pak odvozují jednotlivé role v archivu. Navrhovaná oprávnění pro systém dlouhodobé archivace jsou následující:

Tabulka 1: Systémová oprávnění Zdroj: autor

Přístup	Popis
Modify	Modifikování parametrů.
Browse	Prohlížení souborů a metadat.
Audit	Povolení auditu objektů.
View report	Zobrazení stavu archivu (archivovaných souborů).
Remove Item	Odebrání objektu z monitoringu.
Users	Přidávat nebo odebírat další uživatele.
Download Item	Stažení archivovaných objektů.

Závěr

V tomto příspěvku byl představen univerzální přístup k zachování integrity archivu elektronických dokumentů. Za použití technik jako je řetězení otisků a konstantní ověřování integrit svěřených objektů, zajišťuje efektivní správu celého systému dlouhodobé archivace. Článek na druhou stranu zatím neřeší zpochybnění důvěryhodnosti již archivovaných dokumentů v případě změny hlavní hashovací funkce, ale na způsobu řešení autor intenzivně pracuje.

Tento obecný princip byl také použit v roce 2011 v nabídce projektu národního digitálního archivu České republiky a to v části věnující se jeho bezpečnosti.

Literatura

BULDAS, Ahto, Peeter LAUD, Helger LIPMAA a Jan VILLEMSON, 1998: *Time-stamping with Binary Linking Schemes.* USA, Santa Barbara: Lecture Notes in Computer Science, vol. 1462, Springer, pp. 486-501. Srpen 1998.

DOSTÁLEK, Libor a kol., 2006: *Velký průvodce infrastrukturou PKI a technologií elektronického podpisu.* Praha: Computer press, 536 stran. Říjen 2006. ISBN: 80-251-0828-7.

SMORUL, Michael, Sangchul SONG a Joseph JAJA, 2009: *An Implementation of the Audit Control Environment (ACE) to Support the Long Term Integrity of Digital Archives*. North Carolina, Chapel Hill: DigCCurr. University of North Carolina.

JEL Classification: M15, L86

Summary

Ensuring Long-term Archivation System Integrity

This article introduced universal approach to integrity preservation of the long-term archive. Using techniques like hash linking and constant integrity monitoring of archived documents, ensures effective management of whole archival system. However, this article does not solve main hash function replacement yet, but author intensively works on a solution to this issue.

This general principle was used in 2011 in national digital archive of the Czech Republic tender, in the section devoted to its security.

Keywords: long-term archivation, document, hash, security, integrity, metadata.

Koncept stochastické dominance a interaktivní metoda vícekriteriálního programování jako analytické nástroje investičního rozhodování

Adam Borovička
adam.borovicka@vse.cz
Doktorand oboru Ekonometrie a operační výzkum
Školitel: prof. RNDr. Ing. Petr Fiala, CSc., MBA (pfiala@vse.cz)

Abstrakt: V příspěvku můžeme identifikovat dvě roviny. Jedna se týká teoretického představení konceptu stochastické dominance a interaktivní metody vícekriteriálního programování. Ve druhé pak aplikujeme zmíněné kvantitativní přístupy ke stanovení „optimální" struktury portfolia investičních instrumentů.

Princip stochastické dominance využíváme pro redukci poměrně početné množiny investičních příležitostí, vybrané varianty se budou potenciálně podílet na konečné skladbě investičního portfolia. Aplikace konceptu stochastické dominance je podmíněna stochastickým charakterem sledovaného problému. Jako náhodná veličina vystupuje výnosnost podílových fondů IS ČS, které jsou pro praktickou analýzu vybrány. Nedílnou součástí je zkoumání pravděpodobnostního rozdělení výnosů jednotlivých podílových fondů pomocí příslušných statistických testů. Prostřednictvím generátoru náhodných čísel vytváříme několik možných scénářů výnosů, s kterými počítáme v závěrečném procesu tvoření samotného portfolia. V proceduře konstrukce portfolia používáme interaktivní metodu vícekriteriálního programování, která je výpočetně založena na maximalizaci pozitivní odchylky od aspiračních úrovní zvolených investorem pro jednotlivé kriteriální funkce popisující výnos, riziko a náklady spojené s investicí. Investor dále sleduje finanční ukazatele typu Sharpeho indexu či modifikované durace, stejně tak diverzifikaci portfolia.

Po všech procedurách včetně postupných úprav nabízeného řešení analytikem podle preferencí rozhodovatele dostáváme výslednou skladbu portfolia podílových fondů.

Klíčová slova: portfolio, rozhodování, stochastická dominance

Úvod

V příspěvku můžeme nalézt dvě zásadní poslání. První je tvořeno představením konceptu stochastické dominance a interaktivní metody vícekriteriálního programování, druhé pak praktickým uplatněním popsaných přístupů na kapitálovém trhu při tvorbě investičního portfolia podílových fondů[1].

Jelikož je investor vystaven poměrně velkému množství investičních instrumentů, koncept stochastické dominance přispěje k redukci možných investičních příležitostí, tedy dojde ke stanovení užší skupiny podílových fondů, které mohou potenciálně tvořit výsledné investiční portfolio. Jelikož bereme v potaz stochastický charakter zkoumaného problému, uvažujeme při výstavbě portfolia několik scénářů. Stanovení definitivní struktury podílových listů[2] probíhá pomocí interaktivní metody vícekriteriálního programování, tedy s aktivní participací potenciálního investora na výsledné skladbě investice.

Koncept stochastické dominance[3]

Princip stochastické dominance umožňuje alternativní přístup k měření (pojetí) rizika v rozhodovacích procesech, které se mnohdy neobejdou bez přítomnosti stochastických prvků.

[1] *Podílový fond* je vnitřní organizační jednotka investiční společnosti bez právní subjektivity (Valach, 2006, s. 240).
[2] *Podílový list* je cenný papír, který představuje podíl podílníka na majetku podílového fondu, a se kterým jsou spojena další práva vyplývající se zákona (zákon ÚZ č. 851, 2011, s. 355).
[3] Celá tato kapitola je čerpána z publikace (Zelený, 2005, s. 400 - 412).

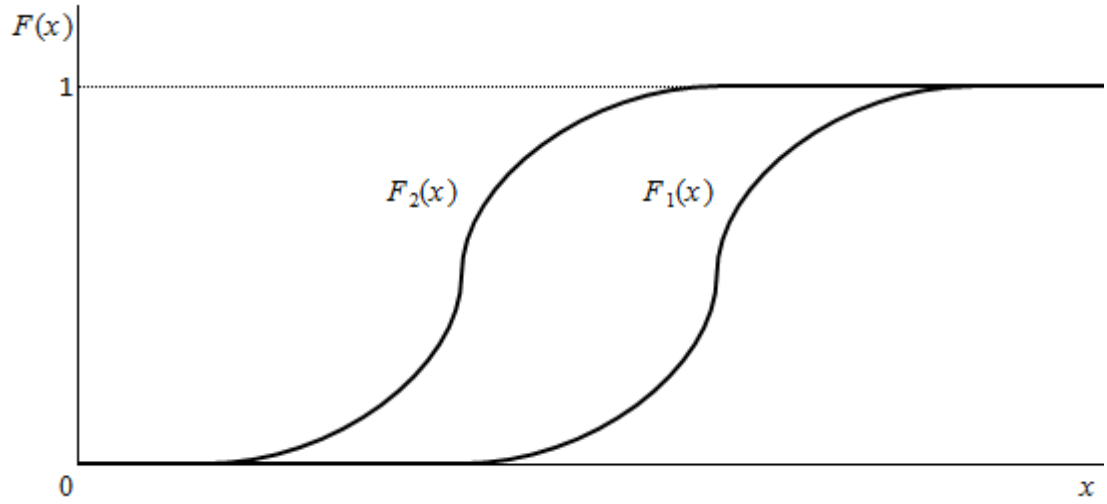
Pro následující účely definujeme *n* náhodných veličin $X_1, X_2, \ldots, X_n$, které mohou nabývat různých hodnot *x*. V případě povědomosti o distribuční funkci jednotlivých proměnných $F_i(x)$ říkáme, že X_1 *stochasticky dominuje* X_2, jestliže platí buď

$$\forall x \quad F_1(x) \leq F_2(x) \tag{1}$$

vyjadřující první stupeň stochastické dominance, nebo

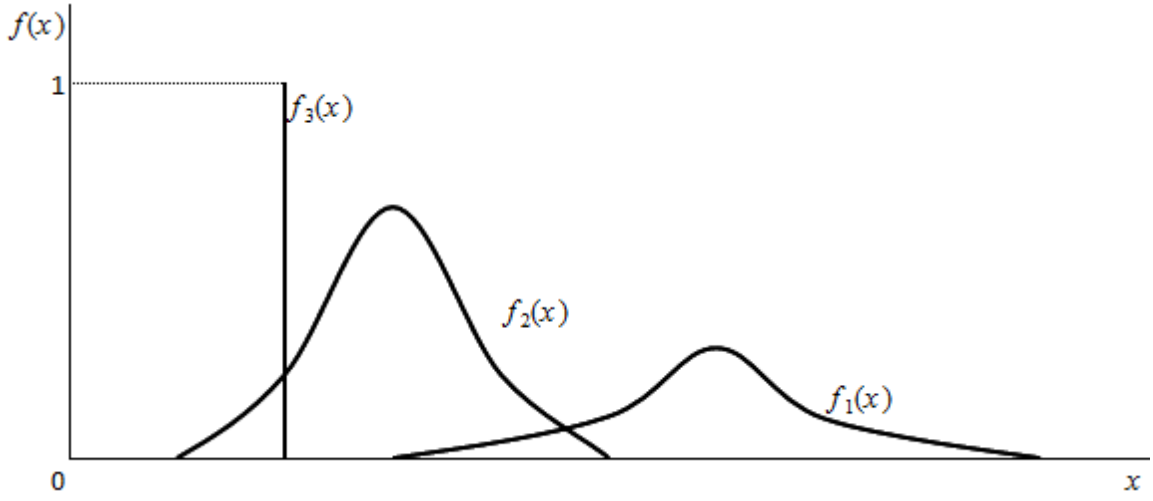
$$\forall x \quad \int_{-\infty}^{x} F_1(y)dy \leq \int_{-\infty}^{x} F_2(y)dy \tag{2}$$

popisující druhý stupeň stochastické dominace. Pro ilustraci zobrazíme na následujícím grafu (Obrázek 1) první stupeň stochastické dominance. Pokud X_1 dominuje X_2, racionální rozhodovatel vybírá méně riskantní variantu X_1, bez ohledu na její charakteristiku variability.



Obrázek 1: Veličina X_1 stochasticky dominuje veličinu X_2

Na dalším grafu (Obrázek 2) se podíváme na problém rizikovosti z pohledu hustot pravděpodobnosti příslušných veličin.



Obrázek 2: Veličiny X_2 a X_3 jsou stochasticky dominovány veličinou X_1

Při pohledu na riziko optikou stochastické dominance veličiny X_2 a X_3 jsou stochasticky dominovány proměnnou X_1, ač vykazuje nejvyšší variabilitu. Tedy rozhodovatel volí takovou alternativu, u které je

pravděpodobnost nenabytí předem stanovené hodnoty náhodné veličiny nejmenší. Je tedy zřejmé, že pravděpodobnostní rozdělení s největším rozptylem stochasticky dominuje ostatní.

Aplikace principu stochastické dominance

Zavedeme množinu nezávislých proměnných výnosů x_i spojených s určitými investičními instrumenty i, kde i = 1, 2, …, n. Pro každou náhodnou veličinu X_i vyjadřující výnos i-té investiční příležitosti je pro hodnoty x_i definován interval, tedy

$$x_i \in \langle a_i; b_i \rangle \subseteq \langle a; b \rangle, \tag{3}$$

kde
$\langle a; b \rangle$ - interval, do kterého spadají všechny hodnoty *x*.

Označme si $f_i(x), F_i(x), \overline{x_i}$ a σ_i popořadě jako hustotu pravděpodobnosti, distribuční funkci, střední hodnotu a směrodatnou odchylku proměnných výnosů i-tého investičního instrumentu.

V porovnání dvou množin proměnných výnosů můžeme vyjádřit vztah stochastické dominance jako

$$1 \succ 2 \Leftrightarrow \begin{Bmatrix} F_1(x) \le F_2(x) & pro\ všechna\ x \in \langle a,b \rangle \\ & a \\ F_1(x) < F_2(x) & pro\ některá\ x \in \langle a,b \rangle \end{Bmatrix}, \tag{4}$$

kde
$\succ$ znamená „stochasticky dominuje“,
$\Leftrightarrow$ vyjadřuje výraz „právě tehdy“.

Tedy investor v duchu popsaného principu preferuje investiční instrument 1 s výnosy vyjádřené náhodnou veličinou X_1 před investicí 2 (X_2).

Pravidlo zajišťující první stupeň stochastické dominance je velmi citlivé na nízké výnosy. V tomto duchu mohou být velmi snadno odmítnuty investiční instrumenty s nízkými výnosy. Je tedy vyžadováno takové opatření, které by nežádoucí jev napravovalo. Navrhujeme tedy určité přiměřené hypotézy racionálního chování, které budou formulovány prostřednictvím konkrétních pravidel.

Rozhodovací pravidlo č. 1

Investiční instrument 1 bude preferován před investičním instrumentem 2, jestliže minimální dosažitelná úroveň výnosu u investice 1 a_1 je větší než minimální dosažitelná úroveň výnosu u instrumentu 2 a_2. Platí tedy

$$a_1 > a_2 \Rightarrow 1 \succ 2. \tag{5}$$

Investor většinou stanovuje minimální úroveň požadovaného výnosu. Tuto hladinu pro *m-tého* investora označíme jako r_m. Poté můžeme psát první minimalizační účelovou funkci

$$R_{im}^{(1)} = P_i(x < t_m) = P_i(x < \max\{L; r_m\} = P_i\left(x < \max\left\{\max_i a_i, r_m\right\}\right), \tag{6}$$

kde
$R_{im}^{(1)}$ - první komponenta rizika R_i,
t_m - individuální efektivní prahová hodnota spojená s výnosem,
L - nejvyšší hodnota z minimálních úrovní požadovaných výnosů u jednotlivých investičních variant.

Optikou výše zmíněného vztahu nejméně riziková investiční příležitost je taková, u které je zajištěna nejmenší pravděpodobnost, že realizovaný výnos nabude hodnot menších než individuální efektivní prahová hodnota výnosu t_m.

Rozhodovací pravidlo č. 2:

Investiční instrument 1 bude preferován před investičním instrumentem 2, jestliže maximální dosažitelná úroveň výnosu u investice 1 b_1 je větší než maximální dosažitelná úroveň výnosu u instrumentu 2 b_2. Platí tedy

$$b_1 > b_2 \Rightarrow 1 \succ 2. \tag{7}$$

Můžeme stanovit druhou účelovou funkci

$$1 - R_i^{(2)} = P_i(x \geq S), \tag{8}$$

kde
$S = \max_i b_i$.

Výraz $(1 - R_i^{(2)})$ maximalizujeme, popř. minimalizujeme $R_i^{(2)}$, což reprezentuje druhou část rizikové složky. Z pohledu druhé minimalizační kriteriální funkce platí, že nejméně riziková investiční alternativa je taková, která minimalizuje pravděpodobnost nerealizování nejlepší úrovně výnosu.

Jak jsme tedy popsali, riziko se skládá v tomto konceptu ze dvou dimenzí. Celkové riziko *i-té* investice vyjádříme dvousložkovým vektorem $R_i = (R_{im}^{(1)}, R_i^{(2)})$. Obě komponenty a jejich simultánní uvažování při analýze rizika je plně konzistentní s konceptem prvního stupně stochastické dominance. Jinými slovy, investor aplikující stochastickou dominanci nemůže vybrat takovou alternativu, která by byla v rozporu s rizikem R_i.

Nyní prezentujeme třetí pravidlo racionální chování, které je založeno na znalosti očekávané hodnoty výnosů.

Rozhodovací pravidlo č. 3

Investice 1 bude preferována před investicí 2, jestliže očekávaná hodnota výnosu první investice $\overline{x_1}$ je větší než očekávaná hodnota výnosu investice druhé $\overline{x_2}$. Formálně zapsáno

$$\overline{x_1} > \overline{x_2} \Rightarrow 1 \succ 2. \tag{9}$$

Pokud se dostaneme do situace rizikové indiference mezi oběma alternativami, pak se můžeme plně spolehnout na ukazatel očekávaného výnosu.

Nyní sestavíme vektor ze třech výše definovaných účelových funkcí pro *i-tý* investiční instrument a *m-tého* investora, který budeme nazývat jako *klasifikační vektor portfolia*[4]

$$PRV_{im} = (1 - R_{im}^{(1)}, \overline{x_i}, 1 - R_i^{(2)}). \tag{10}$$

Je čas definovat nedominovanou variantu v souvislosti se složkami klasifikačního vektoru. *Nedominovaná* varianta je taková dostupná varianta, u které není možné zvýšit jednu z komponent výše zmíněného vektoru bez snížení alespoň jedné z ostatních. Všechny nedominované varianty tvoří množinu nedominovaných variant. Ideální variantou rozumíme takovou, která bude nabývat maximálních hodnot všech tří komponent vektoru. Jelikož většinou tato varianta není dosažitelná, snažíme se zvolit takovou nedominovanou variantu, která bude ideální alternativě co možná nejblíže.

V případě znalosti střední hodnoty $\overline{x_i}$ a směrodatné odchylky σ_i všech pravděpodobnostních rozdělení, můžeme modifikovat výpočet hodnot *L*, *S* a tedy i t_m. Podstata určení bude ve vzdálenosti od průměru vyjádřená směrodatnou odchylkou multiplikovanou konstantou *k*, resp. *k'*. Tedy

[4] *PRV* = Prospect (or Portfolio) Ranking Vector

$$L = \max_i(\overline{x_i} - k'\sigma_i),$$
$$S = \max_i(\overline{x_i} + k\sigma_i), \tag{11}$$
$$t_m = \max(L; r_m),$$

kde
k'a k - konstanty volené rozhodovatelem, které mohou být stanoveny na nulové úrovni či z praktického hlediska spíše na úrovních odpovídajícím kladným hodnotám.

Klasifikační vektor s částečnou informací

V situaci částečné informace vycházíme z předpokladu, že investor zná průměr a rozptyl, ale vůbec nemá povědomost o pravděpodobnostech vyžadovaných při výpočtu obou složek rizika. K vyřešení situace poslouží forma Tchebyshevovy nerovnosti, jejíž použití má ale negativní dopad na hodnověrnost řešení. Stanovíme si hodnoty k_i

$$k_i = \frac{S - \overline{x_i}}{\sigma_i}, \tag{12}$$

kde
$S = \max_i(\overline{x_i} + k\sigma_i),$

$\overline{x_i} = \frac{\sum_{j=1}^{s}(x_j)}{s}$ - průměrný výnos *i-tého* investičního instrumentu,

$\sigma_i = \sqrt{\frac{\sum_{j=1}^{s}(x_j - \overline{x_i})^2}{s}}$ - směrodatná odchylka výnosů *i-tého* investičního instrumentu.

Dále definujeme hodnotu k'_{im} dle vztahu

$$k'_{im} = \frac{t_m - \overline{x_i}}{\sigma_i}, \tag{13}$$

kde
$L = \max_i(\overline{x_i} - k'\sigma_i),$
$t_m = \max(L; r_m).$

Ke stanovení prvního a třetího člena klasifikačního vektoru využijeme z teorie pravděpodobnosti konceptu Cramerovy nerovnosti. Pak bude pro třetí komponentu vektoru platit

$$1 - R_i^{(2)} = P(x_i \geq S) = P(x_i \geq \overline{x_i} + k_i\sigma_i) \leq \frac{1}{k_i^2 + 1}. \tag{14}$$

Pro první část rizika platí

$$R_{im}^{(1)} = P(x_i \leq \overline{x_i} + k'_{im}\sigma_i) \geq \frac{1}{k'^2_{im} + 1}. \tag{15}$$

První komponentu klasifikačního vektoru lze vyjádřit ve tvaru

$$1 - R_{im}^{(1)} = P(x_i \geq \overline{x_i} + k'_{im}\sigma_i) \leq 1 - \frac{1}{k'^2_{im} + 1}. \tag{16}$$

Nakonec stanovíme klasifikační vektor portfolia pro *i –tý* investiční instrument (*i-tou* variantu) *m-tého* investora (*m-tého* rozhodovatele) pomocí limitů na pravých stranách vztahů (14) a (16)[5]

$$PRV_{im} \overset{P}{\Leftrightarrow} \left(\frac{1}{k_{im}^{'2}+1}, \overline{x_i}, \frac{1}{k_i^2+1} \right), \tag{17}$$

kde

$\overset{P}{\Leftrightarrow}$ označuje pravděpodobnostní implikaci.

V analogii se stochastickou dominancí můžeme stanovit pravidla k určení dominance ve smyslu PRV. Tak tedy varianta *p* dominuje variantu *q* v duchu PRV právě tehdy, je-li

$$\begin{aligned} &[a] \quad p \underset{PRV_1}{\succ} q \overset{P}{\Leftrightarrow} \frac{1}{k_{pm}^{'2}+1} \geq \frac{1}{k_{qm}^{'2}+1}, \\ &[b] \quad p \underset{PRV_2}{\succ} q \Leftrightarrow \overline{x_p} \geq \overline{x_q}, \\ &[c] \quad p \underset{PRV_3}{\succ} q \overset{P}{\Leftrightarrow} \frac{1}{k_p^2+1} \geq \frac{1}{k_q^2+1}, \end{aligned} \tag{18}$$

kde alespoň jedna nerovnost musí být splněna jako ostrá[6].

Jak jsme si mohli povšimnout, výše zmíněný koncept dominance velmi závisí na hodnotách stanovených investorem (rozhodovatelem) r_m, k a $k^{'}$. Jako minimální akceptovatelný výnos *m-tým* investorem označujeme r_m. Blíže se však zaměříme na další dvě zmíněné hodnoty. Velikost konstant *k* a *k'* závisí na důvěře investora ve výskyt extrémních hodnot výnosů mezi různými dostupnými variantami. Čím budeme uvažovat větší hodnoty pozitivně odchýlených od průměru, tím budou hodnoty *k* a *1/k'* nabývat vyšších hodnot. Naopak u pesimistického pohledu budou hodnoty *1/k* a *k'* vysoké. „Hyperoptimistický" investor volí $k = \infty$ a $k' = 0$, naopak „hyperpesimistický" stanovuje hodnoty $k' = \infty$ a $k = 0$.

Interaktivní metoda vícekriteriálního rozhodování

K nalezení uspokojivé struktury investičního portfolia využijeme interaktivní přístup vícekriteriálního programování. Vstupní informací od rozhodovatele jsou váhy jednotlivých kritérií, které jsou určeny pomocí Saatyho metody (více viz Saaty, 1980, s. 17-21, či Fiala, 2008, s. 53-54). Dále musí rozhodovatel stanovit aspirační úrovně pro kritéria, tedy minimální požadované úrovně pro maximalizační kritéria a maximální požadované úrovně pro minimalizační kritéria. Analytik[7] nabízí řešení, které rozhodovatel (investor) posuzuje a případně žádá vylepšení hodnoty některého z kritérií.

Nyní si interaktivní metodu vícekriteriálního programování blíže popíšeme. Budeme řešit úlohu, která je založena na maximalizaci vzdálenosti hodnot kritérií od jejich aspiračních úrovní, což vyjádříme pomocí eukleidovské metriky. Tedy

$$\begin{aligned} z &= \sqrt{\sum_{i=1}^{k} [w_i (f_i^*(x) - f_i(x))]^2} \rightarrow \max \\ x \in X^1 &= \left\{ x \in R^n;\, f_i(x) \geq f_i^*(x), i \in A;\, f_i(x) \leq f_i^*(x), i \in B;\, A \cup B = \{1,2,\ldots,k\} \right\} \\ x \in X^2 &= \left\{ x \in R^n;\, g_l(x) \leq b_l, l = 1,2,\ldots,m;\, x_j \geq 0,\, j = 1,2,\ldots,n \right\}, \end{aligned} \tag{19}$$

kde

$f_i(x)$ - *i-tá* kriteriální funkce,

---

[5] Platí pro $k_{im}^{'} > 0$.

[6] Nerovnosti platí pro $k_{pm}^{'}, k_{qm}^{'} > 0$.

[7] *Analytik* podporuje rozhodovací proces z hlediska zpracování preferenčních informací rozhodovatele a předkládá řešení.

w_i - váha *i-té* kriteriální funkce,
$f_i^*(x)$ - limitní (aspirační) úroveň *i-tého* kritéria,
x - n-složkový vektor proměnných,
$g_l(x)$, resp. b_i - levá, resp. pravá strana některých vlastních omezení, kterých je m,
A - množina obsahující indexy kritérií maximalizační povahy,
B - množina obsahující indexy kritérií mimimalizační povahy.

Množina X^1 spoluvytvářející množinu přípustných řešení obsahuje vyjádření požadovaných minimálních či maximálních úrovní (podle typu kritéria) kriteriálních funkcí. Množina X^2 obsahuje ostatní vlastní a všechny nevlastní omezení modelu. Sjednocením obou množin dostáváme celkovou množinu přípustných řešení dané úlohy. Podle konkrétní situace se samozřejmě model může rozšířit o další proměnné například binárního charakteru.

Než se pustíme do živé interakce mezi rozhodovatelem a analytikem, podíváme se na přístup ke stanovení aspiračních úrovní. Sestavíme takový matematický model, který bude optimalizovat hodnotu konkrétní kriteriální funkce na množině X^2

$$\begin{gathered} f_i(x) \to \max(\min) \\ x \in X^2 = \left\{ x \in R^n; g_l(x) \le b_l, l = 1,2,\ldots,m; x_j \ge 0, j = 1,2,\ldots,n \right\}. \end{gathered} \tag{20}$$

Optimální hodnota příslušného kritéria je vodítkem pro stanovení aspirační úrovně. Jinými slovy udává u maximalizačního kritéria její horní hodnotu a u minimalizačního kritéria její dolní hodnotu. Je evidentní, že rozhodovatel bude volit aspirační úrovně podstatně nižší, resp. vyšší, které by nevylučovaly v rozhodovací úloze existenci přípustného řešení. V případě stochastického problému o několika scénářích se stává vodítkem pro stanovení aspiračních úrovní průměrná hodnota účelových funkcí všech úloh.

Po sestrojení modelu analytik nabízí optimální řešení úlohy. Pokud se rozhodovateli nelíbí ani jedna hodnota vektoru proměnných x v souvislosti s dosaženými hodnotami jednotlivých kritérií, proces končí nezdarem, rozhodovatel musí přehodnotit své preference. V případě nespokojenosti jen s některými kritérii dochází k úpravě[8]

$$\begin{gathered} z = \sqrt{\sum_{i=1}^{k} [w_i (f_i^*(x) - f_i(x))]^2} \to \max \\ x \in X^1 = \left\{ x \in R^n; f_i(x) \le f_i^*(x), i \in A; f_i(x) \ge f_i^*(x), i \in B; A \cup B = \{1,2,\ldots,k\} \right\} \\ x \in X^2 = \left\{ x \in R^n; g_l(x) \le b_l, l = 1,2,\ldots,m, x_j \ge 0; j = 1,2,\ldots,n \right\} \\ x \in X^3 = \left\{ x \in R^n, f_i^N(x) \ge \overline{f_i^N} + \Delta_i, i \in A^N, A^N \subseteq A; f_i^N(x) \le \overline{f_i^N} - \Delta_i, i \in B^N, B^N \subseteq B \right\}. \end{gathered} \tag{21}$$

Množina X^3 obsahuje požadované úpravy rozhodovatelem vybraných kritérií. U konkrétních maximalizačních kritérií $f_i^N(x)$ požaduje zvýšit neuspokojivou hodnotu $\overline{f_i^N}$ o Δ_i, u vybraných minimalizačních kritérií $f_i^N(x)$ chce naopak snížit neuspokojivou hodnotu $\overline{f_i^N}$ o Δ_i. Množina A^N zahrnuje všechny indexy maximalizačních kritérií, s jejichž hodnotami není rozhodovatel spokojen, množina B^N pak indexy všech minimalizačních kritérií, s jejichž hodnotami není spokojen.

K úpravě řešení nemusí dojít jen v případě, že rozhodovatel není striktně nespokojen s řešením, nýbrž i v situaci, kdy chce mít informativní představu, jak by se řešení změnilo, pokud by si dovolil požádat o vylepšení hodnoty některých kriteriálních funkcí. Jak je patrné, rozhodovatel neklade při vylepšování příslušných hodnot žádné požadavky na velikost možného zhoršení ostatních kritérií, které jsou

[8] Například některá kritéria dosahují „pouze“ svých aspiračních úrovní, což může být pro rozhodovatele ne zcela uspokojivé.

z pohledu investora hodnotově v pořádku. Tímto způsobem algoritmus pokračuje až do okamžiku, kdy je rozhodovatel spokojen či smířen s podobou řešení.

Stochastické programování

Pokud nechceme akceptovat v mnoha případech zjednodušení problému, zahrnujeme do modelu náhodnostní charakter. Optimální řešení, resp. hodnota účelové funkce zkoumaného problému pak závisí na realizaci přítomných náhodných veličin (Hadley, 1964, s. 158-159).

Prakticky budeme řešit stochastické modely v duchu optimalizace pomocí metody Monte Carlo, kdy generujeme hodnoty konkrétních pravděpodobnostních rozdělení příslušných náhodných veličin, čímž dostaneme různé scénáře. Každému scénáři odpovídá určitá hodnota účelové funkce modelu. Generování scénářů omezíme předem stanoveným počtem. Vybereme řešení s největší hodnotou této funkce (Dlouhý a kol., 2007, s. 54). K identifikaci pravděpodobnostních rozdělení náhodných veličin poslouží *Kolgomorovův-Smirnovův test* dobré shody (více viz Hindls a kol., 2006, s. 157-158).

Cílové programování – nedělitelnost aktiv

Mnohdy se může stát, že výsledná struktura portfolia nakonec nemůže být zcela dodržena, protože obchod s příslušnými investičními instrumenty probíhá pouze ve standardizovaných jednotkách o aktuální tržní hodnotě. Jinými slovy musíme dodržovat podmínku nedělitelnosti aktiv. Požadavku obchodu ve standardizovaných jednotkách vyhovuje následující model, který upravuje „optimální“ strukturu portfolia

$$
\begin{aligned}
z &= \sum_{i=1}^{n} (d_i^+ + d_i^-) \rightarrow \min \\
c_i b_i + d_i^- - d_i^+ &= M x_i^* \quad i = 1, 2..., n \\
\sum_{i=1}^{n} c_i b_i &\leq M \\
d_i^+ d_i^- &= 0 \quad i = 1, 2..., n \\
d_i^+, d_i^- &\geq 0 \quad i = 1, 2..., n \\
b_i &\in Z_0^+ \quad i = 1, 2..., n,
\end{aligned}
\tag{22}
$$

kde

c_i - cena za standardizovanou jednotku *i-tého* investičního instrumentu,

b_i - počet standardizovaných jednotek *i-tého* investičního instrumentu,

M - celková zamýšlená investovaná částka,

x_i^* - podíl *i-tého* investičního instrumentu na portfoliu (skutečně investované částce),

d_i^-, resp. d_i^+ - záporná, resp. kladná odchylka od původního podílu *i-tého* investičního instrumentu na portfoliu.

Lze si povšimnout, že modelovaný proces je založen na konceptu cílového programování (vice viz Kosmidou et al., 2004, s. 85-90). Nevýhodou modelu je skutečnost, že nová struktura portfolia se někdy může poměrně významně lišit od původní, která byla stanovena na základě interaktivní rozhodovací procedury zahrnující určité investorovy požadavky na portfolio. Nová skladba portfolia tak může původní požadavky porušovat. Zde tedy evidujeme určité zjednodušení situace, která by si ještě zasloužila hlubší propracování. V případě praktické aplikace s vybranými podílovými fondy popsaný problém významně nenastává, protože podílové listy vybrané pomocí principu stochastické dominance se obchodují po kusech a jejich cena se pohybuje kolem jedné koruny české.

Sestavení portfolia podílových fondů

Potenciální investor se rozhoduje část svých volných peněžních prostředků vložit do otevřených podílových fondů[9]. Jelikož je dlouholetým klientem České spořitelny, rozhodl se pro investici do otevřených podílových fondů Investiční společnosti České spořitelny. Jednoho dne se tedy vydal na jednu z poboček banky za investičním poradcem, aby s ním prodiskutoval možnosti investice.

Investor vybírá ze čtyř skupin podílových fondů, a to *smíšených*, *dluhopisových*, *akciových* a fondů *peněžního trhu*. V nabídce Investiční společnosti České spořitelny jsou v těchto skupinách následující fondy (Tabulka 1).

Tabulka 1: Nabídka otevřených podílových fondů České spořitelny

Fondy peněžního trhu	Smíšené fondy	Dluhopisové fondy	Akciové fondy
Sporoinvest	*Osobní portfolio 4 Plus* *Fond řízených výnosů* *Konzervativní MIX* *Vyvážený MIX* *Dynamický MIX* *Akciový MIX*	*Sporobond* *Trendbond* *Bondinvest* *Korporátní dluhopisový* *High Yield dluhopisový*	*Sporotrend* *Global Stocks* *Top Stocks*

Cílem investora je sestavit takové portfolio, které by splňovalo jeho požadavky na něj kladené.

Požadavky a preference investora

Zaprvé si investující subjekt přeje, aby v portfoliu byly zastoupeny všechny čtyři skupiny podílových fondů, přesněji aby byl vybrán z každé skupiny jeden zástupce. Zároveň ale musí platit, aby investice do jednoho podílového fondu nepřesáhla 30 % portfolia. Investovaná částka nepřesáhne 100 000 Kč. Dále sleduje určité investiční ukazatele, jako je modifikovaná durace[10] či Sharpeho index[11] výsledného portfolia. Investor stanovuje maximální velikost modifikované durace pro celé portfolio na hladině 2,5, Sharpeho index pak na minimální hodnotě 0,1.

Dále investor stanovuje hodnotící kritéria. Bude jimi výkonnost (výnos), riziko dle (Sharpe, 1970, s. 42) a náklady portfolia. Důležitost kritérií investor vyjadřuje pomocí vah stanovených Saatyho metodou (Tabulka 2).

Tabulka 2: Váhy jednotlivých kritérií

Výnos	Riziko	Náklady
0,642	0,292	0,066

Rozhodovací procedura

Nejdříve redukujeme počet potenciálních investic do otevřených podílových fondů pomocí konceptu stochastické dominance. Podle Kolgomorova-Smirnovova testu[12] můžeme konstatovat, že výnosy většiny podílových fondů mohou být popsány normálním pravděpodobnostním rozdělením[13]. Pouze u fondů Sporoinvest, Korporátní dluhopisový, High Yield dluhopisový, Fond řízených výnosů a

[9] *Otevřený podílový fond* je takový fond, jehož spravující investiční společnost je povinna podílové listy na žádost podílníka odkoupit do určitého termínu za aktuální hodnotu majetku fondu připadající na jeden podílový list (Veselá, 2011, s. 267).

[10] *Modifikovaná durace* kvantifikuje úrokové riziko (více viz Málek, 2003, s. 22, či Veselá 2011, s. 617-623).

[11] *Sharpeho index* měří výkonnost konkrétního investičního instrumentu s ohledem na jeho rizikovost (více viz Gordon a kol., 1989, s. 404).

[12] *Kolgomorovův-Smirnovův tes*t byl proveden v softwaru *Crystal Ball* v rámci prostředí MS EXCEL.

[13] *Normální rozdělení* je spojité pravděpodobnostní rozdělení se dvěma parametry, střední hodnotou μ a rozptylem σ^2.

Sporotrend se jeví jako vhodnější rozdělení logistické[14]. Po stanovení všech vstupních údajů $(k = k^{'} = 1, r_m$ dle průměrných výnosů) využijeme princip stochastické dominance v každé skupině podílových fondů, abychom redukovali velkou množinu investičních příležitostí. V případě znalosti pravděpodobnostních rozdělení sledovaných veličin vychází vybraná množina totožně jako v situaci částečné informace. Vybrali jsme tedy následující podílové fondy (Tabulka 3).

Tabulka 3: Podílové fondy vybrané pomocí principu stochastické dominance

Fondy peněžního trhu	Smíšené fondy	Dluhopisové fondy	Akciové fondy
Sporoinvest	*Akciový MIX*	*High Yield dluhopisový*	*Sporotrend* *Top Stocks*

Aplikace interaktivní metody vícekriteriálního programování

Jak již bylo zmíněno výše, nachází se v modelu náhodná veličina – (mesíční) výkonnost (výnos) podílových fondů. Vygenerujeme si deset možných scénářů pomocí generátoru náhodných čísel příslušných pravděpodobnostních rozdělení[15] a budeme řešit dle (19) matematický model[16], který bude základem pro použití interaktivní metody vícekriteriálního programování. Tedy

$$z = \sqrt{\left[0,642\left(1,17 - \sum_{i=1}^{5} v_i x_i\right)\right]^2 + \left[0,292\left(6,32 - \sqrt{\sum_{i=1}^{5}\sum_{j=1}^{5} x_i x_j \sigma_{ij}}\right)\right]^2 + \left[0,066\left(2,69 - \sum_{i=1}^{5} n_i x_i\right)\right]^2} \rightarrow \max$$

$$\left.\begin{array}{c} \sum_{i=1}^{5} v_i x_i \geq 1,17 \\ \sqrt{\sum_{i=1}^{5}\sum_{j=1}^{5} x_i x_j \sigma_{ij}} \leq 6,32 \\ \sum_{i=1}^{5} n_i x_i \leq 2,69 \end{array}\right\} X^1$$

$$\left.\begin{array}{c} \dfrac{\sum_{i=2}^{3} d_i x_i}{\sum_{i=2}^{3} x_i} \leq 2,5 \\ \sum_{i=1}^{5} s_i x_i \geq 0,1 \\ \sum_{i=1}^{5} x_i = 1 \\ x_i \leq 0,3 y_i \quad i = 1,2,\ldots,5 \\ y_i = 1 \quad i = 1,2,3 \\ \sum_{i=4}^{5} y_i = 1 \\ x_i \geq 0 \quad i = 1,2,\ldots,5 \\ y_i \in \{0,1\} \quad i = 1,2,\ldots,5 \end{array}\right\} X^2,$$

[14] *Logistické rozdělení* je spojité pravděpodobnostní rozdělení se dvěma parametry, střední hodnotou a parametrem měřítka.

[15] Ke generování náhodných čísel byly využity funkce programu *SPSS* a *MS EXCEL*.

[16] Všechny použité optimalizační modely byly řešeny v modelovacím systému *LINGO*.

kde

x_i - podíl *i-tého* podílového fondu na portfoliu[17],

y_i - (ne)investice do *i-tého* podílového fondu,

v_i - měsíční výnos (výkonnost) *i-tého* podílového fondu,

n_i - náklady spojené s investicí do *i-tého* podílového fondu,

d_i - modifikovaná durace *i-tého* podílového fondu obsahující dluhopisy,

s_i - Sharpeho index *i-tého* podílového fondu,

σ_{ij} - kovariance výnosů *i-tého* a *j-tého* podílového fondu.

Účelová funkce maximalizuje součet čtverců odchylek od aspiračních úrovní jednotlivých kritérií, množina X^1 obsahuje omezení na splnění minimálních (maximálních) požadovaných úrovní kritérií, množina X^2 pak zahrnuje požadavky na velikost modifikované durace u podílových fondů obsahující dluhopisy a Sharpeho index portfolia, stejně tak na diverzifikaci portfolia a počet podílových fondů. Dále obsahuje nezbytné podmínky nezápornosti, resp. binarity pro příslušné proměnné. Aspirační úrovně jsou stanoveny na třetině optimálních hodnot maximalizačních kritérií a pěti třetinách optimálních hodnot minimalizačních kritérií dle konkrétní úlohy (20).

Zakončení procedury

Po prvotní optimalizaci úlohy dostáváme následující strukturu portfolia (Tabulka 4).

Tabulka 4: Skladba portfolia před úpravami

Podílový fond	Podíl	Kritéria	Hodnoty
Akciový MIX	***10 %***	Výnos	8,4 %
Sporoinvest	***30 %***		
High Yield	***30 %***	Riziko	5,95 %
Sporotrend	***30 %***		
Top Stocks	***0 %***	Náklady	1,59 %

Budoucího investora velmi těší výnos investice, který vysoce převyšuje zvolenou aspirační úroveň. Zaměřuje se tedy na riziko, které by si přál snížit alespoň na hodnotu 5,5 %. Analytik nabízí upravené řešení (Tabulka 5).

Tabulka 5: Výsledné portfolio

Podílový fond	Podíl	Kritéria	Hodnoty
Akciový MIX	***14 %***	Výnos	7,47 %
Sporoinvest	***30 %***		
High Yield	***30 %***	Riziko	5,5 %
Sporotrend	***26 %***		
Top Stocks	***0 %***	Náklady	1,59 %

Kromě požadované změny rizika se snížil o více než jeden procentní bod výnos. Tato změna je pro investora přijatelná a ještě žádá snížení o 0,5 procentního bodu rizika. Při tomto snížení opět rapidně klesá výnos (cca na úroveň 6,4 %). Investor se tedy rozhoduje, že na tuto změnu již nepřistoupí a spokojí se s řešením předešlým (viz Tabulka 5), které samozřejmě dle (22) respektuje nedělitelnost obchodovaných aktiv. Je třeba připomenout, že výsledné řešení vykazuje pravděpodobnostní charakter.

[17] $i=1$ (Sporoinvest), $i=2$ (Akciový MIX), $i=3$ (High Yield dluhopisový), $i=4$ (Sporotrend), $i=5$ (Top Stocks).

Závěr

Cílem článku bylo představit koncept stochastické dominance, navrhnout interaktivní metodu vícekriteriálního programování s praktickým příkladem na závěr. Princip stochastické dominance využíváme jako nástroj hodnocení variant vhodný pro redukci velmi mohutné množiny alternativ. Pro stanovení „optimální" struktury portfolia využíváme metodu vícekriteriálního programování, kdy v průběhu dochází postupně k úpravě řešení až do podoby rozhodovatelem přijatelné. V duchu praktické aplikace neopomineme zahrnout do rozhodovací procedury stochastický charakter. Na závěr stanovujeme výslednou skladbu portfolia podílových fondů Investiční společnosti České spořitelny v podobě: 14 % Akciový MIX, 30 % Sporoinvest, 30 % High Yield dluhopisový a 26 % Sporotrend.

Literatura

DLOUHÝ, M., J. FÁBRY, M. KUNCOVÁ a T. HLADÍK: *Simulace podnikových procesů*. Brno: Computer Press, 2007, ISBN 978-80-251-1649-4.

FIALA, P.: *Modely a metody rozhodování*. Praha: Oeconomica, 2008, ISBN 978-80-245-1345-4.

GORDON, J. A. a W. F. SHARPE: *Fundamentals of Investments*. New Jersey: Prentice Hall, Inc., 1989, ISBN 0-13-340167-7.

HADLEY, G.: *Nonlinear and Dynamic programming*. USA: Addison-Wesley Publishing Company Inc., 1964, ISBN 0-201-02664-3.

HINDLS, R., S. HRONOVÁ, J. SEGER a J. FISCHER: *Statistika pro ekonomy*. 7. vydání. Praha: Professional Publishing, 2006, ISBN 80-86946-16-9.

KOSMIDOU, K. a C. ZAPOUNDIS: *Goal Programming Techniques for Bank Asset Liability* Management. Norwell: Luwer Academic Publishers, 2004, ISBN 1-4020-8104-9.

MÁLEK, J.: *Risk Management (Vybrané kapitoly)*. Praha: VŠE, 2003, ISBN 80-245-0633-5.

SAATY, T. L.: *The Analytic Hierarchy Process*. USA: McGraw-Hill, Inc., 1980, ISBN 0-07-054371-2.

ČESKO. Sbírka zákonů ÚZ č. 851 *Cenné papíry, Kapitálový trh*. Ostrava: Sagit, 2011, ISBN 978-80-7208-871-3.

SHARPE, W. F.: *Portfolio Theory and Capital Markets*. USA: McGraw-Hill, Inc., 1970, ISBN 07-056487-6.

VALACH, J.: *Investiční rozhodování a dlouhodobé financování*. 2. přepracované vydání. Praha: Ekopress, 2006, ISBN 80-86929-01-9.

VESELÁ, J.: *Investování na kapitálových trzích*. 2. rozšířené a aktualizované vydání. Praha: Wolters Kluwer ČR, 2011, ISBN 978-80-7357-647-9.

ZELENÝ, M.: *Human Systems Management - Integrating Knowledge, Management and Systems*. New Jersey: World Scientific Publishing, 2005, ISBN 978-9810249137.

JEL Classification: C44, C61, G11

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu IGA F4/16/2011 „Modely operačního a finančního managementu".

Summary

The Stochastic Dominance Approach and Interactive Multi-objective Programming Method as Analytic Tools of Investment Decision Making

In the article we can identify two levels. The first one is related to theoretical introduction of the stochastic dominance approach and interactive multi-objective programming method, in the second we apply mentioned quantitative approaches for making „optimal" portfolio structure of investment instruments.

We use the draft of stochastic dominance for reduction of relatively huge set of investment opportunities. The selected alternatives will share in final capital portfolio shape. The application of the stochastic dominance principle is determined by stochastic character of studied problem. The yield rate of shares funds is stated as random

variable. We choose the shares funds of Investment company Česká spořitelna for practical application. The integral part of process is research of probability distribution of yield rate of particular shares fund via appropriate statistic tests. By means of the random number generator we create several possible scenarios which will be available for making final portfolio structure. For finding of „optimal“ portfolio form we use interactive multicriteria programming method whose computational algorithm is based on maximization of positive deviation from aspiration levels chosen by investor (decision maker) for separate objective functions (criteria) as are yield rate, riskiness and costs connecting with investment. The investor further takes into account some financial indicators as Sharpe index or modified duration, the portfolio diversification as well.

After all procedures including successive revision of offered solution by analyst according to decision maker preferences we obtain final portfolio form of shares funds.

Keywords: portfolio, decision making, stochastic dominance.

Hallův model permanentního příjmu

Milan Bouda
xboum00@vse.cz
Doktorand oboru ekonometrie a operační výzkum
Školitel: Prof. RNDr. Václava Pánková CSc., (pankova@vse.cz)

Abstrakt: Článek se zabývá odhadem parametru sigma, který lze v Hallově modelu permanentního příjmu interpretovat jako elasticitu dlouhodobé substituce. Článek je rozdělen na dvě části. První je teoretická část a ta se zabývá nejprve teorii spotřebitele a úrokovou mírou. Dále je pak rozebrána ekonomická interpretace parametru „sigma“. Dále je zde teoreticky popsáno očekávání úrokové míry a hypotéza racionálních očekávání, na které je založen moderní makroekonomický přístup. Na počátku praktické části jsou popsány datové zdroje, které posloužily k odhadu modelu. Jedná se o kombinaci dvou datových zdrojů a to dat z ČSÚ, který má údaje o spotřebě domácností a dat z ČNB, která disponuje údaji o úrokových měrách. V další kapitole je formulován matematický model Hallova modelu permanentního příjmu tak, aby vycházel z teoretických předpokladů, které uvádí Hall ve svých dvou článcích věnujících se této problematice. Model je následně simulován v Dynare. Jedná se o volně šiřitelný toolbox, který lze spustit v komerčním Matlabu či ve volně dostupném softwaru Octave. Tato simulace má za úkol ověřit, zda mnou vytvořený matematický model může být v dlouhodobé rovnováze. Po tomto ověření lze přistoupit k odhadu parametrů modelu pomocí Bayesovské metody, která nám umožňuje zvolení apriorních pravděpodobností, které fungují jako váhy při odhadování aposteriorních pravděpodobností, které se tak vyhnou extrémním a odlehlým hodnotám. V této kapitole je tedy proveden odhad parametru, který lze interpretovat jako elasticitu dlouhodobé substituce. Vedlejším produktem modelu jsou také předpovědi. Nejprve bodová předpověď spotřeby a úrokové míry a dále průměrná předpověď těchto proměnných na 4 období. Pomocí elasticity dlouhodobé substituce lze určit, jak moc je česká společnost konzumně orientovaná. Tomuto problému se věnuje poslední část této práce.

Klíčová slova: Hall, dynare, elasticita dlouhodobé substituce, racionální očekávání.

Úvod

Jednou z nejvýznamnějších determinant, které ovlivňují citlivost úspor a spotřeby na úrokovou míru, je elasticita dlouhodobé substituce, to je míra citlivosti růstu spotřeby k reálné úrokové míře. Ta určuje, jak se jednotlivci rozhodují mezi současnou a budoucí spotřebou. Tedy mezi spotřebou a tvorbou úspor. Detailní studie z 20. století provedené v USA nedávají žádné přesvědčivé důkazy o tom, že elasticita dlouhodobé substituce je pozitivní. Následné objevy týkající se této problematiky jsou zcela opačné. Tuto náhlou změnu lze vysvětlit tím, že dnes využíváme vhodnější odhadové metody, než tomu bylo dříve.

Vysoká reálná úroková míra vyvolá u lidí odklad spotřeby, všechno ostatní zůstane na stejné úrovni. Závažnost dlouhodobé elasticity substituce je jedna z hlavních otázek makroekonomie. Jestliže jsou spotřebitelé donuceni odložit svou spotřebu mírným nárůstem úrokové míry, tak (1) pohyby úrokové míry mají za následek pokles spotřeby kdykoliv ostatní komponenty agregátní poptávky rostou. Celková produkce není příliš ovlivněna změnami těchto komponent; (2) ztráta mrtvé váhy ze zdanění úroků je velice významná.

V této práci odhadneme parametry reprezentativní osobní užitkové funkce raději než parametry funkcí spotřeby a úspor. Jak již poukázal Lucas (1976), není zde nic, co bychom mohli správně nazvat funkcí spotřeby či úspor: vztah mezi spotřebou, příjmem a reálnou úrokovou mírou závisí na širším makroekonomickém kontextu a nemůže být nezávislý na čase, třebaže se spotřebitelé vždy snaží maximalizovat stejnou užitkovou funkci. Techniky, které zde předvedu, jsou více robustní a respektují tento druh nestability více než standardní ekonometrické modely spotřeby nebo úspor.

Základní myšlenkou tohoto článku je, že spotřebitelé plánují změny v jejich spotřebě z roku na rok a to v množství, které závisí na jejich očekávání reálné úrokové míry. Současná úroveň spotřeby se liší od plánované a to tak, že je to naprosto nepředvídatelná náhodná proměnná, která obsahuje všechny dostupné informace na další rok, které samozřejmě ještě nebyly zahrnuté do plánování v předchozím roce. Pokud dojde ke změně očekávání reálné úrokové míry, tak by zde měla být také odpovídající

procentní změna spotřeby. Citlivost spotřeby na změny očekáváné reálné úrokové míry měří právě dlouhodobé elasticita substituce.

Teorie spotřebitele a úroková míra

Základní model s rozdělenou spotřebou a výnosem z jednoho aktiva se objevil již ve výzkumech Breedena (1977, 1979).

Rozložená spotřeba c_t je zde zlogaritmována. Výnos, který nám přinese aktivum r_{t-1} v časovém intervalu $t-1$ až t má normální rozdělení s kovarianční maticí, která se v čase nemění. To můžeme znázornit následujícím lineárním vztahem:

$$\overline{c}_t = \sigma \overline{r}_{t-1} + c_{t-1} + k. \tag{1}$$

Očekávanou změnu zlogaritmované spotřeby zde dostaneme jako, parametr σ krát očekávaný výnos investice plus konstanta. Nejjednodušší racionalizace tohoto modelu je založena na hypotéze, že spotřebitel maximalizuje očekávanou hodnotu dlouhodobě oddělitelnou užitkovou funkcí:

$$\sum e^{-\delta t + \left[1-(1/\sigma)\right]c_t}. \tag{2}$$

Jedna z mnoha věcí, které musí dnešní spotřebitel čelit: utratit méně za spotřebu v roce $t-1$ a investovat tuto úsporu do jednoho aktiva a spotřebovat to v období t. Učiňme zde předpoklad, že jedna investovaná jednotka v období až má stochastický výnos $e^{r_{t-1}}$ v roce t. Podmínka prvního řádu má pro velmi malý odklad spotřeby v období $t-1$ až t následující tvar (za předpokladu že se rozhodujeme v období):

$$E_{t-1}\left[e^{r_{t-1}-(1/\sigma)c_t} - e^{\delta-(1/\sigma)c_{t-1}} \right] = 0. \tag{3}$$

Rovnice 3 je přesnou matematickou formulací tohoto principu: mezní míra substituce by se měla rovnat poměru cen současné a budoucí spotřeby. Podrobnosti viz. Hall (1988 či 1978).

Nejprve tedy předpokládejme, že reálná úroková míra r_{t-1} je podmíněna dostupností této informace v období $t-1$. Řídí se normálním rozdělením se střední hodnotou $\overline{r}_{t-1}$. Druhý předpoklad je takový, že spotřebitelovo pravidlo pro zpracování nové informace ohledně příjmu a úrokových měr činí rozdělení spotřeby lognormálním, podmíněné informací dostupnou v minulém roce. Tedy $\log c_t$ je normální se střední hodnotou $\overline{c}_t$. Vzhledem k tomu, že nová příchozí informace v období t má vztah jak k aktuálnímu výnosu z investic splatných v období t, tak na spotřebitelovu spotřebu v tom roce. Proto jsou dvě náhodné proměnné r_{t-1} a c_t korelovány.

Aplikací dlouhodobé alokační podmínky za předpokladu lognormality dostáváme vztah mezi očekávanou hodnotou zlogaritmované spotřeby v čase t a spotřebou v čase $t-1$ a střední hodnotou reálné úrokové míry. Přesně tak je to uvedeno ve vzorci 1. Vysoká hodnota parametru σ znamená, že když je očekávána vysoká úroková míra, tak spotřebitel dobrovolně odloží svou současnou spotřebu ve prospěch budoucí spotřeby.

Interpretace parametru σ

V modelu z finanční teorie, který je založený na maximalizaci očekávané hodnoty dlouhodobé aditivní užitkové funkce z rovnice 2 je parametr σ interpretován jako reciproká hodnota koeficientu relativní averze k riziku. Ačkoliv je parametr σ také označován jako dlouhodobé elasticita substituce.

Pokud má spotřebitel vysokou averzi k riziku, tak musí mít také nízkou elasticitu dlouhodobé substituce. Abychom dostali jednoznačné výsledky z této práce je potřeba eliminovat toto automatické propojení. Empirické poznatky dokládají, že dlouhodobé substituce je velice nízká nebo dokonce nula a to

neodporuje žádnému širokému přesvědčení akademické obce o spotřebitelově chování. Na druhou stranu je zde odpovídající závěr, že koeficient relativní averze k riziku je blízký nekonečnu a to není kompatibilní výrok s tím, co pozorujeme na spotřebiteli. Ten není ochoten přistoupit na příliš vysoké riziko. Spíše naopak, funkce vyjadřující vztah spotřebitele k riziku je pro většinu racionálních spotřebitelů konkávní.

Hall ve svém článku dále probírá nejrůznější důvody proč je parametr σ právě elasticitou substituce a ne reciprokou hodnotou koeficientu relativní averze k riziku. Cituje také řadu autorů, kteří ve svých statích dokázali jeho domněnku. Například Kreps a Porteus (1978), Attanasio (1987), Epstein a Zin (1987a, 1987b) a Weil (1987).

V dalších setkáních proto budu parametr σ interpretovat jako dlouhodobé elasticitu substituce a vyhnu se jakýmkoli závěrům ohledně averze k riziku.

Očekávání reálné úrokové míry

Ukážeme si dva způsoby jak měřit očekávanou reálnou úrokovou míru. Budeme sledovat změny spotřeby po dobu, pro kterou jsou dostupná data z průzkumů proti očekávaným změnám cen. Očekávanou reálná úrokovou míru získáme tak, že vezmeme nominální úrokovou míru a očistíme ji o daně a očekávanou změnu cen. Můžeme zde také použít reálné výnosy z akcií, protože tyto data jsou také k dispozici.

Druhý přístup se vztahuje k podmíněné střední hodnotě reálné úrokové míry, $\overline{r}_{t-1}$, a pozorovaným hodnotám proměnných známým spotřebiteli v čase, kdy volí svou úroveň spotřeby c_{t-1}. Připomeňme, že $\overline{r}_{t-1}$ je střední hodnota subjektivního rozdělení reálné úrokové míry typického spotřebitele, který se rozhoduje na následující období. Očekávání specifikujeme pomocí hypotézy racionálních očekávání. Subjektivní rozdělení je potom jakousi distribuční funkcí lineárních kombinací pozorovaných proměnných.

Hypotéza racionálních očekávání

Je založena na předpokladu, že ekonomické subjekty formují svá očekávání tak, že přitom využívají efektivně veškeré podstatné informace a to jak minulé, tak současné a jejich prognózy. Technicky získáme očekávanou proměnnou, tak že použijeme metodu instrumentálních proměnných a pomocí všech dostupných vysvětlujících proměnných se pokusím vymodelovat očekávání. Podrobně o tomto odhadu pojednává Hušek (2007). Odhad nelineárního modelu racionálních očekávání pomocí zobecněné momentové metody uvádějí Hansen a Singleton (1982).

Škola racionálního očekávání je ekonomická teorie, vznikla v USA na počátcích 70. let 20. století, základní hypotézu racionálních očekávání sepsal již v roce 1961 americký ekonom John Muth. Rozbory jeho prací provedli a na jeho práci se o deset let později, v roce 1971, odkázali američtí ekonomové Robert Lucas a Thomas Sargent. Tyto práce se dají pokládat za počátek školy racionálního očekávání. Rozdílně od monetarismu se tato škola profilovala více makroekonomicky, z tohoto důvodu je označována za „novou klasickou makroekonomii“. Na rozdíl od koncepce adaptivních očekávání vychází z teorie celkové ekonomické rovnováhy, také sází na pružnost mezd, z toho vyplývá i chápání tržní ekonomiky jako naprosto stabilní systém, který se může zhroutit pouze vnějším zásahem. Zásahy státu jsou tak, naopak od Friedmanova monetarismu, neregulérní. Závěrem je to, že stát, který svojí hospodářskou politikou provádí pouze očekávané a ohlášené úkony, hospodářství nepoloží, pokud ovšem provede neočekávaný krok, tak může svojí ekonomiku položit „na lopatky“. Škola racionálního očekávání patří k největším kritikám keynesiánského systému, monetaristé se snažili zase diskutovat, o tom jestli a proč je vše keynesiánské zlé. Kritiky školy racionálního očekávání byli například Sanford Grossman a Joseph Stiglitz.

Data

V této kapitole jsou popsány časové řady, které jsem použil v této práci. Dále jsou zde odkazy na jejich zdroj.

$c_t =$ logaritmy reálné (v cenách roku 2000) spotřeby domácností předmětů krátkodobé spotřeby. Jedná se o čtvrtletní data, která jsem získal ze stránek ČSÚ. Data spotřeby jsou dostupná od Q1 2006 až po Q2 2011.

$i_t =$ jedná se o průměrnou úrokovou míru, kterou vypláceli banky rezidentům ČR za vklad se splatností od jednoho roku do dvou let včetně. Data jsou získána ze stránek ČNB a jejího systému časových řad (ARAD). Časová řada je napozorovaná od Q1 2004 až po Q3 2011.

V modelu jsou dále použity další proměnné, ty už však neobsahují reálně napozorované hodnoty. Jedná se o kapitál k_t, R je konstanta, která nám udává míru opotřebení kapitálu (roční odpis), $d_t =$ roční dividenda, $b_t =$ investice.

Model

Při formulaci matematického modelu Hallova modelu permanentního příjmu kladu důraz především na to, že chci najít rovnovážný stav modelu a to v dlouhodobém pohledu. Při sestavování jednotlivých rovnic jsou použity poznatky zmíněné v předchozích kapitolách tohoto článku. Spotřebitelé řeší následující rozhodovací problém:

$$\max_{\{c_t,k_t\}_{t=0}^{\infty}} -\frac{1}{2}\sum_{t=0}^{\infty}\beta^t\left(c_t-b_t\right)^2,$$

s ohledem na níže uvedené omezující podmínky:

$$\begin{aligned}
&c_t + k_t = Rk_{t-1} + d_t,\\
&d_t = \rho d_{t-1} + (1-\rho)\mu_d + \varepsilon_d,\\
&b_t = \rho_b b_{t-1} + (1-\rho_b)\mu_b + \varepsilon_b,\\
&i_t = k_t - k_{t-1},\\
&\mu_{ct} = \mu_{ct+1,}\\
&\mu_{ct} = b_t - c_t.
\end{aligned}$$

V účelové funkci domácnosti řeší problém jak co nejefektivněji rozdělit svůj současný příjem mezi spotřebu a investice. Přičemž investice může mít za následek zvýšení spotřeby v příštím roce. Omezující podmínky jsou konstruovány tak, aby systém mohl dosáhnout rovnováhy. První omezující podmínka nám říká, že současná úroveň spotřeby a kapitálu se rovná kapitálu v loňském období a dividendě, kterou získáme za investice do akcií a cenných papírů. Druhá a třetí rovnice nám vyjadřují závislost výše dividend a investic na jejich o rok zpožděných hodnotách. Dále je v každé z těchto rovnic zahrnuta jakási složka racionálního očekávání. Tyto složky můžeme subjektivně ovlivňovat podle toho, jaká panuje ve společnosti nálada. Čtvrtá rovnice je identita a znázorňuje následující fakt a to že reálná úroková míra je vysoká právě tak, aby její výše pokryla znehodnocení kapitálu. Pátá rovnice je rovněž identita a říká nám, že dnešní mezní užitek ze spotřeby se musí rovnat našemu očekávanému užitku v příštím roce. To je podmínka, která musí platit a spotřebitel se podle této podmínky rozhoduje, zda bude více spotřebovávat dnes či zda zainvestuje a bude se v budoucnu těšit z větší spotřeby. Poslední šestá rovnice je také identita a znázorňuje rovnost mezního užitku ze spotřeby a rozdílu investic (úspor) a spotřeby.

Simulace a odhad parametrů modelu

V první části této kapitoly provedeme simulace modelu, který je uveden v předchozí kapitole. Simulaci provedu v Dynare. Jedná se o doplněk, který lze spustit v Matlabu či v softwaru Octave. Dynare je speciální SW na řešení nejrůznější ekonomických problémů jako jsou DSGE (dynamic stochastic general equilibrium) a OLG (overlapping generations) modely. Tyto modely v sobě zahrnují racionální očekávání subjektů. Dynare má svou vlastní syntaxi a Matlab využívá pouze k výpočtu. Syntaxe je velice podobná běžnému zápisu ekonometrického modelu, tak jak jí známe z literatury. Tento doplněk do Matlabu je všem lidem volně dostupný a je používán národními bankami po celém světě, které si tak mohou simulovat dopady monetární a fiskální politiky na jejich ekonomiky.

Simulace

Provedená simulace nám pomůže ověřit fakt, zda mnou namodelovaný systém je v dlouhodobé rovnováze a existuje zde tzv. rovnovážný stav. Dynare nabízí mnoho nejrůznějších výstupů a charakteristik jakými jsou například: hodnoty jednotlivých proměnných v rovnovážném stavu, souhrn druhů proměnných a jejich počty, matice kovariancí exogenních šoků, obdobu redukovaného tvaru u modelů simultánně závislých rovnic (policy and transition functions), momenty simulovaných proměnných, korelaci simulovaných proměnných, autokorelaci simulovaných proměnných (zpožděných až o pět období) a nakonec jediným grafickým výstupem této simulace jsou funkce odezvy na jednotkové exogenní šoky. Při praktickém provedení této simulace jsem dospěl k závěru, že model formulovaný v kapitole 6 má dlouhodobý rovnovážný stav. Z výsledků je patrná velmi vysoká korelace mezi spotřebou a objemem kapitálu v ekonomice a také mezi investicemi a úrokovou mírou. Funkce odezvy nám poskytly důležitou informaci, že jednotkový exogenní šok je hned v dalším období vstřebán a po další období nijak neovlivňuje výši spotřeby, investic a úrokové míry. Vyčtené výstupy jsou uvedeny níže. Jedná se o výstup, tak jak ho poskytne Dynare.

STEADY-STATE RESULTS:

c	10
k	100
mu_c	20
b	30
d	5
in	0

EIGENVALUES:

Modulus	Real	Imaginary
0	0	0
0	0	0
1	1	0
1.5	1.5	0

MODEL SUMMARY:

Number of variables:	6
Number of stochastic shocks:	2
Number of state variables:	3
Number of jumpers:	1
Number of static variables:	2

MATRIX OF COVARIANCE OF EXXOGENOUS SHOCKS:

Variables	e_d	e_b
e_d	1.000000	0.000000
e_b	0.000000	1.000000

POLICY AND TRANSITION FUNCTIONS:

	c	k	mu_c	b	d	in
Constant	10.000000	100.000000	20.000000	30.000000	5.000000	0
k(-1)	0.050000	1.000000	-0.050000	0	0	0
e_d	0.047619	0.952381	-0.047619	0	1.000000	0.952381
e_b	0.952381	-0.952381	0.047619	1.000000	0	-0.952381

MOMENTS OF SIMULATED VARIABLES:

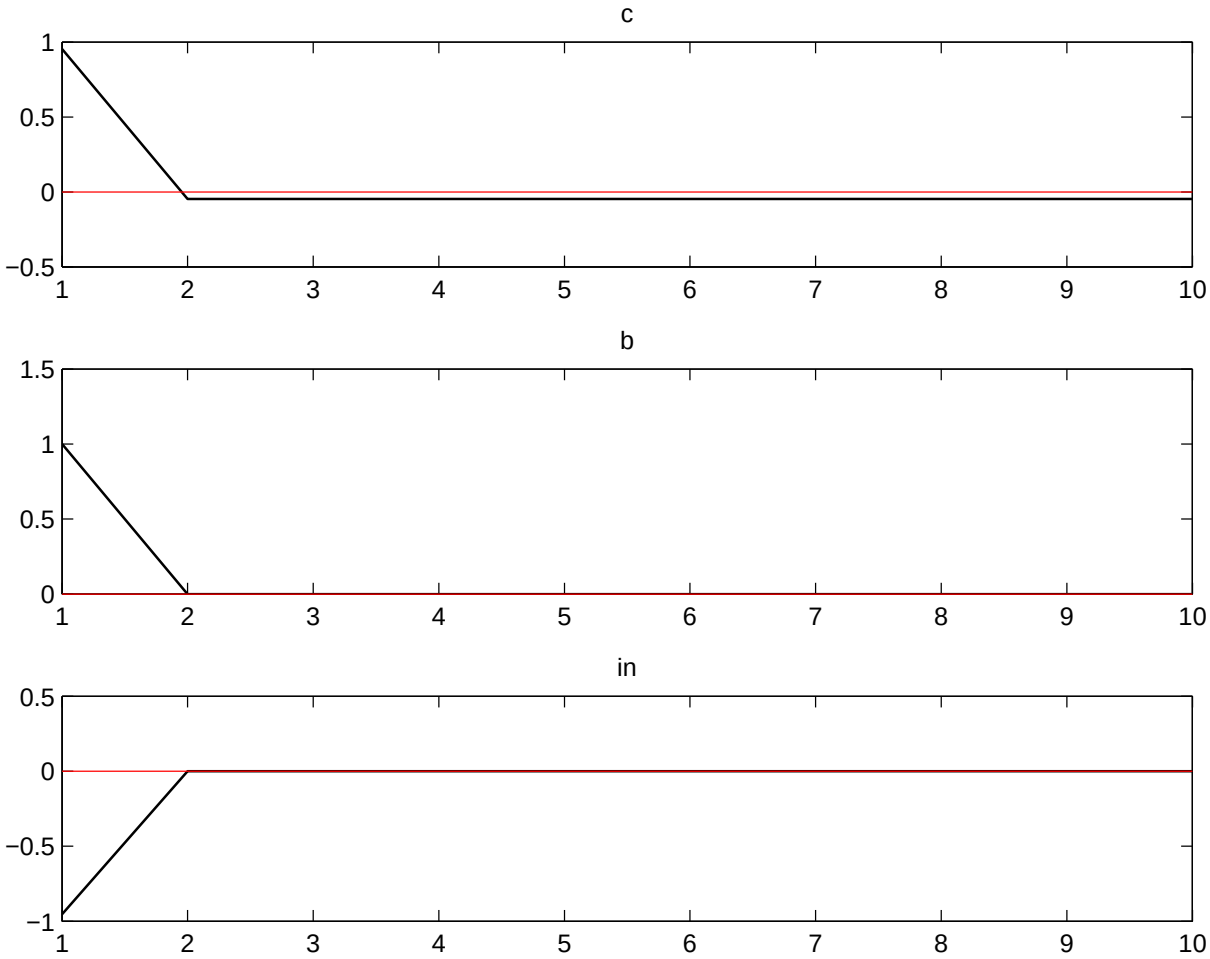
VARIABLE	MEAN	STD. DEV.	VARIANCE	SKEWNESS	KURTOSIS
c	14.579215	3.341132	11.163163	0.074314	-1.175023
k	191.963586	63.624032	4048.017393	0.094085	-1.412176
mu_c	15.401821	3.181202	10.120043	-0.094085	-1.412176
b	29.981036	0.988255	0.976647	0.020340	-0.025504
d	5.017984	0.991149	0.982377	0.040744	0.058554
in	0.035189	1.334007	1.779575	0.021528	0.149622

CORRELATION OF SIMULATED VARIABLES:

VARIABLE	c	k	mu_c	b	d	in
c	1.0000	0.9553	-0.9553	0.3059	0.0222	-0.2001
k	0.9553	1.0000	-1.0000	0.0106	0.0238	0.0094
mu_c	-0.9553	-1.0000	1.0000	-0.0106	-0.0238	-0.0094
b	0.3059	0.0106	-0.0106	1.0000	-0.0015	-0.7066
d	0.0222	0.0238	-0.0238	-0.0015	1.0000	0.7087
in	-0.2001	0.0094	-0.0094	-0.7066	0.7087	1.0000

AUTOCORRELATION OF SIMULATED VARIABLES:

VARIABLE	1	2	3	4	5
c	0.9191	0.9178	0.9165	0.9126	0.9142
k	0.9996	0.9993	0.9991	0.9989	0.9987
mu_c	0.9996	0.9993	0.9991	0.9989	0.9987
b	0.0358	0.0241	0.0082	-0.0355	-0.0153
d	0.0141	-0.0078	0.0048	-0.0026	-0.0072
in	0.0186	0.0179	-0.0114	-0.0356	-0.0278



Obrázek 1: Funkce odezvy

Odhad parametrů modelu

K odhadu parametrů je použito Bayesovské metody. Která spočívá v nalezení největší posteriorní hustoty pravděpodobnosti parametrů. Bayesovskou metodou odhadujeme kompletní DSGE model na rozdíl od metody GMM, která je založena na konkrétním rovnovážném stavu jako je např. Eulerova rovnice spotřeby. Bayesovská technika dále umožňuje zvolit, v praxi je to i nutné, apriorní pravděpodobnosti, které fungují jako váhy při odhadování rozdělení aposteriorních pravděpodobností, které se tak vyhnou extrémním a odlehlým hodnotám. Dále umožňuje zahrnout šoky, které jsou interpretovány jako chyby pozorování ve strukturálních rovnicích. Podrobněji o této metodě pojednávají An a Schorfheide (2006).

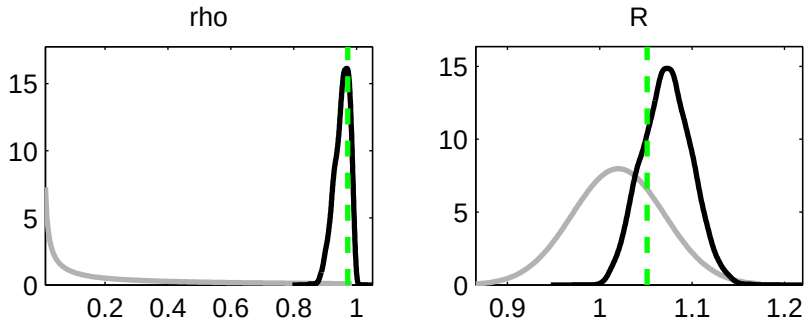
Nejzajímavějším výsledkem je odhad parametru, který můžeme interpretovat jako dlouhodobé elasticitu substituce. Hall při svých výzkumech dokázal, že tato elasticita je velice blízká nule a tedy lidé nereagují nijak elasticky na změny úrokové míry. Otázkou je samozřejmě volba úrokové míry. Pro někoho to může být roční výnos z akcií ČEZu a pro jiného zase třeba spořící konto. V této práci jsem zvolil průměrnou úrokovou míru, kterou vypláceli banky rezidentům ČR za vklad se splatností od jednoho roku do dvou let včetně. Nejedná se sice o příliš vysoké výnosy, ale je to dostupné všem a není k tomu potřeba žádné speciální vzdělání jako při obchodech na burzách atd. Při spuštění odhadu v Dynare získáme následující výstup.

Tabulka 1: Odhad parametrů modelu

	prior mean	post. mean	conf. interval		prior	pstdev
rho	0,1	0,9534	0,9148	0,9912	beta	0,2
R	1,02	1,0722	1,0284	1,1144	norm	0,05

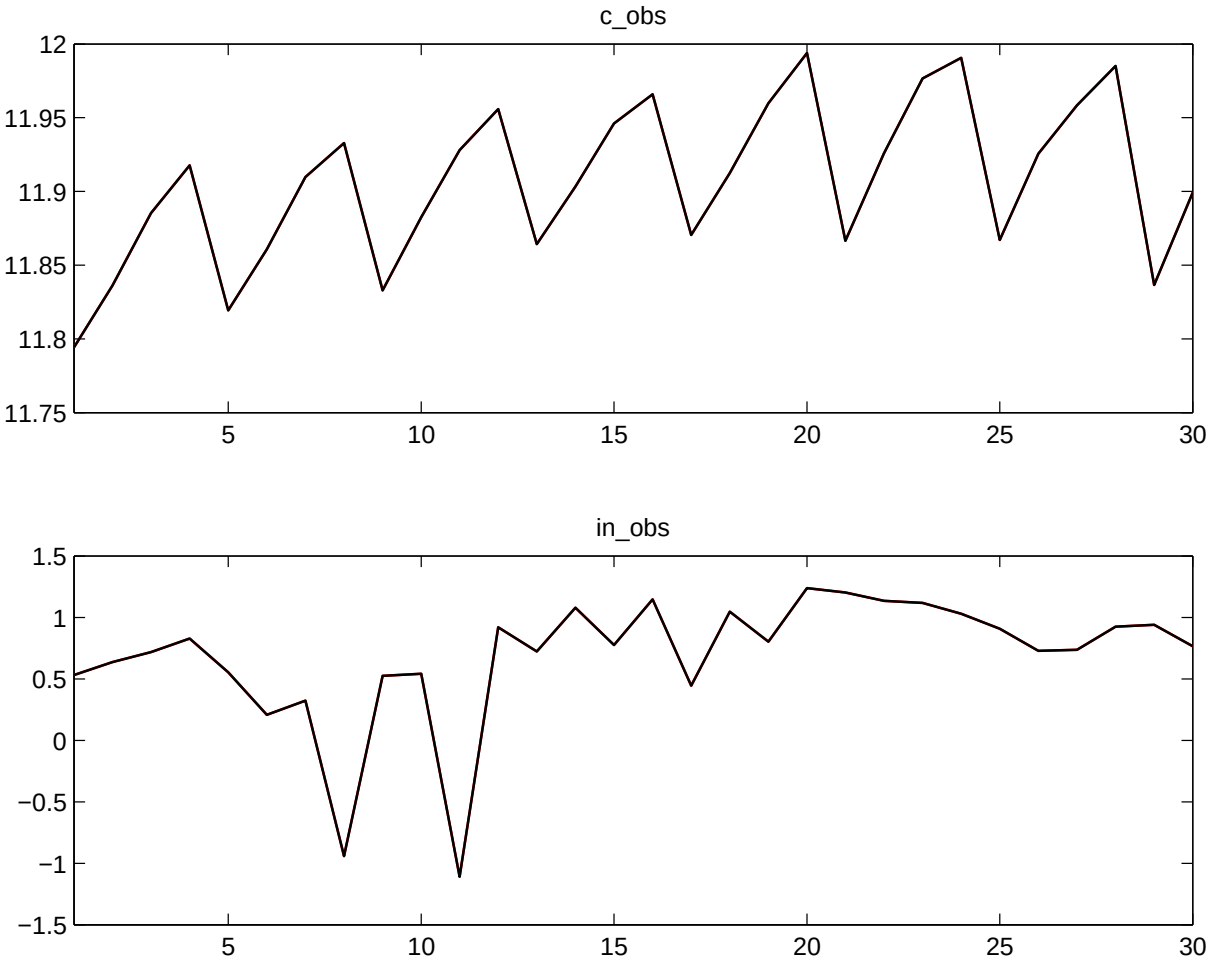
Naším cílem je stanovit hodnotu elasticity dlouhodobé substituce. Tu vypočteme podle tohoto předpisu: $\sigma = (1-\rho)$, hodnota tohoto výrazu je velice blízká nule stejně jako v USA, když tento výzkum prováděl Hall. Nutno poznamenat, že Hallovi vycházel parametr od 0,066 do 0,346. V této práci je hodnota parametru 0,0466. Nutno poznamenat, že Hall kombinoval nejrůznější úrokové míry a odhadové techniky. V jeho výzkumu jsem však nenarazil na techniku použitou v této práci.

Dynare nabízí při odhadu celou řadu grafů. Následující obrázek rozložení apriorních a posteriorních pravděpodobností, které jsou špičatější.



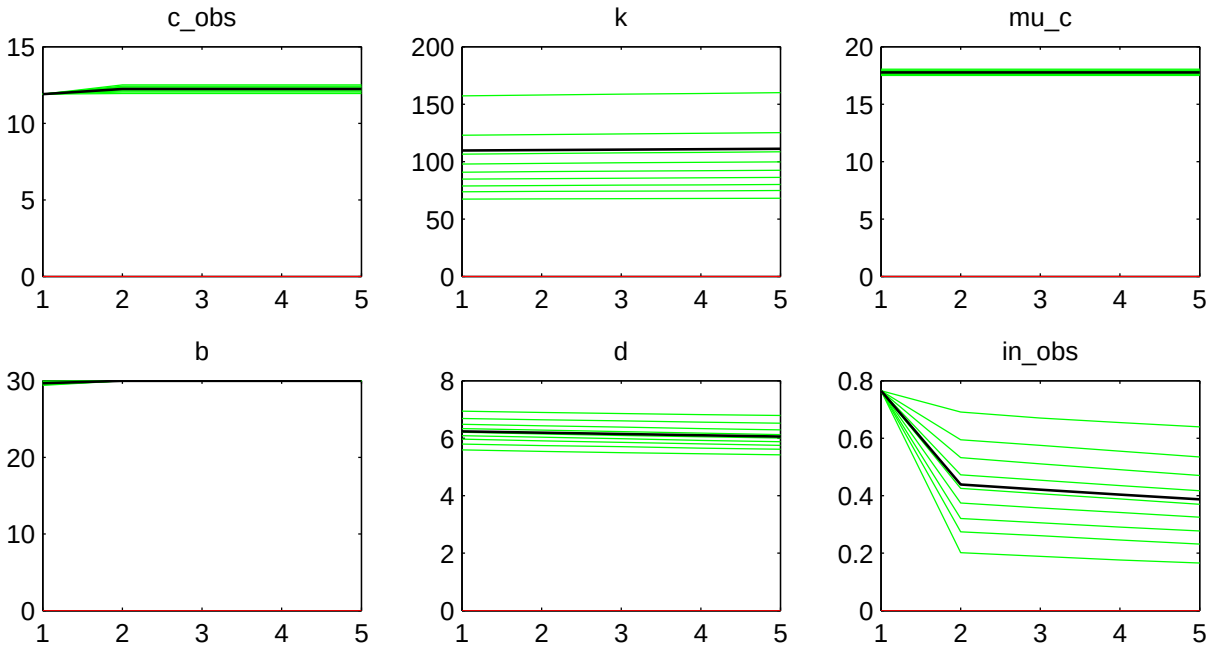
Obrázek 2: Apriorní a posteriorní rozložení pravděpodobností

Na dalším obrázku je možno zhlédnout logaritmy pozorovaných proměnných (spotřeby a úrokové míry).



Obrázek 3: Logaritmy pozorovaných proměnných

Dalším zajímavým výstupem jsou předpovědi. Protože v našem modelu je jedna ku-předu hledící proměnná, tak můžeme provést předpověď proměnných. Zajímat nás však budou pouze grafy c_obs a i_obs, u kterých máme napozorované hodnoty a jsou u nich hodnoty odpovídající realitě.



Obrázek 4: Bodová předpověď na následující 4 čtvrtletí

Závěr

V této práci je formulován Hallův model permanentního příjmu jako DSGE model a jeho následná simulace, která má za úkol ověřit, zda takto formulovaný model má dlouhodobý rovnovážný stav. Model byl samozřejmě kalibrován a to způsobem, že parametry modelu jsem přebral z podobných prací. Rovnovážného stavu bylo tedy dosaženo a poté jsme aplikací dvou exogenních šoků získali funkce odezvy. S jejich pomocí můžeme zkoumat, jak se model chová při jednorázovém, nečekaném šoku. Model se vrátil do rovnováhy hned v následujícím období a nedochází zde vůbec k žádnému rozkolísání.

Dále je zde odhadnutý parametr, který je interpretován jako dlouhodobé elasticita substituce. Jinak řečeno, jde o to, jak moc jsou čeští spotřebitelé ochotni vzdát se současné spotřeby statků krátkodobé spotřeby v jejich budoucí prospěch. Hodnota parametru, tak jak jsem ho odhadl v této práci je velice blízká nule a to lze interpretovat následujícím způsobem. Český spotřebitel tedy nemění svou výši budoucí (plánované) spotřeby podle toho jaká je současná tržní úroková míra. Tento jev může mít několik interpretací. Pro opravdu robustní odhad je potřeba více pozorování. V porovnání s USA, kde mají k dispozici dokonce i předválečná data, máme velmi málo pozorování. Máme k dispozici kvartální data o spotřebě a to pouze v době po revoluční. V USA mají téměř celé století. Dále jsem jako tržní úrokovou míru zvolil, průměrnou úrokovou míru, kterou vypláceli banky rezidentům ČR za vklad se splatností od jednoho roku do dvou let včetně. Výše a variabilita této úrokové míry je velice malá oproti akciovým výnosům. Dalším faktorem proč je dlouhodobé elasticita substituce tak nízká, může být fakt, že české domácnosti mají tak omezený rozpočet ohledně své spotřeby, že v něm prostě není žádný prostor pro budoucí plánování. Jejich omezení současné spotřeby by musel vykompenzovat vysoký očekávaný výnos. Toho nám v naší historii a na těchto datech nebylo dopřáno. Pokud vás ani jeden z těchto argumentů nepřesvědčil proč zpochybnit tento odhad, tak je zde ještě poslední možnost a to že české domácnosti jsou více konzumně orientované než ty americké v letech do roku 1988. V tomto článku však chybí porovnání s ostatními státy světa. To se však v dohledné době změní a dojde k rozšíření zde uvedené problematiky o aplikovanou část, kde použiji několik úrokových měr a provedu srovnání několika států. V navazujícím článku se dozvíte odpovědi na následující otázky.

Které národy jsou nejvíce spořivé a myslí na svou budoucnost? Které z nich si umí užívat nejvíce života v současnosti a budoucnost je naopak vůbec netíží?

Literatura

ADJEMIAN Stéphane a kol., 2011. *Dynare: Reference Manual, Version 4*. Dynare Working Papers, 1, CEPREMAP.

AN a SCHORFHEIDE Frank, 2006. *Bayesian analysis of DSGE models.* Working papers 06-5, Federal Reserve Bank of Philadelphia.

BREEDEN, Douglas T. *Changes in Consumption and Investment Opportunities and the Valuation of Securities.* Stanford Univ 1977. Ph.D. dissertation.

BREEDEN, Douglas T., 1979. *An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities.* Journal of Financial Economics 7, s. 265-296.

HALL, R. E., 1988. *Intertemporal substitution in consumption.* Journal of Political Economy 96(2), 339-57.

HALL, R. E., 1978. *Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and evidence.* Journal of Political Economy 86(6).

HANSEN, Lars a SINGLETON, Kenneth J., 1983. *Stochastic Consumption, Risk Aversion, and the Tempoval Baehavior of Asset Returns.* J.P.E. 91. s 249-265.

HUŠEK, Roman, 2009. *Aplikovaná ekonometrie. Teorie a praxe.* Vysoká škola ekonomická v Praze. Nakladatelství Oeconomica. Praha. ISBN 978-80-245-1623-3.

LUCAS, Robert E., Jr. 1976. *Econometric Policy Evaluation: A Critique.* Carnegie-Rochester Conference Series on Public Policy 1.

WEIL, Philips, 1989. *The Equality Premium Puzzle and the Risk-Free Rate Puzzle.* Journal of Monetary Economics 24. North-Holland, s.401-421.

JEL Classification: P46, E21.

Summary

Hall‘s permanent income model

The article discusses the estimation of sigma parameter, which can be in Hall’s permanent income model interpreted as the intertemporal elasticity of substitution. The article is divided into two parts. The first part is a theoretical part and deals with the theory of consumers and interest rates. Next, there is economic interpretation of parameter sigma. Moreover, I am theoretically describing expectations of interest rates and hypothesis of rational expectations. In these modern methods is based macroeconomic approach. At the beginning of the practical section are described data sources that were used to estimation. There is used a combination of two data sources provided by ČSU, which has data of consumption and by ČNB, which has data of interest rates. In the next chapter is formulated mathematical model of Hall’s permanent income model. It is based on theoretical assumptions that Hall publishes in his two articles dealing with this issue. This model is simulated in Dynare. It is open-source toolbox that can be used in commercial software Matlab or in open-source Octave. This simulation checks that my model can be in long-term equilibrium. After this verification we can proceed to estimation parameters of this model by using the Bayesian methods. This method allows us to select priors that are used as weights in estimating the posteriors. It helps us to avoid extreme values. In this chapter is sigma parameter estimated and next it is interpreted as the intertemporal elasticity of substitution. Byproduct of the model is also predictions. There are forecasts of consumption and interest rates. First their point forecast for four seasons and secondly their mean forecast. Using the intertemporal elasticity of substitution we can determine how much are people in Czech Republic consumerist-oriented. This issue is discussed in the last part of this work.

Keywords: Hall, dynare, elasticity of intertemporal substitution, rational expectation.

Optimization of Harvesting Natural Resources

Viktor Chrobok
viktor.chrobok@vse.cz
Doktorand oboru ekonometrie a operační výzkum
Školitel: doc. Mgr. Ing. Martin Dlouhý, Dr., MSc, (dlouhy@vse.cz)

Abstract: The paper focuses on a modification of the predator-prey model in case of harvesting one of the populations. Firstly there is a description of the basic model then there is introduced a profit maximizing harvester. There are some essential assumptions when dealing with this kind of the optimization such as long term stability and prevention of natural decrease of harvested species. The optimization algorithm is based on the general character of the system and provides universal technique to deal with similar problems. Since the explicit form of the differential equations cannot be calculated an estimate of the predator-prey development was needed. The estimate was done by non-linear least squares and brought robust results applicable to various datasets. The described method was demonstrated on an illustrative example when harvesting prey population. At the end are given most important differences between prey and predator harvesting cases.

Keywords: Predator-Prey Model, Optimization, Mathematical Bioeconomics, non-linear least squares.

Introduction

The Predator-Prey model describes interactions between two species, where one of them is a prey to the other. Some examples of such populations are foxes and rabbits, lynxes and hares or various fish populations. The paper is focused on the case of introducing a third party to the system, which is harvesting one or both populations. The main assumption of this text is that this harvesting is done with a goal of maximizing the profit. There will be given fundamental characteristics of the basic model and description profit maximizing algorithm.

The model is described by differential equations, which were firstly proposed independently by Alfred J. Lotka in 1925 and Vito Volterra in 1926. It is an useful tool for analysing development of biological systems in time, which is mainly the subject of mathematical biology. However when a harvesting modification is added, there arises also a problem from economist's point of view. Description of harvesting in the Predator-Prey model is currently quite rare and incomplete. Some mention about the topic can be found in Clark (1976)0, other resources covering this topic are currently unknown to the author of this paper.

Methodology

The main tool used for modelling Predator-Prey systems is mathematical analysis. The system is described by a following pair of first-order, non-linear, differential equations:

$$\frac{dx}{dt} = (a - by) * x \wedge \frac{dy}{dt} = (-m + nx) * y,$$

$$t \geq 0, x(0) = x_0, y(0) = y_0, \qquad (1)$$

where x is the amount of the prey population, y is the amount of the predator population, t is time, a is a growth rate of the prey population, b shows how much of the prey population is killed by the predator population, m is a death rate of the predator population, n shows how much of the predator population is renewed by consuming the prey population. All parameters a, b, m, n are essentially positive constants and since x and y are representing the amount of prey and predator populations they are non-negative numbers.

It can be easily proved that for a positive number of both populations, there exists just one equilibrium point, $[x_e, y_e] = \left[\frac{m}{n}, \frac{a}{b}\right]$.The system given by (1) can be solved by a separation of variables to get the following formula:

$$\left(\frac{y}{y_0}\right)^a e^{-b(y-y_0)} = \left(\frac{x}{x_0}\right)^{-m} e^{n(x-x_0)}, \qquad (2)$$

which implies that the system is not converging to the equilibrium, neither diverging from it, but oscillating; i.e. if $[x_0, y_0] \neq [x_e, y_e]$ then the system is periodically moving on a closed curve and gets back to $[x_0, y_0]$ after a specific period of time.

Note that (2) is describing the system as an implicit function and that it is impossible to get explicitly xor y as functions of time from (1). A linearization of the system given by (1) is commonly used to get those results because of this shortcoming. The linearization is based on the application of Taylor's theorem, whose description could be found for example in Hájková (2003)0. Linearization of the basic system given by (1) is provided in Doucet (1993).

Figures included in this paper were obtained by using Maple software. Techniques of plotting graphs there are not analytically precise, but based on the methods of numerical mathematics. For more reference see Lee (2004), for a brief description of theoretical background Pánková (2003) could be used.

There are also mentioned some concepts of intermediate microeconomics in the following text, which could be found for example in Varian (2006)0.

Graphical Representation of the Basic Model

Let's choose specific values of parameters described above and show a behaviour of the Predator-Prey system. Since both harvesting modifications provided below are based on the basic model it will be very useful in the further analysis.

For the following figures were chosen these values of the parameters:

$$a = 5; b = 0.02; m = 10; n = 0.01; x_0 = 1000; y_0 = 500. \qquad (3)$$

For plotting the results was used Maple v. 12, the stepsize for all figures in this paper was set to 0.01 time units. Those parameters of the model shall be used till the end of this paper, the differences in the initial values will be always especially stated. The figure showing the development of the system given by (1) with parameters and initial conditions given by (3) are plotted below.

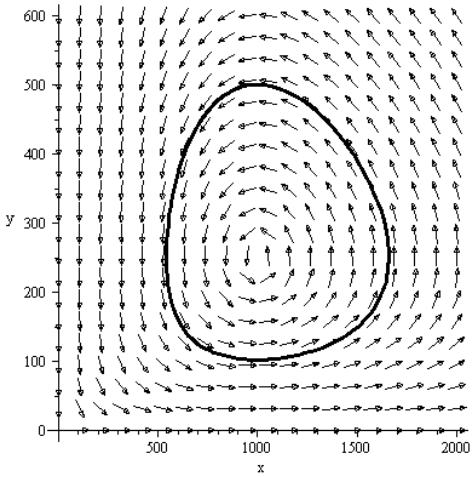


Figure 1. Basic Predator-Prey Model ([*x*, *y*] plot)

The arrows in the Figure 1 show the direction of the derivatives, which means that the model moves anti-clockwise. The figure very clearly shows the oscillating character of the model. Note the different scale on the axes. The curve is generally non-symmetrical.

Introducing Optimization Process

We believe that the model was properly analyzed from a mathematician's point of view so far. At this point, it is necessary to mention that a significant amount of harvesters would just maximize their profit in the given situation. Modifying the basic model in this case is mathematically much more complicated than in all previous cases. So it will be unavoidable to make some simplifications and

assumptions in order to solve any given optimization problem. We will start by summarizing some essential facts about the optimization then try to solve the problem more properly and give some empirical evidence.

Non-standard Issues Concerning Optimization

Firstly a harvester should consider if they are seeking a long run or a short run profit. Short run profit could be easily made by simply harvesting everything available. Optimizing for a long run profit is far more complicated.

It is obvious that it is non-profitable for a harvester when the harvested population is naturally decreasing. So a smart harvester would never let the predator population exceed its equilibrium value and would rather harvest the prey himself than leave it to the predators in case of prey-harvesting. On the other hand they would not let predators die of hunger when they could harvest and sell them instead of letting them starve. We must find an intersection of the two mentioned sets of points in order of harvesting both populations. This is an area where both populations are growing i.e. the prey population is higher than its equilibrium value and there are less-than-equilibrium predators.

The best situation for the harvester is when the harvested population is increasing as much as possible, which is when the amount of the predator population is in its minimum if harvesting the prey and when the amount of the prey is in its maximum for the predator-harvesting. So for prey-harvesting, it would be most profitable to harvest a little amount of the prey around the predators' minimum which could lead to the significant destabilization of the system if it is done without any respect to biological circumstances. Harvesting just predators is much more stable since their number is dependent on the number of the prey.

We should also consider some transaction costs[1] for each harvesting so it would be cheaper to harvest less often in larger amounts as in the real world. This is neglecting the strategy in the previous paragraph so it would be necessary to find some balance.

To make it more realistic, interest rate should be also added and the fact that the consumption today is better than consumption tomorrow. This fact is in conflict with the previous paragraph so it is clearly showing that one should find a balanced solution rather than an extreme one.

According to the previously presented facts, it is obvious that optimal harvesting will be done discretely. Therefore a new problem arises because the development of the system and all interactions are done continuously. Another complication is that these interactions cannot be given explicitly. It means that we cannot avoid simplifying a model's constraints.

One should note that by harvesting the length of the cycle could be changed which could bring more profit in the future. So it could be better to give up some profit to just make an "ideal" system which would bring more profit in the future.

General Assumptions

There are some assumptions valid for the optimization of harvesting for both species:

- **Long-term stability of the system**

  Since we suppose that a harvester is seeking a long-term profit they should prevent an extinction of any species[2].This condition is fulfilled if and only if the minimal amount of any species in the whole system is high enough to keep the system working according to the differential equations by which it is described. As was mentioned above if the system is far from the equilibrium the model could break which is usually followed by the extinction.

[1] We could obviously introduce more sophisticated costs structure, i.e. fixed and variable costs or find some concave relationship between the harvested amount and costs of harvesting. But a variable part would not influence optimization algorithm described below and cost function would bring unnecessary complications.

[2] We do not assume that it is possible to extinguish all predators just because a harvester needs prey. This would lead to a simple one-species-optimization.

- **No natural decrease of harvested species**
  Any harvester would definitely not let their prey[3] naturally die when they could harvest it themselves[4].

These two assumptions suggest that a harvester should let the system move on just one curve given by unique initial conditions. As was mentioned above in case of harvesting the prey there is no rationale behind letting the predator population rise above its equilibrium value. In the case of harvesting the predator population a profit-maximizing harvester would never let the prey population decrease below its equilibrium value.

There are two essential facts needed in the following optimization algorithm:

- **Amount of population which could be harvested is rising in the distance of the cycle from the equilibrium**
  As we know from the previous chapters the tangent of the curve showing both populations is zero when the predator population is in its minimum. The longer is the distance from the equilibrium the flatter is the curve in the neighborhood of the predator's minimum. This intuitively means that a harvester could harvest more prey during the same time period on the further curve without any crucial consequences. A similar relationship exists for the predator-harvesting.
- **Relationship between maximal possible harvesting amount and time is concave**
  This means that a harvester could increase their profit by harvesting more often. The proof of this statement for prey-harvesting is based on the fact that the closer to the equilibrium predator population is the fewer amounts for harvesting are available. The proof for predator-harvesting is based on a very similar fact.

Optimization Algorithm

We already know that there are reasons to harvest as often as possible in small amounts, i.e. a nature of the model, and on the other hand transaction costs were introduced for one harvesting, which is forcing a harvester to harvest less often. We also know that it is most profitable to harvest when the model is far from the equilibrium. We will introduce a condition on the minimal amount of population needed to keep the model working[5]. This is an exogenous variable in the optimization process.

We would not assume any interest rate because then the nominal profits gain in the different periods would not be equal which would hugely complicate the optimization process.

Since we consider theoretically infinite number of periods we can assume that at the beginning a harvester would influence the system to move it to the cycle where a maximal profit could be reached. This curve is defined by the minimum limit described above. The costs of this movement will be offset by higher profit in the future. A limit of an average profit for infinite periods will not be affected.

It is necessary to have the same units in the optimization. We will not introduce any money but just define transaction costs as some fixed amount of the harvested population. The next important relationship we need to know is a function describing how many individuals could be harvested and how much time it will take. This relationship will be described precisely in the following text.

The very first equation will be the profit equation, which tells that the profit is calculated as a difference between revenues and costs. Both are measured in the numbers of prey or predator individuals. We will calculate just a profit in one period since we have previously explained that there is no difference between the periods so we will work with just one time period. Revenues are calculated as a product of the amount of harvested population during one harvesting and a number of

[3] Note that the "prey" could be even a predator from the basic model in this situation.
[4] It could be profitable to let a harvested population naturally decrease if the transaction costs are extremely high and a gain is offset by a loss caused by transaction costs.
[5] Concrete value of this constant depends on the equilibrium value and on the concrete population, i.e. it would be different for fish or mammal populations. In the following text are examples with cycles given by 30 %, 50 % and 70 % of the predator equilibrium value. Finding the proper value of this constant is the subject of biology so we will not try to do it in this paper.

harvestings during one time period. Costs are calculated as a product of transaction costs for one harvesting and a number of harvestings during one time period. The whole relationship is summarized in the next equation.

$$\pi = n * x_h - n * TC, \qquad (4)$$

for the prey-harvesting and

$$\pi = n * y_h - n * TC, \qquad (5)$$

for the predator-harvesting. Where π is the profit, n is the number of harvestings in one time period, x is the amount of prey harvested in one harvesting, y is the amount of predator harvested in one harvesting and TC are the transaction costs.

If one considers knowing the relationship between the amount of harvesting and time needed to harvest it, and the length of one time period, which is one, it is easy to determine n^* when x^* respectively y^* is known. The relationship between time needed to move the system after one harvesting to another one and the amount of harvested population could be quantified as following:

$$t_h = f(x_h, y_{min}), \qquad (6)$$

for harvesting the prey and

$$t_h = f(y_h, y_{min}), \qquad (7)$$

for harvesting the predator. The index h means that the unit is connected to one harvesting. Since one time period lasts $t = 1$ we get:

$$n = \frac{1}{t_h}. \qquad (8)$$

Insert (6) resp. (7) and (8) to (4) resp. (5) and get:

$$\pi = \frac{1}{f(x_h, y_{min})} * x_h - \frac{TC}{f(x_h, y_{min})} = \frac{x_h - TC}{f(x_h, y_{min})}, \qquad (9)$$

resp.

$$\pi = \frac{1}{f(y_h, y_{min})} * y_h - \frac{TC}{f(y_h, y_{min})} = \frac{y_h - TC}{f(y_h, y_{min})}.$$

It is obvious that the difference between the profit and the loss is given by the relationship between the optimal amount of harvested population and the transaction costs. Since we consider y_{min} given the only variable which could be optimized is x_h resp. y_h. The first derivative of (9) is:

$$\pi' = \frac{f'(x_h, y_{min}) * (x_h - TC) - f(x_h, y_{min})}{f^2(x_h, y_{min})},$$

resp.

$$\pi' = \frac{f'(y_h, y_{min}) * (y_h - TC) - f(y_h, y_{min})}{f^2(y_h, y_{min})}.$$

If the function of profit is smooth then an extreme exists in the point where the first derivative is zero. This point is a maximum if and only if the first non-zero even-order derivative is negative. The most important factor here is the character of the function f. This will be subject of our further analysis in the following parts of the paper.

Hence the second derivative is:

$$\pi'' = \frac{2 * f'(x_h)[f'(x_h) * (x_h - TC) - f(x_h)] - f(x_h) * [f''(x_h) * (x_h - TC)]}{f^3(x_h)},$$

$$\pi'' = \frac{2 * f'(y_h)[f'(y_h) * (y_h - TC) - f(y_h)] - f(y_h) * [f''(y_h) * (y_h - TC)]}{f^3(y_h)}.$$

Since the profit function and its derivative are smooth, the profit is maximal if and only if the optimal value x_h^*, resp. y_h^* fulfils following condition:

$\pi'(x_h^*) = 0 \;\&\; \pi''(x_h^*) < 0$,

$\pi'(y_h^*) = 0 \;\&\; \pi''(y_h^*) < 0$.

Description of the Data

The time cannot be given explicitly as a function of parameters of the basic model or the actual amounts of populations. However as was mentioned above the relationship containing a time factor is needed for the optimization.

Because of the necessity of estimating the function $f(x_h, y_{min})$ representative data were selected from two different models. For each model three possible starting points were chosen, each of them has x coordinate equal to the equilibrium value and the other coordinate is given by the limit applied on the minimum number of predator population – 30 %, 50 % and 70 % of the equilibrium value. From this starting point x and y coordinates were collected at every hundredth of unit time.

The first analyzed model is given by:

$\frac{dx}{dt} = (5 - 0.02 * y)x,$

$\frac{dy}{dt} = (-10 + 0.01 * x)y,$

$t \geq 0, x(0) = 1000, y_{30}(0) = 75, y_{50}(0) = 125, y_{70}(0) = 175.$

The second one is given by:

$\frac{dx}{dt} = (2 - 0.005 * y)x,$

$\frac{dy}{dt} = (-5 + 0.0025 * x)y,$

$t \geq 0, x(0) = 1000, y_{30}(0) = 120, y_{50}(0) = 200, y_{70}(0) = 280.$

This data had to be modified in order to get the required relationship. Since the modification algorithm is different for harvesting different populations it is described in the special chapters.

For simplicity we shall look closely to optimizing the harvesting of the prey in this part. A similar approach could be used when harvesting predator population, however there are some differences which will be described at the end of this paper. We should keep in mind previously mentioned facts and develop our theories more precisely by using the empirical evidence. The important factor here is time, which was not analyzed in depth before. The problem is that time cannot be given explicitly since the basic equations cannot be solved. So we will estimate the function $f(x_h, y_{min})$. Firstly the modification of the data is needed.

Estimating the Needed Function

The aim of the modification is to find two points with similar y coordinate and with the x coordinates in the different half-planes given by the prey equilibrium value. The y coordinate must be always lower than y_e. When we have two points fulfilling the previous criteria we can compute 2 main characteristics:

- **time (dt)** – how much time it does take till the system moves from the point with lower x value to the other one,
- **harvested prey (dx)** – absolute value of the difference between the x coordinates.

Now we have at least some rough image of the function $f(x_h, y_{min})$. The aim of the following is to find some sufficient estimate of this function. There are generally two ways to do this. The first one is based on numerical mathematical methods and the second one is using an econometrical approach. We have chosen the second one because of the following reasons:

- attainability of software for econometrical modelling,
- economic purpose of this paper,
- econometrical methods are based on the numerical mathematics, so one should get the same results using both methods after concerning some assumptions.

The main criterion for the estimate is minimal residual sum of squares. Note that there are no stochastic variables in the data; our only objective is to find a sufficient estimate of the non-elementary

function. It is possible to use econometrical methods here, but one should be very careful about interpreting some results such as confidence intervals etc.

It is obvious that a relationship between dt and dx is non-linear. Let's assume that for a given model and y_{min} the function $f(x_h)$ is quadratic. Since a zero time change brings a zero change in the prey population the model should not contain any intercept but the intercept of a low value could significantly improve the model and a harvester would not need any estimate for dt values very close to zero. So we will decide on the basis of econometrical criteria.

Selection of the Regression Model

It was determined that the best fit for the particular model and y_{min}obtained by OLS is given by the following formula:

$dt = \alpha + \beta * dx^2 + \gamma * dx + \varepsilon.$

The results of OLS regressions for all possible models in the given situation are in the following table

Table 1: Results of all needed regressions

	const	sq_dx	dx	p-value (dx)	R-squared	AIC
M1_y70	0,033987	1,46E-06	0,000134	.	0,987335	-209,73
M1_y50	0,039685	4,25E-07	0,000114	**	0,988236	-240,727
M1_y30	0,033091	1,43E-07	0,000139	***	0,991099	-258,135
M2_y70	0,066321	9,30E-07	0,000264	***	0,987376	-359,106
M2_y50	0,054342	2,49E-07	0,000228	***	0,992081	-429,245
M2_y30	0,055766	9,20E-08	0,000196	***	0,992709	-444,363

In the table, estimates are shown for all three parameters for each model, p-value of dx[6], R-squared and AIC. "M1_y70" refers to the Model 1 with $y_{min} = 0.7 * y_e$, where y_e is an equilibrium value of predator population etc. According to the values of the R-squared and AIC one could say that the model is describing the reality better with lowering y_{min}. Evaluating the influence of the growth and interaction rates is not so easy since there all four parameters of the ODE system are changed.

Non-linear Regression

We believe that knowing just one function, which could be generally used for any model and y_{min}, would make the optimization much easier. This is unfortunately computationally harder than the previous estimate but we will show the algorithm how to estimate the function after knowing the possibilities. Our case could illustrate a situation when a harvester is deciding between two possible harvesting locations and some external authority could set one of the three harvesting limits. When knowing the probabilities of setting each limit, prices of the harvested species and transaction costs for each system the estimated function f could help to find harvester's optimal decision.

We shall estimate a relationship given by:

$dt = \alpha * (m1 + \varphi_1 * m2) * (y70 + \omega_{11} * y50 + \omega_{21} * y30) + \beta * dx^2 * (m1 + \varphi_2 * m2) * (y70 + \omega_{12} * y50 + \omega_{22} * y30) + \gamma * dx * (m1 + \varphi_3 * m2) * (y70 + \omega_{13} * y50 + \omega_{23} * y30) + \delta * dy * (m1 + \varphi_4 * m2) * (y70 + \omega_{14} * y50 + \omega_{24} * y30) + \varepsilon,$

where $m1, m2, y70, y50, y30$ are dummies, i.e. for M1_y70 is $m1 = 1, y70 = 1$. The parameters φ and ω are showing the influence of the model or limit selection. Other variables have their usual meaning.

In order to use non-linear least squares (NLS) method the initial conditions for all parameters are needed so firstly, it is necessary to estimate them. Estimate the values of $\alpha, \beta, \gamma, \delta$ as an average of

[6] All other p-values are evaluated as "***" i.e. lower than 0.01, "**" means lower than 0.05, "." Means higher than 0.1.

these values for M1-y70 and M2_y70. Since we consider the parameter β as the most significant one according to its p-value we will estimate φ and ω for NLS using values of β by using these formulae:

$$\hat{\varphi} = \frac{\frac{\beta_{M2_y70}}{\beta_{M1_y70}} + \frac{\beta_{M2_y50}}{\beta_{M1_y50}} + \frac{\beta_{M2_y30}}{\beta_{M1_y30}}}{3},$$

$$\hat{\omega}_1 = \frac{\frac{\beta_{M1_y50}}{\beta_{M1_y70}} + \frac{\beta_{M2_y50}}{\beta_{M2_y70}}}{2},$$

$$\hat{\omega}_2 = \frac{\frac{\beta_{M1_y30}}{\beta_{M1_y70}} + \frac{\beta_{M2_y30}}{\beta_{M2_y70}}}{2}.$$

So the initial conditions $(\alpha, \beta, \gamma, \delta, \varphi, \omega_1, \omega_2)$ are given by $(0.047, 5.497E\text{-}07, 0.00018, -0.0002, 0.622, 0.279, 0{,}099)$. Let's consider that the initial values $\varphi_i, \omega_{1i}, \omega_{2i}$ are the same for each parameter (i.e. $\varphi_1 = \varphi_2 = \varphi_3 = \varphi_4$ etc.). Now proceed to the estimate by NLS. All parameters are significant in the this model. It is necessary to multiply each variable by the dummy and its parameter separately.

According to the AIC the influence of the dummies is not same for all parameters. It does make sense and brings valuable results to differ the influence of the dummies in the model.

Hence the estimated function is:

$$\widehat{dt} = 0.0278 * (m1 + 2.49 * m2) * (y70 + 0.796 * y50 + 0.83 * y30) +$$
$$+0.00000132 * dx^2 * (m1 + 0.714 * m2) * (y70 + 0.264 * y50 + 0.0982 * y30) +$$
$$+0.000211 * dx * (m1 + 1.175 * m2) * (y70 + 0.912 * y50 + 0.781 * y30) \quad (10)$$

We will demonstrate the quality of the model estimated by NLS by the following figures.

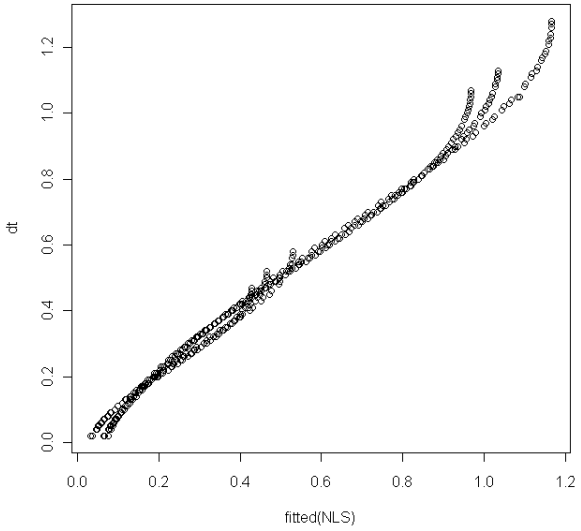


Figure 2: dt vs. fitted (dt)

It is clear that if we exclude observations gained from the points very close to the predator or prey equilibrium values the model works very well. The model's deficiency for high values is not important till the transaction costs are very high so a harvester wants to harvest when the predator population is in its equilibrium. If one gets x_h^* higher than $x_{max} - x_{min}$ they cannot use this estimate and they must use dt and dx for x_{max} and x_{min} calculated previously as optimal values[7].

[7] This could not be always an optimal solution. The optimality of the solution could be tested by selecting some little lower dt and dx and calculating a profit for this values. If the newly calculated profit is lower, then the optimal solution was found, else some heuristics should be used.

Next figure shows residuals of the model.

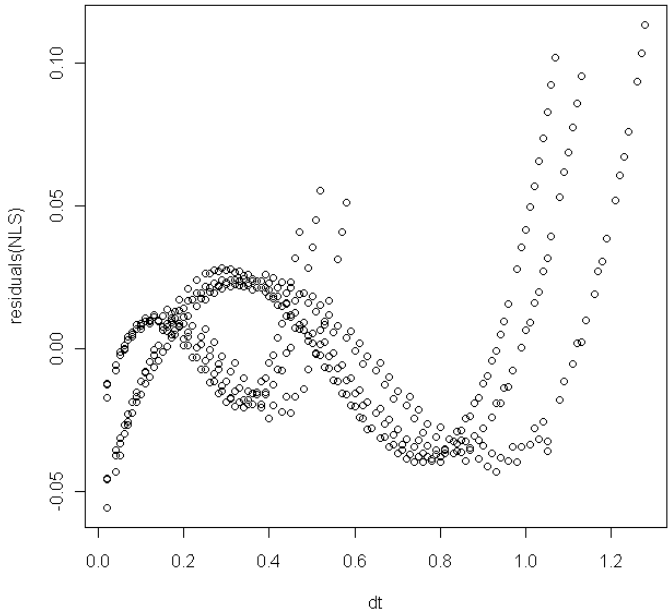


Figure 3: Residuals vs. dt

Figures 2 and 3 confirm that the data does not have a stochastic character. Next time, it could be better to not consider observations very close to the predator equilibrium. Based on the previously presented facts the estimated model can be used unless the system is close to the predator equilibrium.

Optimization Example

Firstly use the fact that $f(x_h)$ is a quadratic function, which could be written as $p * x_h^2 + q * x_h + r$. Insert this to EQ 6 and get optimal x_h:

$$x_h^* = \frac{p*TC+\sqrt{(p*TC)^2+p*(qTC+r)}}{p}, \qquad (11)$$

Consider a Predator-Prey system given by Model 1 where the predator population cannot decrease below one half of its equilibrium value and a prey-harvesting, profit-maximizing harvester who has transaction costs 50 species of prey population per one harvesting. Insert the values to (10) and get:

$$\widehat{dt} = (-0.022 + 0.00000035 * dx^2 + 0.00019 * dx). \quad (12)$$

After realizing that $dx = x_h$ the values from (11) could be inserted to (12) to get:

$$x_h^* \cong 355.$$

So one can easily find other necessary values and get that in this case a harvester should optimally harvest when there are about 1185 preys and 133 predators in the system. They should harvest 355 preys, i.e. till the prey population declines to approximately 830. It would take approximately 0.15 time unit so a harvesting should be done 7 times per year. The optimal revenues are 23,670 preys, costs 335 and the profit is 23,335 per one time period.

Optimization of Harvesting the Predator Population

The algorithm for optimizing harvesting of the predator population is very similar to the previously described. However the basic difference is that before the other population (predators) was taking a profit from a harvester. Now the non-harvested population is needed to keep the harvested one alive. This brings at least one significant fact which should be considered during optimization.

Consider a situation when both populations are rising and a harvester harvests small amounts of predators (i.e. harvestings do not affect a rising character of both populations). In this model the prey

population could be continuously rising to infinity without breaking any condition on the minimal limit. At the same time some amount of the predator population is harvested, so a harvester is making a profit. The further the system is from the equilibrium the higher is the predators' growth rate since they have a lot of prey in the system.

Calculating a profit in this situation cannot be done by using estimations since the optimum is at infinity. A rational solution of this complication is to consider that a harvester must harvest on the curve given by y_{min} as in the previous situation because of the computing problems and because it is closer to the real world.

Conclusion

The paper described a modification of the predator prey model with a harvester who is maximizing a profit from harvesting. We suggested an algorithm to optimize harvesting by maximizing the profit. Proper optimization is mathematically impossible due to the character of the ODE system utilized. We have successfully used econometric tools to estimate the system's behavior so we were able to quantify a profit function. Using the described algorithm could crucially help any entrepreneur dealing with similar systems.

References

CHROBOK, V. 2009. *Harvesting in the Predator-Prey Model*, Praha: VŠE.

CLARK, C. W. 1976. *Mathematical Bioeconomics: the optimal management of renewable resources*, New York: John Wiley.

DOUCET, P., SLOET, P. B. 1993. *Mathematical Modeling in the Life Sciences*, 2nd edition, Trowbridge: Redwood Books.

HÁJKOVÁ, V., JOHN, O., ZELENÝ, M. 2003. *Matematika*, Praha: Karolinum.

LEE, H. J., SCHIESSER, W. E., 2004. *Ordinary and Partial Differential Equation Routines in C, C++, Fortran, Java, Maple, and MATLAB*, London: Chapman & Hall/CRC.

PÁNKOVÁ, V. 2003. *Nelineární optimalizace pro ekonomy*, Praha: Professional Publishing.

REKTORYS, K., 1973. Přehled užité matematiky, Praha: SNTL.

VARIAN, H. R. 2006. *Intermediate Microeconomics: a modern approach*, New York: W. W. Norton & Company.

JEL Classification: C02, C61

Shrnutí

Optimalizace těžby přírodních zdrojů

Článek je zaměřen na modifikaci modelu dravec-kořist pro případ těžby jedné z populací. Nejdříve je uveden popis základního modelu, do kterého je později zavedena třetí strana, která maximalizuje těžbu jedné z populací. Jsou zde uvedeny základní předpoklady o chování systému, jako je dlouhodobá stabilita systému a zamezení přirozenému úhynu populací. Optimalizační algoritmus je založen na obecném charakteru systému a poskytuje univerzální nástroj pro řešení podobných problémů. Jelikož z příslušných diferenciálních rovnic nelze vyjádřit explicitně velikost populace v závislosti na času, tak bylo nezbytné tento vztah odhadnout. Odhad byl proveden metodou nelineárních nejmenších čtverců a přinesl robustní výsledky aplikovatelné i na jiné množiny výchozích bodů. Popsaná metoda byla ukázána na ilustrativním příkladu zabývajícím se těžbou populace kořisti, na konci jsou uvedeny zásadní rozdíly mezi těžbou populace dravců a populace kořisti.

Klíčová slova: model dravec-kořist, optimalizace, matematická bioekonomie, metoda nelineárních nejmenších čtverců.

Lineární regresní model s intervalovými výstupy – vybrané charakteristiky a jejich výpočet

Miroslav Rada
miroslav.rada@vse.cz
Doktorand oboru ekonometrie a operační výzkum
Školitel: prof. RNDr. Jan Pelikán, CSc. (pelikan@vse.cz)

Abstrakt: Článek je motivován studiem množiny potenciálních MNČ-odhadů intervalové regrese. Je známo, že pro bodové vstupy a intervalové výstupy má tato množina speciální kombinatorickou strukturu, konkrétně se jedná o zonotop – speciální typ polyedru reprezentovatelného jako Minkowského součet konečného počtu úseček nebo také jako lineární obraz krychle. Článek je zaměřen na rozbor dalších možných prakticky použitelných reprezentací a charakteristik zonotopu.

Hlavním výstupem článku je metaalgoritmus RRR, který nabízí jednotný algoritmický rámec pro řešení tří základních problémů spojených se zonotopy – enumerace fazet, enumerace vrcholů a výpočet objemu. Algoritmus umožňuje současné řešení všech tří úloh a tím úsporu části výpočetního úsilí. Prezentovaný algoritmus je kompaktní pro enumeraci fazet a výpočet objemu.

Klíčová slova: zonotop, intervalová regrese, enumerace vrcholů, enumerace fazet, rekurze, objem

Intervalová regrese a zonotopy

Intervalový regresní model s intervalovými výstupy a možnosti vyjádření jeho řešení

Pod pojmem *intervalový regresní model* se pro účely tohoto textu rozumí dvojice $(X, \boldsymbol{y})$ taková, že X je matice o rozměru $n \times p$ s hodností p a $\boldsymbol{y} = [y^d, y^h]$ je intervalový vektor rozměru $n \times 1$. *MNČ-řešení intervalového regresního modelu* je množina

$$B(X, \boldsymbol{y}) = \left\{\beta : \beta \in \mathbb{R}^p; \beta = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}y; y^d \leq y \leq y^h\right\}. \quad (1)$$

Pod pojmem intervalové regrese je možné si představit klasickou lineární regresní rovnici $\mathrm{E}(y) = X\beta_0$, kde β_0 je neznámý vektor regresních parametrů, přičemž přesné hodnoty pozorování závislé proměnné y nejsou známy; je známo pouze to, že patří do intervalů $\boldsymbol{y}$. $B(X, \boldsymbol{y})$ je tedy řešením problému *„Jsou dána přesná vstupní data ovlivňující nepřesně (intervalově) zadanou velikost výstupu – jaké jsou možné MNČ-odhady závislosti mezi vstupem a výstupem?“*.

Je možné si povšimnout, že nad intervalovým vektorem $\boldsymbol{y}$ se nepředpokládá žádné pravděpodobnostní rozdělení. Pokud by nějaké rozdělení nad $\boldsymbol{y}$ zadané bylo, bylo by možné se na množinu $B(X, \boldsymbol{y})$ dívat jako na obor hodnot rozdělení jednotlivých MNČ-řešení.

Obecnější a ucelenější studii intervalové regrese, zejména zobecnění na podobu $(\boldsymbol{X}, \boldsymbol{y})$, kde jak matice vstupů, tak vektor výstupů jsou zadány intervalově, podává včetně možných aplikací a relevantní literatury práce Černý, Antoch, Hladík (2011).

Implicitní definice množiny $B(X, \boldsymbol{y})$ pomocí rovnice (1) neposkytuje příliš jasnou představu o její podobě. Je nicméně zřejmé, že se jedná o omezený polyedr, a jak bude patrné z dalšího textu, jde o polyedr se speciálními vlastnostmi, které do určité míry usnadňují jeho vyjádření pomocí „rozumnějších“ zápisů. Obecný omezený polyedr je možné popsat např. jako konvexní obal množiny vrcholů (platí pro omezené mnohostěny, nazvěme tuto množinu *vrcholová reprezentace polyedru*) nebo jako průnik množiny poloprostorů (*fazetová reprezentace*). Znalost vrcholů polyedru $B(X, \boldsymbol{y})$ jakožto krajních bodů přináší informaci o „extrémních“ hodnotách odhadovaných regresních parametrů a umožňuje vyjádřit všechny jejich hodnoty jako konvexní kombinaci vrcholů. Fazetová (poloprostorová) reprezentace může být užitečná např. pro optimalizaci nad MNČ-řešením, poskytuje také snadnou kontrolu, zda zadaný vektor regresních parametrů patří do MNČ-řešení intervalového regresního modelu (ačkoli

tato otázka může být snadno zodpovězena vyřešením úlohy lineárního programování, viz Černý, Rada (2010)), apod.

Vzhledem k tomu, že velikost vrcholové, resp. fazetové reprezentace obecného polyedru může být exponenciální funkcí velikosti vstupu (velikostí vstupu rozumějme např. dimenzi polyedru nebo počet parametrů v případě intervalové regrese), může být účelné „náročný“ popis polyedru nahradit méně náročným, byť méně přesným, geometrickým objektem. Zajímavé výsledky se týkají zejména aproximace polyedrů elipsoidy.

Další užitečnou charakteristikou polyedru může být jeho objem – v případě množiny $B(X, \boldsymbol{y})$ je to přirozená míra „neurčitosti“, kterou do klasického regresního modelu přináší neurčitost intervalových pozorování.

Článek je organizován následovně. Nejprve jsou popsány speciální vlastnosti $B(X, \boldsymbol{y})$ jako zonotopu – speciálního typu polyedru a následně jsou rozebrány známé výsledky týkající se jeho reprezentace. Stěžejní částí článku je druhá kapitola, kde je navržen algoritmus řešící problémy enumerace vrcholů a fazet zonotopu a výpočtu objemu zonotopu.

Zonotop

Minkowského součet dvou množin $A \in \mathbb{R}^d$ a $B \in \mathbb{R}^d$ je operace

$$A + B \coloneqq \{a + b : a \in A, b \in B\}.$$

Zonotop je konvexní polyedr definovaný jako Minkowského součet konečného počtu úseček. Zonotop Z v $\mathbb{R}^d$ určený m úsečkami je tedy $Z = L_1 + L_2 + \cdots + L_m$, kde $L_i = [a_i, b_i]$ je úsečka s krajními body $a_i, b_i \in \mathbb{R}^d$ pro $i = 1, \dots, m$. Alternativně je možné psát

$$Z = \{s\} + G_1 + G_2 + \cdots + G_m, \tag{2}$$

přičemž $s = a_1 + a_2 + \cdots + a_m$ a $G_i = [0, g_i]$, $g_i = b_i - a_i$ pro všechna $i = 1, \dots, m$. Vektor s je *posunutí* zonotopu a vektory g_i se nazývají *generátory* zonotopu. Zonotop je s využitím předchozího značení možné jednoznačně popsat pomocí s a seznamu generátorů $g_1, \dots, g_m$. Takový zonotop se značí $\mathcal{Z}(s; g_1, \dots, g_m)$. Dodejme, že *dimenze* zonotopu je dimenze afinního obalu seznamu generátorů a značí se $\dim(\mathcal{Z})$.

Zonotopy mají zajímavé geometrické vlastnosti:

a) Jsou středově symetrické. Tato vlastnost plyne z faktu, že úsečky jsou středově symetrické a Minkowského součet zachovává středovou symetrii.

b) Krychle v $\mathbb{R}^d$ je speciální případ zonotopu s generátory navzájem kolmými.

c) Každý zonotop je lineární obraz nějaké krychle s dimenzí nejvýše m, kde m je počet generátorů zonotopu. Pro obecný zonotop Z v $\mathbb{R}^d$ je tedy možné psát

$$Z = \{x : x = Gy, y_d \le y \le y^h, y \in \mathbb{R}^m\}, \tag{3}$$

kde matice G má rozměr $d \times m$. Důkaz tohoto tvrzení je možné najít např. v Černý, Rada (2010).

d) Každé lineární zobrazení zonotopu je samo o sobě zonotopem. Tato vlastnost bezprostředně plyne z předchozích dvou tvrzení, neboť při zobrazení $f(x) = G'x$ bude obraz zonotopu Z vyjádřen jako

$$Z' = \{G'x : x \in Z\} = \{z : z = G'Gy, y_d \le y \le y^h\}.$$

e) Podoba zonotopu není závislá na pořadí jeho generátorů, tj. pro libovolné $i \in \{1, \dots, n\}$ platí

$$\mathcal{Z}(s; g_1, \dots, g_n) = \mathcal{Z}(s; g_i, g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_n).$$

f) Mějme zonotop $Z = \mathcal{Z}(s; g_1, g_2, \dots, g_n)$, přičemž existují indexy $i, j \in \{1, \dots, n\}$ takové, že $g_i = \alpha g_j$ a $\alpha \in \mathbb{R}$. S využitím tvrzení e) je možné bez újmy na obecnosti předpokládat, že $i = 1$ a $j = 2$. Potom platí, že

$$\mathrm{Z} = \begin{cases} \mathcal{Z}(\mathrm{s}; \mathrm{g}_1 + \mathrm{g}_2, \mathrm{g}_3, \dots, \mathrm{g}_\mathrm{n}) & \text{pro } \alpha \ge 0, \\ \mathcal{Z}(\mathrm{s} + \mathrm{g}_2; \mathrm{g}_1 - \mathrm{g}_2, \mathrm{g}_3, \dots, \mathrm{g}_\mathrm{n}) & \text{pro } \alpha < 0. \end{cases} \tag{4}$$

Generátor g_2 nazveme *redundantním*. Redundantní generátory je možné postupně odstraňovat použitím převodu (4). Z původní reprezentace zonotopu $s; g_1, \dots, g_n$ tak vznikne reprezentace $s'; g'_1, \dots, g'_{n'}$, přičemž $n' \leq n$. Proceduru, která odstraní všechny redundantní generátory, pojmenujme ji red:

(5) $$\left(\mathrm{s}'; \mathrm{g}'_1, \dots, \mathrm{g}'_{\mathrm{n}'}\right) = red(\mathrm{s}; \mathrm{g}_1, \dots, \mathrm{g}_\mathrm{n}).$$

Zaveďme následující značení: symbolem $A_{\cdot i}$ se značí i-tý sloupec dané matice A.

Připomeňme nyní tvrzení c) z předchozího výčtu a předpokládejme zonotop Z z tohoto tvrzení. Pokud položíme $s = Gy^d$ a $g_i = G_{\cdot i}(y_i^h - y_i^d)$, je možné zonotop Z napsat pomocí vzorce (2), kde $G_i = [0, g_i]$.

Pokud se nyní vrátíme k intervalovému regresnímu modelu $(X, \boldsymbol{y})$ a jeho MNČ-řešení $B(X, \boldsymbol{y})$, pak je možné s využitím předchozích definic a tvrzení psát, že MNČ-řešení se dá vyjádřit jako

$$B(X, \boldsymbol{y}) = \mathcal{Z}\left(Qy^d; Q_{\cdot 1}\left(y_1^h - y_1^d\right), \dots, Q_{\cdot n}\left(y_n^h - y_n^d\right)\right),$$

kde $Q = \left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}$.

Reprezentace MNČ-řešení jako různých popisů zonotopu

Zápis množiny $B(X, \boldsymbol{y})$ jako zonotopu otevírá možnost použít k jejímu popisu algoritmů specifických pro zonotopy, přičemž jako vstup pro tyto algoritmy je k dispozici popis pomocí generátorů.

Pro algoritmy formulované nad zonotopy je důležitá věta o mezích počtů fazet a vrcholů zonotopu (viz Ziegler (2004)), podle které pro počet fazet (označme ho f_{d-1}) a počet vrcholů (značený f_0) zonotopu dimenze d o m generátorech platí

$$f_{d-1} \leq 2\binom{m}{d-1} \text{ a } f_0 \leq 2\sum_{i=0}^{d-1}\binom{m-1}{i}.$$

Navíc, nerovnosti výše jsou pro určité zonotopy splněné jako rovnosti, a proto tyto meze není možné obecně zlepšit. Není proto možné pro obecný zonotop navrhnout algoritmus, který by v čase polynomiálním v dimenzi zonotopu spočítal ze zadaných generátorů jeho fazetovou nebo vrcholovou reprezentaci.

Pro problémy podobného typu, tj. kde velikost výstupu je exponenciální funkcí velikosti vstupu, má nicméně smysl preferovat algoritmy, pro které se velikost vypočteného výstupu zvětšuje polynomiálním tempem. Například pro enumeraci vrcholů jisté třídy zonotopů o m generátorech a dimenzi d je zřejmě možné navrhnout triviální algoritmus, který bude postupně testovat obrazy všech vrcholů krychle, která je jeho inverzním obrazem ve smyslu vzorce (3), a vypíše ty vrcholy krychle, které jsou vrcholy zkoumaného zonotopu (testovatelné pomocí lineárního programování). Takový algoritmus zpracuje 2^m bodů, přičemž pro danou třídu zonotopů může být vrcholy např. jen 2^d z nich. Algoritmus tedy pro každý vypsaný vrchol zonotopu musí ověřit řádově 2^{m-d} vrcholů krychle. Jistě by bylo účelné navrhnout takový algoritmus, který na vypsání nového vrcholu bude potřebovat čas nejvýše polynomiální v počtu celkem vypsaných vrcholů, nebo dokonce čas polynomiální ve velikosti vstupu.

Další vhodnou vlastností algoritmů s exponenciální časovou složitostí je *kompaktnost*. Algoritmus je kompaktní, pokud je jeho paměťová složitost polynomiální ve velikosti vstupu. (Ne)užitečnost (ne)kompaktního algoritmu je možné demonstrovat na algoritmu pro výpočet fazetové reprezentace d-dimenzionálního zonotopu – je možné navrhnout triviální algoritmus, který postupně projde všechny (d-1)-tice generátorů zonotopu a spočte, jakou fazetu tyto generátory určují. V případě, že je nově spočtená fazeta originální (tj. nebyla dosud vygenerována jinou (d-1)-tící generátorů, přidá ji do seznamu fazet, jinak ji zahodí. Takový algoritmus má pro svůj běh nutně exponenciální paměťové požadavky. Užitečnější by byl zřejmě algoritmus, u kterého by bylo implicitně zaručeno, že výstup nebude obsahovat duplicitní fazety a u kterého by paměťové požadavky bylo možné omezit polynomem závislým jen na vstupu.

Pro získání fazetové reprezentace zonotopu je možné použít Seymourův algoritmus pro generování nadrovin matroidů (Seymour (1994)), který je polynomiální ve velikosti vstupu a výstupu, nicméně není kompaktní.

Pro výpis všech vrcholů zonotopu existuje kompaktní algoritmus polynomiální ve velikosti vstupu a výstupu, viz Avis, Fukuda (1996). Problém enumerace vrcholů zonotopu má

Obecně se předpokládá (viz Ziegler (2004)), že výpočet objemu zonotopu je exponenciálně časově složitý problém. Zonotop Z dimenze d o m generátorech je možné rozdělit do rovnoběžnostěnů tvořených jednotlivými d-ticemi generátorů. Jeho objem je poté roven součtu objemů těchto rovnoběžnostěnů, tedy (předpokládejme bez újmy na obecnosti, že Z má plnou dimenzi)

(6) $$vol(Z) = \sum_{1 \le i_1 < i_2 < \cdots i_{d-1} < i_d \le m} \left|\det\left(g_{i_1} \cdots g_{i_d}\right)\right|.$$

RRR Algoritmus

Reduction-and-Reconstruction-Recursion algoritmus je metaalgoritmus umožňující řešit (přinejmenším) následující úlohy týkající se zonotopů:

a) výpočet objemu,
b) enumerace vrcholů,
c) enumerace fazet.

Algoritmus při svém běhu volá procedury $BasicCase$ a $Combine$, jejichž konkrétní podoba je vázána na úlohu, k jejímuž řešení se RRR algoritmus používá.

Na vstupu algoritmu se očekává zonotop zadaný seznamem generátorů v $\mathbb{R}^d$ a posunutím, tedy $Z = \mathcal{Z}(s; g_1, \dots g_n)$. Je možné si povšimnout, že pro každou permutaci π prvků $\{1, \dots, n\}$ je $\mathcal{Z}(s; g_1, \dots g_n) = \mathcal{Z}(s; g_{\pi(1)}, \dots g_{\pi(n)})$. Bez újmy na obecnosti můžeme předpokládat, že ze seznamu generátorů zonotopu $g_1, \dots, g_n$ jsou generátory $g_1, \dots, g_{\dim(Z)}$ lineárně nezávislé vektory, a mimo to také navíc předpokládejme, že dimenze Z je d. Řekněme, že procedura red definovaná vztahem (5) kromě redundantních generátorů také změní pořadí tak, aby prvních d generátorů bylo lineárně nezávislých.

Podoba řídící procedury, která se rekurentně vyhodnocuje, je následující:

[1] **Function** $RRR(\hat{s}; \hat{g}_1, \dots, \hat{g}_{\hat{n}})$
[2] $(s; g_1, \dots, g_n) \coloneqq red(\hat{s}; \hat{g}_1, \dots, \hat{g}_{\hat{n}})$
[3] **if** $n = \dim(g_1, \dots, g_n)$ **then**
[4] $OUTPUT \coloneqq BasicCase(s; g_1, \dots, g_n)$
[5] **else**
[6] **choose** i **such that** $g_n \notin H_i$
[7] $DATA_1 \coloneqq RRR(s; g_1, \dots, g_{n-1})$
[8] $DATA_2 \coloneqq RRR(P(s, g_n); P(g_1, g_n), \dots, P(g_{n-1}, g_n))$
[9] $OUTPUT \coloneqq Combine(s; g_1, \dots, g_n; DATA_1, DATA_2)$
[10] **end**.

Procedura $P(v, g)$ je projekční funkce, promítající vektor $v \in \mathbb{R}^d$ rovnoběžně s generátorem g do vhodně zvolené nadroviny. V textu předpokládejme nadrovinu

(7) $$H_i = \{\xi : \xi \in \mathbb{R}^d, \xi_i = 0\}.$$

Základní idea algoritmu spočívá v postupném výpočtu zonotopu Z pomocí posloupnosti zonotopů $Z_d, \dots, Z_n$, kde $Z_i = \mathcal{Z}(s; g_1, \dots, g_i)$ pro i splňující $d \le i \le n$.

První člen posloupnosti Z_d je vzhledem k předpokladu, že prvních d generátorů Z je lineárně nezávislých, kombinatoricky jednoduchý zonotop odpovídající d-dimenzionálnímu rovnoběžnostěnu, pro nějž je výpočet požadovaného výstupu triviální. Výpočet výstupu pro Z_d řeší procedura $BasicCase$.

Při konstrukci každého dalšího, řekněme i-tého, zonotopu Z_i se použije již spočtený výstup pro zonotop Z_{i-1} a zkoumá se, jakým způsobem se výstup změní přidáním generátoru g_i. Algoritmus je

založen na pozorování, že přidání nového generátoru k zonotopu vede k „roztažení“ jeho střední části. Toto roztažení je dáno definicí Minkowského součtu – pokud je dané x prvkem množiny Z_{i-1}, pak, vzhledem k $Z_i = Z_{i-1} + G_i$ musí být $x' = x + g_i$ prvkem Z_i. Jak bude patrné dále, pokud bude známa konkrétní kombinatorická podoba této „střední části“, bude možné vypočítat výstup pro zonotop Z_i. Za „střední část“ nyní prohlasme zobrazení zonotopu Z_{i-1} rovnoběžně s generátorem g_n do nějaké nadroviny (zobrazení jednotlivých bodů pomocí projekční funkce $P(v, g)$). Při znalosti výstupu algoritmu pro oba zonotopy, tj. pro Z_{i-1} a jeho obraz, jsou tyto výstupy zpracovány pomocí procedury $Combine$ do výstupu pro zonotop Z_i.

Aplikace 1: Výpočet objemu

Pro výpočet objemu definujme procedury $BasicCase$ a $Combine$ následujícím způsobem:

$$\begin{aligned} \mathrm{BasicCase}(\mathrm{s}; \mathrm{g}_1, \dots, \mathrm{g}_\mathrm{n}) &= |\det(\mathrm{g}_1, \dots, \mathrm{g}_\mathrm{n})| \\ \mathrm{Combine}(\mathrm{s}; \mathrm{g}_1, \dots, \mathrm{g}_\mathrm{n}; \alpha_1; \alpha_2) &= \alpha_1 + \alpha_2 |\mathrm{g}_{\mathrm{ni}}| \end{aligned} \tag{8}$$

Symbol g_{ni} v předchozím vzorci označuje i-tou složku vektoru g_n, index i označuje index proměnné ξ_i v definici nadroviny (7), do které procedura P promítala zonotop Z_{n-1}.

Podoba procedury $Combine$ vychází přímo ze vzorce (6). Na objem Z_{i-1} je možné se dívat jako na součet determinantů matic, které je možné vytvořit z generátorů $g_1, \dots, g_{i-1}$. Druhý sčítanec procedury $Combine$, tedy $\alpha_2|g_{ni}|$ interpretujme následovně. Pomiňme numerické potíže a předpokládejme, že determinant matice A počítáme převodem A na dolní trojúhelníkový tvar a vynásobením prvků na diagonále. Promítání pomocí funkce P je vlastně anulování jedné složky všech sloupců matic tvořených těmi d-ticemi generátorů, které obsahují generátor g_i. Hodnota α_2 je součet všech subdeterminantů matic rozměru $(d-1) \times (d-1)$ tvořených promítnutými generátory $P(g_1, g_i), \dots, P(g_{i-1}, g_i)$.

Výhodou takového postupu oproti přímému výpočtu pomocí vzorce (6) je fakt, že jsou v rámci funkce red implicitně vynechávány nulové determinanty.

Aplikace 2: Enumerace vrcholů

Pro triviální případ enumerace vrcholů rovnoběžnostěnu je možné psát

$$BasicCase(s; g_1, \dots, g_n) = \left\{ v : v = s + \sum_{i=1}^{n} c_i g_i, c \in \{0,1\}^n \right\}.$$

Horní hranice zonotopu Z pro generátor g je množina

$$UB(Z, g) := \{x \in Z : \max_\alpha \{\alpha : x + \alpha g \in Z\} = 0\}.$$

Dolní hranice $LB(Z, g)$ zonotopu Z pro generátor g je definována analogicky. Vzhledem ke středové symetrii zonotopu je možné psát $LB(Z, g) := UB(Z, g)$.

Při přechodu od Z_{i-1} k Z_i se vytvoří nová horní hranice zonotopu pro g_i posunutím původní o g_i, pro každý vrchol v patřící do horní hranice Z_{i-1} bude součástí vrcholového popisu Z_i vrchol $v + g_i$. Vrcholy Z_{i-1} ležící na spodní hranici jsou i vrcholy Z_i. Množinu vrcholů zonotopu Z_{i-1} můžeme rozdělit na tři disjunktní podmnožiny: množinu vrcholů patřících zároveň do horní i spodní hranice E (extrémní body obrazu Z_{i-1}), množinu vrcholů patřících pouze do horní hranice H, a množinu vrcholů patřících pouze do dolní hranice D. Vrcholy D a H jsou si navzájem středově symetrickými obrazy.

[1] **Function** $Combine(s; g_1, \dots, g_n; VertexSet_1, VertexSet_2)$
[2] $E := \{v, v + g_n : v \in VertexSet_1, P(v, g_n) \in VertexSet_2\}$
[3] $D := \{v : v \in VertexSet_1, v \in LB(Z(s; g_1, \dots, g_{n-1}), g_n)\}$
[4] $H := \{v + g_n : v = c - 2v', v' \in D, c = s + \sum_{i=1}^{n-1} g_i/2\}$
[5] $OUTPUT := E \cup H \cup D$
[6] **end**.

Aplikace 3: Enumerace fazet

Mějme proceduru $Facet(s; g_1, \dots, g_n; j_1, \dots, j_{d-1})$, která pro daný zonotop vypočítá nadrovinu (a její poloprostor), ve kterém leží fazeta (a zonotop sám) určená generátory s indexy $j_1, \dots, j_{d-1}$. Při znalosti takové procedury budeme jako adekvátní výstup enumerace fazet zonotopu Z považovat množinu d-1-prvkových indexních množin, kterou označíme $J(Z)$, vzhledem ke středové symetrii každý prvek takové množiny bude generovat dvě fazety.

S reprezentací výstupu pomocí indexních množin souvisí určitá komplikace – při návratu z na vyšší úroveň rekurze je nutné mít informaci, kterým generátorům zonotopu vyšší úrovně odpovídají generátory zonotopu nižší úrovně. V jednotlivých procedurách bude proto předáván seznam indexů původních generátorů $j_1, \dots, j_n$, jimž odpovídají generátory aktuálně zpracovávaného zonotopu. V případě, že bude v rámci procedury red zmenšován počet generátorů, bude zachován menší z indexů.

Procedura $BasicCase$ má opět triviální podobu – stěny rovnoběžnostěnu dimenze n jsou určeny $(n-1)$-ticemi generátorů.

$$BasicCase(s; g_1, \dots, g_n; j_1, \dots, j_n) = \{\{j_1, \dots, j_n\} \backslash \{j_i\} : 1 \leq i \leq n\}$$

Co se týče procedury $Combine$, využijeme pozorování, že $J(Z_{i-1})$ je podmnožinou $J(Z_i)$. Nově vzniklé fazety jsou určeny přidaným generátorem g_i a generátory určujícími fazety obrazu Z_{i-1}, tedy

$$Combine(s; g_1, \dots, g_n; IndexSet_1, IndexSet_2; j_1, \dots, j_n) =$$
$$= IndexSet_1 \cup \{set \cup j_n : set \in IndexSet_2\}.$$

Pro vypsání fazet ve smyslu poloprostorové reprezentace zonotopu je pak nutné na všechny indexní množiny spustit proceduru $Facet$.

Výše uvedená formulace podobně jako při výpočtu objemu implicitně vylučuje některé indexní množiny odpovídající duplicitním fazetám. Formulace algoritmu výše přesto může duplicitní fazety generovat. Pro odstranění duplicit je možné vždy před „přidáním“ individuální indexní množiny $J = \{j_1, \dots, j_n\}$ do výsledku otestovat, zda existuje takový generátor $g_k, k < \max(J)$, že $g_k \cup J$ tvoří množinu lineárně závislých vektorů. Pokud ano, pak bude daná fazeta vygenerována jinou indexní množinou a množina J se do výstupu nepřidává.

Závěr

Práce se zabývá množinou MNČ-odhadů lineárního regresního modelu s intervalovými výstupy. Tato množina má podobu specifického mnohostěnu – zonotopu. V práci jsou krátce diskutovány charakteristiky zonotopu jako množiny MNČ-odhadů intervalového regresního modelu, a následně navrhnuty algoritmy pro výpočet vybraných charakteristik. Jmenovitě se jedná o enumeraci vrcholů, enumeraci fazet a výpočet objemu množiny MNČ-odhadů.

Hlavní výsledek práce je metaalgoritmus RRR, který pro zonotop zadaný seznamem generátorů spočte množinu fazet, množinu vrcholů, a objem zonotopu. Algoritmus je založený na použití rekurzivního algoritmického rámce, který umožňuje počítat výše zmíněné výstupy simultánně s minimálním přidaným úsilím oproti tomu, když je počítán pouze jeden výstup. Tato vlastnost je prakticky použitelná zejména v případech, kdy je nutné spočítat objem zonotopu.

Časová složitost algoritmů je vzhledem k exponenciální velikosti produkovaného výstupu nutně exponenciální. Z pohledu paměťové složitosti je algoritmus kompaktní pro výpočet objemu a pro enumeraci fazet.

Použitý algoritmický rámec je založený na sofistikovaném procházení jednotlivých d-tic generátorů. Proces procházení a jeho efektivita závisí na počátečním uspořádání seznamu generátorů. Je otázkou, zda je možné a smysluplné hledat taková seřazení generátorů, která povedou ke snížení doby běhu algoritmu.

Otázkou hodnou pozornosti je hledání dalších aplikací RRR na úlohy spojené se zonotopy.

Literatura

AVIS, D., FUKUDA, K. Reverse search for enumeration. *Discrete Applied Mathematics* 65 (1996) 21–46.

ČERNÝ, M, ANTOCH, J. HLADÍK, M. *On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data* [online]. Výzkumná zpráva. Vysoká škola ekonomická v Praze, Fakulta informatiky a statistiky. 2011 [cit. 30.1.2012]. Dostupné z: http://nb.vse.cz/~cernym/plr.pdf.

ČERNÝ, M., RADA, M., A note on linear regression with interval data and linear programming, In: *Quantitative Methods in Economics: Multiple Criteria Decision Making XV*, Kluwer, Iura Edition, Slovakia, 2010, s. 276–282.

SEYMOUR, P. D. A note on hyperplane generation. *Journal of Combinatorial Theory*, Series B 61, 1994, s. 88-99.

ZIEGLER, G. *Lectures on Polytopes*, Springer, Germany, 2004.

JEL Classification: C29, C65.

Dedikace: Publikace vznikla s podporou projektu IGA VŠE F4/18/2011.

Summary

Crisp input-interval output linear regression model – some characteristics and their computation

In the first part of this paper, we connect the OLS-solution of linear regression with interval output to the well-known object of computational geometry – a zonotope. The main result of this paper is in the second part. We propose algorithms for the facet enumeration, vertex enumeration and volume computation over zonotopes given by the generator description. The algorithms are based on the common recursive frame which allows solving the problems simultaneously with minimal effort added to the computation of one of them. This is practically usable especially when the volume of the zonotope is required.

Space complexity of both the facet enumeration and volume computation is polynomial in input size. The proposed algorithm is compact for both problems, but not for the vertex enumeration.

The algorithmic frame is based on traverse through subsets of generators. The process of traverse depends on the initial ordering of generators. There is an interesting question whether it's possible and meaningful to find a good ordering which would shorten the run time.

Additional effort could also be dedicated to further applications of the algorithmic frame in the area of zonotopes and computational geometry.

Keywords: interval regression, zonotope, recursion, vertex enumeration, facet enumeration, volume.

Dealing with Fiscal Data Uncertainty (DSGE-VAR Approach)

Jakub Ryšánek
jaac@atlas.cz
PhD. student, Econometrics and Operational Research
Supervisor: prof. Ing. Roman Hušek, CSc. (husek@vse.cz)

Abstract: Fiscal data in both developed and emerging economies do not provide direct information that could be easily used in macroeconomic analyses. Fiscal variables that the government has under its control, such as tax rates, do not seem to be in line with other fiscal measures, such as aggregate tax revenues. Discrepancies of this kind put any empirical business cycle study with a fiscal block into a vague position, with the results being counterintuitive. To answer the question of what is the response of the gross domestic product in terms of both amplitude and timing on the shocks in the fiscal area, I work with a vector autoregressive model (VAR) that includes several domestic variables. To lower the effect of data uncertainty in the fiscal area, I form fairly strict prior assumptions about the model parameters based on a structural dynamic stochastic general equilibrium (DSGE) model as suggested by Leeper *et al.* (2010). This Bayesian analysis fully coincides with the DSGE-VAR approach taken in Del Negro and Schorfheide (2004). The results of this analysis are more plausible comparing to standard output of traditional VARs. A positive shock to taxation results in an initial drop of the economic activity and is followed by gradual increase as the government expenditures increases in response to increased tax revenue. Furthermore, as a by-product, I derive estimated quarterly figures for those fiscal variables that are officially published only on annual basis.

Keywords: vector autoregressive model (VAR), DSGE model, fiscal policy, Bayesian analysis.

Introduction

Fiscal data, as they are published, usually do not create a basis for macroeconomic analysis for which there are several reasons even in developed countries. In the first place, complete government statistics are available on annual basis, while quarterly government accounts are available only for certain government variables. To study the business cycle properties of an economy, one commonly works with data based on quarterly frequency. Especially in the case of the Czech Republic, it would be very inconvenient to work with annual data because of short data sample. Relevant data for both government sector and the rest of the economy exist since 1995.

Second, there are distinct discrepancies among the development of government aggregate variables on one hand and the government control variables on the other hand. Traditional econometric exercises of models that include fiscal variables therefore result in insignificant estimates of the model parameters or the estimates are biased. The methodology of the data management in official agencies also induces an uncertainty to the fiscal data. According to the Government Finance Statistics (GFS) of the International Monetary Fund, data are stored in IMF's database on the cash basis, as are some of the fiscal data from the Czech Statistical Office. For example, this is the case of monthly collection of taxes. On the other hand, data based on the accrual principle better reflect position of the economy in the business cycle. To bypass the discrepancies among various fiscal variables, Mendoza (1994) and Jones (2002) suggest a computation of tax rates that are more in line with the economic development. Such approach is also taken in this paper and discussed further.

Third, revenues and expenditures of the government sector are affected very much by one-time operations that have only a temporary effect (if any effect) on the rest of the economy. It is important to analyze government variables net of such operations since the Czech Republic has gone through a tedious process of privatization in the recent history.

The modeling approach adopted in this paper builds on Del Negro and Shorfheide (2004) who use a structural DSGE model to simulate artificial time series. Next, they fit a VAR on these artificial time series and resulting estimates of the VAR parameters serve as a prior for final BVAR model which is estimated on real data. Del Negro and Shorfheide (2004) use hierarchical priors in that the parameters

of the DSGE model themselves come from a prior distribution and posterior estimates serve as priors for the BVAR. Markov chain Monte Carlo techniques are then used to draw from joint posterior distribution of both DSGE and BVAR parameters. For simplicity, I depart from their approach in the following way. Instead of estimating the model parameters jointly, I separate the DSGE and BVAR parts. First, I impose subjective prior distribution on the structural DSGE parameters and compute the posterior estimates. I then take posterior mode values of the parameters to simulate artificial time series of interest and fit an unrestricted VAR on these series. Second, I use the estimated distributions of the VAR parameters and the covariance matrix of the random disturbances to form a normal-gamma prior based on preceding computations into the final BVAR, which is estimated on real data.

In the subsequent parts of this paper, I discuss the main building blocks of the structural DSGE model, empirical BVAR model and its prior settings. Data section includes description of all data sources used in this paper and calibration of tax rates based on Mendoza (1994) and Jones (2002). Posterior distributions of estimated parameters as well as computational pitfalls of the approach taken in this paper are discussed in a separate appendix.

DSGE model

The structural part of the model closely follows Leeper *et al.* (2010) who incorporate a block of fiscal variables into a DSGE model of the US economy. They use a neoclassical growth model that includes households, firms and the government. The economy is closed and no monetary authority is assumed. To capture more closely evidence stemming from real data, the model also features standard rigidities, such as investment adjustment costs, variable capacity utilization and external habit formation as shown further.

Households

The representative household maximizes its utility function

$$E_0 \sum_{t=0}^{\infty} \beta^t u_t^b \left[\frac{(c_t - hC_{t-1})^{1-\gamma}}{1-\gamma} - u_t^l \frac{l_t^{1+\kappa}}{1+\kappa} \right],$$

which is an increasing function of consumption (c_t) and a decreasing function of hours worked (l_t), γ, κ are positive parameters denoting risk aversion and inverse Frisch elasticity of the labor supply, respectively, and $\beta \in (0,1)$ is a discount factor. E is an expectation operator. Consumption is taken relative to aggregate lagged consumption, which incorporates the external habit formation in consumption as described in Fuhrer (2000). Stationary preference shocks u_t^b and u_t^l follow AR(1) processes:

$$\ln(u_t^b) = \rho^b \ln(u_{t-1}^b) + \varepsilon_t^b, \quad \varepsilon_t^b \sim N(0, \sigma_b^2),$$
$$\ln(u_t^l) = \rho^l \ln(u_{t-1}^l) + \varepsilon_t^l, \quad \varepsilon_t^l \sim N(0, \sigma_l^2).$$

The representative household maximizes its utility subject to a flow budget constraint, which is given by

$$(1+\tau_t^c)c_t + i_t + b_t = (1-\tau_t^l)w_t l_t + (1-\tau_t^k)R_t^k v_t k_{t-1} + R_{t-1} b_{t-1} + z_t.$$

The expenditure side consists of consumption including the excise tax[1] (or the value added tax), τ_t^c, investment in physical capital, i_t, and 1 period risk-free government bonds, b_t. All this can be financed by labor income, $(1-\tau_t^l)w_t l_t$, capital income, $(1-\tau_t^k)R_t^k v_t k_{t-1}$ – both net of labor and capital income taxes), government bonds plus interest from the previous period, $R_{t-1}b_{t-1}$, and lump-sum transfers from the government, z_t. Capital income is dependent on the variable capacity utilization, v_t. If this variable equals 1, then all physical capital in the economy is put to use.

[1] On the expenditure side, I follow Leeper *et al.* (2010) who assume taxes only in the consumption sector.

However, the process of transforming investment into capital is costly because the law of motion for capital features adjustment costs:

$$k_t = (1-\delta(v_t))k_{t-1} + \left[1 - s\left(\frac{u_t^i i_t}{i_{t-1}}\right)\right] i_t.$$

where k_t is the time t stock of physical capital, δ is the depreciation rate dependent on the variable capacity utilization and $s(.)$ is the investment adjustment cost term which has nonzero effect as long as there is a difference in the level of investment in two consecutive periods. Leeper *et al.* (2010) do not specify exact functional form of the investment adjustment cost, they only state that $s(1)=s'(1)=0$ and $s''(1)>0$. Fernández-Villaverde *et al.* (2011) assume that both increase and reduction of investment relative to preceding period result in the same amount of adjustment cost. Their functional form,

$$\frac{\eta}{2}\left(\frac{u_t^i i_t}{i_{t-1}} - 1\right)^2,$$

satisfies the above assumptions about $s(.)$ and its derivatives, η can be treated as an unknown parameter. Shock to the marginal efficiency of investment, u_t^i, follows an AR(1) process:

$$\ln(u_t^i) = \rho^i \ln(u_{t-1}^i) + \varepsilon_t^i, \quad \varepsilon_t^i \sim N(0, \sigma_i^2).$$

Following Schmitt-Grohe and Uribe (2008), the functional form of $\delta(v_t)$ is quadratic in v_t:

$$\delta(v_t) = \delta_0 + \delta_1(v_t - 1) + \frac{\delta_2}{2}(v_t - 1)^2.$$

Firms

The representative firm rents capital and labor from the household and generates output according to the Cobb-Douglas production function with constant returns to scale:

$$y_t = u_t^a (v_t k_{t-1})^\alpha l_t^{1-\alpha},$$

where $\alpha \in (0,1)$, y_t is the total output and u_t^a denotes a stationary neutral technology shock to productivity that follows an AR(1) process:

$$\ln(u_t^a) = \rho^a \ln(u_{t-1}^a) + \varepsilon_t^a, \quad \varepsilon_t^a \sim N(0, \sigma_a^2).$$

Firms strive for maximization of their profit which is the difference between the total output and the total costs payable to the households, or

$$y_t - (w_t l_t + R_t^k v_t k_{t-1}),$$

where w_t is the wage rate and R_t^k is the rental rate. Since we assume perfect competition among firms, the wage and rental rates always equal to the respective marginal productivities of labor and capital:

$$w_t = \frac{(1-\alpha)y_t}{l_t}, \quad R_t^k v_t = \frac{\alpha y_t}{k_{t-1}}.$$

Fiscal policy

The government collects taxes from private consumption, labor income and capital income. It can also issue one-period bonds. On the expenditure side, previously issued bonds need to be repaid, lump-sum transfers to the households are made and actual government expenditures also include the government investment. The government budget constraint therefore is

$$B_t + \tau_t^k R_t^k v_t K_{t-1} + \tau_t^l w_t L_t + \tau_t^c C_t = R_{t-1} B_{t-1} + G_t + Z_t.$$

Leeper *et al.* (2010) do not assume any forward-looking behavior of the government based on rational expectations. Instead, they use a set of backward-looking rules for the evolution of all variables that the government has under its control – i.e. actual government expenditures and all types of tax rates. Fernández-Villaverde et al. (2011) use similar decision rules for the fiscal variables, but also include persistence terms in each of the equations. The corresponding equations with such extension are

$$G_t = \varphi_g Y_t + \gamma_g B_{t-1} + \varepsilon_t^g, \tag{1}$$

$$\tau_t^k = \varphi_k Y_t + \gamma_k B_{t-1} + \phi_{kc}\varepsilon_t^c + \phi_{kl}\varepsilon_t^l + \varepsilon_t^k, \tag{2a}$$

$$\tau_t^l = \varphi_l Y_t + \gamma_l B_{t-1} + \phi_{lc}\varepsilon_t^c + \phi_{lk}\varepsilon_t^k + \varepsilon_t^l, \tag{2b}$$

$$\tau_t^c = \varphi_c Y_t + \gamma_c B_{t-1} + \phi_{cl}\varepsilon_t^l + \phi_{ck}\varepsilon_t^k + \varepsilon_t^c, \tag{2c}$$

Choice of the $\varphi_i, i = \{g,k,l,c\}$ parameters that measure responses of the fiscal variables on the gross domestic product (equations 1 and 2) can be made so that the government has incorporated an automatic stabilizer (if such response parameters are negative). Unlike in the original paper, I assume that all tax rates depend on the output of the economy.

Solution and estimation

Optimal decision rules for the representative household can be computed by taking first derivatives of the utility function subject to the budget constraint and the law of motion for the evolution of the physical capital. This results in a system of nonlinear equations, some of which are forward-looking in that they include the expectations of future variables. One can further log-linearize these equations around the steady state of the economy via first order Taylor expansion. The same can be done with the representative firm. The rest of the model is already linear. There are standard algorithms that transform the system of linearized equations into backward-looking laws of motion that form the solution of the model – see e.g. Blanchard and Kahn (1980), or Sims (2001), among many others.

Since most of the modeled variables have their reasonable counterparts in real data[2], it is possible to add a block of measurement equations and inspect the solution of the model with help of the Kalman filter, as described in Hamilton (1994). Kalman filtering also provides the value of the logarithm of the likelihood function for a given parameterization of the model, which is useful when estimating the model parameters using Bayesian techniques.

I use the standard methodology along the lines of Lancaster (2008) and Koop (2006) for estimation of the model parameters. Prior distributions of a subset of the model parameters get combined with the likelihood function based on the observed data, which results in the posterior distributions for particular parameters. First, I use numerical optimization techniques to seek for the posterior modes, next, I draw from the posterior distributions with help of Markov chain Monte Carlo simulators around these modes. Prior distributions are adopted from the original paper and a comparison of priors and posteriors can be found in the appendix (figure 2).

Some of the observed variables are known only with annual frequency or not known on the entire history. Kalman filtering creates consistent estimates of all variables on a quarterly basis. Unlike usual interpolation procedures, such as cubic splines, the Kalman filter returns interpolated series which are consistent with the underlying economic model. This way I derive a compact set of time series of the same length, which can then be explored with conventional econometric techniques.

Empirical analysis

Once we process the data through the Kalman filtering routine when working with the structural DSGE model of previous section, we can then create a set of simulated time series of implied consolidated tax rate, aggregate tax revenue of the government, actual government expenditures and

[2] Data treatment is discussed further in a separate section.

total output of the economy[3]. Stacking these variables into a vector $y_t = \{\tau_t, tax_t, G_t, Y_t\}$, we can set up for an unrestricted vector autoregressive model (VAR) of the form

$$y_t = \Gamma_0 + \Gamma_1 y_{t-1} + ... + \Gamma_p y_{t-p} + u_t, \quad (3)$$

where Γ matrices are unknown regression parameters, p is the lag length and u_t is a vector of i.i.d. normally distributed error terms. Ordinary least squares solution (OLS) based on equation (3) gives us the estimates of the parameters, their covariance matrix and the covariance matrix of the residuals, all of which can be used as a prior for the final Bayesian VAR (BVAR) model in which we combine information stemming from the real data and subjective prior assumptions.

Since OLS assumes normality of all parameters and the shocks in all equations must have strictly positive variances, it makes sense to form a Normal-inverse Wishart prior based on OLS results. Wishart distribution is a multivariate generalization of gamma distribution which has a zero bound. Thus, it forms a convenient prior distribution for the error terms. As Zellner (1971) points out, such prior is conjugate because the posterior distributions come from the same family of prior distributions. Furthermore, conditional distribution for the parameters and error covariance matrix can be deduced analytically. Exact formulas can be found in Del Negro and Schorfheide (2004).

To answer the question how shocks in the fiscal area affect the real economy, we can construct the impulse response functions (IRF) based on the estimated BVAR model. Uncertainty about the regression parameters can be assessed with help of the posterior distributions, which results in IRF with appropriate confidence bands. Identification scheme used for the purpose of this paper follows standard Cholesky decomposition with ordering of variables $\tau_t \rightarrow tax_t \rightarrow G_t \rightarrow Y_t$. The response of the economy after a shock to taxation can be found in the appendix, figure 3.

Data

The data sample from 1996Q1 to 2010Q4 includes time series of the Czech Republic on real consumption, real investment, real hours worked, real government spending, real government debt and tax income of the government split into labor tax income, capital tax income and consumption tax income. All real variables are derived as nominal variables deflated by the deflator of the GDP. Most of the fiscal data are not known with quarterly frequency but with annual frequency only. The Kalman filtering setup that solves this data problem is presented in the appendix.

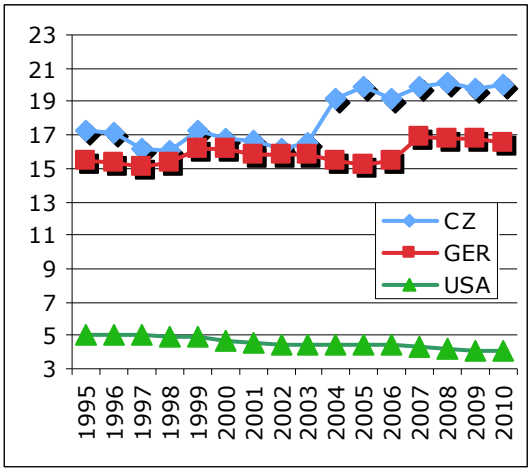
The structural DSGE model uses full set of the data for estimation of the model parameters, while the empirical BVAR model, which takes the outcomes of the DSGE model as its priors, processes only aggregate tax income, real output and real government expenditures.

All variables are detrended with help of Hodrick-Prescott filter and only the business cycle information is retained for the consequent analysis.

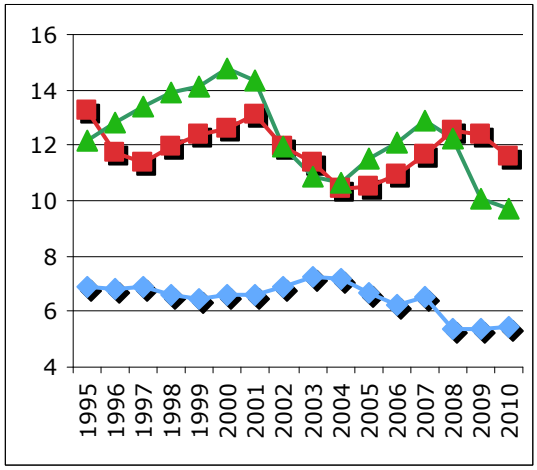
Mendoza (1994) and Jones (2002) propose a way how to compute time series of the effective tax rates that are consistent with the national accounts. The reason for this is that reliable database for the tax policy on the history is not available and the state tax revenues mostly do not coincide with the setting of various tax rates. Mendoza (1994) makes computations based on the OECD database and Jones (2002) uses the NIPA tables for the US economy. They both provide the same results. The OECD database is a very convenient data source because it contains data for all member countries, including the Czech Republic, and allows for an easy cross-country comparison. Figure 1 depicts annual time series for consumption tax rates, labor tax rates and capital tax rates. I also include the household tax for total income, which is used to compute the labor and capital tax rates.

[3] The last three variables are also known from the observed data, however, to impose prior assumptions on the final BVAR model, at this stage of the computation we work with simulated series from the DSGE model.

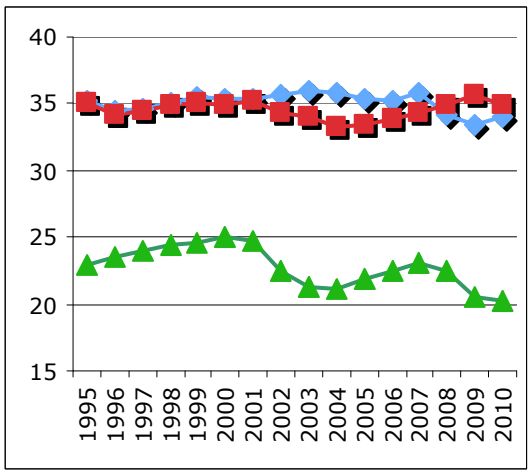
Consumption tax rate



Tax rate on total income



Labor tax rate



Capital tax rate

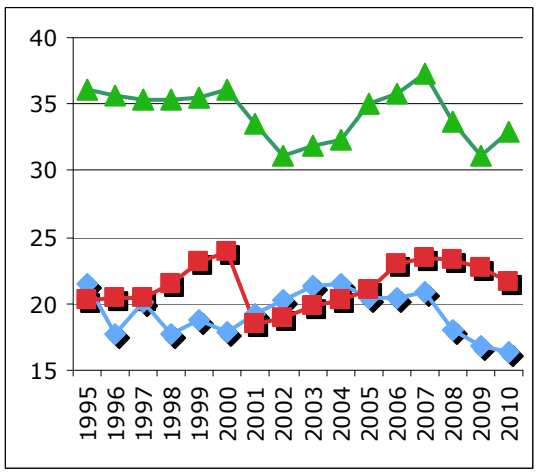


Figure 1: Effective tax rates [Units %] – CZ = Czech Republic, GER = Germany.
Source: OECD database. Methodology: Mendoza (1994).

With the exception of the average tax rate on total income, implied Czech rates resemble those in Germany. Labor income of employees is taxed at a higher rate comparing to the USA, on the other hand implied taxing of the corporate profits is higher in the USA comparing to both Germany and the Czech Republic.

Conclusion

Empirical contradictions in fiscal data make it very difficult for any kind of traditional econometric analysis. The existence of one-time transactions makes the analysis even more challenging and a relevant data adjustment is a necessity. Discrepancies between tax rates, aggregate tax revenues and other government variables were identified in this paper. Calibration of tax rates consistent with the macroeconomic development, as suggested by Mendoza *et al.* (1994) and Jones (2002), was computed for all taxes that are important in the Czech Republic – i.e. value added tax (consumption tax), income tax of employees (labor tax) and corporate profit tax (capital tax).

Annual frequency of most of the relevant fiscal data was then overcome by a model driven interpolation into quarterly figures based on Kalman filtering and the structural DSGE model. In next stage, simulated data were used to construct basic prior assumption on the correlations and cross-

correlations among fiscal variables, following Del Negro and Schorfheide (2004). Final relationships among these variables were studied in a Bayesian manner with help of the VAR model.

It turns out from the impulse response functions, that shocks to the consolidated tax rate initially lower the economic activity. As these shocks fade away, the economic activity increases in line with higher tax revenue of the government and its increased government spending.

References

BLANCHARD, O. and KAHN, C. (1980): The Solution of Linear Difference Models under Rational Expectations, *Econometrica*, Vol. 48, No. 5, 1305-1312.

DEL NEGRO, M. and SCHORFHEIDE, F. (2004): Priors from General Equilibrium Models for VARs, *International Economic Review*, 45(2), 643–73.

FERNÁNDEZ-VILLAVERDE, J., GUERRÓN-QUINNTANA, P., KUESTER, K. and RUBIO-RAMÍREZ, J. (2011): Fiscal Volatility Shocks and Economic Activity, *Working Papers* 11-32, Federal Reserve Bank of Philadelphia.

FUHRER, J. (2000): Optimal Monetary Policy in a Model with Habit Formation, *Computing in Economics and Finance 2000* 306, Society for Computational Economics.

HAMILTON, J. (1994): *Time Series Analysis*, Princeton University Press, Hoboken NJ.

JONES, J. (2002): Has Fiscal Policy Helped Stabilize the Postwar U.S. Economy?, *Journal of Monetary Economics*, Vol. 49, 709-746.

KOOP, G. (2006): *Bayesian Econometrics*, Wiley, West Sussex, UK.

LANCASTER, T. (2008): *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing, Oxford, UK.

LEEPER, E., PLANTE, M. and TRAUM, N. (2010): Dynamics of Fiscal Financing in the United States, *Journal of Econometrics*, Vol. 156, 304-321.

MENDOZA, E., RAZIN, A. and TESAR, L. (1994): Effective Tax Rates in Macroeconomics: Cross-Country Estimates of Tax Rates on Factor Incomes and Consumption, *NBER Working Paper Series*, 4864, National Bureau of Economic Research.

SCHMITT-GROHE, S. and URIBE, M. (2008): What's news in business cycles? *NBER Working Paper Series*, 14215, National Bureau of Economic Research.

SIMS, C. (2001): Solving Linear Rational Expectations Models, *Journal of Computational Economics*, 20, 1_20.

ZELLNER, A. (1971): *An Introduction to Bayesian Inference in Econometrics*, John Wiley & Sons, New York.

JEL classification: C11, C63, H30, H50.

Appendix

Adjusted Kalman filter

Usual Kalman filtering problem can be defined in the state space as

$$\begin{aligned} x_t &= F(\phi)x_{t-1} + G(\phi)\varepsilon_t, \;\; \varepsilon_t \sim N(0, \mathrm{E}), \\ y_t &= A(\phi)x_t \;\; + \omega_t, \qquad \omega_t \sim N(0, \Omega), \end{aligned} \tag{4}$$

where x_t is the vector of state variables, y_t is the vector of observed data. All model matrices depend on the parameterization ϕ. While structural shocks, ε_t, are usually contemporaneously correlated, the measurement errors, ω_t, are independent of each other.

Since we work with fiscal data that are known only with annual frequency, we need to introduce some unobserved variables that will be used to estimate the quarterly figures of appropriate series. Then we linearly map these unobserved variables on the data we have in hand. Of course, we need to distinguish between stock and flow variables:

Stocks: $$y_t = \frac{1}{4}\left(x_t + x_{t-1} + x_{t-2} + x_{t-3}\right) + \omega_t,$$

Flows: $$y_t = x_t + x_{t-1} + x_{t-2} + x_{t-3} + \omega_t,$$

This is not enough for the Kalman filter to work properly because the structure of the state space evolves always one period ahead. Hence, if we move, say, to a second quarter of a year, such aggregation of quarterly estimates would be mapped on an annual figure that we do not have for the second quarter.

For that matter, we can adjust the standard Kalman filter algorithm to account for time-varying measurement errors so that the matrix Ω in equation (4) now becomes time dependent, or Ω_t. If we are in a period of aggregation we impose some justifiable measurement error, if we move away from that period, we impose a sufficiently large measurement error so that the Kalman filter computations are not influenced by the measurement equation:

$$y_t = \frac{1}{4}\left(x_t + x_{t-1} + x_{t-2} + x_{t-3}\right) + \omega_t \quad \begin{cases} std(\omega_t) = \infty, \; t = within\, year, \\ std(\omega_t) << \infty, \; t = end\, of \; year. \end{cases}$$

Priors and posteriors

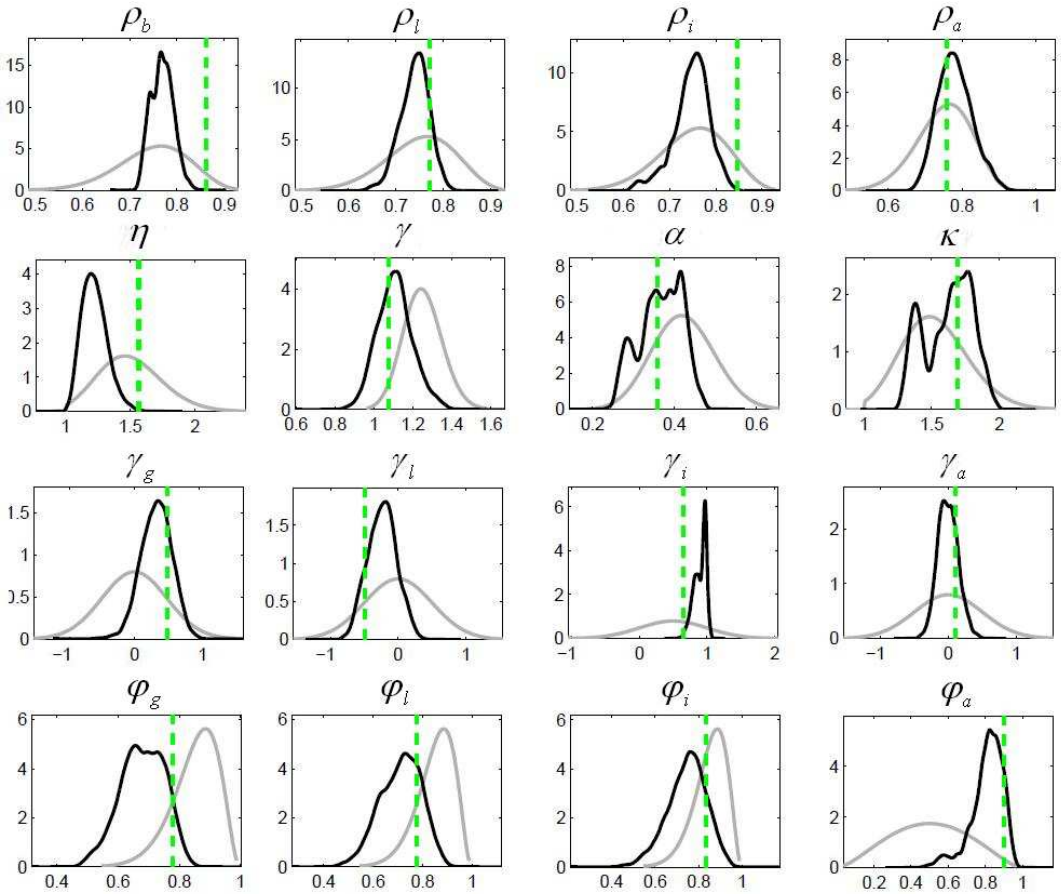


Figure 2: Prior and posterior distributions of estimated parameters from the DSGE model (grey = priors, black = posteriors, green = initial mode values for posterior sampling to start from).

Impulse response functions

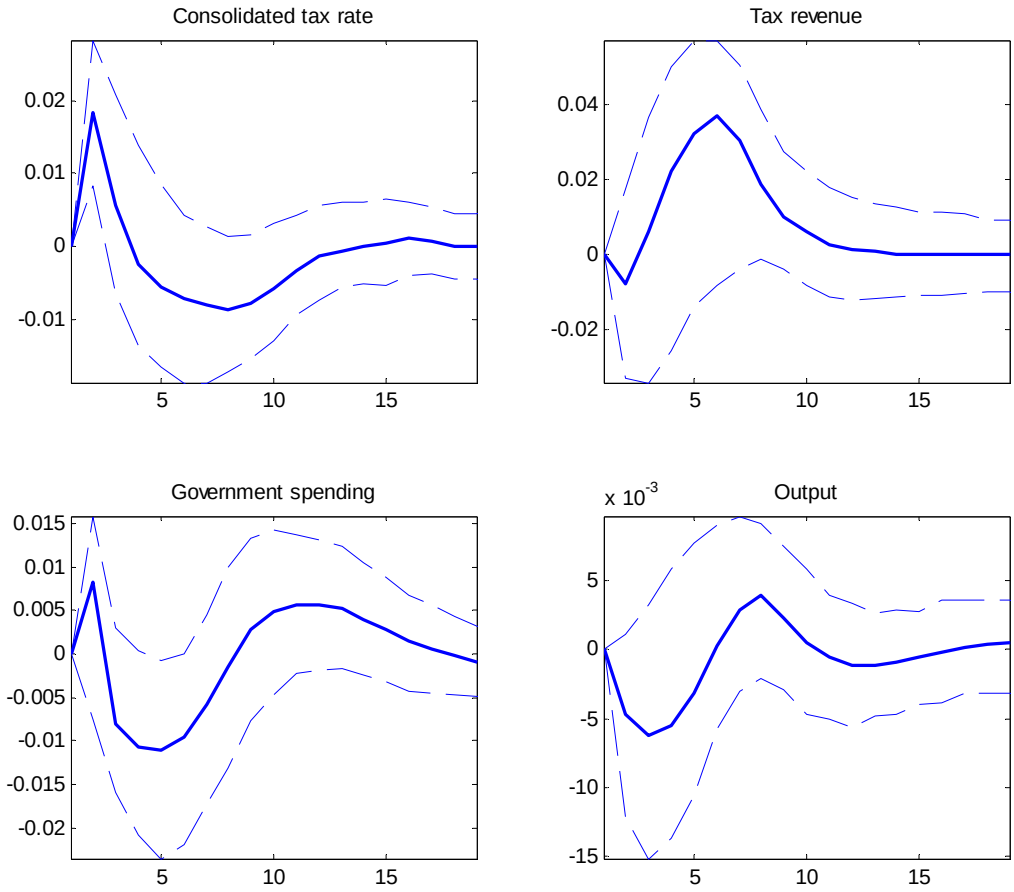


Figure 3: Response of the real variables after a positive shock to the consolidated tax rate.

Shrnutí

Zpracování datové nejistoty ve fiskální oblasti (přístup DSGE-VAR)

Fiskální data v konvergujících, ale i rozvinutých, ekonomikách neposkytují informace přímo použitelné v rámci makroekonomické analýzy. Fiskální veličiny, které má vláda pod kontrolou, např. daňové sazby, zpravidla nejsou v souladu s ostatními fiskálními veličinami, např. daňovými výnosy. Díky těmto nesrovnalostem jsou výsledky jakékoliv empirické studie založené na čisté analýze dat velmi neintuitivní. K zodpovězení otázky, jak reaguje hrubý domácí produkt z hlediska načasování a amplitudy na šoky ve fiskální oblasti, pracuji s vektorovým autoregresním modelem (VAR), který obsahuje blok základních domácích veličin rozšířený o fiskální veličiny. Za účelem minimalizace efektů datové nejistoty ve fiskální oblasti na konečné výsledky zavádím bayesovským stylem apriorní omezení na modelové parametry. Základem pro apriorní nastavení parametrů je výstup ze strukturálního modelu všeobecné rovnováhy (DSGE), který navrhují Leeper *a kol.* (2010), což reflektuje DSGE-VAR postup navržený autory Del Negro a Schorfheide (2004). Empirické výsledky jsou oproti čistému VAR přístupu mnohem realističtější. Pozitivní šoky do konsolidované míry zdanění nejprve snižují ekonomickou aktivitu, která postupně nabírá tempo s tím, jak roste daňový výnos státu, a tedy i vládní výdaje. Mimo to jako vedlejší produkt konstruuji odhadované čtvrtletní časové řady pro ty fiskální veličiny, které jsou oficiálně publikované pouze s roční frekvencí.

Klíčová slova: vektorový autoregresní model (VAR), DSGE model, fiskální politika, bayesovská analýza.

Kvantitativní dopady úrokových sazeb

Michal Řičař
michal.ricar@vse.cz
Doktorand oboru ekonometrie a operační výzkum
Školitel: prof. Ing. Roman Hušek, CSc. (husek@vse.cz)

Abstrakt: Článek se zaměřuje na zkoumání, zdali klíčový měnově-politický nástroj ČNB v podobě 14denní repo sazby fakticky působí na klíčové makroekonomické komponenty - nezaměstnanost, HDP, spotřebu, investice -, jednotlivé sektory ekonomiky a v neposlední řadě aktiva v podobě výstavby nemovitostí, akciového indexu PX či poskytnutých úvěrů a míry defaultů. Tržní distorze na trhu aktiv jasně prokázaly, že jsou schopny značně postihnout reálnou ekonomiku, taktéž finančně-ekonomická krize 2008 a násl. poukázala, že transmisní mechanismus byl značně oslaben a jeho principy nejsou plně rozpoznány a možné defekty zřetelně diagnostikovány. Současný výzkum by se tak měl orientovat na podrobné zkoumání vztahů mezi úrokovou mírou CB a možnými faktory finančních a hospodářských nepokojů a nikoli primárně pouze na veličiny, jež mohou působit na cenovou hladinu.

Moderní ekonometrické techniky VAR poskytují kvalitní a snadno interpretovatelnou platformu, na jejímž základě je možno odhalit základní vazby mezi zkoumanými proměnnými a tudíž poukázat na další možný směr výzkumu. Tato práce jich využívá a doplňuje o základní statistické nástroje k determinaci příčinných vazeb, jejich následností.

Provedený výzkum poukázal na skutečnost, že změna úrokové sazby v malé otevřené ekonomice není schopna adekvátně zajistit finanční stabilitu, neboť její vliv je na aktiva nejednoznačný, a to i díky faktu, že ČR je stále konvergující ekonomikou. Jednoznačnou výzvou do budoucna tak je dále prohlubovat znalosti o transmisním mechanismu a zejména relevantně nastavit regulační rámec CB k zajištění nejen cenové, ale i finanční stability s podporou stability hospodářské.

Klíčová slova: úroková sazba, monetární politika, cenová stabilita, finanční stabilita, VAR.

Úvod

Úrokové sazby v tržní ekonomice jsou klíčovým měnově-politickým nástrojem centrální banky, jež aplikuje režim cílování inflace, k zajištění cenové stability se snahou o stabilitu makroekonomickou. Krize posledních let však prokázala, že orientace měnového tvůrce primárně na cenovou stabilitu, která je zpravidla aproximováno spotřebitelským indexem, jenž opomíjí vývoj cen aktiv - zejména nemovitostí, akciových trhů, komodit ad. -, není postačující podmínkou k zajištění stability finanční a hospodářské. Současná odborná diskuse v kontextu režimu cílování inflace se zaměřuje zejména na otázky spjaté s preemptivním působením centrální banky při nadměrném růstu cen aktiv či úvěrové expanze v případě, kdy není dle prognostického modelu CB přímo ohrožen inflační cíl. Poslední finančně-ekonomická krize jednoznačně vyvrátila názorový proud, jenž akcentoval cenovou stabilitu jakožto postačující podmínku stability finanční. Nutnost detailnějšího studia vývoje aktiv a nezbytnost implementace determinace fundamentálních cen, resp. signifikantních odchylek od těchto cen, se jeví nanejvýš žádoucí, neboť bubliny na těchto trzích značně ovlivnily jak finanční, tak reálnou stabilitu. Jako eminentní se tak jeví zjistit, do jaké míry je měnově-politický nástroj CB v podobě úrokové míry schopen ovlivňovat jak vývoj na trzích aktiv, tak reálnou ekonomiku. V případě, že nelze empiricky podložit jednoznačný vliv, lze vyřknout hypotézu, že rozšíření působnosti CB tak, aby byla schopna zajistit stabilnější finanční a makroekonomické prostředí, přičemž by se minimalizovala nadměrná kulminace a koncentrace rizik, by bylo vhodné. K tomuto účelu by však bylo pravděpodobně nezbytné využít regulační nástroje. S ohledem k současnému vývoji v Evropské unii je aktuálnost změn v dohledu a regulaci značná, přičemž pro relevantní výsledky a plynoucí užitek v konečném důsledku pro občany je zajisté nezbytné jednak synchronizovat implementační rámec nových pravidel, jednak jej optimalizovat tak, aby byl maximálně účinný napříč členskými státy.

Následující text nejprve stručně charakterizuje režim cílování inflace s diskusí nad současnou úlohou CB, poté je představena užitá ekonometrická teorie s empirickou metodikou postupu a s charakteristikou užitých dat, na jejichž základě jsou učiněny závěry o úrokové míře ČNB a jejího vlivu na zkou-

mané ekonomické veličiny a tedy i o schopnosti současného nastavení monetární politiky působit na finanční a hospodářskou stabilitu.

Režim cílování inflace

Navzdory faktu, že z didaktických důvodů mnohdy ekonomická literatura nabízí deskripci vzniku peněz přes peněžní multiplikátor, lze úlohu centrální banky v režimu cílování inflace omezit na nastavování klíčového měnově-politického nástroje v podobě definované, zpravidla krátkodobé, úrokové míry. Tato politika s sebou přináší tzv. endogenitu peněz, neboť CB může pevně stanovovat buďto úrokovou míru peněžního trhu nebo přímo množství peněz v ekonomice. Jestliže režim cílování inflace volí prvou alternativu, je poptávka komerčních bank při zvolené ceně peněz (úrokové míře) vždy uspokojena ze strany CB, čímž je kvantitativně determinována i cenová hladina v ekonomice. Výše cenové hladiny je však ovlivněna s určitým zpožděním, jež vyplývá z cenových strnulostí, bez nichž by nebylo možno provádět jakoukoli měnovou politiku, neboť by se změna nominální úrokové míry okamžitě plně reflektovala do cen, a tedy reálné veličiny - na jejichž základě se spotřebitelé a investoři rozhodují - by se nezměnily. Z vysledovaných vazeb v rámci transmisního mechanismu se předpokládá, že změna v úrokové míře se promítne do reálné ekonomiky se zpožděním od 3 měsíců do 1 roku dle toho, zdali se jedná o spotřebitelský kanál, jenž je změnou ovlivněn, či podnikatelský kanál (Jílek, 2004).

V případě ČNB je za měnově-politický nástroj zvolena 14denní repo sazba, čímž je nastavena krátkodobá úroková sazba. Dlouhodobé sazby, od nichž se odvíjejí ceny pro subjekty ekonomiky mimo finanční sektor (hypotéční, spotřebitelské, podnikatelské aj. úvěry, vnitřní výnosové procento dluhopisů, akcií atd.), se formují na základě očekávání účastníků finančního trhu o budoucím vývoji ekonomiky, resp. rozhodování CB. Nově nastavená cena peněz ze strany CB taktéž způsobuje mezinárodní úrokový diferenciál, čímž je ovlivněn měnový kurz; v opačném případě by bylo možno využít arbitráže.

Výše uvedené naznačuje určitou posloupnost fungování měnové politiky, jež se označuje jako transmisní mechanismus, s jehož využitím se CB snaží ovlivňovat inflaci, resp. zajistit hladký hospodářský vývoj. Transmisní mechanismus je v zásadě rozčleněn dle subjektů ekonomiky, tedy jak již bylo uvedeno kanál spotřebitelský, podnikatelský, měnového kurzu ad. Režim cílování inflace je specifický v tom, že se jedná o tzv. vpřed-hledící měnovou politiku s vyhlášeným cílem inflace ve střednědobém horizontu (u ČNB se jedná zpravidla o 1 rok), tedy skutečnost, kdy CB cíluje vyhlášenou inflační prognózu pro daný časový horizont, již získává na základě vlastního ekonometrického modelu. V současné době jsou využívány dynamické stochastické modely všeobecné rovnováhy, jejichž výhodou oproti historickým modelům je obohacení o mikroekonomické základy (např. racionální očekávání apod.). Mimořádně důležitá je tak důvěryhodnost prognóz, neboť dle nich se rozhodují ekonomické subjekty. Tyto modely však zpravidla abstrahují od fungování finančních trhů, neboť se předpokládá jejich bezproblémová funkce, taktéž nemají implementovány determinanty fundamentálních cen aktiv (nemovitosti, akciový trh, komodity apod.) či vývoj měnových agregátů, které jsou sice implicitně inkorporovány v nastavené úrokové sazbě, avšak konkrétní výše není v tomto režimu explicitně zahrnuta do prognóz.

Finančně-ekonomická krize z let 2008, 2009 poukázala na jednoznačnou slabinu současných modelů ekonomiky využívaných CB, neboť právě finanční sektor stál za šokem, jenž se postupně přelil z finanční oblasti do reálné ekonomiky. Nelze tak tvrdit, že cenová stability s předvídatelnou měnovou politikou je postačující podmínkou pro stabilitu finanční a hospodářskou.

Současný režim cílování inflace tedy úvěry, ceny aktiv a peníze uvažuje pouze do té míry, do níž nepřímo ovlivňují inflační prognózu. Dle Zamrazilové (2011) nelze polemizovat o tom, že finanční nestabilita je peněžním jevem. Vzhledem k tomu, že expanze úvěrů je silně spjata s vývojem cen aktiv, měly by tyto veličiny být pečlivě sledovány CB, neboť bubliny na realitním trhu za současného využití finančního inženýrství mohou eskalovat do značných finančně-hospodářských nestabilit, jak bylo předesláno. Nicméně shoda mezi odbornou veřejností nadále panuje v tom, že primárním cílem CB musí být cenová stabilita, široká diskuse je vedena zejména nad regulačním rámcem k tomu, aby byla lépe ukotvena finanční, resp. hospodářská stabilita na základě činnosti CB. Hlavními argumenty proti využívání měnové politiky jakožto prevence krizí jsou zejména rizika spjatá se zhoršením kontroly inflace

a rozkolísáním produktu, přičemž rozpoznání, zdali se jedná o nadměrný růst cen aktiv či úvěrovou expanzi, resp. nepřiměřený vývoj akciového trhu, je dle Bernanke (2010) možné až ex post, k čemuž se přiklánějí i další (Gerlach, 2010, Greenspan, 2010). Na druhé straně však volná měnová politika FED pravděpodobně přispěla k finančně-ekonomické krizi 2008, 2009, neboť jak ukazuje Taylor (2008), kdyby se bývala CB (FED) držela empiricky vysledovaného pravidla (tzv. Taylorova pravidla) a nereagovala na probíhající ekonomický útlum z tzv. Dot-com-bubble, bublina na realitním, resp. úvěrovém trhu by se nevytvořila. S uvedeným souvisí tzv. risk-taking channel měnové transmise, kdy ekonomické subjekty akceptují vyšší riziko, a tedy uvolňují úvěrové standardy, z důvodu dlouhodobé nízké úrokové sazby, která může přispět k finančním a hospodářským nestabilitám (Zamrazilová, 2011).

Včasné rozpoznání bubliny např. na trhu nemovitostí je však velice obtížnou záležitostí, neboť cena se může lišit od kvantifikované ceny fundamentální, za níž stojí zejména růst peněžní zásoby, růst úvěrů použitelných pro financování bydlení, růst objemu stavební výroby, vývoj úrokových sazeb, demografické faktory, velikost samotného trhu nemovitostí a omezenost nabídky nemovitostí v dané lokalitě (Komárek, 2011), avšak nemusí nutně implikovat vznik bubliny s následnou rezonancí do ekonomiky. Z toho důvodu Komárek (2011) doporučuje sledovat, zdali se nerovnovážný vývoj projevuje i na ostatních trzích aktiv (měnový kurz, akcie ad.), přičemž akcentuje, že výhodné je sledování úvěrové aktivity, měnových agregátů, vnímání rizika a formování rizikové prémie finančními trhy, neboť bubliny vznikají v komplexním ekonomickém prostředí.

Ekonometrický model

Práce využívá VAR modelů, resp. Grangerovy kauzality, jež je bezprostředně s těmito technikami spjata. Pro soubor *n* časových řad proměnných $y_t = (y_{1t}, y_{2t,}..., y_{nt})'$ lze VAR model řádu *p* VAR(*p*) zapsat jako

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + ... + A_p y_{t-p} + u_t,$$

kde

A_i – je matice koeficientů o rozměru *n*x*n*,

$u_t = (u_{1t}, u_{2t}, ..., u_{nt})'$ – je nepozorovatelný i.i.d. proces náhodných složek.

Uvažujme jednorozměrný model VAR(1) o dvou proměnných, jejž je možno vyjádřit v podobě

$$y_t = b_{10} - b_{12} z_t + c_{11} y_{t-1} + c_{12} z_{t-1} + \varepsilon_{yt},$$

$$z_t = b_{20} - b_{21} y_t + c_{21} y_{t-1} + c_{22} z_{t-1} + \varepsilon_{zt},$$

při $\varepsilon_{it} \sim i.i.d(0, \sigma_{\varepsilon i}^2)$ a $\text{cov}(\varepsilon_y, \varepsilon_z) = 0$.

V maticovém vyjádření pak strukturní tvar zapíšeme

$$\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix} \begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix},$$

resp.

$$BX_t = \Gamma_0 + \Gamma_1 X_{t-1} + \varepsilon_t.$$

K tomu, aby bylo možno vhodně pracovat s VAR systémem, je výhodné přejít na neomezený redukovaný tvar, který v roli regresorů obsahuje pouze predeterminované, resp. endogenní zpožděné proměnné (Hušek, 2007):

$$B^{-1}BX_t = B^{-1}\Gamma_0 + B^{-1}\Gamma_1 X_{t-1} + B^{-1}\varepsilon_t,$$

tedy

$$X_t = A_0 + A_1 X_{t-1} + e_t,$$

kde

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = \frac{1}{(1-b_{21}b_{12})} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{bmatrix},$$

přičemž ε je i.i.d., čímž *e* má charakteristiky $(0, \sigma_i^2)$ a $E(e_{it}) = 0$.

Rozptyl náhodných složek

$$Var(e_{1t}) = E(e_{1t}^2) = \frac{E(\varepsilon_{yt}^2 + b_{12}^2 \varepsilon_{zt}^2)}{\Delta^2} = \frac{\sigma_y^2 + b_{12}^2 \sigma_z^2}{\Delta^2}$$

je v čase nezávislý, totožný závěr můžeme učinit pro rozptyl e_{2t}.

Kovarianci náhodných složek neomezeného redukovaného tvaru lze vyjádřit

$$COV(e_{1t}, e_{2t}) = E(e_{1t} e_{2t}) = \frac{E[(\varepsilon_{yt} - b_{12}\varepsilon_{zt})(\varepsilon_{zt} - b_{21}\varepsilon_{yt})]}{\Delta^2} = \frac{-(b_{12}\sigma_z^2 + b_{21}\sigma_y^2)}{\Delta^2} \neq 0,$$

čímž je zřejmé, že náhodné složky jsou sériově zkorelované, a tedy šoky ve standardním VAR systému jsou korelované. Z toho důvodu se pro účely IR analýzy využívá VAR model v ortogonalizované podobě, podrobněji viz Hušek (2007), Arlt (2009).

Grangerova kauzalita

Výhodou analýzy více časových řad je mimo jiné to, že je možno zkoumat jejich vzájemné vztahy. Granger[1] definoval koncept kauzality s využitím VAR modelů. Vycházel z teze, že pokud řada $\{\mathbf{X}_t\}$ ovlivňuje řadu $\{\mathbf{Y}_t\}$, pak by řada $\{\mathbf{X}_t\}$ měla zlepšit prognózu $\{\mathbf{Y}_t\}$, *vice versa.*

Nechť $\{\mathbf{X}_t\}$ je *m*-rozměrný náhodný proces a $\{\mathbf{Y}_t\}$ je *n*-rozměrný náhodný proces. Označme množinu, jež obsahuje veškeré dostupné informace v čase *t*, $\mathbf{\Omega}_t$. Dále uvažujme, že $\mathbf{\Omega}_t \backslash \{\mathbf{X}_{t-s}, s \geq 0\}$ znamená všechny dostupné informace s výjimkou informací obsažených v minulosti a přítomnosti procesu $\{\mathbf{X}_t\}$. Nechť $\mathbf{Y}_t(h \backslash \mathbf{\Omega}_t)$ je předpověď procesu $\{\mathbf{Y}_t\}$ konstruovaná v čase t, založená na informacích v množině $\mathbf{\Omega}_t$. Proces $\{\mathbf{X}_t\}$ *kauzálně působí v Grangerově smyslu* (dále jen „*kauzálně působí*“) na proces $\{\mathbf{Y}_t\}$, pokud pro minimálně jeden z horizontů předpovědi *h* = 1,2,… je rozdíl

$$\mathrm{MSE}[\mathbf{Y}_t(h \backslash \mathbf{\Omega}_t \backslash \{\mathbf{X}_{t-s}, s \geq 0\}] - \mathrm{MSE}[\mathbf{Y}_t(h \backslash \mathbf{\Omega}_t)]$$

pozitivně definitní matice. Pokud proces $\{\mathbf{X}_t\}$ kauzálně působí na proces $\{\mathbf{Y}_t\}$ a proces $\{\mathbf{Y}_t\}$ kauzálně působí na proces $\{\mathbf{X}_t\}$, potom se proces $\{\mathbf{Z}_t\} = (\{\mathbf{X}_t\}', \{\mathbf{Y}_t\}')'$ označuje jako *systém zpětné vazby*. Obdobné závěry lze učinit pro jednorozměrné procesy. V praxi je velmi častým problémem dostupnost veškerých informací informační množiny $\mathbf{\Omega}_t$, z toho důvodu se mnohdy omezujeme na informace dané současností a minulostí procesu $\{\mathbf{Z}_t\} = (\{\mathbf{X}_t\}', \{\mathbf{Y}_t\}')'$ a v převážné většině případů uvažujeme lineární předpovědi.

Hušek (2007) na straně 245 uvádí, že není možno kauzalitu v Grangerově smyslu považovat za příčinnou závislost, neboť se jedná o testování, zda změna jedné proměnné předcházení změně jiné proměnné (Řičař, 2010).

Užitá data a metodika postupu

Veškerá data byla obdržena z Českého statistického úřadu a ČNB za období IQ 1996 - IIIQ 2011 (63 pozorování) a následně převedena na přírůstky za užití logaritmizace. Ve všech testech, resp. modelech byla vždy testována úroková sazba ČNB v podobě 14denní repo sazby (R) s danou veličinou, a to s úvěry domácnostem (UVER_DOM), úvěry nefinančním podnikům (UVER_POD), úvěry finančním podnikům (UVER_FIN_POD), podíl míry defaultů na poskytnutých úvěrech (UVER_DEFAULT),

[1] Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. 1996, Econometrica 37(3), 424-438. Dostupné z www: http://www.econometricsociety.org/abstract.asp?ref=0012-9682&vid=37&iid=3&aid=0012-9682(196908)37:3%26lt;424:ICRBEM%26gt;2.0.CO;2-L&s=-9999

nová výstavba nemovitostí (BYTY), index PX (PX), nezaměstnanost (U), hrubý domácí produkt (HDP), spotřeba domácností (C), investice (I) a přidané hodnoty sektorů dle ESA95 a členěné dle NACE- Zemědělství, lesnictví a rybářství (ZEM), Průmysl, těžba a dobývání (TEZ), Zpracovatelský průmysl (ZPR), Stavebnictví (STA), Obchod, doprava, ubytování a pohostinství (OBCH), Informační a komunikační činnosti (INF), Peněžnictví a pojišťovnictví (PEN), Činnosti v oblasti nemovitostí (NEM), Profesní, vědecké, technické a administrativní činnosti (VED), Veřejná správa a obrana, vzdělávání, zdravotní a sociální péče (VER).

S využitím programu Eviews 6.0. byla pro každou časovou řadu nejprve zkoumána stacionarita za užití Dickey-Fullerova testu, Phillips-Perronova testu a KPSS testu. Vzhledem ke skutečnosti, že klíčová proměnná R byla detekována jako stacionární, nebylo nutno testovat kointegrační vektory, resp. časové řady, jež byly identifikovány jako nestacionární, byly první diferencí převedeny na řady stacionární, jejichž následnou stacionaritu potvrdily uvedené testy. Dále bylo přistoupeno ke korelačním maticím s proměnným obdobím za účelem analyzovat, jakým způsobem se vyvíjí vzájemný vztah v časové evoluci. Následně byly sestavovány VAR modely, a to vždy s R a danou proměnnou. V prvém kroku byla testována vhodná délka VAR systému za užití AIC, SQ a HQ informačních kritérií. Po determinaci optimální délky byl proveden test Grangerovy kauzality, resp. Waldův test blokové exogenity, díky čemuž byla obdržena informace o vzájemném příčinném vztahu proměnných v systému, tedy spíše charakter endogenní či exogenní, jenž dobře slouží k představě o následnostech v ekonomickém celku. V neposlední řadě byl VAR systém podroben analýza výkonnosti z hlediska zkoumané proměnné v kontextu úrokové sazby k určení, zda je R vhodným predikátorem budoucího vývoje, resp. zda významněji působí na generování časové řady, tedy jaká je intenzita vlivu.

Hlavní statistiky jednotlivých VAR modelů jsou uvedeny v následujícím textu.

Kvantitativní dopady úrokové míry ČNB

Korelační analýzu s proměnným obdobím je možno nalézt v tabulce 1 (symbol "*" označuje signifikantní vztah na 5% hladině významnosti), na jejímž základě je možno učinit závěr, že veličiny v období 1996 - 2011 mají odlišné charakteristiky než v letech 2004 - 2011, resp. 2006 - 2011. Zatímco tedy mezi obdobím 2004 - 2011 a 2006 - 2011 nebyly nalezeny signifikantní rozdíly, analýza korelace 1996 - 2011 vykazuje významně odlišné výsledky.

Tabulka 1: Korelace proměnných s R v čase Zdroj: ČSÚ, ČNB, 2011

Proměnná	Korelace s R 1996Q1-2011Q3	Korelace s R 2004Q1-2011Q3	Korelace s R 2006Q1-2011Q3
UVER_DOM	0,05	0,11	0,15
UVER_POD	0,33*	0,57*	0,74*
UVER_FIN_POD	0,09	0,23	0,35
UVER_DEFAULT	-0,29	0,10	0,10
BYTY	0,15	0,16	0,19
PX	0,15	0,32	0,35
U	-0,27*	-0,30	-0,38
HDP	0,42*	0,53*	0,55*
C	0,24*	0,42*	0,49*
I	0,15	0,29	0,31
ZEM	0,09	0,17	0,22
TEZ	0,19	0,37*	0,45*
ZPR	0,16	0,41*	0,50*
STA	0,15	0,11	0,10
OBCH	0,29*	0,31	0,29
INF	0,14	0,13	0,40*
PEN	0,11	-0,02	-0,04
NEM	0,02	0,09	0,13
VED	0,11	0,40*	0,48*
VER	0,06	-0,08	-0,10

Pro komplexní pohled na evoluci vlivu úrokové míry na danou veličinu jsou uvedeny výsledky VAR a Grangerovy kauzality i za 1996 - 2011, jež shrnuje tabulka 2, neboť z ekonomické teorie i praxe je patrné, že režim cílování inflace je zaměřen na horizont delší nežli 3 měsíce, a tedy lze předpokládat, že VAR systém může detekovat významné vazby až k t(-4), k čemuž je nutno disponovat dostatečně dlouhou časovou řadou.

Hodnoty v tabulkách u Grangerovy kauzalita udávají hladinu významnosti, s níž je možno předpokládat, že zpožděná proměnná předchází vysvětlované proměnné, a to nejprve zdali úroková míra působí v Grangerově smyslu na danou proměnnou a poté inverzně. Následující sloupce jsou věnovány VAR systémům, kdy první sloupec označuje determinované nejvhodnější zpoždění vzhledem k informačním kriteriím (LAG), druhý sloupec vyhodnocuje koeficient determinace, jenž kvantifikuje schopnost VAR modelu vysvětlit fluktuace zkoumané proměnné (R^2), zbylé čtyři sloupce udávají hodnoty koeficientů pro úrokovou míru v daném zpoždění (R(-1), R(-2)) a zkoumané veličiny (X(-1), X(-2)), kde symbol "*" značí signifikantní koeficient na 5% hladině významnosti.

Tabulka 2: Grangerova kauzalita a VAR modely 1996Q1-2011Q3 Zdroj: Vlastní modelace

Proměnná X 1996Q1-2011Q3 (63)	Grangerova kauzalita		VAR					
	R → X	X → R	LAG	R^2	R(-1)	R(-2)	X(-1)	X(-2)
UVER_DOM	0,84	0,91	1	0,42	-0,01	-	-0,64*	-
UVER_POD	0,83	0,42	1	0,22	0,02	-	0,44*	-
UVER_FIN_POD	0,42	0,01	2	0,10	0,18	0,01	-0,11	-0,33*
UVER_DEFAULT	0,06	0,07	1	0,31	-0,18*	-	0,39*	-
BYTY	0,14	0,10	2	0,43	0,38*	-0,19	-0,47*	-0,62*
PX	0,07	0,31	2	0,21	0,03	-0,25*	0,35*	-0,09
U	0,26	0,07	2	0,34	-0,11	0,00	0,10	0,44*
HDP	0,08	0,27	1	0,17	0,03*	-	0,11	-
C	0,26	0,86	2	0,21	0,01	0,01	-0,09	0,33*
I	0,15	0,04	2	0,10	0,00	0,09*	-0,02	0,17
ZEM	0,03	0,34	3	0,44	0,35*	0,22	-0,73*	-0,69*
TEZ	0,00	0,23	2	0,25	0,07*	0,05	-0,43*	-0,21
ZPR	0,01	0,09	1	0,16	0,09*	-	-0,31*	-
STA	0,00	0,27	2	0,26	-0,04	0,13*	-0,29*	-0,06
OBCH	0,10	0,89	1	0,05	0,05	-	-0,13	-
INF	0,01	0,22	1	0,30	0,06*	-	0,38*	-
PEN	0,74	0,42	1	0,09	-0,04	-	-0,29*	-
NEM	0,34	0,03	2	0,04	0,05	0,00	-0,05	0,05
VED	0,09	0,32	2	0,11	0,02	0,07*	-0,20	-0,02
VER	0,16	0,53	1	0,04	0,02	-	-0,07	-

Srovnáním výsledků analýz za období 2004 - 2011 a 2006 - 2011 lze za relevantní výsledky považovat analýzy za 2004 - 2011, neboť oproti 2006 - 2011 nebyly nalezeny významné odchylky ve VAR systémech či Grangerově kauzalitě s výhodou většího počtu pozorování. Tabulka 3 shrnuje výsledky analýzy pro období 2004 - 2011.

Tabulka 3: Grangerova kauzalita a VAR modely 2004Q1-2011Q3 Zdroj: Vlastní modelace

Proměnná 2004Q1-2011Q3 (31)	Grangerova kauzalita		VAR					
	R → X	X → R	LAG	R^2	R(-1)	R(-2)	X(-1)	X(-2)
UVER_DOM	0,46	0,79	2	0,62	0,03	-0,01	0,13	0,65*
UVER_POD	0,09	0,62	1	0,28	0,07	-	0,26	-
UVER_FIN_POD	0,02	0,03	1	0,16	0,17*	-	0,01	-
UVER_DEFAULT	0,31	0,00	2	0,65	-0,12	0,04	0,20	0,52*
BYTY	0,09	0,47	2	0,56	0,41*	-0,07	-0,41*	-0,75*
PX	0,76	0,63	1	0,17	-0,05	-	0,44*	-
U	0,00	0,00	1	0,39	-0,34*	-	0,08	-

Proměnná 2004Q1-2011Q3 (31)	Grangerova kauzalita		VAR					
	R → X	X → R	LAG	R^2	R(-1)	R(-2)	X(-1)	X(-2)
HDP	0,01	0,14	1	0,25	0,05*	-	0,03	-
C	0,07	0,97	2	0,38	0,00	0,03*	-0,05	0,31
I	0,05	0,09	2	0,25	-0,02	0,17*	-0,04	0,16
ZEM	0,59	0,50	2	0,27	0,28	0,01	-0,53*	-0,43*
TEZ	0,03	0,67	2	0,25	0,08	0,08	-0,44*	-0,26
ZPR	0,15	0,53	1	0,08	0,07	-	-0,19	-
STA	0,20	0,60	2	0,16	-0,02	0,10	-0,27	-0,02
OBCH	0,00	0,63	1	0,32	0,11*	-	-0,33*	-
INF	0,18	0,99	1	0,14	0,04	-	0,25	-
PEN	0,55	0,10	1	0,08	0,05	-	-0,25	-
NEM	0,03	0,70	1	0,16	0,09*	-	-0,14	-
VED	0,10	0,93	2	0,16	0,03	0,10	-0,17	-0,26
VER	0,30	0,40	2	0,15	0,02	0,02	-0,05	-0,25

Zkoumané veličiny lze rozčlenit do 3 pomyslných kategorií, a to (1) kategorie aktiv, jež se skládají z finančních aktiv, kam spadají veškeré úvěry, dále z nemovitostí a indexu PX, za (2) druhou kategorii lze považovat základní makroekonomické veličiny, tedy nezaměstnanost, HPD, spotřebu a investice a (3) jako poslední kategorii je možno označit výkonnost jednotlivých odvětví.

Prvá kategorie poukazuje posilující vliv úrokové míry v čase na jednotlivé komponenty této kategorie. Radikální vývoj je možno zaznamenat zejména u poskytnutých úvěrů podnikům, kde se korelační koeficient v čase významně mění a je vždy signifikantní. Uvedené implikuje, že kvartální změna v úrokové míře se v současnosti reflektuje do změny ve výši podnikových úvěrů v intenzívnější formě nežli tomu bylo v minulosti. Lze tedy usuzovat, že transformační proces ekonomického prostředí ČR k západním ekonomikám s sebou přináší i podstatný vliv úrokové míry, a to i díky faktu, že ČNB během tohoto období optimalizovala komunikaci s finančním trhem a tedy kvalitativně zlepšila prováděnou monetární politiku. Z VAR analýzy však vyplývá, že zpožděná hodnota změny v R není vhodným predikátorem současné změny podnikových úvěrů, přitom totožné závěry lze učinit i pro ostatní finanční aktiva. Taktéž Grangerova kauzalita poukazuje na zmíněnou skutečnost se současným potvrzením závěrů z korelační analýzy, tedy posilující vliv úrokové míry na jednotlivé veličiny v čase. Zajímavé výsledky poskytuje analýza defaultů, z níž plyne poměrně silná vazba na jejich minulou hodnotu a nepřímou závislost na úrokové míře. Snížením R je tedy možno očekávat zvýšený podíl defaultů z poskytnutých úvěrů. Tento závěr koresponduje s uvedenou tezí risk-taking channel měnové transmise, tedy skutečnost, kdy nižší úroková sazba vede ke snižování úvěrových standardů a vyšší rizikové expozici. V kontextu ČR se také pravděpodobně jedná o kontinuální vliv rozvoje ekonomického prostředí s tendencemi spotřebního chování konvergujícího k západním ekonomikám.

Změny v úvěrech jsou tedy v zásadě inertní na předcházející změny v úrokové míře a mimo podnikové úvěry i na změny v aktuálním období. Taktéž se jeví, že změny v těchto veličinách se významněji nepromítají do změn v úrokové míře, což reflektuje skutečnost současného predikčního systému ČNB, jenž poskytnuté úvěry domácnostem a podnikům přímo nezahrnuje do prognózované inflace, resp. nastavované úrokové míry. Na druhé straně si je možno povšimnout, že změny v defaultech určitým způsobem na změny v úrokové míře působí, resp. lze se domnívat, že korespondují s ekonomickým vývojem, na nějž úroková míra reaguje. Taktéž úvěry finančním podnikům jsou oboustranně signifikantní v Grangerově kauzalitě, což však plyne ze samotného fungování režimu cílování inflace, kdy ČNB uspokojuje výši poptávky po penězích na základě stanovené úrokové míry, a tedy jiný závěr z provedených testů by byl rozporuplný.

Z dosavadní provedené analýzy je tedy možno říci, že byl potvrzen kritizovaný stav současného nastavení způsobu řízení měnové politiky, kdy je primárně hleděno na cenovou stabilitu, aniž by se zohledňoval vývoj ve finančních aktivech, úvěrech. V případě ČR je však nutno poznamenat, že vývoj úrokové míry a úvěrů je značně ovlivněn již zmíněným konvergenčním procesem, a poskytované úvěry jsou přímo odvislé od tržních změn. Faktem však zůstává, že nízká úroková sazba přispívá k vyšší míře defaultů.

V případě kvartálních změn ve výstavbě nemovitostí je patrný vliv předchozí hodnoty R s nejvyšším koeficientem determinace se signifikantním vlivem úrokové míry z kategorie aktiv, přičemž pro oba analyzované časové úseky je koeficient determinace i koeficient R(-1) poměrně vysoký. Grangerova kauzalita naznačuje, že od roku 2004 je vliv veden od změny v úrokové míře směrem ke změně ve výstavbě nemovitostí, zatímco koeficient korelace signalizuje minimální vliv současné změny úrokové míry na současnou změnu v nemovitostech. Při bližším pohledu na znaménka specifikovaných VAR modelů za obě období je však poměrně náročné ekonomicky interpretovat kladné působení zpožděné změny v úrokové míře na současnou změnu ve výstavbě nemovitostí, neboť ekonomická teorie předpokládá opačné působení, tedy zvýšení úrokové míry by teoreticky mělo mít za následek pokles výstavby z důvodu výdajového (dražší hypotéční úvěry) i podnikového (nákladnější investiční výdaje) kanálu transmisního mechanismu. Tento rozpor navozuje domněnku, že model by měl být rozšířen tak, jak uvádí Komárek (2011) a v roli regresorů by měly vystupovat další proměnné, díky nimž by vliv úrokové míry byl ekonomicky interpretovatelný. Uvedené VAR modely však splňují svůj účel a poukazují na určitý vliv úrokové míry na nemovitosti a tedy nezbytnost sledování ČNB (obecně CB) tohoto aktiva, neboť společně s úvěrovou expanzí je schopno značně rozkolísat finanční i hospodářskou stabilitu.

Index PX je dle provedených testů nezávislý na vývoji úrokové míry, což je pravd. odrazem skutečnosti značně omezeného počtu obchodovaných titulů a celkově stále nízké úloze tohoto trhu v kontextu celkových podmínek ČR. I přesto je však vhodné doporučit simultánní sledování veškerých komponent kategorie aktiv ze strany měnového tvůrce v případě, že mimo primární cíl v podobě cenové stability je zaměřen na stabilitu finanční a hospodářskou. Tudíž i v tomto případě je nutno akcentovat nutnost relevantního nastavení kompetentního regulačního rámce.

Druhá definovaná kategorie naznačuje poměrně stabilní vývoj korelace v čase s výjimkou spotřeby, jež v pozdějším období vykazuje větší svázanost s úrokovou mírou, již lze opět interpretovat jako důsledek kvalitnější a důslednější komunikace ze strany ČNB, resp. stabilizace provádění měnové politiky. Grangerova kauzalita za období 2004 - 2011 poukazuje na jednoznačný vliv úrokové míry na všechny zkoumané veličiny, přitom současná změna spotřeby i investic je signifikantně ovlivněna změnou v úrokové míře s půlročním odstupem, což koresponduje s ekonomickou teorií. Nezaměstnanost je v této kategorii velmi zajímavou položkou, neboť disponuje nejvyšším koeficientem determinace při současném signifikantním koeficientu zpožděné úrokové míry, jenž nabývá záporné hodnoty. Snížením úrokové míry tak je možno v krátkodobém horizontu očekávat vzrůst nezaměstnanosti, což lze označit za důsledek režimu cílování inflace, jenž reaguje na ekonomický útlum, a tedy vzrůst nezaměstnanosti, resp. očekávaný pokles cenové hladiny pod stanovený inflační cíl, snížením úrokové míry s půlročním až ročním předstihem. Tento vztah je také znám pod termínem Phillipsovy křivky pro krátké období. Závěr VAR modelu je tak korektní i z ekonomického hlediska.

Testy za druhou kategorii tedy jednoznačně potvrdily účinnost měnové politiky v režimu cílování inflace a vliv úrokové míry je patrný napříč všemi zahrnutými veličinami v této kategorii, přičemž pro bližší zkoumání faktického vlivu je nutno provést podrobné dekompozici transmisního mechanismu, vazby ve finančním systému s návazností na reálnou ekonomiku, pouze tehdy lze učinit korektní a smysluplné závěry o dalším směřování měnové politiky.

Poslední kategorie obsahující odvětví české ekonomiky pregnantně determinuje rostoucí vliv úrokové míry na dílčí sektory, a to zejména na celkový vývoj v průmyslu, zpracovatelském průmyslu a ve výzkumné činnosti. Grangerova kauzalita poukazuje spíše na působení úrokové míry na odvětví, resp. změna v úrokové míře s čtvrtročním až půlročním předstihem předznamenává změnu ve výkonnosti daného sektoru, nicméně dílčí koeficienty determinace signalizují, že modely by bylo vhodné dále specifikovat. Vliv úrokové míry je však jako v předešlém případě patrný. Čímž je opětovně potvrzena účinnost inflačního cílování.

Závěr

Z provedeného výzkumu vyplynulo, že vliv úrokové míry, jakožto hlavního měnově-politického nástroje ČNB, resp. CB, je signifikantním činitelem pro ovlivňování jak cenové hladiny, tak reálné ekonomiky. Finanční aktiva, nemovitosti či index PX sice nenaznačily významnější provázanost se změ-

nami v úrokové míře, navzdory tomu je však nanejvýš žádoucí jejich pečlivé sledování ze strany CB, neboť krize z let 2008, 2009 jasně prokázala, že nadměrná úvěrová expanze společně s vývojem na trhu nemovitostí může rezonovat z finanční nestability v nestabilitu hospodářskou, byť byla CB sledována cenová stabilita.

Postupně sílící vliv úrokové míry na zkoumané veličiny v čase dokazuje účinnost zvoleného režimu cílování inflace a částečně potvrzuje vhodnou komunikaci a celkový kvalitativní posun v činnost ČNB. Budoucí výzkum by se tak měl ubírat směrem analýzy transmisního mechanismu při současné snaze o implementaci determinace bublin do ekonometrických modelů k zajištění dalšího zlepšování prognóz a k hlubšímu porozumění ekonomických souvislostí.

Diskuse napříč Evropskou unií ohledně regulačního rámce centrálních bank je tak nanejvýš aktuální a výsledky této práce poukazují na nezbytnost vhodného nastavení pravidel ve finančním systému tak, aby byla dosažena finanční a hospodářská stabilita, přičemž měnová politika CB by měla zůstat primárně zaměřena na stabilitu cenovou.

Literatura

ARLT, Josef., ARLTOVÁ, Markéta, 2007. *Ekonomické časové řady.* Praha : Grada Publishing, a.s., 2007. ISBN 978-80-247-1319-9.

BERNANKE, Bens, et. al., 1999. *Infaltion Targeting: Lessons from the International Experience.* Princeton University Press, 1999.

HUŠEK, Roman, 2007. *Ekonometrická analýza.* Praha : Vysoká škola ekonomická v Praze, Nakladatelství Oeconomica, 2007. ISBN 978-80-245-1300-3.

GERLACH, Stefan et. al., 2009. Are the golden years of central banking over? Paper for the 38 Economics Conference of the Austrian National Bank on "Central Banking after the Crisis: Responsibilities, Strategies, Instruments". Vienna, 2010.

GREENSPAN, Alan, 2010. The Crisis. Greenspan Associates LLC. April 2010.

JÍLEK, Josef, 2004. *Peníze a měnová politika.* Praha : Grada Publishing, a.s., 2004. ISBN 80-247-0769-1.

KUBICOVÁ, Ivana, KOMÁREK, Luboš, 2011. Možnosti identifikace bublin cen aktiv v české ekonomice. *Politická ekonomie*. 2011, Vol. 59, No. 2, pp. 164-183.

ŘIČAŘ, Michal, 2010. *Ekonometrická analýza monetární a fiskální politiky ČR s dopady na podnikovou sféru*. Praha, 2010. Diplomová práce. Vysoká škola ekonomická v Praze, Fakulta podnikohospodářská.

TAYLOR, B. John, 2008. The Financial Crisis and the Policy Responses: An Empirical Analysis of What Went Wrong. November 2008.

ZAMRAZILOVÁ, Eva, 2011. Měnová politika: Staré lekce, nové výzvy. *Politická ekonomie*. 2011, Vol. 59, No. 1, pp. 3-21.

JEL Classification: C10, C32, E43

Summary

Quantitative impacts of interest rates

The performed research showed that the influence of interest rate as the main monetary policy instrument of the CNB, a central bank generally, is a significant factor to influence a price level and the real economy. Although financial assets, property and the PX index don't indicate significant coherence with changes in the interest rate, it is highly recommended to observe them carefully from the point of a CB, since the crisis of 2008, 2009 clearly demonstrated that excessive credit expansion together with the development of the property market can resonate from financial instability to the global economic instability, although a CB could precisely observe price stability.

Gradually rising influence of the interest rate on the analyzed variables in time has proven the effectiveness of the inflation targeting regime, and partly confirmed the appropriate communication and qualitative shift in the overall activity of the CNB. Thus future research should follow deeper analysis of the transmission mechanism with simultaneous trying to implement the determination of bubbles into the econometric models to ensure continued improvement of forecasts and better understanding of the economic relationships.

Discussion across the European Union about the regulatory framework of central banks is highly actual and the results of this work showed neediness for a suitable set of rules in the financial system thus could be obtained the financial and economic stability, while the CB monetary policy should remain primarily focused on the price stability.

Keywords: interest rate, monetary policy, price stability, financial stability, VAR.

Modelling generator constraints for the self-scheduling problem

Jiří Šumbera
sumbera@jikos.cz
PhD. student, Econometrics and Operational Research
Supervisor: doc. Mgr. Ing. Martin Dlouhý, Dr., MS, dlouhy@vse.cz

Abstract: This paper deals with the self-scheduling problem faced by a powerplant when faced with a single price scenario. Powerplant's characteristics affecting its dispatching behaviour are described, such as maximum and minimum stable generation, minimum up and down times, time-dependent start-up costs and ramping rates. These characteristics are then translated into constraints applicable within a mixed-integer linear programming framework. Several formulations are presented which could be used in different modelling approaches. A case study involving a newly build gas-fired powerplant in Malženice is undertaken. The powerplant is dispatched against the hourly prices of the Slovak day-ahead short-term market (XMtrade®/ISOT) for 2011 delivery. Fuel is priced according to the official Slovak gas reference price and full pass-through of EU ETS carbon dioxide emissions price is assumed. Technical characteristics are based on publicly available information for a similar unit in Ireland. Several scenarios aimed at investigating the impact of individual constraints and horizon length were constructed. The results confirm the need to model detailed unit-commitment characteristics such as start-up costs using mixed-integer linear programming approaches since approaches based on Linear programming grossly overestimate solution value. The results also illustrate that selection of a simplified formulation may result in much higher computation speeds without sacrificing result accuracy. Finally, the results also demonstrate that the length of the horizon has a marginal impact on the solution quality, while having a significant influence on the computation speed.

Keywords: Power generation dispatch, Power generation scheduling. Mixed-integer linear programming, Start-up costs.

Introduction

The self-scheduling problem (SSP) is a profit maximization problem faced by a generator, which sells power in a day-ahead market. The simple version of SSP is assumes a small (price-taking) generator along with its technological (and commercial) constraints and a single price forecast. The two simplifying assumptions both deal with the price: the single price forecast bears no uncertainty and the bids of a price-taker do not influence the resulting prices. Consequently, once the price uncertainties are removed, the problem is reduced to accurately implementing the various constraints faced by the generator. In such case, the optimal solution to the problem depends only on the technical characteristics of the generator. On the contrary in the real world forecasts are always uncertain implying a range of possible outcomes. Similarly, under specific conditions such as peak load, no generator is too small to have no market power. To deal with the above problems approaches such as robust programming (Landry, 2009) and oligopolistic models (Conejo, 2002), are necessary.

Similar but much larger problem is the unit commitment problem (UCP) faced by the Independent System Operator (ISO). This problem consists of selecting the generating units to be put into service, i.e. committed, during a scheduling period and deciding how much will each of them generate. The committed units must meet the system load and reserve requirements at minimum operating cost, subject to a variety of constraints. The individual generators have to meet their own technical constraints as well as the system-wide load and reserve constraints. In its most complex form, the UCP is highly non-linear due to quadratic losses within the network and non-linear relationships between fuel consumption and generation.

Historically a wide range of solution methods has been applied to both UCP and SSP. The methods include dynamic programming, Lagrangian relaxation, branch and bound, expert systems, fuzzy theorem, the hop field method, simulated annealing, tabu search, genetic algorithms, neural networks, evolutionary programming and others. A bibliographical survey of various approaches is provided by

(Padhy, 2004). The major limitations of the numerical techniques are the problem dimensions, large computational time and complexity in programming. Recently, due to advances in mixed-integer linear programming (MILP) solvers, formulations to both the UCP and SSP have been proposed (García-González, 2006).

The structure of the paper is as follows. The next section reviews the general technical characteristics of powerplants together with their causes. In the following section the SSP problem is formulated as MILP problem within which a powerplant's characteristics are captured as constraints. When several methods of formulating a constraint are possible, all are presented. The next section presents a case study for a small SSP and its results. The summary concludes and outlines areas of further research.

Powerplant technical and economic characteristics

Operating limits

Each generator can produce electricity within a certain range of values. The upper and lower limits - the *maximum capacity* P^{max} and the *minimum stable generation* P^{min} – typically refer to the limits of sustained generation under normal operating condition.

Production costs

The cost of production function is typically expressed as a quadratic function of the generation level p

$$c^p = \alpha + \beta p + \gamma p^2, \quad [1]$$

even though there is no fundamental reason for a quadratic relationship – it is merely a convenient approximation. In reality, the best a powerplant operator can do is to measure the steady-state fuel consumption corresponding to a certain generation level. Thereby the set of pairs [p, c^p] is created, which is then approximated/extrapolated using a convenient method. Traditional UCP methods such as Lagrangian relaxation use a quadratic function or a higher order polynomial, while solution methods based on MILP use piecewise linear function(s). The production cost function need not be constant in time due to several reasons: the cost of the fuels is changing and the efficiency of the unit can be changing with the season (due to efficiency depending on the difference between the operating and the outside temperature).

Start-up and shutdown costs

As described in the section on operating limits, the unit can either produce within its operational range (unit is online) or have no output at all (unit is offline). The transitions between the offline and online states are known as the start-up and shutdown. During the start-up of a thermal powerplant additional costs are incurred. The start-up, which can last up to several hours, typically involves the heating up of the equipment in order to produce high pressure steam to drive the steam turbine. During this period fuel is consumed while no power is produced. In the case of some fuels such as lignite, the burning process has to be stabilised by fuel oil (lignite does not combust well at low temperatures) and thus additional costs are incurred.

In addition, during the start-up and shutdown, the unit is subjected to thermal stresses, which can lead to fatigue and possible permanent damage of the unit's components requiring future maintenance or repair. The sum of the average additional maintenance costs resulting from each start-up together with the total cost of the fuel used to bring the unit to within its operating limits is known as the start-up cost, C^u.

Contrary to the start-up process, when the change in temperature depends on the amount of heat supplied from the fuel and lost to the surroundings, during a shutdown heat exchange is due to losses only and follows an exponential decay function with the limit being the temperature of the surroundings, which is typically reached in the range of 48-60 hours. Therefore, the start-up cost depends exponentially (Christober, 2011) on the number of hours since the last shutdown, $h^d_{t,}$

$$c_t^u = C^{u,c}\left(1-\exp\left(\frac{h_t^d}{H}\right)\right). \quad [2]$$

Cold start refers to a start-up of a powerplant that is at ambient temperature. The corresponding time it takes for the unit to cool down to this temperature level is known as the *cold start time*. Should the unit be restarted sooner, the amount of fuel needed to bring the unit to operational temperature would be lower. Depending on the time between the shutdown and the subsequent start-up, one can refer to a *hot start* (typically 6-8hrs and less) or a *warm start* (6-48hrs). These times and costs would normally refer to empirically determined values from typical start-up times: hot start – immediately after the unit is shut down, warm start – after the unit is shutdown over night, cold start – after the unit is shut down for the weekend.

The total start-up cost of a generator is the sum of fuel costs (depend on shutdown duration) and other start-up costs (typically a constant reflecting the increased maintenance)

$$C^u{}_t = FO^u{}_t FP_t + C^{u,other}. \quad [3]$$

Minimum up and down times

These constraints refer to the minimum time the unit has to be on once it starts up (MUT) and the minimum time it has to be off, once a shutdown occurs (MDT). MDT constraints arise due to necessary maintenance after a unit has been shut down, e.g. in the case of nuclear powerplants there is a technically mandated minimum down time of 15-24 hours (Schmid, 2010), while for flexible gas units no such a requirements exists. MUT constraints typically reflect the need to minimise thermal stresses in the equipment which could otherwise arise.

Ramp rates and start ramp rates

As has been previously suggested, for thermal units, rapid changes in temperature or output may lead to increased maintenance costs. Consequently, safe ramp up and ramp down rates are provided by the manufacturer. On occasions when the prices are high enough, the plant's operator may choose not to respect them, effectively exchanging the increased future maintenance cost for higher revenue in the short term. In addition, the physical design of a unit do provide an upper bound on these ramp rates, however these may be so high, that they need not be considered in the model. For example the ramp rate might allow for an increase from zero to full output in less then one model period. Such a constraint will never be binding.

Other constraints

Other constrains typically arise from commercial contracts or legislation. For example, the total generation within a horizon may be limited from both above and below due to constraints on the fuel. Commercial contracts may impose minimum and maximum daily or yearly offtakes (typically long term gas contracts). Similarly, emission caps may limit the total quantity of emissions ($SOX/NOX/CO_2$) generated (emissions are directly related to fuel consumption). Finally a broad category of constraints may apply to a particular problem. For example a generator may not be available on weekends or the transmission grid may limit the total generation of several units connected to the same node of the network.

Sources of revenue

Powerplant's main source of revenue typically comes from the sales of its generation. These can be realized on forward markets, day-ahead markets or markets with a very short lead horizon such as the intraday or balancing markets. In addition, the powerplant may receive additional payments from the TSO for providing network services. These may include ancillary services (services related to maintaining the balance of the system in real time) or capacity payments (payments designed to motivate the construction of otherwise non-profitable peak powerplants). Moreover, the powerplant may be eligible for payments from re-dispatching (situation when a powerplant is dispatched by the

TSO not based on its economic schedule but in order to satisfy grid constraints). In this paper, for simplicity reasons, only sales on the day-ahead market are be considered with the assumption that the full capacity is offered. On the other hand, the above-mentioned additional sources of revenue – if present – can change the economics of the powerplant considerably.

Problem formulation

Nomenclature

Indices

i: generators, $i \in 1..I$

t: periods in the modelling horizon, $t \in 1..T$

k: production cost function segments, $k \in 1..K$

Decision variables (all lower indices i refer to generator i)

p_{it}: Production quantity in period t [MW]

$l_{kit:}$ Production in the k-th segment of the pcw. linear approximation in period t [MW]

c^p_{it}: Production cost in period t [EUR]

c^u_{it}, c^d_{it}: Start-up and shutdown cost of generator i in period t [EUR]

x_{it}: Unit commitment variable in period t (1 if online, 0 otherwise)

y_{it}: Start-up variable in period t (1 if unit started in period t, 0 otherwise)

$y^h_{it}, y^w_{it}, y^c_{it}$: Hot/warm/cold start-up auxiliary variable in period t

$y^{h,exp}_t, y^{w,exp}_t$: Hot/warm start expiration binary variables

z_{it}: Shutdown variable in period t (1 if unit shut off in period t, 0 otherwise)

h^d_{it}: Number of hours generator has been offline in period t since the last shutdown [h]

Parameters (all lower indices i refer to generator i)

α, β, γ: Constant/linear/quadratic coefficient of a quadratic production cost curve

B_i: Linear coefficient of a linear production cost curve (constant efficiency)

A_{0i}, B_{ki}: Constant/linear coefficient of a pcw. linear production cost curve with k segments

$A^{tan}_{ki}, B^{tan}_{ki}$: Constant/linear coefficient of the k-th lower approximation of a production cost curve

H_i: Characteristic time of a cooling function [h]

λ_t: Market price in period t [EUR/MWh]

P^{min}_i, P^{max}_i: Minimum and maximum generation capacity [MW]

P^k_i: Maximum generation capacity of generator up to the k-th segment, $P^0_i = P^{min}_i$, [MW]

$C^{u,h}_i, C^{u,w}_i, C^{u,c}_i$: Hot/warm/cold start-up cost [EUR]

$C^{u,other}_i$: Other start-up cost [EUR]

FO^u_{it}: Fuel offtake at start-up in period t[GJ]

FP_t: Fuel price in period t [EUR/GJ]

H^h_i, H^w_i, H^c_i: Hot, warm and cold start-up time [h]

$T^{min,u}_i, T^{min,d}_i$: Minimum up and down time [h]

RR^u_i, RR^d_i: Ramp rate up and down [MW/h]

Simplest formulation of SSP

The objective function to be maximized is the generator's profit expressed as the difference between revenues and costs

$$\max \sum_{t=1}^{T} \left(\lambda_t p_t - c^p_t - c^u_t - c^d_t \right) . \quad [4]$$

The basic characteristics of a generator is its maximum capacity, which sets an upper bound on generation in any instant

$$p_t \leq P^{max}, \ \forall t, \quad [5]$$

while the simplest formulation of production cost has constant unit costs

$$c^p_t = B p_t, \ \forall t. \quad [6]$$

Given a set of prices λ_t and cost of production characteristics (B) and assuming zero start-up and shutdown costs, c^u_t and c^d_t, the above equations define the whole SSP and can easily by solved using linear programming (LP) methods without the need for binary or integer variables. Provided that the linear cost function coefficient B approximates well the average production cost at maximum capacity, the above LP problem provides an upper bound on profit to all other formulations.

On the other hand, as has been suggested in the previous section, most real world units are much more complex. They cannot run at arbitrarily low output and there are limits on how quickly they can change their output and/or how quickly they can start. Apart from ramping rates, these additional constraints require the MILP framework. As is the case with any LP/MILP problem, additional constraints will result in a decrease of the objective function, i.e. profit.

The basic solution can be described purely in terms of powerplant's generation in each hour. The other quantities can be easily derived. The above formulation results in two possible solution values: either the powerplant generates at full capacity (in case its unit production costs exceed the market price) or the unit does not run (and the system load is served by other powerplants whose production costs allow them to be profitable at given power prices).

Unit commitment

The fact that the unit is either not producing electricity (unit off) or is producing in the admissible range (unit online) has to be modelled using a binary variable x_t:

$$x_t P^{min} \leq p_t, \ \forall t \quad [7]$$

$$p_t \leq x_t P^{max}, \ \forall t. \quad [8]$$

The first equation implements the minimum stable generation, while the second one simultaneously implements the maximum generation limit (thus replacing equation [5]) as well as limits generation to 0, when the unit is offline.

Production costs

Equation [6] describes production cost as a linear function of the production quantity. Better approximation is obtained using a piecewise (pcw.) linear function

$$c^p_t = A_0 + \sum_{k=1}^{K} B_k l_{kt}, \forall t, \quad [9]$$

where B_k are the unit production costs in the k-th segment of the pcw. linear approximation and l_{kt} are the corresponding generation variables. They are bound by the following constraints

$$l_{kt} \leq P^k - P^{k-1}, \ \forall t, \ \forall k. \quad [10]$$

The equation defining the total generation as a sum of generation at each segment is also required

$$p_t = P_0 x_t + \sum_{k=1}^{K} l_{kt}, \forall t \quad . \quad [11]$$

Because the approximation is always greater or equal to the quadratic approximation, it is known as the *upper pcw. linear approximation*. One can also create a *lower pcw. linear approximation* by taking a set of tangents at points P_k (Figure 1b). This approximation does not divide the unit into several virtual smaller ones such as equation [9], but instead creates a lower bound on the production cost directly, using K equations

$$c^p_t \geq A_k x_t + B^{tan}_k p_t, \ \forall t, \ \forall k \in K. \quad [12]$$

The approximation improves with increasing number of segments k, however this also tends to increase the problem size. There are two approaches to this sub-problem: one tries to improve on the quality of the approximation without changing the model size, the other tries to iteratively refine the approximation in the neighbourhood of the solution.

(Wu,2011) deals with the former approach and finds the best segment partitioning method for the upper approximation. Traditionally, the segments $[P^{k-1}, P^k]$ are equidistant. Wu formulates the general

unconstrained optimization problem as a minimum of the sums of squared differences of the arc and chord lengths and solves it using Newton-Rhapson methods for optimal P^k.

(Viana, 2011) works with the lower approximation, starting with just two segments: one corresponding to the tangent at P^0 and the other to the tangent at P^k. After solving the problem with this simple approximation the solution $p^{*1}{}_t$ in period t is found. The method works by adding additional constraints of the type [12] to the problem and resolving. The constraints are only added to the periods when the unit is switched on. The method converges when the delta between the approximation of the cost function using existing tangent envelope and a new envelope created by adding a new constraint is less than a predefined ε.

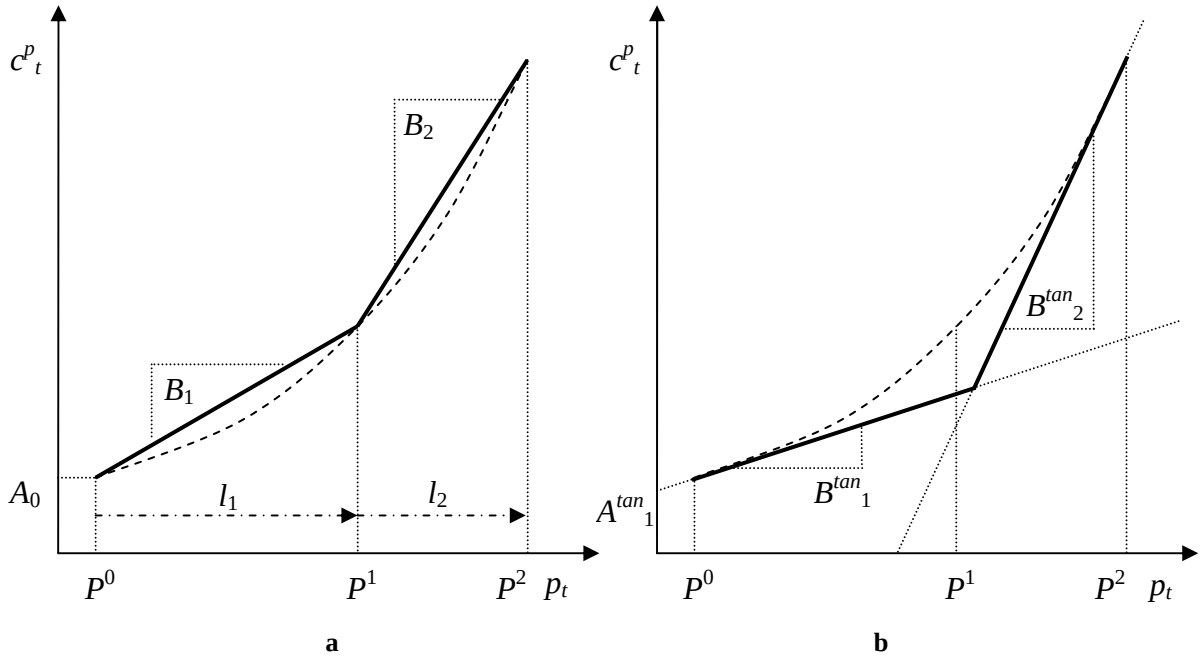


Figure 1: Approximations (bold) of the quadratic production cost function (dashed):

a) Upper pcw. linear approximation using segments. b) Lower pcw. linear approximation using tangents.

Both the lower and upper pcw. approximations work only for convex (monotonically non-decreasing) functions, i.e. only if B_k form an increasing series. Non-convex functions require the use of additional binary variables b_k, which are non-zero only if the level of generation is within the corresponding segment k. The approximation is carried out using a pcw. linear method. The necessary equations are:

$$l_{kt} \leq (P^k - P^{k-1}) b_{kt}, \ \forall t, \quad [13]$$

$$l_{kt} \geq (P^k - P^{k-1}) b_{k+1,t}, \ \forall t, \quad [14]$$

$$b_{kt} \geq b_{k+1,t}, \ \forall t. \quad [15]$$

Equation [15] forces an order in which the binary variables take on the value 1. Equations [13-14] replace equation [10], ensuring that the generation at the previous segment is at its maximum before the next segment is utilised.

Start-up and shutdown costs

Even though they are not strictly necessary, the model formulations are simple if the variables representing a powerplant's start-up and shutdown are explicitly defined. Their lower bounds are defined by how the unit commitment variable, x_t, changes in time:

$$x_t - x_{t-1} \leq y_t \leq 1 \ \forall t, \quad [16]$$

$$x_{t-1} - x_t \leq z_t \leq 1, \ \forall t. \quad [17]$$

These variables need not be defined as binary, but they effectively behave as such. The simplest start cost formulation (Landry, 2009) assumes that there are 3 start-up states (hot/warm/cold) and the start-up costs are a pcw. constant function of the time offline.

$$c_t^u \geq C^{u,h}\left(x_t - x_{t-1}\right), \forall t, \quad [18]$$

$$c_t^u \geq C^{u,w}\left(x_t - \sum_j^{Hw} x_{t-j}\right), \forall t, \quad [19]$$

$$c_t^u \geq C^{u,c}\left(x_t - \sum_j^{Hc} x_{t-j}\right), \forall t. \quad [20]$$

The above formulation implies that the hot/warm/cold start-up costs apply as soon as the unit is down for one/H_w/H_c periods. The formulation is easy to generalize to a higher number of start-up states. Conversely, only one of the equations is necessary to model a single start-up cost applicable after H_x periods.

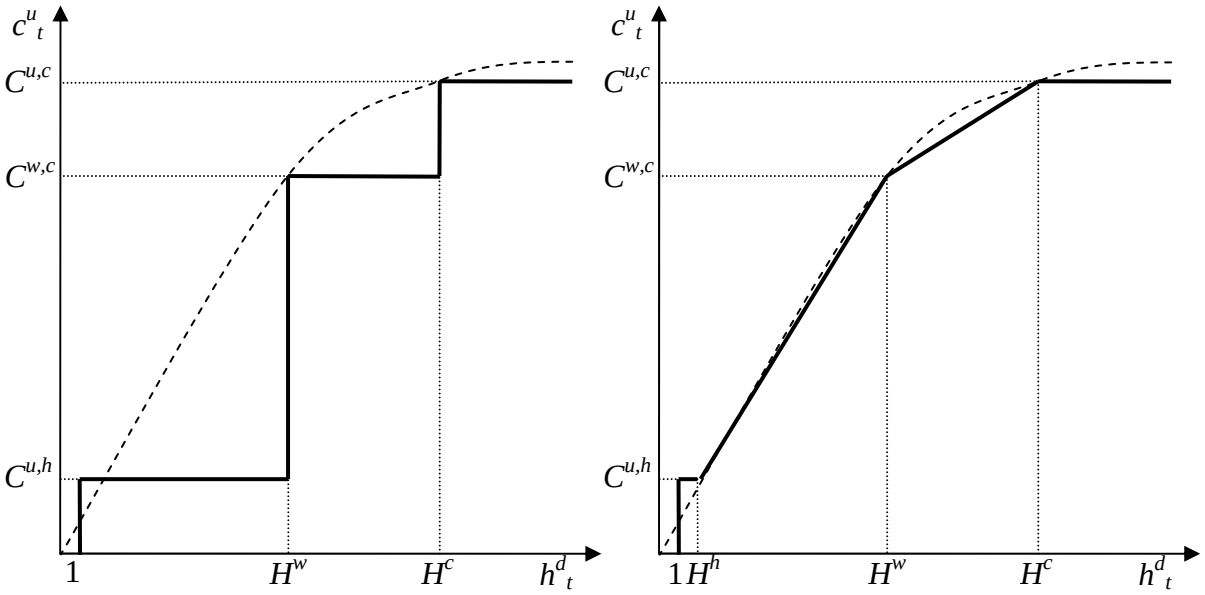


Figure 2: Approximations (bold) of the exponential start-up cost curve (dashed):

a) piecewise constant, b) piecewise linear.

(Viana, 2011) works with two start-up states only, however allows for the start-up costs to vary linearly in between the two start-up times. Power market modelling software *PLEXOS for Power Systems* developed by (Energy Exemplar, 2011) generalizes this approach into multiple start-up states resulting in the following 10 equation formulation for 3 start-up states. Below, the formulation is broken down into 3 groups of equations.

The first group defines the start-ups and the corresponding costs:

$$y^h_t \leq 1, y^w_t \leq 1, y^c_t \leq 1, \forall t, \quad [21]$$

$$y_t = y^h_t + y^w_t + y^c_t, \ \forall t, \quad [22]$$

$$c^u_t \geq C^{u,h} y^h_t + C^{u,w} y^w_t + C^{u,c} y^c_t, \ \forall t. \quad [23]$$

The first equation sets the upper bounds on the auxiliary hot/warm/cold start-up variables. The second defines generator's start-up period *t* as being hot, warm, cold or their mix. The last equation defines start-up costs to be at least equal to an affine combination of the hot/warm/cold start-up costs.

$$h^d_t - h^d_{t-1} \geq 1 - Tx_t, \ \forall t \quad [24]$$

$$h^d_t \leq T(1 - x_t), \ \forall t \quad [25]$$

$$h^d_t - h^d_{t-1} \leq 1, \ \forall t \quad [26]$$

The next group of equations insure that that the counting variable h^d_t, which counts the number of hours that unit has been down, takes on the correct value in period t:

- if the unit is down in period *t*, the counter increases, (equation [24]),

- it the unit is on in period *t*, the counter is reset to zero (equation [25]),
- the maximum increase of the counter is 1 (equation [26]).

Finally, the last set of equations takes care of the linear interpolation. Two new binary variables $y^{h,exp}_t$ and $y^{w,exp}_t$, which reach value 1, when the linear interpolation between the hot/warm and warm/cold start-up times is no longer applicable, i.e. it expires:

$$y^h_t + y^{h,exp}_t \leq 1_t, \ \forall t, \quad [27]$$

$$(H^w - H^h)y^h_t - (H^w - h^d_t) \leq (H^c + T)y^{h,exp}_t, \ \forall t, \quad [28]$$

$$y^w_t + y^{w,exp}_t \leq 1_t, \ \forall t, \quad [29]$$

$$(H^c - H^w)y^w_t - (H^c - h^d_t) \leq (H^c + T)y^{w,exp}_t, \ \forall t. \quad [30]$$

Equation [27] determines the transition from a hot start to a warm start. As long as $h^d_t \leq H^h$, the start is hot. In case of the equality, i.e. $h^d_t = H^h$, y^h_t must take on value 1, so that the expiration variable on the RHS is not activated. When $H^h < h^d_t < H^w$, in order not to activate the expiration variable, the LHS must be at most zero. This is satisfied as long as

$$y^h_t \leq \frac{(H^w - h^d_t)}{(H^w - H^h)}, \quad [31]$$

i.e. y^h_t is inversely proportionate to the to the position of the current hours down h^d_t on the segment $[H^h, H^w]$. Finally, when $h^d_t \geq H^w$, there is no permissible positive value for y^h_t and the expiration variable $y^{h,exp}_t$ on the RHS is activated. Because this variable is binary, equation [26] insures that the hot start variable is indeed 0. The second pair of equations enforces a similar behaviour for the transition from warm to cold.

Minimum up and down times

Minimum up and down time constraints are very simple to formulate. The equation below shows the minimum up constraints for a minimum up time of $T^{min,u}$ periods

$$y_t \leq x_{t+k}, \ \forall t, \ \forall k \in 1..\ T^{min,u}-1. \quad [32]$$

The equation ensures that, if the unit starts in period *t*, it has to be online until the period $t + T^{min,u} - 1$. The formulation for minimum down time is similar:

$$z_t \leq 1 - x_{t+k}, \ \forall t, \ \forall k \in 1..\ T^{min,d} - 1. \quad [33]$$

Ramp rates

Most thermal units have limits on how quickly they can ramp up and down in order to avoid thermal stresses leading to failure or increased maintenance costs:

$$p_t - p_{t-1} \leq RR^u, \ \forall t, \quad [34]$$

$$p_{t-1} - p_t \leq RR^d, \ \forall t. \quad [35]$$

This formulation works for units without a minimum stable generation P^{min} and also if the ramp rates are greater or equal to the minimum stable generation ($P^{min} \leq RR^u$ and $P^{min} \leq RR^d$). In the opposite case, the equations need to be modified to allow for the unit to start-up, e.g.

$$p_t - p_{t-1} \leq (RR_u + P^{min})x_t - P^{min}x_{t-1}. \quad [36]$$

Case study

Case study data

To illustrate the effects of modelling accuracy on the solution, a SSP problem has been set up, which tries to mimic a real world situation as close as possible. German utility EON has commissioned a new 430MW combined cycle gas turbine (CCGT) powerplant near the Slovak town of Malženice in January 2011 (EON, 2011). The SSP will try to model the optimal dispatch of the CCGT assuming a perfect forecast of the actual prices on the relevant market (Slovak day-ahead power exchange XMtrade®/ISOT). Since no information about the relevant gas price was available (commercial fuel

contracts are often confidential), it was assumed that the relevant gas price is the official Slovak reference gas price (Slovakia, 2010), even though information in the press suggests EON's gas price could be lower due to contract renegotiations in 2010. In addition, the full pass-through of the opportunity cost of CO2 emissions under the EU ETS scheme was assumed. The emission allowance prices were taken from (EEX,2011). The inclusion of emission costs does not modify any of the formulations presented above, it only changes the absolute value of fuel prices.

Given the lack of detailed technical data on the Malženice CCGT, data for a similar CCGT commissioned in June 2010 in Aghada, Ireland, was used. Aghada's technical characteristics are captured in the Irish Single Electricity Market modelling study (SEM, 2010) in a format suitable for power market modelling software *PLEXOS*. Since both units are very similar (both have a single-shaft configuration, both have similar P^{max} and both have similar reported efficiencies), the SEM data for the Aghada CCGT data will be used without additional modifications.

The SEM dataset contains the following characteristics for all powerplants:

- Operating range: P^{min}, P^{max} (winter, summer)
- Operating cost: No load cost (A_0), marginal heat rates (equivalent to B_k). Marginal efficiency is 65.6% and the corresponding full load efficiency is 57%
- Ramping rates: RR^u, RR^d
- Minimum up and down times
- Start-up: Fuel offtake at hot, warm, cold start and the corresponding start-up times

The data relevant for the Aghada CCGT is shown in Table 1.

Table 1: Aghada CCGT characteristics Source: SEM, 2010

Property	Value	Units	Property	Value	Units
P^{max}	431.6	MW	H^c	72	hrs
P^{min}	215	MW	$FO^{u,h}$	1200	GJ
P^0	215	MW	$FO^{u,w}$	1800	GJ
P^1	431.6	MW	$FO^{u,c}$	2400	GJ
A_0	354.511	GJ/hr	$T^{min,u}$	4	hrs
B_1	5.497	GJ/MWh	$T^{min,d}$	4	hrs
H^h	1	Hrs	RR^u	22	MW/min.
H^w	12	Hrs	RR^d	22	MW/min.

In addition, the assumption that the unit is initially out and has been so for more than H^c periods was made.

Case study formulation

This data was used within a market modelling tool Plexos with XPRESS-MP as the MILP solver. A base model was set up and additional 6 sensitivity models were run. The base model has all the constraints from equations [4], [7]-[11], [16], [21]-[30], [32] and [33], initialized using data from table 1. In addition, the base model uses a rolling horizon of 6 days, i.e. 144 periods, split into two groups (3+3 days). When the 6 day problem is solved only the solution of days 1-3 is saved and the model is then rerun for the next 6 day step, i.e. days 4-6. This set-up was chosen to accurately capture the longest constraint in the problem formulation, which is a cold start after 3 days. This way the solution of day 3 takes into account the most remote future constraint with direct impact on that day.

The following sensitivities were run:

- Horizon sensitivity 1: model with horizon of 2 days (1+1)
- Horizon sensitivity 2: model with horizon of 14 days (7+7)
- Production cost sensitivity: modelled using constant efficiency equal to P^{max} efficiency
- Start cost sensitivity: only warm start-up costs modelled
- Constraint sensitivity: No MUD, MDT constraints
- Simple SSP (LP): No MUD, MDT, P^{min} constraints, no start-up costs
- Fuel sensitivity: constant fuel price (equal to average 2011 price)

The goal of the first two sensitivities is to determine the impact of the modelling horizon, the next 3 test impact from simplifying and excluding some constraints and the final case tests the sensitivity of the base case to fuel price.

Case study results

All models finished under 3 minutes time on a Pentium M@1.86 GHz with 2 GB of RAM, except the model with 14 day (7+7) horizon, which took almost 20 minutes to finish and was 7 times slower than the base model. On the other hand, the model with a short 2 day (1+1) horizon finished in only 40% of the time. This suggests that solution time increases rapidly with longer horizons. Comparing the different horizons further, it is evident that both costs and revenues increase with longer horizons. The latter increase at slightly faster rate resulting in the objective function (Net profit) value being highest for the 7+7 horizon. However the overall change is less than 0.5%. The number of starts is almost constant.

The results of production cost sensitivity (Scenario Constant Efficiency) are not surprising as the only difference to the base model is higher efficiency when running close to P^{min} levels. Consequently, the model is able to sustain generating at these levels with lower costs resulting in decrease in number of start-ups as well as overall increases in generation, fuel and emission cost and most importantly in higher net profit.

Removing the MUT and MDT constraints(Scenario No MUT, MDT) from the formulation allows the powerplant to be more flexible and either capture price spikes in single periods or return online soon after shutdown. Consequently generation and net profits increase, even if only marginally.

Table 2: Results for Slovak reference price scenarios

Property	Base (3+3)	1+1	7+7	Const Efficiency	Single Start Costs	No MUT,MDT	Simple SSP (LP)	Const Fuel Price	Units
N. of Start-ups	167	168	168	140	159	167	121	149	-
Fuel Price	7.98	7.98	7.98	7.98	7.99	7.98	8.08	8.36	EUR/GJ
Fuel Cost	52021	50266	52332	58585	51377	52001	63099	46269	EUR000
Emission Cost	5295	5111	5329	5981	5211	5293	6194	4058	EUR000
Start-up Cost	2570	2556	2597	2226	2528	2570	0	2452	EUR000
Total Gen. Cost	59887	57933	60258	66791	59116	59864	69293	52779	EUR000
Price Received	65	66	65	64	65	65	65	68	EUR/MWh
Generation	1028	994	1034	1162	1014	1028	1236	872	GWh
Pool Revenue	67108	65178	67474	74518	66293	67086	80641	59054	EUR000
Net Profit	**7222**	**7245**	**7216**	**7727**	**7177**	**7222**	**11349**	**6274**	EUR000
Solution Time	164	68	1187	162	12	157	4	164	s

Allowing the powerplant to be subject to warm starts only (scenario Single Start Costs) decreases the costs of start-ups after long downtimes and increases them after short ones. Given the fact that the base solution has approximately 3 starts per week, it is likely to have both hot and cold starts. For this particular situation, this simplification results in slightly higher average start-up costs. On the other hand, simplifying the start-up cost function yields a significant 14-fold increase in computation speed. This sensitivity essentially replaces equations [21]-[30] with a single, much simpler equation [19].

The Simple SSP (LP) model which assumes no MUD, MDT and P^{min} constraints as well as no start-up costs (and can thus be solved using LP) illustrates the need to include such constraints in the problem formulation. This model's solution has 20% higher generation and 50% higher net profit.

The final sensitivity (scenario Const Fuel Price) illustrates that all of the above conclusions depend heavily on the fuel (and power market) price assumptions. A constant fuel price throughout the year decreases the number of starts and total generation by approximately 10%. The annual 3TWh generation suggested by the powerplant owner (EON, 2011) implies access to gas prices below the Slovak reference gas price.

Conclusion

This paper has described a powerplant's characteristics affecting its dispatching behaviour and translated them into constraints applicable within a MILP framework for solving the simple scheduling problem of generator dispatch against a single price scenario. A case study involving a newly build gas-fired powerplant in Malženice has been undertaken. The results confirm the need to model detailed unit-commitment characteristics such as start-up costs. On the other hand they also illustrate the need for accurate modelling in terms of constraint formulation detail and appropriate horizon selection.

Future extension of the research could focus on improving the model formulations in terms of solution speed without sacrificing model accuracy as well as on the inclusion of additional characteristics required for more accurate modelling, such as the generation during a start-up or powerplant's participation in ancillary services market along with the associated revenues.

References

ARROYO J. M., CONEJO A. J., 2004. Modeling of Start-Up and Shut-Down Power Trajectories of Thermal Units. *IEEE Transactions on Power Systems*. Vol. 19, No. 3, pp. 1562-1568. Available from: http://www.uclm.es/area/gsee/web/antonio.htm

CHRISTOBER, C., RAJAN, A., 2011. An Evolutionary Programming Based Tabu Search Method for Unit Commitment Problem with Cooling-Banking Constraints. *Journal of Electrical Engineering*. Vol. 62, No. 1 / January-February 2011, pp 11-18. Available from: http://versita.metapress.com/content/pj55845873811681/

CONEJO A. J., ARROYO J. M., 2002. Optimal Response of an Oligopolistic Generating Company to a Competitive Pool-Based Electric Power Market. *IEEE Transactions on Power Systems*. Vol. 17, No. 3, pp. 424-430. Available from: http://www.uclm.es/area/gsee/web/antonio.htm

GARCIA-GONZALEZ, J., PARRILLA E., MATEO, A., 2007. Risk-averse profit-based optimal scheduling of a hydro-chain in the day-ahead electricity market. *European Journal of Operational Research*. Vol. 181, Iss. 3, pp. 1354–1369

EEX, 2011. European Energy Exchange AG: Emission Allowances - European Energy Exchange. *www.eex.com* [online]. Available from: http://www.eex.com/en/Download/Market%20Data/EU%20Emission%20Allowances%20-%20EEX

EON, 2011. New E.ON combined gas and steam turbine power station in Malzenice starts commercial operations. *www.eon.com* [online]. Published Date: 10/01/2011. Available from: http://www.eon.com/en/media/news-detail.jsp?id=10089&year=2011

LANDRY, J.C., 2009. A Robust Optimization Approach to the Self-scheduling Problem Using Semidefinite Programming. Waterloo, Ontario, Canada. Master thesis. University of Waterloo, Management Sciences. Available from: http://uwspace.uwaterloo.ca/handle/10012/4961

ENERGY EXEMPLAR, 2011. PLEXOS® Desktop Edition [program]. Version 6.202 R12. Energy Exemplar Pty Ltd. Available from: http://energyexemplar.com/products/plexosdesktopedition/

PADHY, N. P., 2004. Unit Commitment—A Bibliographical Survey. *IEEE Transactions on Power Systems*. Vol. 19, No. 2, pp. 1196-1205.

SCHMID, J., 2010. Renewable Energies and Base Load Power Plants: Are They Compatible? *Renews Special* [online]. Issue 35. Berlin: German Renewable Energies Agency. Available from: http://www.unendlich-viel-energie.de/en/details/article/523/renewable-energies-and-base-load-power-plants-are-they-compatible.html

SLOVAKIA, 2010. Nariadenie vlady Slovenskej republiky z 26. apríla 2010, ktorým sa mení a dopĺňa nariadenie vlády Slovenskej republiky č. 409/2007 Z. ktorým sa ustanovujú pravidlá pre fungovanie trhu s plynom. In: *Zbierka zákonov č. 212/2010*. S 1769-70. Available from: http://www.urso.gov.sk/doc/legislativa/nv_212-2010_sk.pdf

SEM, 2010. Redpoint Validated Forecast Model and PLEXOS Validation Report 2010. *www.allislandproject.org*. Published Date: 12/4/2010. Available from: http://www.allislandproject.org/en/market_decision_documents.aspx?article=cb4ee33b-a83a-47ce-956a-6cff30900495

VIANA, A., PEDROSO J. P., 2011. A new MILP-based approach for Unit Commitment in power production planning. In: *Proc. of the VII ALIO–EURO – Workshop on Applied Combinatorial Optimization, Porto, Portugal*, May 4–6, 2011, pp. 9-12. Available from: http://paginas.fe.up.pt/~agomes/temp/alio_euro_2011/uploads/Conference/ALIO-EURO_2011_proceedings_v2.pdf

WU, L., 2011. A Tighter Piecewise Linear Approximation of Quadratic Cost Curves for Unit Commitment Problems. *IEEE Transactions on Power Systems*. Vol. 26, No. 4, pp. 2581-2583. Available from: http://people.clarkson.edu/~lwu/J15.pdf

XMTrade, 2011. XMtrade®/ISOT: Yearly DM report. *www.isot.sk* [online]. Available from: http://www.isot.sk/portal/Pages/PublishedInformation/YearlyDMReport/Default.aspx?lang=en

JEL Classification: L94.

Shrnutí

Modelování technických omezení elektrárny pro problém dispečinku

V tomto textu byly popsány charakteristiky elektrárny ovlivňující její dispečerské řízení. Tyto charakteristiky byly zachyceny jako omezení v problému smíšeného celočíselného programování, v němž je elektrárna provozována za účelem maximalizace zisku proti jedinému cenovému scénáři. Tento problém byl zachycen na konkrétní případové studii pro nově postavenou paroplynovou elektrárnu v Malženicích na Slovensku. Výsledky studie potvrzují nutnost zachycení technicko-ekonomických omezení, jako jsou například startovací náklady. Výsledky také ilustrují potřebu přesného zachycení těchto omezení spolu s vhodnou volbou modelovacího horizontu.

Budoucí rozšíření výzkumu se bude soustředit na zlepšení formulace modelu pro zrychlení výpočtu bez ztráty kvality a přesnosti řešení. Mezi další možné okruhy rozvoje patří také zachycení dalších charakteristik, jako je například výroba elektřiny při startu či zapojení elektrárny na trhu s podpůrnými službami.

Klíčová slova: Dispečerské řízení elektráren. Plánování provozu elektráren. Smíšené celočíselné programování. Startovací náklady.

Problematika ambientních daní v ekologii z pohledu teorie her

Martina Zouharová
em.zouharova@gmail.com
Doktorandka oboru ekonometrie a operační výzkum
Školitel: Prof. Ing. Josef Jablonský, CSc., (jablon@vse.cz)

Abstrakt: Článek se zaměřuje na problematiku penalizace plošného znečištění pomocí ambientních daní. Vychází ze základních závěrů publikovaných v odborné literatuře, více však akcentuje strategickou interakci, k níž mezi znečišťovateli při tomto typu zdanění dochází. Skrze herně teoretický přístup jsou demonstrovány některé problematické aspekty dříve publikovaných závěrů; článek nevyvrací existenci Nashovy rovnováhy při hledání optimální míry zdanění, poukazuje však na přítomnost většího počtu takových (vzájemně se nedominujících) rovnováh, což značně komplikuje dosažení zamýšleného výsledku v podobě snížení environmentální zátěže.

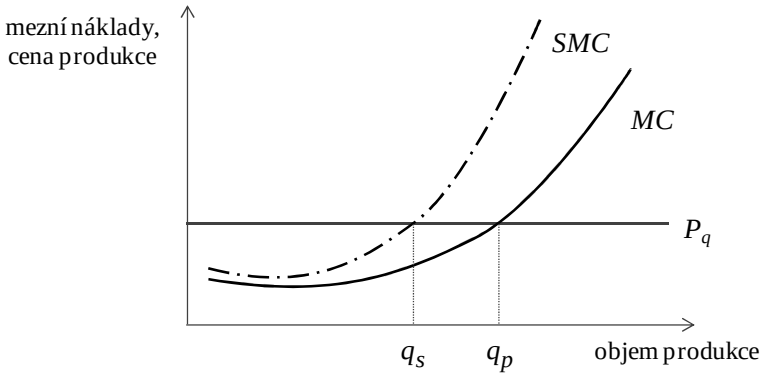
Klíčová slova: ambientní daň, plošné znečištění, teorie her.

Úvod

Rostoucí zatížení životního prostředí vede v posledních desetiletích k stále intenzivnějším snahám o vytvoření regulačních opatření, která by napomohla tento vývoj zmírnit. Obecný trend představuje přechod od prvotně uplatňované *administrativní formy* těchto opatření (závazné předpisy vymezující pravidla chování dotčených subjektů) k regulacím, které jsou založeny na cenovém mechanismu a jako takové odpovídají lépe charakteru tržního hospodářství. Mezi takové nástroje patří *ekologické daně*.

Ekologické daně lze užít v zásadě dvěma způsoby: daňově zvýhodnit to, co považujeme za žádoucí, nebo daňově zatížit, co za žádoucí považovat nelze. Výrazným specifikem tohoto typu daní je požadavek *alokační funkce*. Motivem jejich zavádění tedy není příjem státního rozpočtu, ale snaha o efektivní alokaci zdrojů, kterou trh v případě statků životního prostředí zajistit nedokáže.

Hlavním důvodem selhávání trhu statků životního prostředí je nesprávná kalkulace nákladů znečišťujících firem, které do svých optimalizačních kalkulů zahrnují vlastní mezní náklady (*MC*), nikoli však již mezní náklady plynoucí ze zhoršení životního prostředí (zvýšená míra emisí, hladiny hluku, znečištění podzemních vod atp.). Zdanění má proto tvořit takový stimul, který by přiměl výrobce do jejich optimalizačních úvah zahrnout *celkové společenské mezní náklady* (*SMC*) spojené s výrobou daného statku. Ty jsou dány součtem mezních nákladů výrobce a mezních společenských nákladů ze znečištění. Společensky optimální množství vyráběné komodity zatěžující životní prostředí (q_s) je pak, jak intuice napovídá, menší, než by výrobce vyráběl bez regulačních opatření (q_p). Situaci ilustruje obrázek 1.



Obrázek 1: Rozdíl optimálního množství produkce z pohledu výrobce (q_p) a společnosti (q_s)

Při obecné znalosti všech nákladových funkcí tedy není složité docílit optimální alokace pomocí stanovení daně *t* ve výši rovnající se mezním nákladům na znečištění. Podotkněme však, že znalost všech nákladových funkcí předpokládá nejen dokonalou informovanost regulačního orgánu o nákladech firem, ale rovněž schopnost kvantifikace firmou produkované úrovně znečištění a společenských nákladů s ním spojených. Obé je v praxi problematickým aspektem, s nímž je potřeba se při aplikaci regulační politiky vyrovnat.

Problémy spojené s ekologickými daněmi

V úvodu jsme naznačili, že efekt zavádění ekologických daní je, vzhledem k neznalosti nákladových funkcí, do značné míry diskutabilní. Mezi odborníky se na toto téma vedou spory, v nichž každý své soudy podpírá rozsáhlou argumentací. Zatímco zastánci ekologických daní apelují na pozitivní dopady jak z ekologického, tak ekonomického pohledu (zvýšení efektivnosti odstraněním *mrtvých ztrát*), odpůrci argumentují regresivními následky, ztrátou mezinárodní konkurenceschopnosti atp., blíže viz např. Jílková (2003). Záměrem našeho článku není se vůči těmto argumentům vymezovat, spokojme se proto s faktem, na němž se lze shodnout: na rozdíl od administrativních opatření, která fungují okamžitě a zcela předvídatelným způsobem, je zavádění ekologických daní spojeno s jistou mírou nejistoty ohledně výsledku, k němuž povedou.

Nejistota výsledné reakce představuje jasný motiv pro snahy o ex-ante analýzy očekávaného chování regulovaných subjektů, které by regulačnímu orgánu mohly být alespoň částečným vodítkem k vhodnému učení potřebné míry regulace. Je zřejmé, že přílišná regulace povede k zabrzdění ekonomiky, nedostatečná regulace k neúnosnému zatížení životního prostředí. V případě ekonomiky jsou projevy regrese či expanze relativně dobře pozorovatelné, s životním prostředím je to v mnoha případech však o poznání složitější; podrobný popis teoretických i praktických problémů spojených s vyhodnocením stavu životního prostředí nabízí např. Marriott (1997).

Jeden z těchto komplikovaných případů, na který se v dalším textu zaměříme, představuje plošné znečištění, při němž není možné odlišit individuální emisní příspěvky jednotlivých firem. V takovém případě tradiční pigouovské přístupy zcela selhávají, s čímž je potřeba se vyrovnat. Základní možnost danění je založená na agregovaném množství emisí, jedná o koncept tzv. *ambientních daní*.

Ambientní zdanění plošného znečištění

Ambientní daň představuje specifickou formu ekologické daně spojenou s problematikou plošného znečištění, kdy regulační orgán:

a) je schopen kvantifikovat celkové, prostorově vymezené environmentální zatížení;
b) je schopen identifikovat původce tohoto znečištění;
c) není schopen, popř. pro vysoké administrativní náklady ochoten určit individuální participaci těchto subjektů na celkovém znečištění.

Podstata ambientních daní spočívá v penalizaci či subvenci odchylky celkového znečištění od společensky žádoucí úrovně. V odborné literatuře se zpravidla vychází z konceptu ambientního zdanění navrženého v práci Segersonové (1988), která kombinovala poplatky a subvence následujícím způsobem:

a) subvence přímo úměrná množství emisí, o které emituje subjekt méně, nežli je společensky žádoucí úroveň;
b) daň přímo úměrná množství emisí, o které emituje subjekt více, nežli je společensky žádoucí úroveň;
c) jednorázová pokuta za překročení stanovené úrovně emisí, tj. emisního limitu.

Toto schéma je pak v literatuře různým způsobem modifikováno a dále rozvíjeno. Hansen (1998) navázal zdanění namísto emisí na kvantifikaci jejich konkrétních dopadů, Xepapadeas (1991) diskutoval dopady zavedení penalizace při nesplnění stanoveného limitu vždy pro jednu z náhodně vybraných firem. Hanley, Shogren a White (1997) se pokusili myšlenku ambientního zdanění rozšířit i pro bodové zdroje znečištění. Herriges, Govindasamy a Shogren (1994) pak diskutovali vliv různého vztahu k riziku zúčastněných firem. Další články se zabývají možnostmi případné spolupráce firem (Millock a Salanié, 2005) apod. Obecným rysem je snaha o formulaci optimální daňové sazby (*t*) a případné

jednorázové pokuty (*K*) spojené s překročením vytčeného emisního limitu. Zpravidla se jedná o diskusi tří případů [viz např. Segerson (1988) nebo Hanley, Shogren a White (2007), s. 92]:

a) jak nastavit optimální velikost *t* při $K = 0$,
b) jak nastavit optimální velikost *K* při $t = 0$,
c) jak nastavit optimální velikost *K* pro libovolné *t*.

Pro první případ se Segersonové (1988) podařilo ukázat, že pokud jsou společenské náklady spojené se znečištěním lineární, pak je při $K = 0$ žádoucí nastavit pro všechny firmy shodnou daň rovnou mezním společenským nákladům – racionálně se rozhodující firmy potom budou volit společensky optimální výši znečištění. V případech (b) i (c) Segersonová rovněž doporučuje konkrétní postup pro nastavení optimální míry *t* či *K* s tím, že za předpokladu racionality regulovaných subjektů dojde z jejich strany k volbě množství odpovídajícího odvozené Nashově rovnováze (*Nash equilibrium*, NE); v případě (b) však při takto vystavěné argumentaci narážíme na jisté komplikace, na něž se dále zaměříme.

Modely pro zkoumání dopadů ambientních daní

Jak již bylo naznačeno, ambientní zdanění závisí namísto individuální míry emisí jednotlivých znečišťovatelů na jejich agregovaném množství. Rozhodujícím faktorem pro udělení pokuty *K* je tedy kombinace zvolených úrovní emisí všech dotčených firem. Jazykem teorie her, „výplata jednotlivých hráčů je odvislá od zvolených strategií zbylých hráčů“. Na případ hledání optimálního nastavení pokuty *K* při nulové sazbě *t* se tedy dále pokusíme nahlížet jako na nekooperativní typ konfliktu.

Při diskusi očekávaných strategií zúčastněných hráčů vyjdeme nejprve ze základního jednoduchého modelu dvou firem, které činí prostřednictvím změn objemu vlastní výroby rozhodnutí, zda emitovat hodně či málo. Zjednodušení v podobě diskrétního prostoru strategií není našemu záměru na závadu, neboť i z takto jednoduchého modelu v podobě maticové hry typu 2×2 je podstata našich námitek k závěrům Segersonové dobře patrná.

V dalším modelu přejdeme od diskrétního prostoru strategií ke spojitému. Díky tomuto zobecnění se jednak o něco více přiblížíme realitě, zejména však bude možné spojitý model principielně ztotožnit s modely uváděnými v námi citované literatuře (Segerson, 1988, Hanley, Shogren a White, 2007, s. 90). Jediným podstatným rozdílem je, že námi předkládaný model je zcela deterministický; konkrétněji řečeno, nepracuje s vlivem náhodných exogenních šoků na míru ambientního znečištění.

Diskrétní model 2×2

Uvažujme model dvou firem maximalizujících zisk, z nichž každá může prostřednictvím změn objemu výroby rozhodovat, zda emitovat hodně (*H*, resp. *h*), nebo málo (*L*, resp. *l*). Předpokládejme dále, že emise obou firem podléhají shodnému ambientnímu zdanění. Snahou regulačního orgánu je přimět firmy k chování v podobě jedné z kombinací (*H*,*l*), nebo (*L*,*h*), při níž je dosaženo žádoucího omezení emisí, zároveň však není příliš zatížena ekonomika. Tímto způsobem definovaný problém můžeme zapsat v podobě maticové hry (viz obrázek 2).

	l	*h*
L	a_2 a_1	$a_2 + b_2$ a_1
H	a_2 $a_1 + b_1$	$a_2 + b_2 - K$ $a_1 + b_1 - K$

Obrázek 2: Maticová hra v modelu 2×2, nulová daň *t* a shodná pokuta *K*.

V takto formulované hře představuje a_i úroveň zisku spojenou se strategií sníženého objemu emisí, b_i přírůstky zisku při neomezení emisí a *K* pokutu v případě, že celkové množství emisí překročí stanovený, společensky přijatelný limit (hodnoty koeficientů b_1, b_2, stejně jako pokutu *K* považujeme automaticky za kladné). V této hře lze pro jednotlivé případy dospět k podmínkám Nashovy rovnováhy v podobě:

Tabulka 1: Dosažitelnost možných NE.

NE	podmínky
(L,l)	nelze
(L,h)	$b_2 \leq K$
(H,l)	$b_1 \leq K$
(H,h)	$b_1 \geq K \leq b_2$

Podotkněme, že může nastat více NE současně, jsou-li splněny podmínky pro několik NE najednou. Je zřejmé, že možné kombinace NE závisí na poměru přírůstků účelových funkcí při zvýšení produkce, tj. na vztahu b_1 a b_2. Níže uvedené tabulky shrnují výsledky pro varianty $b_1 = b_2$ (tabulka 2) a $b_1 < b_2$ (tabulka 3).

Tabulka 2: Možné kombinace NE při $b_1 = b_2$.

NE	podmínky
(H,h)	$K < b_1 = b_2$
$(H,h), (H,l), (L,h)$	$K = b_1 = b_2$
$(H,l), (L,h)$	$K > b_1 = b_2$

Tabulka 3: Možné kombinace NE při $b_1 < b_2$.

NE	podmínky
(H,h)	$K < b_1 < b_2$
$(H,h), (L,h)$	$K = b_1 < b_2$
(L,h)	$b_1 < K < b_2$
$(L,h), (H,l)$	$b_1 < b_2 \leq K$

Na základě tabulek 2 a 3 je možné dojít ke dvěma závěrům:

a) pokud má být pokuta stimulem pro snižování emisí, musí být, jak ostatně intuice napovídá, dostatečně vysoká;
b) ani dostatečně vysoká pokuta nedává firmě vzhledem k existenci více NE, které se vzájemně nedominují, návod, jak se zachovat. [Jak poznamenává Binmore (2007, s. 26), z racionálního hlediska lze obhájit i volbu paretovsky dominovaného NE, v případě nedominovaných NE je absence racionální kombinace strategií o to zřejmější.]

Druhý z uvedených závěrů je při snahách o tvorbu účinné environmentální politiky poměrně závažnou komplikací, neboť rapidně zvýší míru nejistoty ohledně reakcí regulovaných subjektů. Specifický případ by představovala situace, kdy by ziskové funkce obou firem byly známy a pokutu by bylo možné nastavit tak, aby náležela intervalu ohraničeného přírůstky b_1 a b_2. V takovém případě by bylo možné předpokládat, že regulační orgán je schopen svou intervencí zajistit společensky žádoucí úroveň emisí na úkor jednoho z účastníků; v kontextu modelu 2×2 by pak existovala jediná NE, která by představovala jediné racionální řešení pro obě firmy (viz předposlední řádek v tabulce 3).

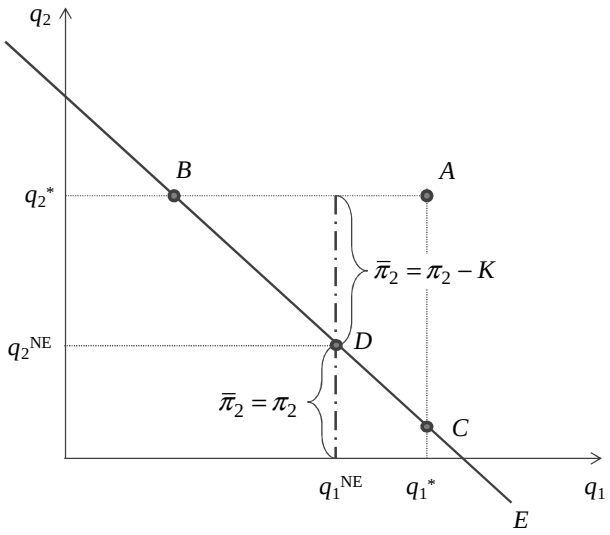
Znalost ziskových funkcí je však v praktických aplikacích diskutabilní. Nejpodstatnějším faktem tak zůstává zpochybnění přínosu závěru o existenci NE v případě jejich většího počtu, kdy lze očekávat snahu firem být *černým pasažérem*, tj. volit strategii odpovídající takovému NE, které se jim jeví jako výhodnější. Volí-li však takovou strategii obě firmy, je zřejmé, že k zamýšlenému omezení environmentální zátěže nedojde.

Model se spojitým prostorem strategií

Opusťme předpoklad diskrétního prostoru strategií. Shodně s předchozím modelem uvažujeme dvě firmy podléhající shodnému ambientnímu danění, kdy překročení emisního limitu E je spojeno se shodnou pokutou pro obě firmy, a sice ve výši K. Lze předpokládat, že snahou firem bude (při dostatečně velkém K) maximalizovat objem produkce, avšak s ohledem na nepřekročení nastaveného li-

mitu. Situaci ilustruje obrázek 3. Na osu *x* vynášíme objem produkce vyráběné první firmou (q_1), na osu *y* množství vyráběné druhou firmou (q_2). Optimální množství vyráběné firmami v případě neexistence pokuty označme jako q_1^*, q_2^*. Této kombinaci odpovídá v grafu bod *A*. Kombinace produkce (q_1, q_2), při nichž jsou emise spojené s produkcí obou firem rovny emisnímu limitu *E*, představuje v grafu shodně označená přímka. Slovy ekonoma by bylo možné přímku *E* označit za *izoemitu* (izoemita znázorněná v grafu má lineární tvar, což není na úkor obecnosti – analýza by probíhala zcela identicky pro jakoukoli ryze klesající izoemitu).

Ukážeme, že NE budou ležet právě na této izoemitě. Vzhledem k předpokladu, že ani jedna z firem nebude chtít vyrábět více, než odpovídá množství q_1^*, resp. q_2^*, je zřejmé, že Nashovým rovnováhám nebude odpovídat celá přímka, ale pouze její část ohraničená v grafu body *B* a *C*. Další text zaměříme na ověření tohoto tvrzení.



Obrázek 3: Kombinace strategií v modelu se spojitým prostorem strategií

Zvolme libovolný bod úsečky *BC*, např. bod *D* se souřadnicemi (q_1^{NE}, q_2^{NE}). Aby byl tento bod NE, musí platit, že při dané volbě výstupu první firmy ve výši q_1^{NE} si druhý hráč nemůže polepšit jinou volbou nežli je q_2^{NE}. Předpokládejme, že před zavedením pokuty měla zisková funkce druhého hráče (π_2) tradiční průběh:

$$\frac{\partial \pi_2}{\partial q_2}(q_2, q_1^{NE}) > 0, \quad \frac{\partial^2 \pi_2}{\partial q_2^2}(q_2, q_1^{NE}) < 0 \qquad \text{pro } q_2 \in (0, q_2^*).$$

Zisková funkce po zavedení pokuty, $\bar{\pi}_2$, bude mít tvar:

$$\bar{\pi}_2(q_2, q_1) = \begin{cases} \pi_2(q_2, q_1), & \text{není-li překročena úroveň emisí } E, \\ \pi_2(q_2, q_1) - K & \text{jinak.} \end{cases} \tag{1}$$

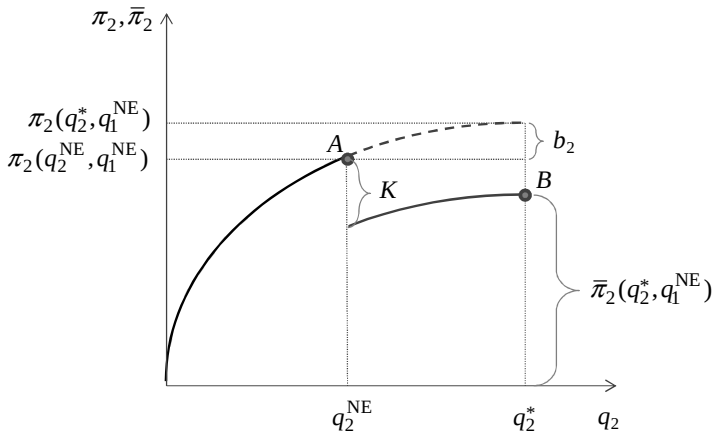
Obrázek 3 ilustruje situaci pro $q_1 = q_1^{NE}$ a $q_2 \in (0, q_2^*)$, tj. pro body na čerchované úsečce. Po zavedení pokuty tedy bude zisk při daném q_1^{NE} růst do q_2^{NE}, a poté se skokově sníží o hodnotu *K* (viz obrázky 4 a 5). Aby si druhý hráč nemohl polepšit volbou jiné nežli rovnovážné strategie (q_2^{NE}), pak musí platit:

$$\pi_2(q_2^*, q_2^{NE}) - K < \pi_2(q_2^{NE}, q_1^{NE}),$$

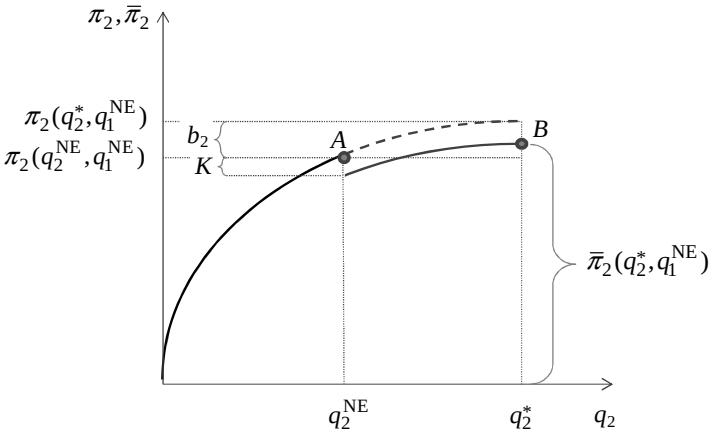
resp.

$$\pi_2(q_2^*, q_1^{NE}) - \pi_2(q_2^{NE}, q_1^{NE}) < K. \tag{2}$$

Jedná se o analogii požadavku, k němuž jsme dospěli v případě modelu 2×2 s diskrétním prostorem strategií; přírůstek zisku spojený s větší výrobou (b_i) musí být menší než hrozící pokuta. Podíváme-li se na obrázek 4, pak splnění podmínky (2) znamená, že bod *A* musí ležet výše nežli bod *B*, v opačném případě (viz obrázek 5) by firma vyráběla shodné množství jako v případě neexistence pokuty.



Obrázek 4: Řez ziskové funkce druhého hráče v q_1^{NE}, $b_2 < K$.



Obrázek 5: Řez ziskové funkce druhého hráče v q_1^{NE}, $b_2 > K$.

Na základě dosud uvedeného lze konstatovat, že všechny Nashovy rovnováhy v případě dvou firem a spojitého prostoru strategií opravdu leží na úsečce *BC* z grafu 1. Vzhledem k námi uvažovanému průběhu ziskové funkce je dále zřejmé, že jiné NE najít nelze. Pokud bychom měli tento závěr více zobecnit, lze říci, že: Nashovou rovnováhu představují všechny kombinace objemů výroby, které vyústí v environmentální zatížení rovné nastavenému limitu. Důležitým problémem tedy zůstává počet těchto rovnováh, který dává tušit velmi diskutabilní úspěšnost případné ambientní daně, resp. pokuty. Jak vidno, v případě modelu se spojitým prostorem strategií jsme tedy dospěli ke zcela analogickým závěrům s předchozím modelem typu 2×2.

Závěr

Za obecnou nevýhodu environmentálních incentiv založených na cenovém mechanismu bývá považována vysoká míra nejistoty ohledně reakcí dotčených subjektů. V případě bodového znečištění tato nejistota vyplývá z neschopnosti kvantifikovat a určit optimální úroveň společensky přijatelného znečištění, na jejímž základě by bylo možné stanovit potřebnou výši daně. V případě penalizace plošného znečištění však nejistota spojená s reakcí regulovaných firem získává ještě další rozměr. Náhle se totiž nedotýká pouze regulačního orgánu, ale všech zúčastněných firem, jejichž zisk je individuálními reakcemi vzájemně determinován. Při pokusech o ex-ante analýzu dopadů tohoto typu zdanění je proto nutné strategické interakci mezi firmami věnovat dostatečnost pozornost.

Ve stěžejní práci týkající se ambientního danění (Segerson, 1988), na jejíchž závěrech staví další zmiňované články či učebnice (Hansen, 1998, Hanley, Shogren a White, 2007 a další), je odvozena Nashova rovnováha pro optimální nastavení jednorázové pokuty. V našem modelu se podařilo ukázat, že nedominovaných rovnovážných řešení existuje v tomto případě celá řada, neboť za Nashovu rovno-

váhu lze považovat všechny kombinace výroby, při nichž jsou agregované emise rovny stanovenému limitu.

Nejednoznačnost řešení značně ohrožuje deklarovanou alokační funkci ambientní daně. Je opodstatněné se domnívat, že pokuta založená na agregovaných údajích nebude dostatečným stimulem pro omezení emisí, a to bez ohledu na její výši.

Námi předložený model se oproti v předchozích oddílech citovaným článkům liší v tom, že neuvažujeme náhodné exogenní vlivy na znečištění. Tento přístup je možné odůvodnit předpokladem, že firmy při svém strategickém rozhodování náhodné vlivy rovněž neuvažují, tj. rozhodují se ve shodě s naším modelem. Analýza existence alternativních NE za přítomnosti náhodné složky by byla technicky náročnější a bude předmětem dalšího výzkumu.

Literatura

BINMORE, Ken. *Playing for Real*. New York: Oxford University Press, 2007. 656 s. ISBN 978-0-19-530057-4.

HANLEY, Nick, Jason F. SHOGREN a Ben WHITE. *Environmental Economics in Theory and Practice*. 2. přeprac. vyd. New York: Palgrave Macmillan, 2007. 459 s. ISBN 978-0-333-97137-6.

HANSEN, Lars Gårn. A damage based tax mechanism for regulation of non-point emissions. *Environmental and Resource Economics*. 1998, **12**(1), s. 99-112. ISSN 0924-6460.

HERRIGES, Joseph A., Ramu GOVINDASAMY, a Jason F. SHOGREN. Budget-balancing incentive mechanisms. *Journal of Environmental Economics and Management*. 1994, **27**(3), s. 275-285. ISSN 0095-0696.

JÍLKOVÁ, Jiřina. *Daně, dotace a obchodovatelná povolení – nástroje ochrany ovzduší a klimatu*. Praha: IREAS, 2003. 156 s. ISBN 80-86684-04-0.

MARRIOTT, Betty B. *Environmental Impact Assessment: A Practical Guide*. New York: McGraw-Hill, 1997. 320 s. ISBN 978-0070404106.

MILLOCK, Katrin a François SALANIÉ. Nonpoint source pollution regulation when polluters might cooperate. *Topics in Economic Analysis & Policy* [online]. 2005, **5**(1) [cit. 2011-12-12]. ISSN 1538-0653. Dostupné z: http://www.bepress.com/bejeap/topics/vol5/iss1/art12.

SEGERSON, Kathleen. Uncertainty and incentives for nonpoint pollution control. *Journal of Environmental Economics and Management*. 1988, **15**(1), s. 87-98. ISSN 0095-0696.

XEPAPADEAS, Anastasios. Environmental policy under imperfect information: Incentives and moral hazard. *Journal of Environmental Economics and Management*.1991, **20**(2), s. 113-126. ISSN 0095-0696.

JEL Classification: H23, Q58.

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu IGA F4/18/2011.

Summary

Ambient Taxation in Ecology from the Game-Theoretical Perspective

Ambient taxation has been proposed by many authors as a means of regulating non-point pollution. This paper stresses the fact that the extant literature fails to fully embrace the issues connected with strategic interaction of the individual polluters. More specifically, although the existence of a socially desirable Nash-equilibrium is undoubted, we argue that the studies presented so far neglect the fact that in typical problems, multiple Nash equilibria exist with no Pareto domination amongst them, which renders the behaviour of individual polluters unpredictable and thus complicates the achievement of the desired policy effect.

Keywords: ambient taxation, non-point pollution, game theory.

Předpovědi extrémních říčních průtoků metodou „peaks over threshold“ a jejich kvalita

Adam Čabla
adam.cabla@vse.cz
Doktorand oboru statistika
Školitel: prof. Ing. Richard Hindls, CSc., dr.h.c.

Abstrakt: Článek se zabývá chováním odhadů metodou peaks over thresholed (POT, špiček nad prahy) při odhadování extrémních říčních průtoků.

Za základní soubor jsou považovány zaznamenané hodnoty průměrných denních říčních průtoků nejmenované řeky v Evropě a z tohoto souboru jsou vytvářeny realistické klouzavé výběrové soubory různých délek (5, 10, 20, 40 a 80 let).

Průměrné denní říční průtoky porušují podmínku nezávislosti a první část výpočtů se zabývá tím, co se stane, když se chvost takovéhoto rozdělení modeluje klasickou metodou odhadu parametrů zobecněného Paretova rozdělení bez předvýběru dat.

Odhadovací metodou je metoda maximální věrohodnosti a prahem je vždy 90% kvantil výběrového souboru.

Pro druhou část výpočtů je potřeba provést výběr dat tak, aby splnily podmínku nezávislosti. Toho je v článku docíleno standardním postupem, při kterém jsou v čase po sobě následující hodnoty přesahující práh brány jako jeden ucelený shluk pozorování a zaznamenáno maximum tohoto shluku. Souhrn těchto maxim na časové ose je poté novým základním souborem.

Práh pro určení těchto shluků byl položen tak, aby jej překročilo právě 1 000 pozorování.

Výsledky ukazují, že metoda POT bez úpravy dat nadhodnocuje odhady extrémních říčních průtoků a metoda po úpravě dat vykazuje výrazně lepší vlastnosti, byť je stále náchylná na výskyt extrémů ve výběrovém souboru.

Klíčová slova: peaks over threshold, zobecněné Paretovo rozdělení, říční průtoky, odhady

Úvod

Teorie extrémních hodnot (*extreme value theory*) je dnes často užívanou částí matematické statistiky, jež se zabývá chováním chvostů rozdělení. Jinými slovy to znamená, že se zabývá stochastickým chováním například vysokých či nízkých (extrémních) teplot, vysokých srážek či ve financích pak maximální ztráty, která může s danou pravděpodobností nastat za daný časový úsek.

Jednou z významných oblastí užití teorie extrémních hodnot v hydrologii je potom pravděpodobnostní odhad nejvyššího průtoku nějakého toku za dané časové období. V takovém případě mluvíme podle délky období o odhadu např. padesátileté, stoleté či tisícileté vody.

Cílem článku je demonstrovat užití jedné ze dvou základních metod teorie extrémních hodnot, nazývanou peaks over threshold (*POT*; *metoda špiček nad prahem*), k odhadu extrémních hodnot průtoku nejmenované evropské řeky na reálných datech z 20. století, přesnost a stabilitu těchto odhadů.

Metoda POT vyžaduje nezávislost jednotlivých pozorování, což je problém, se kterým je potřeba se u závislých pozorování vyrovnat úpravou výběru dat.

Použitá metodologie

Při modelování extrémů existují dvě základní metody. Metoda blokových maxim (*block maxima method)* vychází z hodnot maxim (analogicky i minim) dané veličiny v náhodných intervalech, většinou časových, přičemž rozdělení těchto maxim konverguje podle Fisher-Tippett-Gnedenkova teorému k zobecněnému rozdělení extrémních hodnot (*generalized extreme value distribution - GEV*). Její užití při řešení problému stanovení extrémních hodnot říčních průtoků vychází ze stanovení maximálních

průtoků v jednotlivých letech a následně v odhadu parametrů GEV. Z kvantilové funkce GEV lze potom určit např. stoletou vodu jako 99% kvantil (Vojtěch, 2010).

Výše zmíněná metoda má jednu zásadní nevýhodu, kterou je potenciální ztráta informací o chování chvostu rozdělení, neboť z každého intervalu vybírá pouze jednu hodnotu. V některých případech tato maximální hodnota může být následována dalšími extrémními hodnotami, vyššími než třeba jiná intervalová maxima, které ovšem nejsou do modelu zahrnuty. Tento problém se často vyskytuje právě u modelování extrémních říčních průtoků, proto se zde výrazně častěji užívá metoda POT za předpokladu, že je k dispozici dostatečně dlouhá časová řada průměrných denních průtoků. Jelikož se touto metodou nebudu v práci zabývat, veškerá dále zmíněna metodologie se týká metody peaks over threshold.

Zobecněné Paretovo rozdělení

Metoda POT vyhází z Pickands-Balkema-de Haanova teorému, dle kterého hodnoty stejně rozdělené náhodné veličiny přesahující určitý, dostatečně vysoký, práh (*threshold)* konvergují k zobecněnému Paretovu rozdělení (generalized Pareto distribution – GPD).

Uvažujeme nezávislé, stejně rozdělené náhodné veličiny X s obecně neznámou distribuční funkcí F. Náhodné veličiny, pro které platí X > u, mají podmíněnou přesahovou (excess) distribuční funkci

$$F_u(y) = P(X - u \le y \mid X > u)$$

pro $0 \le \omega_F - u$,

kde X je náhodnou veličinou, u je daný práh, y = x – u jsou pozorování nad prahem a $\omega_F \le \infty$ *je pravý koncový bod distribuční funkce F. Podle Pickands-Balkema-de Haanova teorému pro velkou třídu distribučních funkcí platí:*

$$F_u \to H_{\xi,0,\beta} = 1 - \left(1 + \xi \frac{y}{\beta}\right)^{-1/\xi}$$

pro $u \to \infty$.

Parametr ξ určuje chování chvostu distribuční funkce F a zobecněné Paretovo rozdělení může podle jeho hodnoty nabývat jednoho ze tří následujících rozdělení:

- Pro $\xi < 0$ je to rozdělení Paretovo.
- Pro $\xi = 0$ je to rozdělení exponenciální.
- Pro $\xi > 0$ je to rozdělení beta.

Metoda peaks over threshold

Jak již bylo řečeno výše, za určitých podmínek konverguje rozdělení hodnot náhodné veličiny přesahující dostatečně vysoký práh (tedy *chvost rozdělení*) k zobecněnému Paretovu rozdělení. Toho využívá metoda POT, jejíž základní postup se dá jednoduše shrnout do dvou následujících kroků:

- určení dostatečně vysokého prahu *u*,
- odhad parametrů GPD pro přesahy prahu *u*, tedy rozdíly mezi původními hodnotami a prahem

Se znalostí parametrů je pak zjišťování hodnot odhadů kvantilů rozdělení, odpovídajících jednotlivým extrémním událostem, pouhým dosazováním do kvantilové funkce.

Určení prahu *u* se v praxi často dělá jednou z grafických metod, jejichž přehled je např. v článku (Tanaka a Takara, 2002). Pro značný rozsah odhadů jsem se rozhodl k přímočařejšímu postupu určení prahu. Prahem je v tomto případě 90% kvantil.

K odhadu parametrů zobecněného Paretova rozdělení lze užít různých metod, např. metody CME a de Haan jsou popsány v mém článku (2011), jinými odhadovacími metodami jsou metoda pravděpodobnostně vážených momentů v (Vojtěch, 2011) nebo momentová metoda. Zde užitou metodou odhadu je metoda maximální věrohodnosti, maximalizující věrohodnostní funkci

$$L(\xi,\beta,\mathbf{x}) = \prod_{i=1}^{k}\left[\frac{1}{\beta}\left(1+\xi\frac{x_i}{\beta}\right)^{-1/\xi-1}\right].$$

Samotná maximalizace vede k soustavě dvou rovnic, jež musí být řešeny numericky. V práci nacházená řešení pocházejí z algoritmu Nelder-Meadovy simplexové metody zabudované v numerickém prostředí programu R.

Metoda maximální věrohodnosti vede k odhadům, které jsou konzistentní a asymptoticky vydatné pro $\xi > -1/2$, odhady pro menší hodnoty ξ nebývají asymptoticky vydatné a pro hodnoty $\xi < -1$ jsou často dokonce nenalezitelné (Vojtěch, 2011). Přehled vlastností některých odhadů lze nalézt např. v (Castillo, Hadi, 1997).

V dalším průběhu práce budu demonstrovat výsledky užití metody POT pro data, jež nesplňují podmínku nezávislosti jednotlivých hodnot, jelikož průměrné denní říční průtoky jsou veličiny na sobě závislé. Aby bylo docíleno nezávislosti jednotlivých hodnot, musí být výběr dat částečně upraven.

Výběr dat pro závislé hodnoty v časových řadách

Metoda POT se často užívá v časových řadách, kde nelze předpokládat nezávislost jednotlivých po sobě jdoucích hodnot. Tento problém vede ke zkreslení odhadů klasickou metodou POT a k jeho odstranění vedou v zásadě velice podobné metody výběru dat. Tyto metody mívají následující kroky:

- určení hodnot nad zvoleným prahem,
- určení maxim těchto hodnot v jednotlivých shlucích pozorování přesahujících práh,
- užití těchto vybraných maxim pro odhad parametrů GPD.

Shlukem hodnot se zde míní v čase na sebe navazující pozorování jednotlivých hodnot přesahujících práh. Předpokládá se pak, že takovéto shluky jsou vytvořeny jedním procesem (např. vydatným deštěm) a jejich hodnoty jsou tak na sobě silně závislé zatímco hodnoty mezi shluky jsou nezávislé. Je možné formulovat různá pravidla pro určení, kdy ještě hodnota patří do stejného shluku jako jiná a kdy ne, popřípadě rozhodovat subjektivně. Často se pracuje s nejnižší hodnotou mezi dvěma lokálními maximy hodnot.

Může také přirozeně nastat situace, kdy dva shluky na sebe navazují a nemusí být nutně rozeznatelné. V takovém případě stojí výzkumník před volbou vypustit pozorování a snížit tak potenciálně kvalitu modelu, nebo naopak pozorování přijmout a riskovat zařazení závislých hodnot.

Podobnými postupy se většinou u průtoků řek získávají u 2-3 hodnoty za rok, někteří autoři (Claps et al., 2003) pak uvádějí až 5-6 hodnot.

Ve druhé části práce užívám pro určení modelovaných hodnot následující postup:

- Práh byl určen jako taková hodnota, kterou přesáhne 1 000 pozorování (cca. 2,74 % všech hodnot)
- Za shluky byly určeny všechny na sebe navazující popřípadě osamělé hodnoty nad prahem
- V každém shluku bylo určeno maximum, které bylo vybráno pro modelování
- V některých shlucích byla vybrána dvě maxima, pro která platilo, že nejnižší v čase mezi nimi umístěná pozorovaná hodnota byla alespoň o 1/3 nižší než nižší z maxim.

Data

Použitým datovým souborem jsou denní údaje o průměrném průtoku nejmenované evropské řeky. Průtoky jsou vždy udávány v m^3/s. Tato data jsou k dispozici za roky 1887 – 2008, ze kterých byly vybrány údaje z 20. století, tj. let 1901 – 2000. Toto považuji za základní soubor, byť to může přinášet určité obtíže.

Cílem této práce je zhodnotit vhodnost modelů POT pro modelování extrémních říčních průtoků tak, jak k tomuto modelování dochází ve skutečnosti. Domnívám se, že k tomuto cíli nezbytně patří srovnat odhady se skutečnými hodnotami, byť tyto skutečné hodnoty můžou být zkreslené vůči potenciální funkci, kterou se extrémní říční průtoky řídí.

Následně byly vytvořeny výběrové soubory – pro první, neupravenou, metodu to byly soubory pětileté, desetileté a dvacetileté, pro druhou metodu pak soubory dvacetileté, čtyřicetileté a osmdesátileté. Tyto soubory nebyly vytvořeny náhodným výběrem, ale tak, jak můžou reálně vznikat – tedy se vždy jednalo o po sobě jdoucí hodnoty základního souboru, které na sebe po jednotlivých letech klouzavě navazovaly. Např. pětileté soubory obsahují po sobě údaje z let 1901 – 1905, 1902 -1906, 1903 – 1907 až 1996 – 2000.

Následují stručné charakteristiky základního souboru.

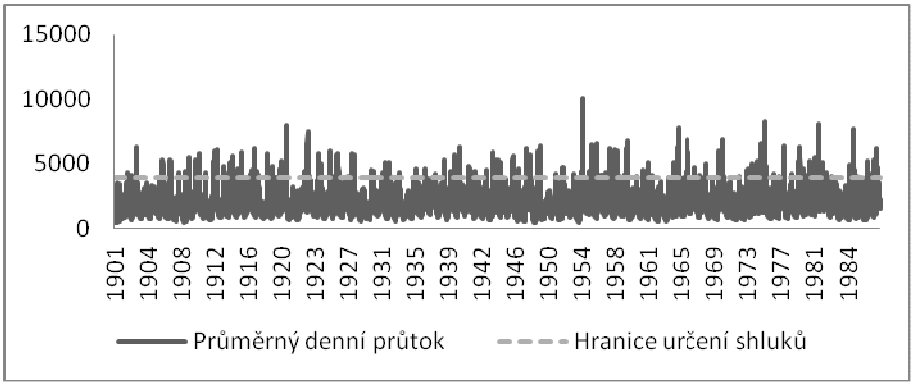
Tabulka 1: Pět největších hodnot maxim jednotlivých shluků Zdroj: vlastní práce

Pořadí	**1.**	**2.**	**3.**	**4.**	**5.**
Hodnota	10 092	9 479	8 267	8 116	7 968
Rok	1954	1991	1975	1981	1920

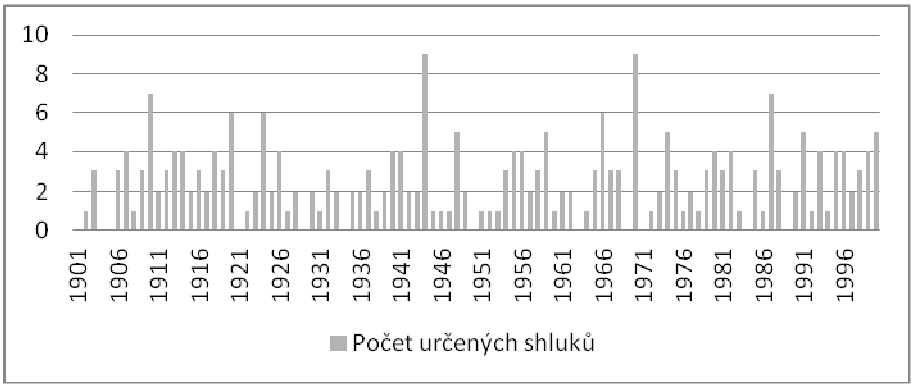
Z první tabulky vyplývají také hodnoty, které jsou v práci brány za *skutečnou* stoletou, padesátiletou a pětadvacetiletou vodu – jsou to postupně 10 092, 9 479 a 8 116. Ve druhé tabulce jsou nastíněny vybrané základní hodnoty základního souboru a v prvním obrázku je zobrazen vývoj hodnot průměrných denních průtoků v čase spolu s hranicí užitou při určování shluků (3 916).

Tabulka 2: Vybrané hodnoty základního souboru Zdroj: vlastní práce

	Hodnota
Počet	36 523
Průměr	1 844,802
$x_{0,25}$	1 209
Medián	1 664
$x_{0,75}$	2 272,5
$X_{0,9}$	2 944



Obrázek 1: Vývoj hodnot průměrných denních průtoků Zdroj: vlastní práce



Obrázek 2: Počet určených shluků v jednotlivých letech Zdroj: vlastní práce

Počet shluků vybraných při druhé užité metodě je 255, což dává v průměr 2,55 hodnot na jeden rok časové řady. Počty vybraných hodnot v jednotlivých letech jsou ukázány ve druhém obrázku.

Provedené experimenty

V první části této kapitoly představím výsledky výpočtu z celých souborů bez úpravy dat, ve druhé části pak výpočty z upravených dat.

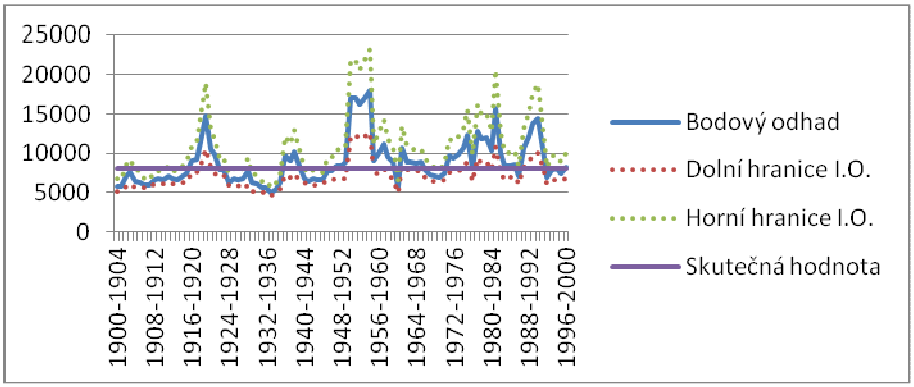
Pro hodnocení zkreslení užívám intervalové odhady z výběrových souborů a počet případů, ve kterých tyto odhady zahrnují hodnotu daného extrémního průtoku v základním souboru.

Ve výpočtech užívám pojmy m-letá voda (m = 25, 50 a 100). Jedná se o minimální průměrný denní průtok v m^3/s, který je pozorován v průměru jednou za *m* let. Odhad se provádí jako odhad p% kvantilu o hodnotě P = 1 – (100/m)/36525

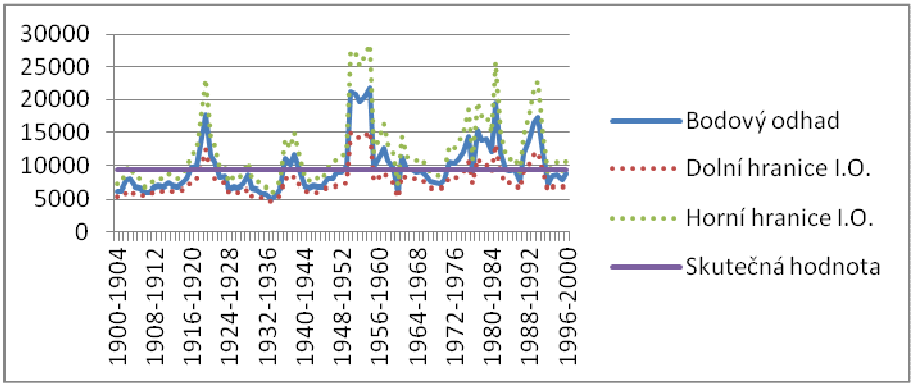
Odhady z neupravených dat

Následující obrázky 2 – 5 představují bodové a intervalové odhady (spolehlivost je 95 %)získané z jednotlivých pětiletých výběrových souborů, obrázky 6 – 8 odhady z desetiletých výběrových souborů a obrázky 9 – 11 odhady z dvacetiletých výběrových souborů. Odhady jsou dělány vždy pro stoletou, padesátiletou a pětadvacetiletou vodu. Zkratka I.O. značí hranice intervalového odhadu.

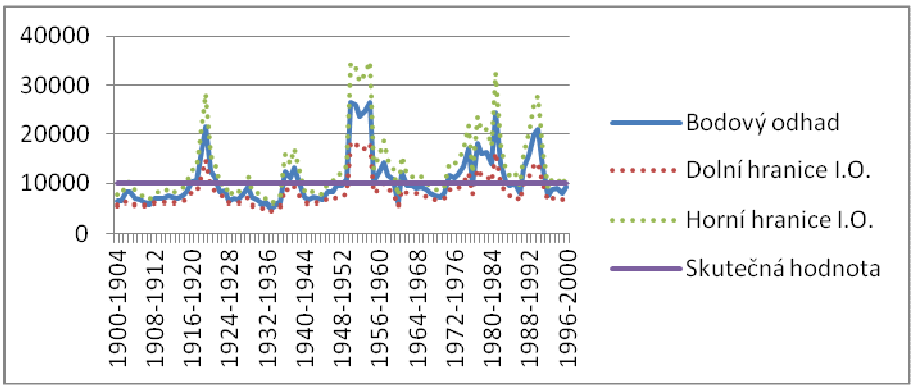
Dále následují tabulky 3 – 5 udávající vývoj přesnosti odhadů pro jednotlivé extrémní průtoky podle rostoucího počtu let výběrových souborů. Zkratka B.O. značí bodový odhad.



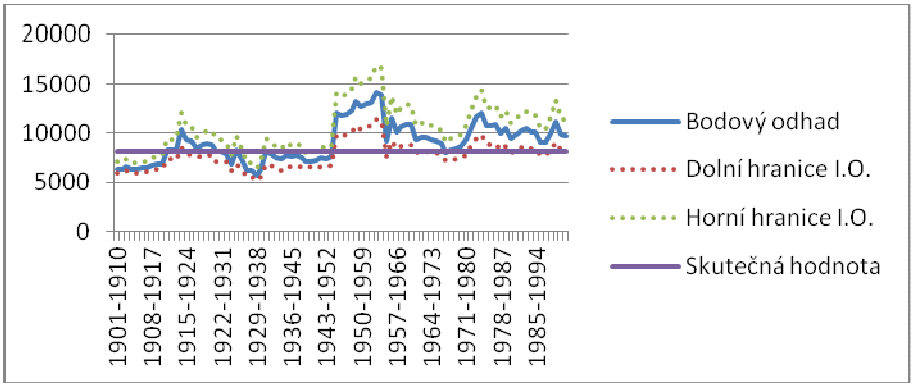
Obrázek 3: Odhad pětadvacetileté vody z pětiletých souborů Zdroj: vlastní práce



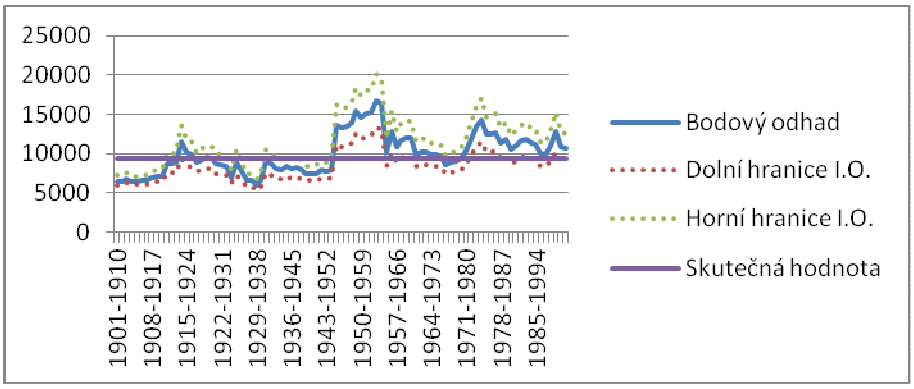
Obrázek 4: Odhad padesátileté vody z pětiletých souborů Zdroj: vlastní práce



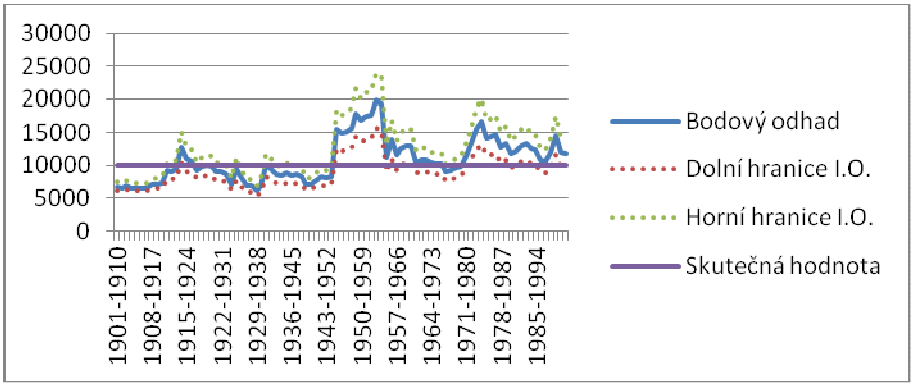
Obrázek 5: Odhad stoleté vody z pětiletých souborů Zdroj: vlastní práce



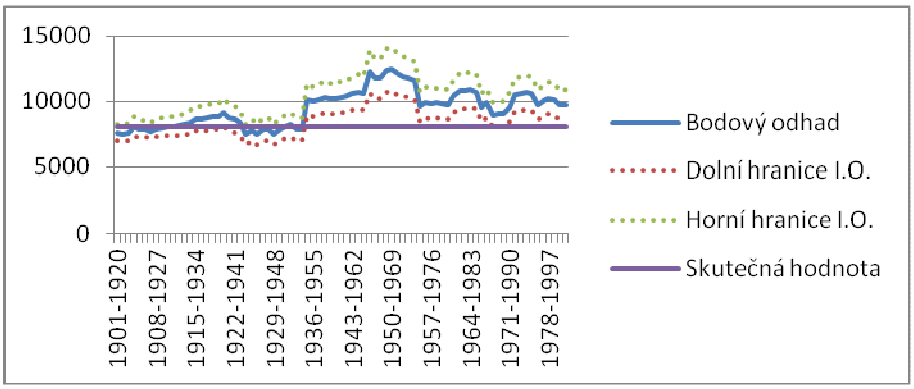
Obrázek 6: Odhad pětadvacetileté vody z desetiletých souborů Zdroj: vlastní práce



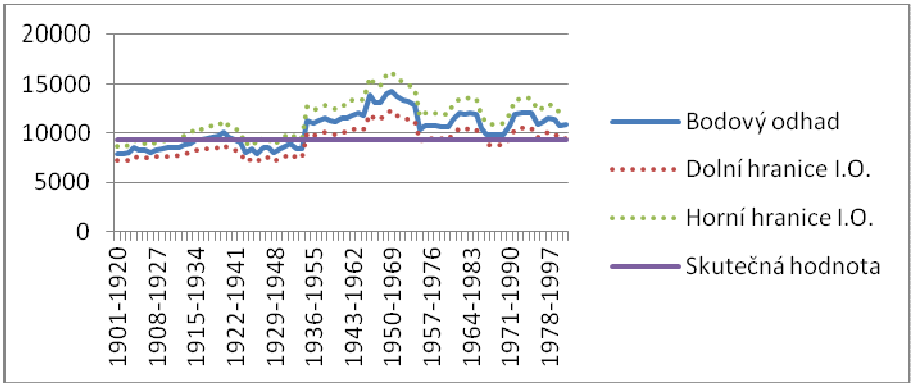
Obrázek 7: Odhad padesátileté vody z desetiletých souborů Zdroj: vlastní práce



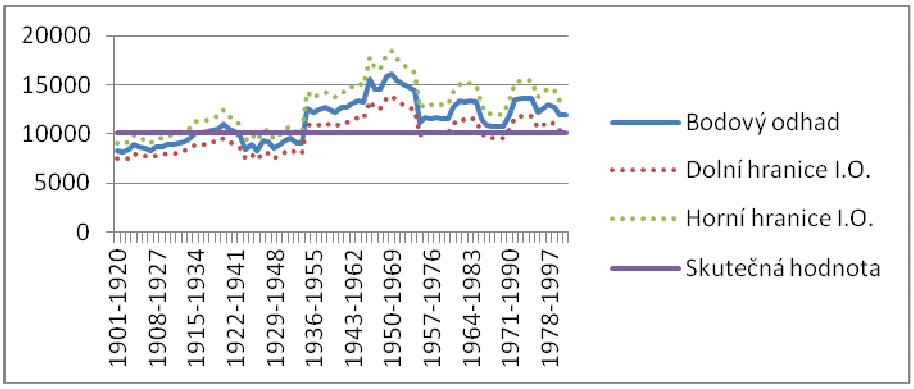
Obrázek 8: Odhad stoleté vody z desetiletých souborů Zdroj: vlastní práce



Obrázek 9: Odhad pětadvacetileté vody z dvacetiletých souborů Zdroj: vlastní práce



Obrázek 10: Odhad padesátileté vody z dvacetiletých souborů Zdroj: vlastní práce



Obrázek 11: Odhad stoleté vody z dvacetiletých souborů Zdroj: vlastní práce

Tabulka 3: Shrnutí přesnosti odhadů pětadvacetileté vody Zdroj: vlastní práce

Velikost výběrů	Počet výběrů	V mezích	Vyšší než I.O.	Nižší než I.O.	Průměrný B.O.	Skutečnost
5 let	96	50	28	19	8 862	8 116
10 let	91	40	16	35	9 083	8 116
20 let	81	37	0	44	9 480	8 116

Tabulka 4: Shrnutí přesnosti odhadů padesátileté vody Zdroj: vlastní práce

Velikost výběrů	Počet výběrů	V mezích	Vyšší než I.O.	Nižší než I.O.	Průměrný B.O.	Skutečnost
5 let	96	44	38	15	9 757	9 342
10 let	91	40	26	25	9 934	9 342
20 let	81	28	12	41	10 383	9 342

Tabulka 5: Shrnutí přesnosti odhadů stoleté vody Zdroj: vlastní práce

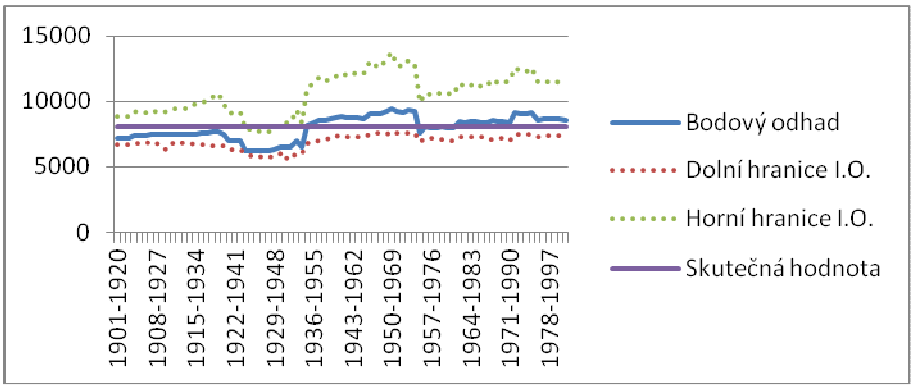
Velikost výběrů	Počet výběrů	V mezích	Vyšší než I.O.	Nižší než I.O.	Průměrný B.O.	Skutečnost
5 let	96	40	40	17	10 766	10 092
10 let	91	36	27	28	10 849	10 092
20 let	81	21	18	42	11 360	10 092

Z obrázků i tabulek je patrné, že odhady konstruované obyčejnou metodou POT jsou silně nestabilní a s rostoucí velikostí výběru míra zkreslení stoupá, což je patrné na klesajícím počtu intervalových odhadů, které správně předpověděly skutečnou hodnotu (třetí sloupec tabulek 3 – 5). Zkreslení je většinou směrem k vyšším hodnotám oproti skutečné hodnotě, což udává pátý sloupec tabulek a velice jednoduchá míra zkreslení porovnávající skutečnost s průměrným bodovým odhadem (šestý sloupec). Obecně tedy můžeme říci, že takovéto odhady jsou nadhodnocené a nezdá se, že by velikost odhadovaného kvantilu měla zásadní vliv na zkreslení.

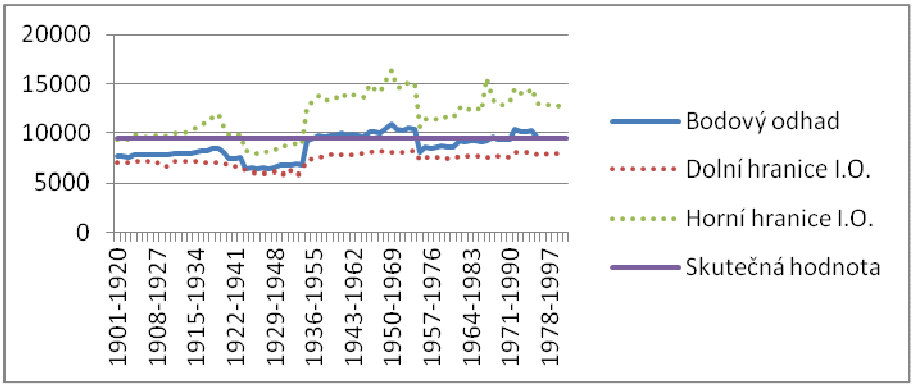
Odhady z upravených dat

Následující obrázky 12 – 14 představují bodové a intervalové odhady získané z jednotlivých dvacetiletých výběrových souborů, obrázky 15 – 17 odhady ze čtyřicetiletých výběrových souborů a obrázky 18 - 20 odhady z osmdesátiletých výběrových souborů. Odhady jsou dělány vždy pro stoletou, padesátiletou a pětadvacetiletou vodu.

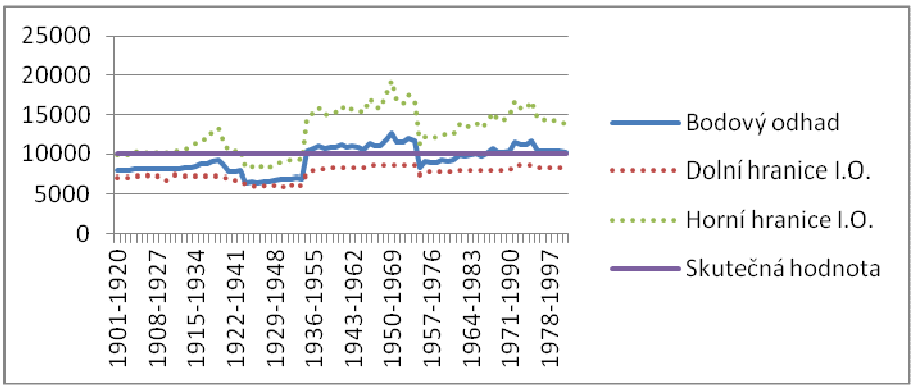
Dále následují tabulky 6 – 8 udávající vývoj přesnosti odhadů pro jednotlivé extrémní průtoky podle rostoucího počtu let výběrových souborů.



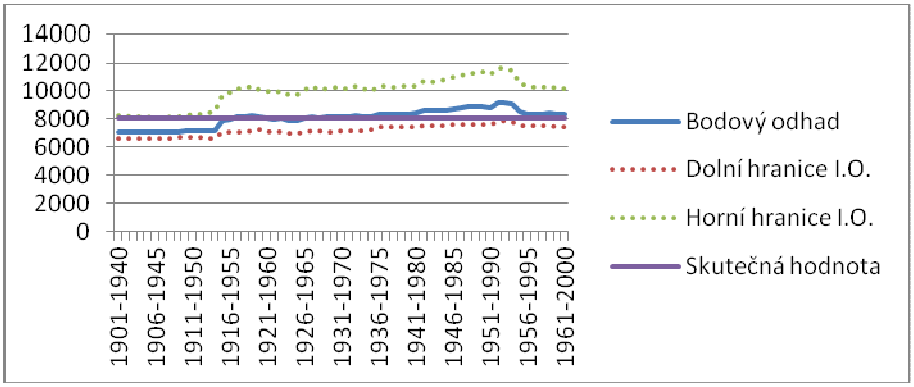
Obrázek 12: Odhad pětadvacetileté vody z dvacetiletých souborů Zdroj: vlastní práce



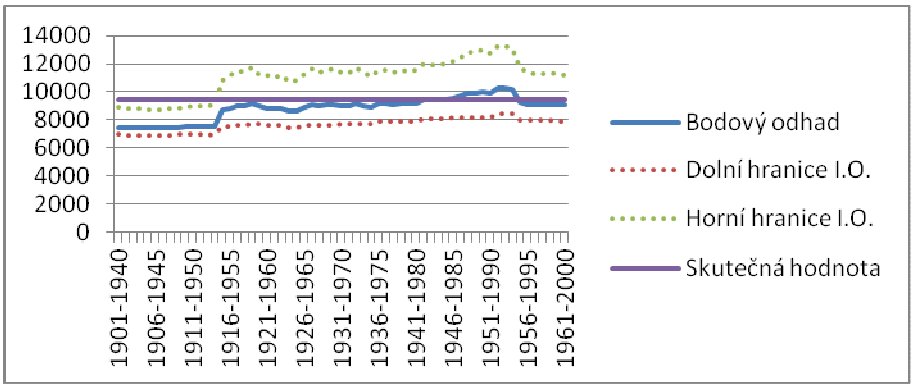
Obrázek 13: Odhad padesátileté vody z dvacetiletých souborů Zdroj: vlastní práce



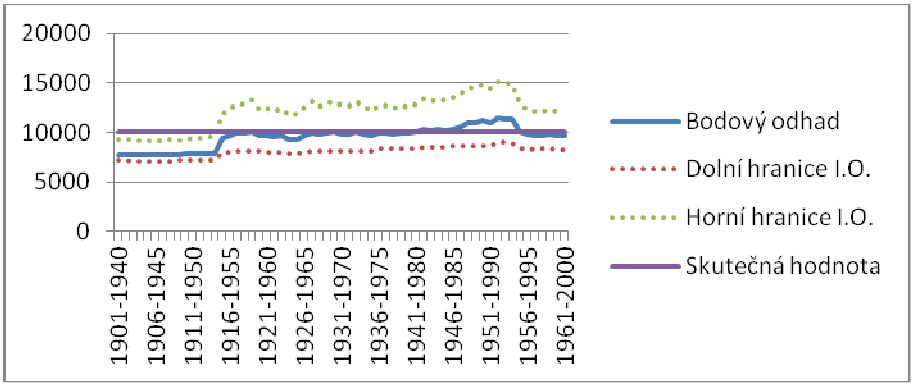
Obrázek 14: Odhad stoleté vody z dvacetiletých souborů Zdroj: vlastní práce



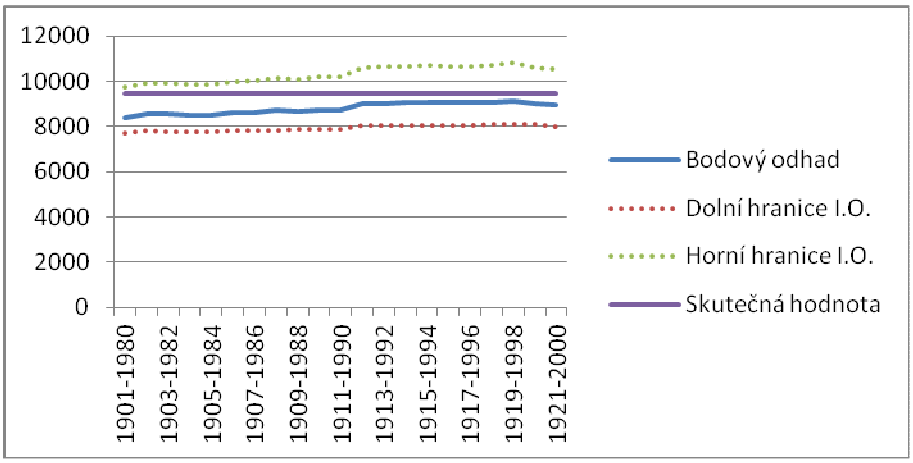
Obrázek 15: Odhad pětadvacetileté vody ze čtyřicetiletých souborů Zdroj: vlastní práce



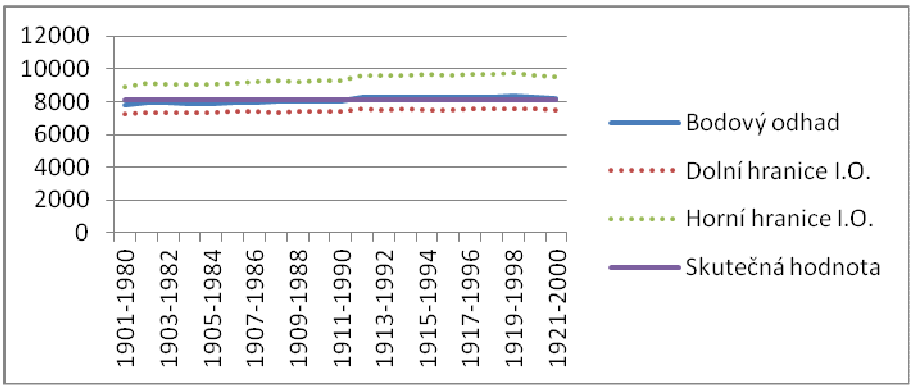
Obrázek 16: Odhad padesátileté vody ze čtyřicetiletých souborů Zdroj: vlastní práce



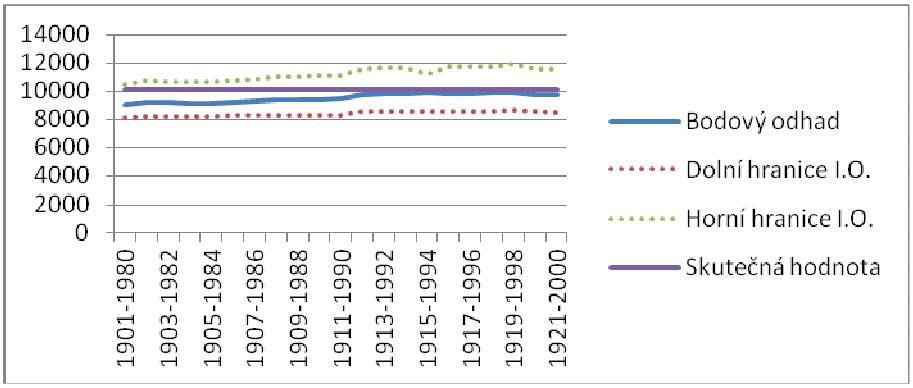
Obrázek 17: Odhad stoleté vody ze čtyřicetiletých souborů Zdroj: vlastní práce



Obrázek 18: Odhad petadvacetileté vody z osmdesátiletých souborů Zdroj: vlastní práce



Obrázek 19: Odhad padesátileté vody z osmdesátiletých souborů Zdroj: vlastní práce



Obrázek 20: Odhad stoleté vody z osmdesátiletých souborů Zdroj: vlastní práce

Tabulka 6: Shrnutí přesnosti odhadů pětadvacetileté vody Zdroj: vlastní práce

Velikost výběrů	Počet výběrů	V mezích	Vyšší než I.O.	Nižší než I.O.	Průměrný B.O.	Skutečnost
20 let	81	75	7	0	7 977	8 267
40 let	61	59	2	0	8 090	8 267
80 let	21	21	0	0	8 101	8 267

Tabulka 7: Shrnutí přesnosti odhadů padesátileté vody Zdroj: vlastní práce

Velikost výběrů	Počet výběrů	V mezích	Vyšší než I.O.	Nižší než I.O.	Průměrný B.O.	Skutečnost
20 let	81	67	14	0	8 645	9 479
40 let	61	47	14	0	8 827	9 479
80 let	21	21	0	0	8 809	9 479

Tabulka 8: Shrnutí přesnosti odhadů stoleté vody Zdroj: vlastní práce

Velikost výběrů	Počet výběrů	V mezích	Vyšší než I.O.	Nižší než I.O.	Průměrný B.O.	Skutečnost
20 let	81	65	16	0	9 331	10 092
40 let	61	47	14	0	9 581	10 092
80 let	21	21	0	0	9 520	10 092

Tentokrát lze z obrázků i tabulek vidět zaprvé vyšší stabilitu a především přesnost odhadů oproti původní metodě odhadu. Nepřesné odhady vyplývají především z datové základny, na obrázcích 12 – 17 je vidět výrazný zlom bodového odhadu v padesátých letech. Jedná se konkrétně o rok 1954, ve kterém došlo ke skutečné stoleté vodě a jehož zahrnutí výrazně změnilo odhady oproti předcházejícím výběrovým souborům, které tento rok neobsahovaly. Vyklouznutí tohoto souboru z výběrů se projevilo propadem odhadu v sedmdesátých respektive devadesátých letech. Nicméně charakter dat je takový, že i přes toto vyklouznutí už jsou odhady přesnější než v počátečních výběrech. Přesnost odhadu měřená průměrným bodovým odhadem je zdá se stabilní již od čtyřicetiletých výběrů, které zahrnují v průměru 100 hodnot. Odhady jsou zdá se mírně podhodnocené, což může být způsobeno způsobem výběru zahrnovaných dat, který je možná příliš přísný a nepouští do souboru některé další hodnoty. U menších výběrových souborů jsou hlavní příčinou první výběry před rokem 1954, nicméně podhodnocení se zdá být přítomno i u osmdesátiletých souborů (obrázky 19 a 20). Tuto domněnku podporuje fakt, že celý tento základní soubor má bodový odhad pro stoletou vodu hodnotu 9 278 (95% interval spolehlivosti pak je 8 356 – 10 679).

K demonstraci velikosti změny odhadu v roce 1954 jsem připravil následující tabulky srovnávající odhady učiněné na základě dat z let 1901 – 1953 a následně z let 1901 – 19543.

Tabulka 9: Změna odhadu extrémních průtoků v roce 1954 Zdroj: vlastní práce

Výběrový soubor	Odhady 25leté vody	Odhad 50leté vody	Odhad 100leté vody	Odhad 500leté vody
1901-1953	7 024	7 364	7 666	8 238
1901-1954	7 632 (+8,7 %)	8 220 (+11,6 %)	8 796 (+14,7 %)	10 088 (+22,5 %)

Pro potvrzení relativní náchylnosti metody POT na výskyt extrémní události jsem na závěr udělal odhad z celého základního souboru a následně odhad z tohoto souboru s přidaným pozorováním o hodnotě 11 500, což představuje přibližný současný ekvivalent hodnoty 10 092 v roce 1954 ve smyslu odhadované hodnoty kvantilu tohoto pozorování z dosud známých odhadů GPD. Změny bodových odhadů extrémních průtoků udává poslední tabulka.

Tabulka 10: Změna odhadu extrémních průtoků v roce 1954 Zdroj: vlastní práce

Výběrový soubor	Odhady 25leté vody	Odhad 50leté vody	Odhad 100leté vody	Odhad 500leté vody
1901-2000	7 987	8 636	9 479	10 736
1901-2000U	8 338 (+4,4 %)	9 136 (+5,8 %)	9 954 (+5,0 %)	11 928 (+11,1 %)

Závěr

Článek měl za cíl analyzovat přesnost odhadů metody POT na reálných datech. Potvrdilo se, že základní metoda bez pečlivějšího výběru dat poskytuje u takovéhoto typu řešeného problému (časové řady s přítomnou závislostí mezi jednotlivými hodnotami) zkreslené odhady a to především směrem k vyšším hodnotám. To nemusí být nutně špatně, jestliže je při rozhodování přítomna averze k riziku, kdy nadhodnocený odhad je lépe hodnocen než podhodnocený, ale z hlediska racionality volby je základním požadavkem na modely jejich přesnost.

Metoda POT vycházející z úpravy dat, která zajišťuje jejich nezávislost, se ukazuje jako přesná, jestliže je přítomno dostatečné množství napozorovaných hodnot, které jsem zde velice zjednodušeně určil jako více než 100, což v tomto konkrétním případě odpovídá 40 letům.

Nicméně stále se nabízí otázka přesnosti odhadu skutečně extrémních hodnot, jestliže ve výběrovém souboru chybí některá z nich. U kratších časových řad se to zde projevilo „efektem roku 1954“, který sám o sobě dokázal výrazně přiblížit odhady realitě.

Stabilita odhadů ve smyslu relativní změny odhadovaných extrémních průtoků je s téměř dvojnásobným počtem pozorování v roce 2000 zhruba dvojnásobná.

Tato práce je založena na skutečných datech, což činí zhodnocení jejich cílů relativně obtížným. Vhodně se proto jeví doplnit toto testování metody POT testy na simulovaných datech, kde o skutečných hodnotách extrémů základního souboru nemůže být pochyb.

Literatura

CASTILLO, Enrique a Ali S. HADI. Fitting the Generalized Pareto Distribution to Data. *Journal of the American Statistical Association*. 1997, **92**(440), 1609 - 1620.

CLAPS, P., LAIO, F. a VILLANI, P. Assessment of extreme flood production mechanisms through POT analysis of daily data. In: *International Conference on Flood Estimation*: *CHR report II*. Berne, Switzerland: Fed. Off. For Water and Geol., 2003, 75 - 85. Dostupné z: http://www.idrologia.polito.it/~claps/Papers/Claps_et_al_Berna_2002.pdf

ČABLA, Adam. Peaks Over Threshold in Modeling of the Czech Household Income Distribution. Prague 22. 09. 2011 – 23. 09. 2011. In: *International Days of Statistics and Economics at VŠE, Prague [CD-ROM]*. Prague : VŠE, 2011, s. 1–10. ISBN 978-80-86175-72-0.

TANAKA, S. a A. TAKARA. A study on threshold selection in POT analysis of extreme floods. In: *The Extremes of the Extremes: Extraordinary Floods*. Oxfordshire, UK: IAHS Publ., 2002, 299 - 304.

VOJTĚCH, Jan. Možnosti využití teorie extrémních hodnot. Praha 11. 02. 2010. In: *Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze.* [online] Praha: Oeconomica, 2010, s. 255–262. ISBN 978-80-245-1647-9. URL:http://fis.vse.cz/studium/doktorske-studium/den-doktorandu/den-doktorandu-2010/.

VOJTĚCH, Jan. *Využití teorie extrémních hodnot při řízení operačních rizik*. Praha, 2011. 159 s. Dizertační práce. Vysoká škola ekonomická v Praze.

JEL Classification: C46.

Summary

Predictions of extremal river discharges using „peaks over threshold“ method and theirs quality

The article deals with the behavior of the estimates of extremal river discharges using method of peaks over threshold (POT).

The results show that the POT method without data arrangement overestimates the extremal discharge estimates. It is not disadvantageous property if the averse-risk opinion is taken, but from the point of view of rational choice the main demand is the accuracy of the estimates.

POT method with the data arrangements seems to go much more accurately depending of course on the amount of sample data. The line was very roughly set to be 100 observations.

Nevertheless the question of accuracy and stability still arises when the sample lacks true extremes. It manifests itself at the shorter series by the „year 1954 effect”.

The article is based on the real dataset which makes the testing of accuracy quite difficult. It than seems to be reasonable to continue the analysis with the simulated data.

Keywords: peaks over threshold, generalized Pareto distribution, river discharges, estimates.

Zobecněné řešení soustavy lineárních algebraických rovnic a zobecněná metoda nejmenších čtverců

Richard Horský
rhorsky@vse.cz
Doktorand oboru statistika
Školitel: prof. ing. Josef Arlt, CSc. (arlt@vse.cz)

Abstrakt: Při klasickém přístupu k soustavám lineárních algebraických rovnic rozhodujeme o existenci řešení pomocí známé Frobeniovy podmínky. Existuje však přístup, při kterém každá soustava lineárních algebraických rovnic má řešení, tzv. zobecněné řešení. Toto řešení také vystupuje ve všeobecně známé metodě nejmenších čtverců. V tomto příspěvku pak získáváme metodu, ve které obecný skalární součin nahrazuje standardní eukleidovský skalární součin a metoda nejmenších čtverců tak získává obecnější podobu.

Klíčová slova: lineární operátor, soustava lineárních algebraických rovnic, skalární součin, Mooreova-Penroseova pseudoinverzní matice, normální řešení soustavy lineárních algebraických rovnic.

Úvod

Zabýváme se otázkou řešení rovnice

$$Ax = b \qquad (1)$$

kde A je lineární operátor definovaný na M-rozměrném eukleidovském prostoru s hodnotami v N-rozměrném eukleidovském prostoru. Takový operátor je možno reprezentovat maticí $\mathbf{A}$ typu (M,N) a rovnici (1) tak interpretovat jako soustavu M lineárních algebraických rovnic pro N neznámých (složek vektoru x). V klasické terminologii se pak matice $\mathbf{A}$ nazývá maticí soustavy a vektor b vektorem pravých stran. Klasické řešení problému (1) zřejmě existuje právě když vektor $b \in \mathrm{R}(A)$, kde $\mathrm{R}(A)$ je obor hodnot operátoru A, neboli je splněna tzv. Frobeniova podmínka o hodnostech. Hodnost matice $\mathbf{A}$ je totiž $\dim \mathrm{R}(A)$.

Častým případem je situace, kdy rovnice (1) představuje nějaký model, přičemž prvky matice $\mathbf{A}$ a složky vektoru b jsou získány měřením. Taková experimentálně získaná soustava může pro nějaké hodnoty měření mít řešení, ale při jiných hodnotách se může stát, že řešení neexistuje. Vzniká potřeba, aby řešení existovalo vždy a nejlépe, aby bylo určeno v nějakém smyslu jednoznačně. Dále budeme hovořit o tzv. zobecněném řešení rovnice (1).

Přirozený způsob, jak v případě, kdy neexistuje klasické řešení rovnice (1) definovat zobecněné řešení, je najít vektor v oboru hodnot operátoru A, který je „nejblíže“ vektoru b. Eukleidovský prostor je lineární prostor se skalárním součinem, tj. také normovaný či metrický prostor, kde metrika je odvozena z normy a norma ze skalárního součinu a prostor je úplný vzhledem k této metrice. Obvykle uvažujeme standardní skalární součin, jehož hodnota je součtem součinů sobě odpovídajících souřadnic násobených vektorů. Z něho odvozená norma je tzv. eukleidovská norma. V dalším textu však nepředpokládáme tento speciální případ a uvažujeme zcela obecný skalární součin. Minimalizace vzdálenosti vektoru b od $\mathrm{R}(A)$ je klasická úloha variačního počtu. Ve skutečnosti ale minimalizujeme kvadrát této vzdálenosti a hovoříme o tzv. cenové funkci. Ukážeme, že problém hledání minima této cenové funkce je ekvivalentní problému určení tzv. pseudoinverzní matice. Tak transformujeme problém variačního počtu na problém z oblasti lineární algebry. Uvedené postupy využívají terminologii funkcionální analýzy a dávají možnost poněkud odlišného pohledu na klasickou problematiku známou z aplikací ve statistice (metoda nejmenších čtverců, lineární regrese).

Základní pojmy a vztahy

Uvažujme lineární operátor

$$A: \mathrm{E_N} \to \mathrm{E_M}, \qquad (2)$$

kde $\mathrm{E_K}$ označuje K-dimensionální eukleidovský prostor. Operátor (2) je reprezentován maticí $\mathbf{A} = (a_{ij})$ typu (M,N). V dalším textu nebudeme striktně rozlišovat mezi operátorem a maticí, která ho reprezentuje. Rovnici (1) je tedy možno interpretovat jako soustavu M lineárních algebraických rovnic pro N neznámých $x = (x_1,...,x_N)$ s maticí soustavy $\mathbf{A} = (a_{ij})$ a vektorem pravých stran $b = (b_1,...,b_M)$. Operátor (2) definuje dvě význačné množiny, *jádro* $\mathrm{N}(A) = \{x \in \mathrm{E_N} : Ax = o_M\}$ a *obor hodnot* $\mathrm{R}(A) = \{y \in \mathrm{E_M} : \text{existuje } x \in \mathrm{E_N} \text{ tak, že } y = Ax\}$. Symbolem o_M označujeme nulový vektor prostoru $\mathrm{E_M}$. Eukleidovský prostor lze opatřit skalárním součinem. Skalární součin je bilineární, symetrická a pozitivně definitní forma. Jejím typickým příkladem je standardní skalární součin (v $\mathrm{E_K}$): $s(x,y)_K = x_1y_1 + ... + x_Ky_K$, kde $x = (x_1,...,x_K), y = (y_1,...,y_k)$. V dalším textu budeme obecný skalární součin v $\mathrm{E_K}$ značit $(x,y)_K$. Norma vektoru je odvozena ze skalárního součinu $\|x\|_K = \sqrt{(x,x)_K}$. Eukleidovský prostor je pak vůči této normě úplný (tzn. cauchyovské posloupnosti jsou konvergentní). Skalární součin nám umožňuje zavést pojem *adjungovaného operátoru*

$$A^*: \mathrm{E_M} \to \mathrm{E_N}, \qquad (3)$$

kde $(Ax, y)_M = (x, A^*y)_N$. Existenci a jednoznačnost operátoru (3) dostáváme jako důsledek Rieszovy věty o reprezentaci spojitého lineárního funkcionálu, viz [1], str. 233 a 237. Na závěr této kapitoly si připomeňme, že platí následující vztahy mezi jádry a obory hodnot operátoru a k němu adjungovaného operátoru:

$$\mathrm{E_N} = \mathrm{N}(A) \oplus \mathrm{R}(A^*) \text{ a } \mathrm{E_M} = \mathrm{N}(A^*) \oplus \mathrm{R}(A). \qquad (4)$$

Sčítance v každém z direktních rozkladů (4) jsou vzájemnými ortogonálními doplňky. Například, je-li $x \in \mathrm{E_N}$, pak existují jediné dva vektory $v, w \in \mathrm{E_N}$ takové, že $x = v + w$, $Av = o_M$, $w = A^*u$ pro nějaké $u \in \mathrm{E_M}$ a $(v,w)_N = 0$.

Zobecněné řešení soustavy lineárních algebraických rovnic

Nejprve definujeme pojem cenové funkce. Funkci

$$F(x) = (Ax - b, Ax - b)_M = \|Ax - b\|_M^2 \qquad (5)$$

nazýváme *cenovou funkcí*. Je to nezáporná funkce $F: \mathrm{E_N} \to \mathrm{E_1}$. Budeme se zabývat otázkou jejího minima. Ukážeme, že existuje alespoň jeden bod, v němž dosahuje funkce (5) svého (absolutního) minima, ale nejprve definujeme:

Zobecněným řešením rovnice (1) budeme rozumět jakýkoli bod $z \in \mathrm{E_N}$, ve kterém funkce (5) nabývá svého minima.

Věta 1. Existuje alespoň jedno zobecněné řešení rovnice (1).

V rovnici (1) je daný operátor *A* a vektor *b*. Podle (4) lze psát $b = d + Ac$, přičemž $A^*d = o_N$. Ale také $x = v + A^*u$, $Av = o_M$. Užijeme-li těchto rozkladů pro výpočet hodnoty cenové funkce (5), dostaneme

$$F(x) = \left(A\left(v + A^*u\right) - d - Ac, A\left(v + A^*u\right) - d - Ac\right)_M = \left(A\left(A^*u - c\right) - d, A\left(A^*u - c\right) - d\right)_M =$$
$$= \left\|A\left(A^*u - c\right)\right\|_M^2 + \|d\|_M^2.$$

Vektor *d* je pevně určen (vektorem *b*). To znamená, že jedině první sčítanec v posledním výrazu je proměnný, závislý na volbě $x = v + A^*u$. Zvolme $x = c = v_c + A^*u_c$, $Av_c = o_M$. Tedy $F(c) = \left\|A\left(A^*u_c - c\right)\right\|_M^2 + \|d\|_M^2 = \|d\|_M^2$.

Vidíme tedy, že cenová funkce (5) dosahuje své minimální hodnoty $\|d\|_M^2$ v bodě *c*. Tato hodnota se nazývá cenou zobecněného řešení a všimněme si, že v případě, že existuje klasické řešení rovnice (1), je jeho cena nulová. Nelze však vyloučit, že zobecněných řešení rovnice (1) existuje více. Označme

$$S = \left\{x \in \mathrm{E}_N : F(x) = \|d\|_M^2\right\}. \quad (6)$$

Množina (6) je neprázdná ($c \in S$), uzavřená (zobrazení *F* je spojité) a konvexní (subaditivita normy) a tudíž v ní existuje prvek s minimální normou. Tento prvek je určen jednoznačně. Podrobný důkaz viz (Taylor, 1973, str. 232).

Věta 2. Existuje jediný vektor $x^* \in S$, pro který platí $\|x^*\|_N = \min\left\{\|x\|_N : x \in S\right\}$. Množina S je lineární (lineál), přičemž jejím směrovým prostorem je jádro operátoru *A*.

Toto jednoznačně určené (podmínkou minimální normy) zobecněné řešení se nazývá *normální řešení* rovnice (1). Z podmínky minimální normy plyne, že vektor normálního řešení musí být ortogonální k jádru *A* a tedy být prvkem oboru hodnot operátoru A^* (viz důkaz Věty 4.). Struktura množiny S implikuje dřívější a obecnější zjištění o její uzavřenosti a konvexitě. Fakt, že vektor $x' - x'' \in \mathrm{N}(A)$ pro dvě zobecněná řešení $x', x'' \in S$ se snadno ověří přímým výpočtem.

Mooreova-Penroseova pseudoinverze

Nyní se budeme zabývat zobecněním pojmu inverzní matice. Symbolem *A* dále označujeme operátor (2) a jak jsme již poznamenali, neodlišujeme operátor od matice, která ho reprezentuje. Uvažujme systém rovností

(i) $AXA = A$,

(ii) $XAX = X$,

(iii) $(AX)^* = AX$,

(iv) $(XA)^* = XA$.

Neznámý operátor *X* je zřejmě reprezentován maticí typu (N,M). Není však zřejmé, zda uvedená soustava rovností má řešení a zda je toto řešení, pokud existuje, určeno jednoznačně. Ve speciálním případě, kdy $M = N$ a operátor *A* je injektivní, tzn. matice, která ho reprezentuje je regulární, je jediným řešením soustavy rovností (i)-(iv) inverzní operátor (matice) A^{-1}.

Věta 3. Pro každý operátor (2) existuje jediný operátor *X* splňující soustavu rovností (i)-(iv).

Nejprve dokážeme jednoznačnost:

Předpokládejme, že *X* a *Y* splňují vlastnosti (i)-(iv). Pak platí

$AX = (AX)^* = X^*A^* = X^*(AYA)^* = X^*A^*(AY)^* = (AX)^*(AY)^* = AXAY = AY$. Z druhé strany

$XA = (XA)^* = A^*X^* = (AYA)^*X^* = (YA)^*A^*X^* = (YA)^*(XA)^* = YAXA = YA$. Užitím předchozích dvou rovností dostáváme $X = X(AX) = X(AY) = (XA)Y = (YA)Y = Y$.

Důkaz existence: Uvažujme operátor

$$T = A^* A : \mathrm{E_N} \to \mathrm{E_N} . \quad (7)$$

Operátor (7) je symetrický a positivně semidefinitní. Označme všechna jeho různá kladná vlastní čísla $\lambda_j,\ j = 1,.....,s$ a jim odpovídající vlastní podprostory $V_j,\ j = 1,...,s$. Je-li operátor (7) singulární, tzn. má netriviální jádro $\mathrm{N}(T)$, pak je jeho vlastním číslem i $\lambda_0 = 0$ s příslušným vlastním podprostorem $V_0 = \mathrm{N}(T)$.

Spektrální projekce

$$P_j : \mathrm{E_N} \to V_j,\ j = 0,...,s , \quad (8)$$

umožňují vyjádřit operátor (7) ve tvaru

$$T = \sum_{j=1}^{s} \lambda_j P_j ,\ TP_0 = 0 . \quad (9)$$

Projekce (8) jsou ortogonální (symetrické) projekce. Platí pro ně

$$P_j P_k = 0 \text{ pro } j \neq k, P_j^2 = P_j, P_j^* = P_j, I - P_0 = \sum_{j=1}^{s} P_j ,\ j,k = 0,...,s . \quad (10)$$

Vyjádření operátoru (7) podle (9) je tzv. spektrální rozklad operátoru a poslední rovnost v (10) je tzv. spektrální rozklad jedničky. Položme nyní

$$X = \sum_{j=1}^{s} \frac{1}{\lambda_j} P_j A^* . \quad (11)$$

Lze přímo ukázat, že operátor (11) je hledaný pseudoinverzní operátor k operátoru (2), tzn. (11) splňuje rovnosti (i)-(iv). Pro ukázku dokážeme vlastnost (i):

$$AXA = A\sum_{j=1}^{s} \frac{1}{\lambda_j} P_j A^* A = A\sum_{j=1}^{s} \frac{1}{\lambda_j} P_j \sum_{k=1}^{s} \lambda_k P_k = A\sum_{j=1}^{s} P_j^2 = A(I - P_0) = A - AP_0 , \quad (12)$$

Druhá rovnost plyne z vyjádření (9), třetí a čtvrtá ze vztahů (10). Nyní už stačí ukázat, že člen AP_0 ve (12) je nulový operátor. Je-li $x \in \mathrm{E_N}$ zcela libovolně zvolený vektor, pak $0 = (TP_0 x, x) = \|AP_0 x\|_M^2$. To znamená, že každý vektor z oboru hodnot operátoru AP_0 má nulovou normu a je tedy nulový, čili operátor AP_0 je konstantní, nulový operátor.

Operátor (11) budeme dále značit symbolem A^+ a nazývat *Mooreův-Penroseův pseudoinverzní operátor*. Je to lineární operátor definovaný v $\mathrm{E_M}$ s hodnotami v $\mathrm{E_N}$. Platí pro něj rovnost

$$A^+ A = I - P_0 , \quad (13)$$

speciálně pro $\mathrm{N} = \mathrm{M}$ a regulární *A* je zřejmě $A^+ = A^{-1}$.

Normální řešení a Mooreova-Penroseova pseudoinverze

V této kapitole ukážeme, že vektor, který je hodnotou Mooreova-Penroseova pseudoinverzního operátoru v pravé straně rovnice (1) je právě ono jediné normální řešení této rovnice.

Věta 4. Vektor $x^+ = A^+ b$ je normálním řešením rovnice (1).

Nejprve ukážeme, že vektor x^+ je prvkem množiny (6), jinými slovy, že tento vektor je zobecněným řešením rovnice (1):

$F\left(x^+\right)=\left(Ax^+-b,Ax^+-b\right)_M=\left\|AA^+Ac-Ac-d\right\|_M^2=\left\|d\right\|_M^2$, přičemž poslední rovnost plyne z vlastnosti (i) pseudoinverzního operátoru.

Dokážeme, že $x^+\in \mathrm{R}\left(A^*\right)$, viz komentář za Větou 2. Stačí si uvědomit, že komplementární projekce $I-P_0$ projektuje právě do $\mathrm{R}\left(A^*\right)$ a $x^+=(I-P_0)x^+$, jak plyne z (13).

Závěr

Při definování zobecněného řešení operátorové rovnice (1) používáme obecný skalární součin v eukleidovském prostoru. Normální řešení zde není výsledkem řešení soustavy získané klasickým postupem variačního počtu s užitím nutné podmínky pro extrém funkce uvnitř množiny aplikované na cenovou funkci. Přirozené zobecňování zde uvedených postupů se nabízí ve směru k Hilbertovým prostorům, tj. nekonečně dimensionálním prostorům se skalárním součinem, úplným vzhledem k normě z tohoto součinu odvozené. Takovým prostorem je třeba prostor $L_2\left(\Im\right)$ (funkcí integrovatelných s kvadrátem v Lebesgueově smyslu v intervalu $\Im$) nebo jeho diskrétní analogie, prostor posloupností sumabilních s kvadrátem, l_2.

Literatura

TAYLOR, A. E. 1973. *Úvod do funkcionální analýzy.* Praha: Academie.

JEL Classification: C000.

Summary

The Generalized Solution of the Systém of Linear Algebraic Equation and The Generalized LSM

It seems as a natural requirment to define a solution of system of linear algebraic equations (SLAE) even if there is no solution in classical sense. Thus we come to the notion of generalized solution and even more to get uniqueness, normal solution of SLAE. The classical outputs of this considerations are well-known in statistics as LSM and linear regression. The selected approach uses the tools of functional analysis and performs another point of view on classical matter. The usage of general scalar product enables the generalization of classical LSM. Moore-Penrose pseudoinverse operator returns the problem from the area of variation calculus back to linear algebra.

Keywords: linear operator, system of linear algebraic equations, scalar product, Moore-Penrose pseudoinverse operator, normal solution.

Použití endogenních vah při srovnání vysokých škol pomocí kompozitního indikátoru

Lenka Hudrlíková
lenka.hudrlikova@vse.cz
Doktorandka oboru statistika
Školitel: doc. Ing. Jakub Fischer, Ph.D. (jfischer@vse.cz)

Abstrakt: Cílem článku je navrhnout alternativní způsob srovnávání vysokých škol, resp. určení endogenních vah u příslušného kompozitního indikátoru. Zvolení nejvhodnějších vah je zásadní krok při agregování ukazatelů, jelikož to má významný vliv na celkový kompozitní indikátor, a tedy na následné srovnání vysokých škol. Všechna současná srovnání vysokých škol používají fixní exogenní váhy založené na názorech odborníků. Vysoké školy nemusí mít primárně stejné cíle a poslání (např. výuka a/nebo věda). Fixní váhy stejné pro všechny vysoké školy tím ztrácejí význam. V článku je navržen endogenní systém vah, při němž škola získá vyšší váhy u ukazatele, kde je relativně lepší, a nižší váhy tam, kde je relativně horší. Existující srovnání vysokých škol také nebere v úvahu prostředí, v němž škola působí. Exogenní charakteristiky, jako školné, jazykové prostředí, velikost školy vyjádřená počtem studentů, či úroveň vědy a výzkumu je v navrhované metodě brána úvahu, za účelem učinit srovnání více správné a užitečné. Metoda „Benefit of doubt" (BoD) přiřadí každé vysoké škole vlastní váhy, jež maximalizují její celkové skóre. Plně neparametrický BoD přístup přináší však více výhod. Za prvé, váhy a následně skóre a pořadí školy jsou založeny pouze na datech. Není třeba předem učinit žádný *a priori* předpoklad či subjektivní hodnocení. Druhá výhoda je robustnost navrhované metody. Za třetí, model dovoluje zahrnutí exogenních proměnných, jež mají vliv na výsledky školy. Výsledné srovnání založené na BoD modelech je rozdílné od původního. Obecně lze říci, že vysoké školy z USA a Velké Británie jsou na horších příčkách v BoD modelech a naopak si velmi polepšily školy ze Švýcarska.

Klíčová slova: srovnání vysokých škol, stanovení endogenních vah, Benefit of doubt model, kompozitní indikátor, terciární vzdělávání

Úvod

Srovnávání je všeobecně velice populární, ačkoliv každé srovnání je do určité míry kontroverzní. Nejinak je tomu u srovnávání vysokých škol (VŠ). Počáteční debata, zda se kvalita VŠ dá vůbec porovnávat či měřit, se zdá v posledních letech již překonána, jelikož srovnání VŠ i díky své popularitě zřejmě nezmizí. Do středu zájmu se dostalo úsilí učinit srovnání co možná nejvíce transparentní, spolehlivé a relevantní (Federkeil, 2008; IHEP, 2007 či Merisotis, 2002). V současné době známá srovnávání VŠ jsou podrobena kritice zejména několika důvodů: (1) výběr souboru dílčích ukazatelů, (2) libovůle při určení váhového systému, (3) nekonzistence mezi různými srovnáními, (4) časté změny metodologie, (5) spoléhání na názory odborníků a veřejného mínění, (6) nedostatečná statistická validita, (7) následné nezdravé soupeření mezi školami (Stella & Woodhouse, 2006). Špatně zkonstruované srovnání pomocí kompozitního indikátoru může snadno poskytnout zavádějící informace, na jejichž základě mohou být učiněny zjednodušené závěry. Příkladem může být mylný závěr, že lepší pořadí při srovnání musí nutně korespondovat se zvýšením kvality školy (Rauhvargers, 2011). Přes zmíněnou kritiku je srovnání VŠ velmi významné, jelikož má vliv na současný systém terciárního vzdělávání. Hazelkorn (2007) poukazuje na silný vliv při rozhodování vedení škol a akademiků ohledně struktury, strategických plánů a marketingu školy. Je samozřejmé, že každá škola chce být ve srovnání s ostatními co nejlepší, školy se tak především zaměřují se na zlepšení v kritériích relevantních při srovnání, a to navíc především na ty, jež mají největší vliv na srovnání (Toffalis, 2011). Toto však rozhodně není účelem srovnávání. Srovnání VŠ je významné i z pohledu trhu práce. Harvey (2008) poukazuje na fakt, že zaměstnavatelé dávají přednost studentům z VŠ, jež jsou na předních místech žebříčků. Politici považují pozici VŠ v mezinárodním srovnání za jednu z měr ekonomické síly a konkurenceschopnosti země. Srovnání VŠ může být, či již v některých zemích je, implementováno do systému alokace veřejných financí. V neposlední řadě pořadí VŠ je velmi sledované zájemci o studium a širokou veřejností (Marginson & van der Wende, 2007). V dnešní době každá škola působí na mezinárodním trhu se vzděláním a žebříčky VŠ mohou velmi ovlivnit poptávku po dané škole.

Vysoké školy jsou činorodé, provádějí vědu a výzkum, učí studenty, získávají zdroje a patenty na inovace atd. Srovnání tedy musí být založeno na vícerozměrném pojetí a následném vyjádření pomocí kompozitního indikátoru. Kompozitní indikátory agregují informace z několika ukazatelů do jednoho souhrnného indikátoru, a z tohoto důvodu jsou často využívány k srovnávacím účelům (JRC/OECD, 2008). Vybrání nejvhodnějších vah je klíčový krok při tvorbě kompozitního indikátoru. Nelze se však vyhnout určité subjektivitě. JRC/OECD (2008) navrhují několik metod, jež lze rozdělit na metody založené na statistických modelech, participační a hybridní metody. Existující žebříčky VŠ používají právě na participačních metodách, tedy jsou odvozeny z teoretického základu a názorů odborníků.

V zahraničí jsou žebříčky vysokých škol známy jako „university ranking" nebo „league tables" (ve Velké Británii). V současné době existuje několik srovnání VŠ a to jak na mezinárodní, tak na národní úrovni. Jeden z prvních a zároveň nejznámějších mezinárodních žebříčků VŠ je *The Academic Ranking of World Universities* (ARWU) tzv. Shanghai Index, který je však zaměřený pouze na vědu a výzkum. Od roku 2004 byl publikován *Times Higher Education Suplement* (THES), který má za cíl více komplexní srovnání škol. V roce 2010 se však rozpadl na dva nezávislé programy. První je *World's Best Universities Ranking* od společnosti Quacquarelli Symonds (QS). Druhý *THE World University Rankings* vznikl ve spolupráci The Times Higher Education a Thompson Reuters. Mezi další populární globální srovnání VŠ patří *Leiden Ranking* zaměřený na bibliometrické indikátory. Specializovaný je také *Webometrics Ranking*, jež měří aktivity školy na webu. Nedostatky v těchto srovnání vedly Evropskou komisi ke snaze o vytvoření nástroje pro srovnání VŠ, který dovolí zachytit různorodost škol. Vznikl projekt *U-Multirank* a *U-Map*, jehož první výsledky by měly být dostupné v roce 2013.

Jelikož se srovnávání VŠ dostává velká pozornost, v roce 2006 skupina The International Ranking Expert Group (IREG) stanovila principy a kritéria pro hodnocení kvality a dobré praxe VŠ, jedná se o tzv. Berlin Principles on Ranking of Higher Education Institutions. Jde o šestnáct základních pravidel, jež diskutují účel a cíle srovnávání, použité váhy, sběr a zpracování dat a prezentaci výsledků srovnání. Berlínské principy mimo jiné říkají, že srovnání by mělo podporovat diverzitu vysokých škol a vzít v úvahu jejich různé cíle. Současné žebříčky škol používající rigidní váhy tento požadavek nesplňují. Jelikož exogenně určené váhy zvýhodňují školy, pro něž se dané váhy nejlépe hodí. Eccles (2002), Van der Wende & Westerheijden (2009) i Zha (2009) poukazují na fakt, že jsou zvýhodněny školy zaměřené na intenzivní výzkum na úkor jiných excelentních škol např. specializovaných škol s úzce zaměřeným studijním programem či vysoké školy zaměřující se na výuku a ne na výzkum.

Dalším z požadavků, jež současná srovnání nesplňují, je ohled na národní specifika a rozdílné systémy vzdělávání. Vysoké školy operují (alespoň v krátkém období) v prostředí daném kulturním, ekonomickým a historickým kontextem vzdělávacích systémů. Při mezinárodním srovnání by tedy exogenní charakteristiky zázemí škol neměly zůstat bez povšimnutí. VŠ jsou silně ovlivněny rozhodnutím vlád resp. ministerstev v oblastech jako je velikost školy, zdroje financování a stupeň specializace. Pokud není velikost VŠ explicitně vyjádřena, vláda používá různé omezení počtu studentů pomocí grantů a stipendií. Zdroje a výši financování také nemají školy zcela ve svých rukou. Výše školného či alespoň jeho horní hranice bývá přesně stanovena. Kde není vláda příliš striktní, zejména u soukromých škol, je školné silně ovlivněno poptávkou a konkurencí na trhu čili školné je exogenní proměnná. Školné má však velký vliv na cíle univerzity (Michaelis, 2004), což přináší některým školám výhodu při srovnávání. Vysoké školy s žádným nebo minimálním školným se musí spoléhat na financování ze veřejného rozpočtu, a tedy často je jejich cílem maximalizovat počet přijatých studentů. Naopak VŠ s vysokým školným mají jako cíl maximalizaci prestiže pomocí výzkumné činnosti. Excelentní výzkum jim může přinést finance od soukromých inovátorských firem a zároveň přinese výhodu při srovnání škol a tedy tím mohou obhájit vysoké školné. Další exogenní charakteristika, jež má velký vliv při mezinárodním srovnání škol je jazyk země, kde daná škola působí. Altbach (2006) se domnívá, že míry určené pro hodnocení kvality VŠ ve své podstatě zvýhodňují velké univerzity z anglicky mluvících center vědy a vzdělanosti, zejména ze Spojených států a Velké Británie. Jelikož angličtina je v současné době lingua franca, neanglické práce jsou publikované méně a také méně citované. Marginson & Van der Wende (2007) dokonce přímo říkají, že Američané mají sklon citovat více Američany. Školy z anglicky mluvících zemí také samozřejmě

lákají mnohem snáze elitní zahraniční studenty, jelikož studenti nemají problém domluvit se i mimo školu.

Srovnání vysokých škol by mělo poskytovat snadnou avšak správnou interpretaci, stimulovat zdravé soutěžení mezi VŠ, rozlišovat a podporovat různé typy a zaměření škol a přispět do debaty o hodnocení kvality vzdělávání. Je zřejmé, že fixní stejné váhy pro všechny školy nemohou tyto požadavky splnit. Přínos tohoto článku spočívá v navrhnutí alternativní metodologie pro srovnání vysokých škol. Idea spočívá v určení endogenních vah, kde by každá univerzita získala vlastní váhy podporující její silné stránky. Navrhovaná metodologie má za cíl eliminovat nedostatky současných systémů. Všechny současné srovnávání používají rigidní váhy, exogenně určené na základě subjektivních názorů. Rigidní váhy nedovolují nebo dokonce odrazují univerzity od specializací (Taylor & Braddock, 2007). Z důvodu diferenciace a specializace použití fixních vah není vhodné (Kälvemark, 2007). Navrhovaná alternativní metoda stanovení pro stanovení endogenních vah spočívá v přiřazení větší váhy ukazatelům, v nichž je škola relativně lepší a naopak nižší váhy kde je daná škola relativně horší. Metoda je zcela neparametrická a není tedy nutné učinit předpoklad o produkční funkci. Základní myšlenka vychází z metody Analýzy obalu dat (DEA; Charnes et al., 1978). V tomto případě se konkrétně jedná o „Benefit of the Doubt" model (BoD), jež dovoluje nově rozšířit původní metodu, aby byla robustní a podmíněná exogenními proměnnými (Cazals et al., 2002, De Witte & Kortelainen 2009).

Existují dva pohledy na exogenní charakteristiky při srovnávání. Na jednu stranu, pokud se na srovnávání díváme „absolutně", exogenní charakteristiky škol bychom neměly brát při srovnání v úvahu. Škola, jež obsadila v žebříčku první místo, má skutečně nejvyšší hodnocení bez ohledu na okolnosti. Na druhou stranu, pokud srovnávání bereme „relativně", měli bychom uvážit rozdílnosti v systému vzdělávání. Takový přístup porovnává pouze podobné školy, tzn. s obdobnými podmínkami. Ve čtvrté části jsou publikovány výsledky pro oba typy srovnávání.

Článek je dále organizován následovně. Druhá část se věnuje ukazatelům pro srovnání VŠ používaným Times Higher Education Suplement (THES), jejichž hodnoty jsou použity pro následnou analýzu. V třetí části je popsána BoD metoda určení vah a skóre vysokým školám. Čtvrtá část obsahuje výsledky a komentáře, posléze následují závěry a perspektivy pro další výzkum.

Data

Pro analýzu a použití navrhované endogenní metody určení vah jsou využita data nejlepších 200 vysokých škol, dle žebříčku publikovanému THES v roce 2009. Tato data jsou vybrány hlavně s ohledem na změnu metodologie v roce 2010. Jelikož jedním z cílů článku je porovnat pořadí VŠ dle původního srovnání a dle BoD modelů, jsou využity stejné indikátory a jejich hodnoty.

První ukazatel „Peer review" je proxy zastupující reputaci školy, jež je kvantifikována pomocí výsledků dotazníkového šetření mezi akademickými pracovníky. Původně má tato proměnná váhu 40 %. Druhý ukazatel „Employer Reputation" (váha 10 %) je založen na názorech zaměstnavatelů, což má odrážet schopnost škol připravit absolventy pro profesní sféru. Třetí ukazatel „Research excellence" (váha 20 %) je měřen počtem citací na jednoho akademického pracovníka. Čtvrtý ukazatel „Teaching excellence" (váha 20 %) je zastoupen podílem učitele na jednoho studenta. Pátý ukazatel „International faculty" (váha 5 %) je vyjádřen podílem zahraničních akademických pracovníků. Šestý ukazatel „International Students" (váha 5 %) je tvořen podílem zahraničních studentů. Detailnější informace lze nalézt na webových stránkách THES. Tabulka 1 obsahuje popisné charakteristiky výběrového souboru 200 nejlépe hodnocených vysokých škol.

Tabulka 1: Popisné statistiky jednotlivých ukazatelů Zdroj: THES 2009, vlastní výpočty

	Peer Review	**Employer Reputation**	**Staff /Student**	**Citations /Staff**	**Intern. Faculty**	**Intern. Students**	**Celkové skóre**
Průměr	75,22	70,37	61,77	66,03	60,51	61,07	70,25
Sm. odchylka	17,86	23,81	24,66	20,94	27,62	26,54	12,22
Minimum	35,00	20,00	9,00	21,00	0,00	13,00	54,00

	Peer Review	Employer Reputation	Staff /Student	Citations /Staff	Intern. Faculty	Intern. Students	Celkové skóre
25% kvantil	61,00	51,00	43,00	50,25	32,00	35,25	60,33
Medián	76,00	75,00	63,00	63,50	61,00	62,00	67,90
75% kvantil	90,75	93,00	84,00	82,75	86,00	86,00	78,88
Maximum	100,00	100,00	100,00	100,00	100,00	100,00	100,00

V tabulce 1 lze vidět, že nejvíce školy excelují v akademické reputaci. Na druhou stranu nejslabší stránkou je podíl zahraničních učitelů na školách. Hodnoty všech ukazatelů jsou standardizovány a převedeny na stupnici 0 až 100. Jelikož se jedná o charakteristiky pouze pro 200 nejlepších škol, medián i průměr jsou nad hodnotou 50.

Jak již bylo zmíněno, zázemí vysoké školy ovlivňuje její výkon. Ukazatelé vyjadřující tyto exogenní charakteristiky jsou: (1) školné, (2) velikost univerzity, (3) vědecko-výzkumná aktivita vzhledem k velikosti a studijním programům a (4) oficiální jazyk země, v níž škola působí.

Určit výši školného jak jednu hodnotu je pro většinu škol téměř nemožné. Jelikož výše školného závisí na mnoha faktorech jako např. na studijním programu, délce jeho trvání, poskytovaných službách, místě školy atd. Školy také často poskytují různé typy finančních podpor a stipendií. Proto jsou v následné analýze uvažovány čtyři kategorie školného: (1) pro domácí a (2) zahraniční studenty, a pro (3) pregraduální a pro (4) postgraduální studenty. Data pochází ze srovnání publikovaného QS v roce 2011. V případech, kde byly hodnoty nedostupné, došlo k nahrazení hodnoty průměrem dané země.

Jedním ze způsobů jak měřit velikost VŠ je pomocí počtu studentů (v přepočtu na FTE). Z důvodu redukce vlivu ročních změn je dále použita kategoriální proměnná se čtyřmi skupinami (méně než 5 000 studentů, 5 000 až 11 999 studentů, 12 000 až 29 999 studentů, více nebo rovno než 30 000 studentů). Tato klasifikace je převzata z QS 2010 stejně jako data. Jak lze vidět v tabulce 2, na téměř 80% vybraných škol je zapsáno více než 12 000 studentů.

Třetí charakteristika je úroveň vědy a výzkumu. Některé školy se orientují na vědu a výzkum zatímco jiné např. na výuku, což také bývá podporováno vládní politikou. Klasifikace škol dle úrovně výzkumu je převzata z QS 2010. Výzkumná aktivita je odhadována na základě počtu dokumentů v databázi Scopus za posledních pět let. Rozlišujeme čtyři kategorie, od žádného či nízkého k velmi vysokému vědeckému výstupu školy. Hranice mezi jednotlivými skupinami jsou závislé na počtu studentů a rozsahu působnosti školy. Toto rozlišení by mělo zlepšit porovnatelnost údajů, jelikož menší instituce jsou schopny produkovat méně vědeckých aktivit než velké instituce.[1] Popisné statistiky v tabulce 2 ukazují, že pouze jedna škola je v nejhorší kategorii, žádná škola není v kategorii střední, a jen 9 % škol patří do skupiny s vysokou úrovní výzkumu. Více než 90 % VŠ má velmi vysokou hodnotu vědeckého výstupu. Což je samozřejmě dáno tím, že ve výběru je jen 200 nejlepší vysokých škol z celého světa.

Nakonec, je třeba zohlednit jazykovou bariéru. Studenti a akademičtí pracovníci jsou nejvíce lákáni školami z anglicky mluvících zemí, jelikož angličtina je lingua franca. V tabulce 2 lze pozorovat, že školy ze zemí kde je angličtina oficiální jazyk (nebo alespoň jedním z nich) mají velké zastoupení. Ve výběru je jen méně než polovina škol ze zemí, v nichž je oficiální jazyk odlišný od angličtiny.

[1] Více o klasifikaci a datech na http://www.topuniversities.com/university-rankings/world-university-rankings/methodology/classifications

Tabulka 2: Popisné statistiky exogenních charakteristik Zdroj: QS 2010, vlastní výpočty

	Dom. školné pregrad. (USD)	Dom. školné postgrad. (USD)	Zahr. školné pregrad. (USD)	Zahr. školné postgrad. (USD)
Průměr	11 367	10 483	17 632	16 803
Směrod. odchylka	12 679	11 732	13 403	12 086
Minimum	0	0	0	0
25% kvantil	3 000	3 000	5 000	5 000
Medián	7 000	7 000	17 000	17 000
75% kvantil	15 000	11 000	27 000	25 000
Maximum	43 000	43 000	43 000	43 000

Velikost (četnosti)		Výzkum (četnosti)		Angličtina (četnosti)	
Malá	5	**Nízký**	1	**Ne**	83
Střední	36	**Střední**	0	**Ano**	117
Velká	112	**Vysoký**	18		
Extra velká	47	**Velmi vysoký**	181		

Uvedené proměnné lze z pohledu vysoké školy považovat za exogenní. Většina vysokých škol totiž trpí nedostatkem autonomie, což je také zdůrazňováno v posledních deklaracích o terciárním vzdělávání. Weber (2006) konstatuje, že autonomie vysoké školy je nutnou, i když ne postačující podmínkou pro excelenci a uzavírá, že nejlepší školy dle standardních hodnocení jsou právě ty relativně autonomní. Menší autonomie škol, znamená větší exogenitu uvedených proměnných.

Metodologie

„Benefit of Doubt“ model

Pomocí hodnot kompozitní indikátoru, resp. skóre lze jednoduše sestavit pořadí vysokých škol. Jednotlivým šesti, v předchozí části popsaným, ukazatelům je třeba přiřadit váhy. Proměnné, kde vysoká škola má komparativní výhodu, je přiřazena větší váha než školám s nižší komparativní výhodou či dokonce nevýhodou. Těm školám je naopak u této proměnné přiřazena menší váha. Je zřejmé, že tato metoda vychází z dat a je tedy v rozporu s předchozí praxí, kde váhy byly určeny na základě názorů expertů bez statistického odůvodnění.

Navrhovaná endogenní metoda „Benefit of the Doubt” (BoD) byla poprvé pojmenována v práci dvojice Melyn & Moesen (1991). BoD model je založen na neparametrickém modelu známém pod názvem Analýza obalu dat (Data Envelopment Analysis), který byl prvně formulován v práci Farella (1957), a následně rozpracován v práci Charnes et al. (1978) jako nástroj pro měření efektivity. Nejsou třeba žádné předpoklady o tvaru produkční funkce. Při použití BoD metody má každá škola své vlastní váhy, které maximalizují její silné stránky a naopak minimalizují slabé stránky ve srovnání s ostatními. Tuto ideu lze vyjádřit pomocí vzorce pro lineární programování (Cherchye et al., 2007):

$$s_c(y) = \max_{w_{c,i}} \sum_{i=1}^{q} w_{c,i} y_{c,i}, \qquad (1)$$

za podmínek

$$\sum_{i=1}^{q} w_{c,i} y_{j,i} \leq 1, \qquad j = 1,...,c,...,n, \tag{1a}$$

$$w_{c,i} \geq 0, \qquad j = 1,...,q, \tag{1b}$$

kde
$s_c(y)$ – hodnota celkové skóre pro školu c,
$w_{c,i}$ – váha pro i-tý ukazatel pro školu c,
$y_{c,i}$ – hodnota i-tého ukazatele pro školu c.

Skóre $s_c(y)$ pro každou z c škol je získáno na základě porovnání výsledků ukazatele y_i (i=1,…,q) s ostatními školami ve výběru Y. Proměnné $y_{c,i}$, v nichž si hodnocená škola vede výborně mají větší váhu $w_{c,i}$ než proměnné, v nichž si škola vede hůře. Rovnice (1) obsahuje omezení zajišťující normalizaci (1a) a non-negativitu (1b). Váhy tedy musí být nezáporné se skóre mezi 0 a 1. Relativně nejlepší školy získají skóre 1.

Robustní „Benefit of Doubt“ model

Hlavní nevýhodou popsaného BoD modelu je, že odlehlá pozorování mohou mít nepříznivý vliv na výsledné skóre. Pro překonání tohoto nedostatku, Cazals et al. (2002) navrhl postup, v němž je výběr pro porovnání škol nahrazen podvýběrem Y_c^m o velikosti $m < q$. BoD skóre hodnocené školy je určeno relativně k tomuto menšímu referenčnímu výběru. Tato robustní technika redukuje vliv odlehlých pozorování, jelikož odlehlé pozorování nutně nemusí být v referenčním výběru. Aby metoda byla robustní je třeba porovnání v různých podvýběrech opakovat B krát, přičemž referenční výběr je $\mathrm{Y}_c^{b,m}$ (b=1,…,B). Robustní skóre je poté určeno jako aritmetický průměr ze všech B porovnání. Toto lze vyjádřit pomocí vztahu:

$$s_c^{b,m}(y) = \max_{w_{c,i}} \sum_{i=1}^{q} w_{c,i}^b y_{c,i}^b, \tag{2}$$

za podmínek

$$\sum_{i=1}^{q} w_{c,i}^b y_{j,i} \leq 1, \qquad j = 1,...m \forall y_{j,i} (j = 1,...m) \in \mathrm{Y}_c^{b,m}, \tag{2a}$$

$$w_{c,i}^b \geq 0, \qquad i = 1,...,q, \tag{2b}$$

$$w_{c,i}^b \in W_e, \qquad i = 1,...,q,\ e \in E. \tag{2c}$$

Připomeňte, že skóre získané pomocí standardní metody BoD leží mezi 0 a 1. V případě robustní metody to neplatí. Pokud je skóre větší než 1, jedná o tzv. super-efektivitu. Jinými slovy škola, která je super-efektivní je lepší než průměr m škol v referenčním výběru.

Robustní a podmíněný „Benefit of Doubt“ model

V první části bylo diskutováno, že zázemí může výrazně ovlivňovat působení a výsledky vysoké školy. Další rozšíření modelu spočívá v zahrnutí těchto exogenních charakteristik z. Cazals et al. (2002) a Daraio & Simar (2005) navrhli do robustní techniky zahrnout exogenní proměnné. Hlavní myšlenka spočívá ve volbě referenčního setu o velikosti m s danou pravděpodobností, místo náhodného vybírání jak je tomu u robustní metody. Pravděpodobnost je určena neparametickou jádrovou funkcí u proměnných z. Referenční set $\mathrm{Y}_c^{b,m}$ je nahrazen $\mathrm{Y}_c^{b,m,\bar{z}}$, kde $\bar{z}$ představuje pozorování s podobnými hodnotami jako z. Pozorování, která jsou podobná hodnocenému pozorování, jsou vybrány častěji než odlišná. Odhad celkového skóre pro každou školu vychází ze vzorce (2-2c). Výsledné podmíněné skóre $S_c^m(y|z)$ bere v úvahu exogenní proměnné z a je robustní vůči odlehlým a

vybočujícím pozorováním. Samozřejmě je zajímavé pozorovat vliv proměnných na BoD skóre a potažmo na pořadí škol. K identifikování tohoto vlivu je využita práce De Witte & Kortelainen (2009), která je rozšířením práce Daraio & Simar (2005). Navrhují neparametrickou regresi podílu podmíněného a nepodmíněného skóre $S_c^m(y|z) / S_c^m(y)$. Pokud charakteristiky mají příznivý (nepříznivý) vliv na hodnocení, regresní křivka bude rostoucí (klesající). K testování významnosti, De Witte & Kortelainen (2009) navrhnuli neparametrický bootstrap test podobně jako Racine et al. (2006). Tyto testy jsou lze považovat za neparametrický ekvivalent standardních t-tesů v OLS regresi. V následující části uvedené výsledky byly vypočteny pomocí programu R. Výsledky odpovídají $m = 100$, jelikož alternativní hodnoty získané pro $m = 50, 60, 70, 80$ a 200 nepřinesly významně jiné výsledky. Hodnota B byla zvolena 2000.

Výsledky

Soubor 200 nejlepších vysokých škol dle THES-QS v roce 2009 byl použit pro vytvoření pořadí těchto škol dle tří různých neparametrických robustních BoD modelů – jeden nepodmíněný a dva podmíněné. V tabulce 3 je pořadí škol pro originální THES-QS a v dalších sloupcích jsou pořadí v BoD modelech. Z důvodu značné rozsáhlosti je uvedeno pořadí pouze pro prvních 10 škol.

Tabulka 3: Pořadí vysokých škol Zdroj: THES 2009, vlastní výpočty

THES - QS	Vysoká škola	BoD 1	BoD 2	BoD 3	THES -QS	Vysoká škola	BoD 1	BoD 2	BoD 3
1	Harvard University (US)	4	18	5	**6**	University of Oxford (UK)	27	38	40
2	University of Cambridge (UK)	8	21	11	**7**	University of Chicago (US)	23	35	35
3	Yale University (US)	12	24	18	**8**	Princeton University (US)	7	20	6
4	University College London (US)	6	19	7	**9**	Massachusetts Institute of Technology (US)	2	17	3
5	Imperial College London (UK)	13	25	23	**10**	California Institute of Technology (US)	1	16	4

Dominance škol ze Spojených států a Velké Británie je jednoznačná. Při endogenním určení vah je však pořadí škol rozdílné. Je zřejmé, že endogenní váhy nejsou příznivě nakloněny školám, které jsou na předních příčkách originálního THES žebříčku. Harvard, jež v původním žebříčku byl na prvním místě, dle nepodmíněné BoD klesl na 4. místo (u podmíněných modelů dokonce na 18. resp. 5. místo). Cambridge klesla z 2. Na 8. místo (u podmíněných modelů na 21., resp. 11. místo) a Yale klesl v žebříčku z 3. až na 12. místo (u podmíněných modelů na 24., resp. 18. místo). Výsledky odpovídají tomu, že tyto školy pracují ve významně příznivém prostředí. V tabulce 4 je pořadí prvních pěti vysokých škol dle jednotlivých BoD modelů.

Tabulka 4: Pořadí vysokých škol

	BoD nepodm. model 1	BoD podm. model 2	BoD podm. model 3
1	California Institute of Technology (US)	Ecole Polytechnique Fédérale de Lausanne (CH)	ETH Zurich (Swiss Federal Institute of Technology) (CH)
2	Massachusetts Institute of Technology (US)	ETH Zurich (Swiss Federal Institute of Technology) (CH)	Ecole Polytechnique Fédérale de Lausanne (CH)
3	ETH Zurich (Swiss Federal Institute of Technology) (CH)	Technical University of Munich (DE)	Massachusetts Institute of Technology (US)
4	Harvard University (US)	Heidelberg University (DE)	California Institute of Technology (US)
5	University of Geneva (CH)	University of Karlsruhe (DE)	Harvard University (US)

Pozoruhodný je výrazný pozitivní posun u švýcarských univerzit. Díky endogenním váhám, jež dovolují specializaci, univerzity ze Švýcarska významně zlepšily svoji pozici oproti původnímu THES-QS srovnání. V modelu 1, jež nebere v úvahu exogenní proměnné charakterizující zázemí školy, ETH Zurich původně na 21. příčce je na 3. příčce a University of Geneva původně na 72. místě je na 5.místě. První příčka nyní patří California Institute of Technology, což je jedna z pouze pěti univerzit s méně než 5 tisíci studenty. Také další počtem studentů malé či střední školy benefitují z BoD srovnání díky tomu, že jejich silné stránky jsou váženy více.

V podmíněném modelu 2 jsou zahrnuty proměnné školné pro (1) domácí pregraduální studenty, (2) domácí postgraduální studenty, (3) zahraniční pregraduální studenty a (4) zahraniční postgraduální studenty a navíc alternativní proměnná (5) škola je/není z anglicky mluvící země. V tomto případě je nejlepších 15 škol z Evropy, což je výrazná změna oproti předchozím modelům. Vysoké školy ze Švýcarska obsadily první dvě místa, na dalších třech místech jsou školy z Německa. V těchto zemích je školné nulové nebo malé (ve všech kategoriích je školné mezi 0 a 1500 USD).

Třetí model byl vytvořen, aby byl co nejvíce v souladu s Berlínskými principy a co bral v úvahu charakter a zázemí školy. Samozřejmě je toto omezeno dostupnými daty. Model 3 tedy zahrnuje (1) lokaci v anglicky mluvící zemi, (2) velikost školy a (3) velikost vědecko-výzkumného výstupu. V modelu 3, jsou na prvních příčkách dvě školy ze Švýcarska a následovány americkými školami. Všech pět škol, které jsou klasifikovány jako malé, významně vylepšily svoji pozici. Především technologické školy benefitují z tohoto modelu, tedy lze říci specializované školy s rozsáhlým kvalitním výzkumným výstupem.

Pořadí dle jednotlivých modelů naznačuje, které proměnné má významný vliv (pozitivní či negativní) u podmíněných BoD modelů. Pro ověření těchto hypotéz použijeme neparametrický bootstrap test. Výsledky v tabulce 5 ukázaly, že anglicky mluvící prostředí má příznivý a významný vliv při srovnání. V souladu s předchozími zjištěními, velikost školy má negativní, ale významný vliv na podmíněnou BoD metodu. Větší škola značí nižší pozici při podmíněném srovnání. Toto je v kontrastu s tvrzeními v literatuře, že srovnávání zvýhodňuje větší školy (např. Altbach 2006). Věda a výzkum má příznivý vliv, což značí větší výzkum lepší hodnocení v nepodmíněném srovnávání. Toto zjištění podporuje námitky ohledně zvýhodňování vědecko-výzkumně orientovaných škol (např. Sadlak, Merisotis, Liu 2008) ve stávajících srovnávání. Vyšší školné pro domácí studenty má negativní vliv na pozici školy ve srovnání, ačkoliv to není významně odlišné od nuly. Školné pro zahraniční studenty má významný pozitivní vliv. V souladu s předchozími závěry anglický jazyk má významný pozitivní vliv na pozici školy při srovnávání.

Tabulka 5: Směr a vliv exogenních charakteristik na robustní podmíněné BoD modely

	BoD model 2		**BoD model 3**	
	Směr	**p-hodnota**	**Směr**	**p-hodnota**
Domácí školné pregrad.	Negativní	1,00		
Domácí školné postgrad.	Negativní	0,61		
Zahraniční školné pregrad.	Pozitivní	0,01		
Zahraniční školné postgrad.	Pozitivní	< 2e-16		
Angličtina	Pozitivní	< 2e-16	Pozitivní	< 2,22e-16
Velikost			Negativní	< 2,22e-16
Věda a výzkum			Pozitivní	< 2,22e-16

Pro prozkoumání robustnosti výsledků jsou v tabulce 6 uvedeny Speramanovy korelační koeficienty. Významná korelace mezi jednotlivými modely, značí, že pokud je pozice školy dobrá v jednom modelu, tak je tomu tak i podle jiného modelu. Všechny pozorované korelace jsou vyšší než 0,55. Více než původnímu žebříčku THES, je nepodmíněné BoD srovnání podobné podmíněným BoD srovnání.

Tabulka 6: Spearmannova korelační matice

	THES	BoD Model 1	BOD Model 2	BOD Model 3
THES	1,000			
BoD nepodm. Model 1	0,653(**)	1,000		
BOD podm. Model 2	0,556(**)	0,797(**)	1,000	
BOD podm. Model 3	0,642(**)	0,963(**)	0,775(**)	1,000

** Korelace je statisticky významná na hladině významnosti 0,0.

Závěr

Navrhovaná metoda pro určení vah je v souladu s tzv. Berlínskými principy pro srovnávání vysokých škol. Na rozdíl od současného systému vah, navržený odpovídá požadavku zachycení rozdílnosti institucí terciárního vzdělávání a zahrnutí vlivu zázemí a prostředí, v němž školy působí. Při použití metody „Benefit of Doubt" jsou váhy přiřazené jednotlivým ukazatelům ve smyslu čím lepší hodnota ukazatele, tím větší váha a naopak. BoD přístup však přináší mnohem více výhod. Za prvé, při aplikování metody není třeba učinit a priori předpoklady, jelikož váhy jsou odvozeny z dat. Za druhé, je možné použít robustní metodu. Za třetí, metoda dovoluje zahrnout do analýzy další exogenní proměnné charakterizující zázemí, které má na školu podstatný vliv.

Pořadí vysokých škol dle různých modelů váhového systému jsou rozdílná. Dle výsledků lze vyvodit závěr, že prostředí anglicky mluvící země skutečně hraje významnou pozitivní roli při srovnání vysokých škol. Velikost školy resp. počet studentů má spíše negativní vliv, zatímco úroveň vědeckého výstupu má pozitivní vliv. Nakonec, vliv školného pro domácí studenty se zdá nevýznamný. Avšak školné pro zahraniční studenty má příznivý vliv pro hodnocení školy. Tyto výsledky potvrzují domněnky o stranění a zkreslení srovnání vysokých škol ve prospěch jednoho určitého typu.

Literatura

ALTBACH, P. (2006). The Dilemmas of Ranking. International Higher Education, 42, pp. 2–3.

CAZALS, C., FLORENS, J. P., SIMAR, L. (2002), Nonparametric Frontier Estimation: A Robust Approach, Journal of Econometrics 106 (1), pp. 1-25.

CHARNES, A. COOPER, W. W., RHODES, E. (1978), Measuring the efficiency of decision making units, European Journal of Operational Research 2, pp. 429-444.

CHERCHYE, L., MOESEN, W., ROGGE, N., VAN PUYENBROECK, T. (2007), An introduction to 'benefit of the doubt' composite indicators, Social Indicators Research 82, pp. 111-145.

DARAIO, C., SIMAR, L. (2005), Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach, Journal of Productivity Analysis 24 (1), pp. 93-121.

DE WITTE, K., KORTELAINEN, M. (2008), Blaming the exogenous environment? Conditional efficiency estimation with continuous and discrete environmental variables, CES Discussion Paper Series DPS 08.33, MPRA Paper 14034.

ECCLES, Ch. (2002). The Use of University Rankings in the United Kingdom, Higher Education in Europe, 27: 4, pp. 423-432.

FARRELL, M. J. (1957), The measurement of productive efficiency, Journal of the Royal Statistical Society, Series A, CXX, Part 3, pp. 253-290.

FEDERKEIL, G. (2008). Rankings and Quality Assurance in Higher Education, Higher Education in Europe, 33:2-3, pp. 219-23.

HARVEY, L. (2008). Rankings of Higher Education Institutions: A Critical Review, Editorial, Quality in Higher Education, Vol. 14, No. 3, pp. 187-207.

HAZELKORN, E. (2007) The Impact of League Tables and Ranking Systems on Higher Education Decision Making, Higher Education Management and Policy, Volume 19, No. 2, OECD

INSTITUTE FOR HIGHER EDUCATION POLICY (IHEP) (2007). College and university ranking systems. Global perspectives and American challenges. Washington, DC: Institute for Higher Education Policy.

JRC/OECD (2008) Handbook on constructing composite indicators. methodology and user guide, OECD Publishing ISBN 978-92-64-04345-9

KÄLVEMARK, T. (2007). University ranking systems: A critique. Technical report, Irish Universities Quality Board, 2007. Online.
http://www.urank.se/Dokument/Torsten_Kalvemark_University_Ranking_Systems_A_Critique.pdf

MARGINSON, S., and van der WENDE, M. C. (2007). To rank or to be ranked: The impact of global rankings in higher education. Journal for Studies in International Education. Vol. 11 (3–4), pp. 306–329.

MERISOTIS, J. P. (2002). On the Ranking of Higher Education Institutions, Higher Education in Europe, vol XXVII, no 4, pp 361-363.

MELYN, W., MOESEN, W. (1991), Towards a Synthetic Indicator of Macroeconomic Performance: Unequal Weighting when Limited Information is Available, Public Economics Research Paper, 17, CES, KULeuven.

MICHAELIS, P. (2004) Education, Research and the Impact of Tuition Fees A Simple Model of the Jahrbuch für Wirtschaftswissenschaften / Review of Economics Publication Info, University of Augsburg, Department of Economics. Online.
http://www.wiwi.uni-augsburg.de/vwl/institut/paper/265.pdf

RAUHVARGERS, A. (2011). Global university rankings and their impact, (Brussels: European University Association). Online.
http://www.eua.be/pubs/Global_University_Rankings_and_Their_Impact.pdf

SADLAK, J., MERISOTIS, J. and LI, N. C. (2008) University Rankings: Seeking Prestige, Raising Visibility and Embedding Quality – the Editors' Views, Higher Education in Europe, Vol. 33, No. 2/3, 195-199

STELLA, A. and WOODHOUSE, D. (2006) Ranking of Higher Education Institutions. Occasional Publications Series No: 6, Melbourne, AUQA.

TAYLOR, P., BRADDOCK, R. (2007). International University Ranking Systems and the Idea of University Excellence, Journal of Higher Education Policy and Management 29:3, pp. 245-260.

TOFALLIS, Ch. (2011). A different approach to university rankings. Higher Education, Online First, 31th March 2011

van der WENDE, M., WESTERHEIJDEN, D. (2009). Rankings and Classifications: The Need for a Multidimensional Approach, Higher Education Management and Policy – Volume 22/3 OECD

WEBER, L. (2006). University autonomy, a necessary, but not sufficient condition for excellence – being a paper presented at IAU / IAUP Presidents' symposium. Chiang Mai, Thailand 8-9 December

ZHA, Q. (2009): Diversification or Homogenization in Higher Education: A Global Allomorphism Perspective, Higher Education in Europe, 34:3-4, pp. 459-479.

JEL Classification: C14, C25, C43, I21

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu IGS F4/19/2011 „Jednofaktorová a vícefaktorová produktivita v kontextu meziodvětvových tabulek a kompozitních indikátorů"

Summary

The usage of endogenous weights in university ranking

This paper contributes to the growing literature on university rankings by proposing an alternative methodology to rank universities. A first issue (and corresponding contribution) comes from the weighting. As universities have multiple outputs, outputs have to be aggregated. Choosing the most appropriate weights is a crucial step in the aggregation of outcome variables as they have obviously a significant effect on the overall composite indicator and the university ranking. All previous university rankings use rigid and exogenous weights based on stakeholder or expert opinions. As higher education institutions have different goals and missions, they are internally differentiated such that a fixed weights for all universities are meaningless. This paper proposes an endogenous weighting system where higher weights are given at outputs the university is relatively good in, and lower weights to outputs the university is relatively less performing. As a second issue, the current rankings do not account for the operational and legal environment where universities are working in. To make rankings more accurate and useful, background of the universities should not be unnoticed.

Benefit of Doubt method which is proposed is in line with Berlin Principles on Ranking of Higher Education Institutions. Each university is given „benefit of doubt". It means each university is assigned its own weights which maximize its overall score. In fact the BoD approach brings much more advantages. First, applying BoD method there is no need for a priori assumptions or subjective value judgements. Weights, subsequently score and ranking are data-based. Second advantage springs from the robustness of the method. Third, method allows to account for exogenous variable which has influence on performance of university. It is also examined the influence of the background exogenous variables on the BoD scores and, thus, on the ranking of universities.

The analysis relies on a sample of 200 universities, corresponding to the universities used in the 2009 Times Higher Education Supplement (THES). Relying on the THES data, similar outcome variables are used. Data for background characteristics include tuition fees, size of university, research output (relative to size and number of faculty) and origin in English speaking country. The chosen variables can be considered an exogenous.

In the paper there are examined unconditional BOD model and two conditional models. Resulting ranking based on unconditional BoD model differs from original THES ranking. Also the correlations between BoD models are higher in comparison to original model and BoD models. Still significant shifts are found in conditional models in comparison to the others models. Nevertheless there are several common patterns. In general universities from USA and UK are lagging in BoD models. A remarkable shift is observed for Swiss Universities. The results indicate that working in an English environment matters. Universities located in English speaking countries obtain a higher position in the ranking. The size of the university has an unfavorable significant influence on ranking, while research output has a significant favourable effect. Finally, domestic tuition fees did not have a significant influence on the ranking of universities, while international tuition fees have favorable significant effect. It was proved claims about biases and distortions in favor of one type of university are justifiable. Rankings with rigid weighting scheme reward some type of universities more. Proposed method to set weights and to obtain ranking is more proper because each university can maintain its strength which are even rewarded. Since university could not complain about unfair weighting, any other weighting scheme would have generated lower overall scores.

Keywords: University ranking, Endogenous weights, Benefit of doubt model, Composite indicator, Higher education.

Měření efektivnosti v oblasti sociálních služeb

Věra Jeřábková
vera.jerabkova@vse.cz
Doktorandka oboru statistika
Školitel: doc. Ing. Jakub Fischer, Ph.D. (fischerj@vse.cz)

Abstrakt: Významný problém fungování veřejného sektoru spočívá v zajištění efektivnosti veřejných výdajů. Vytvoření modelu efektivnosti financování a poskytování sociálních služeb není jednorázovým tématem, nýbrž jde o dynamický proces, který vyžaduje pravidelně hodnotit změněné sociálně ekonomické podmínky. Aby mohla být tato efektivnost změřena, musíme být schopni změřit spotřebu a kvalitu, resp. výsledek, poskytovaných služeb. Současně však musíme uvažovat vliv celé řady kritérií a faktorů.

Tento příspěvek je zaměřen na měření efektivnosti nejrozšířenějších pobytových služeb sociální péče, a to domovů pro seniory. Efektivnost zde bude zkoumána z hlediska typu poskytovatele registrované služby. Aplikovány budou celkem tři přístupy: (1) poměrové ukazatele, (2) statistické odhady nákladových funkcí, a (3) kalkulace tzv. „produkčních mezer“.

Klíčová slova: sociální služby, efektivnost, nákladová funkce, ukazatele výkonnosti, Cobb-Douglasova produkční funkce.

Úvod

Sociální služby jsou legislativně vymezeny zákonem č. 108/2006 Sb., o sociálních službách a jsou chápány jako činnost nebo soubor činností, které zajišťují pomoc a podporu osobám za účelem sociálního začlenění nebo prevence sociálního vyloučení. Jedná se tak nejen o služby, které jsou spojeny s péčí o staré, neschopné či nemohoucí občany, ale i o služby, které pomáhají člověku řešit vzniklou nepříznivou sociální situaci, jejíž příčiny mohou být různé. Poskytování sociálních služeb je veřejnou službou, která bezprostředně reaguje na měnící se potřeby občanů.

V oblasti sociálních služeb neexistují standardy či návody pro měření efektivnosti. Jedná se o složitou a značně problematickou oblast, neboť jednotlivec si spotřebu sociálních služeb nevolí dobrovolně, ale je ni odkázán, a to z důvodu zdravotního stavu, věku, či sociální havárie. Mezi stěžejní problémy oblasti sociálních služeb pak patří neexistence minimálních kvantitativních standardů pro jednotlivé kraje, nedodržování principu rovných podmínek při poskytování dotací, neproplácení ošetřovatelské péče v pobytových zařízeních zdravotními pojišťovnami (Jeřábková, 2009) a neúčelnost příspěvku na péči.

Problematiku efektivnosti sociálního systému a sociálních služeb tak není možné řešit jako integritu všech segmentů. Při hodnocení efektivnosti systému financování a poskytování sociálních služeb je nezbytné uplatnit jiné pohledy na hodnocení efektivnosti služeb sociální péče, sociální prevence a sociálního poradenství. Rovněž je nutno uplatňovat i jiná hlediska z pohledu státu, krajů, měst a obcí a nestátních neziskových organizací.

Při hodnocení současně musíme uvažovat vliv celé řady kritérií a faktorů, jako například kritéria pro stanovení míry závislosti, strukturu uživatelů jednotlivých typů sociálních služeb z hlediska míry jejich závislosti (tj. výše příspěvku na péči), kritéria pro přiznávání dotací ze strany státu (MPSV) a zřizovatele, přístup zdravotních pojišťoven k financování poskytované zdravotní péče v pobytových zařízeních, dále pak výši průměrných nákladů připadajících na jednoho klienta, strukturu financování těchto nákladů, úroveň vybavenosti jednotlivých regionů sociálními službami, aj.

V předkládaném příspěvku je pozornost věnována pouze službám sociální péče, konkrétně pak péče poskytované v domovech pro seniory, neboť se jedná o nejrozšířenější a nejnákladnější typ poskytovaných sociálních služeb. Je zkoumána zejména nákladová efektivnost z hlediska poskytovatele sociální služby, jímž mohou být MPSV, příspěvkové organizace obcí a krajů, organizační složky obcí, nestátní neziskové organizace a další právnické či fyzické osoby.

Efektivnost a metody jejího měření

Pojem „efektivnost“ vyjadřuje účinnost vložených zdrojů a získaného užitku. V ekonomii je pak systém považován za efektivní, pokud se a) nikomu nemůže dařit lépe, aniž by se někomu jinému dařilo hůře (tzv. Paretovo optimum); b) je dosaženo maximální možné produkce z daného množství vstupů; a c) vyrábí se za nejnižší možné náklady na jednotku.

Efektivnost můžeme definovat jako vztah mezi vstupy a výstupy, resp. poměr mezi vloženými zdroji a dosaženými výsledky. Pro účely kvantitativní ekonomické analýzy se pak mírou efektivnosti rozumí poměr konkrétních vstupů a výstupů sledovaného transformačního procesu (1).

$$míra\ efektivnosti = \frac{výstupy}{vstupy} \qquad (1)$$

Metody, jimiž můžeme efektivnost měřit, lze rozdělit na tři velké skupiny, a to v souvislosti s nároky na odborné znalosti ve statistice, ekonometrii a operačním výzkumu (Dlouhý, 1999). Jsou jimi:

- Tradiční metody založené na ukazatelích informačního systému – za tyto ukazatele jsou považovány různé poměrové ukazatele (např. průměrné náklady na klienta) představující nejužívanější metodu měření efektivnosti;
- Statistická a ekonometrická analýza – zahrnuje různé statistické metody (např. regresní a korelační analýza) a ekonometrické odhady produkčních a nákladových vztahů;
- Jiné metody – například simulační modely či modely založené na lineárním programování. Pro jejich použití je zapotřebí specializovaného softwaru.

V příspěvku je využito těchto tří přístupů: (1) poměrové ukazatele, (2) statistické odhady nákladových funkcí, a (3) kalkulace tzv. „produkčních mezer“.

Poměrové ukazatele jsou považovány za tradiční a nejužívanější metodu měření efektivnosti. Pro prvotní představu o efektivnosti v oblasti sociálních služeb lze totiž snadno vyjít z kalkulace ukazatelů výkonnosti, tj. poměru „output/input“ či „input/output“. Tyto ukazatele podávají základní představu o fungování sledovaného zařízení a rovněž slouží pro porovnání s jinými subjekty (viz např. Slintáková, 2003). Nejprve je ovšem nutné určit vstupní proměnné, jimiž mohou být například celkové náklady (TC), provozní náklady (OC), náklady na mzdy (W), z nepeněžních ukazatelů můžeme použít například počet zaměstnanců (S) v organizaci. Zatímco však stanovení vstupních proměnných není tak obtížné, u výstupních proměnných je situace, zejména v oblasti sociálních služeb, mnohem horší.

Výstupní proměnná, která by zohledňovala náročnost služby spolu s její kvalitou, se totiž v oblasti sociálních služeb statisticky nesleduje. U sociálních služeb se statisticky neeviduje, jakou částku za konkrétní službu či provedený ošetřovatelský výkon daná osoba zaplatila. Rovněž není dostupný ukazatel, který by zohledňoval kvalitu poskytnutých služeb (např. počet klientů, kteří jsou spokojeni s péčí). Je proto nutné vyjít z proxy proměnné, která by byla porovnatelná mezi jednotlivými zařízeními, což u pobytových služeb může být například proměnná „počet lůžek“ (B).

Pro hodnocení efektivnosti domovů pro seniory mohou být využity například tyto ukazatele výkonnosti:

- TC/B – celkové náklady na jednotku výstupu (tj. počet lůžek)
- OC/B – provozní náklady na jedno lůžko
- W/B – mzdy na jedno lůžko
- S/B – přepočtený počet pracovníků na jedno lůžko
- B/TC – output (tj. počet lůžek) vytvořený z jedné jednotky celkových nákladů (v mil. Kč)
- B/S – output vytvořený jedním pracovníkem (přepočtené počty)

Druhým přístupem vyjádření efektivity poskytovaných sociálních služeb je poměřit skutečné náklady k nákladům očekávaným, které jsou vypočteny na základě odhadnutého modelu nákladové funkce (Dlouhý, 1999). Nákladová funkce je jedním z ekonomických modelů, který popisuje vztah mezi náklady a výstupem, tj. úrovní produkce. Jednou z metod odhadu nákladové funkce je pak využití regresní analýzy. Jelikož pro oblast sociálních služeb neexistují odhady nákladových funkcí, je možno vyjít z nákladových funkcí určených pro oblast zdravotnictví. Spotřeba zdravotní péče a spotřeba soci-

ální péče jsou považovány za komplementy a může zde tak existovat podobnost i při měření efektivnosti. Nákladová funkce se odhaduje z jednorovnicových modelů, které jsou obvykle ve tvaru:

$$TC = \beta_0 + \beta_1 \cdot Q, \quad (2)$$

$$TC = \beta_0 + \beta_1 \cdot Q + \beta_2 \cdot Q^2, \quad (3)$$

$$TC = \beta_0 \cdot Q^{\beta_1}, \quad (4)$$

V tomto příspěvku je využito logaritmické regrese pro odhad provozních nákladů (OC) a mzdových nákladů (W), kde za úroveň produkce (výstup, Q), tj. vysvětlující proměnnou, je uvažován počet lůžek. Odhad nákladových funkcí pak vychází z tvaru:

$$OC = \beta_0 \cdot B^{\beta_1}, \quad (5)$$

$$W = \beta_0 \cdot B^{\beta_1}, \quad (6)$$

kde
OC – provozní náklady
W – mzdové náklady
B – počet lůžek
β_0, β_1 – parametry

Poslední uvedený přístup k hodnocení efektivnosti veřejných a soukromých domovů pro seniory se týká kalkulace tzv. „produkční mezery“ domovů pro seniory, kdy se porovnává aktuální výstup s tzv. „virtuálním“ výstupem. Za aktuální výstup se považuje počet lůžek v zařízení a virtuální výstup je odhadnut (zvlášť pro veřejné i soukromé domovy pro seniory) na základě Cobb-Douglasovy produkční funkce (Jablonský, 2004). Za jediné faktory produkce se v tomto případě uvažují provozní náklady (OC) a přepočtený počet zaměstnanců (S). Pro oblast sociálních služeb je tudíž možné vycházet z modifikovaného tvaru:

$$Y = a \cdot OC^{\alpha} \cdot S^{\beta}, \quad (7)$$

kde
Y – „virtuální“ počet lůžek
a, α, β – parametry modelu
OC – provozní náklady
S – přepočtený počet zaměstnanců

Logaritmickou transformací mocninné regresní funkce se pak získá tvar:

$$\ln Y = \ln a + \alpha \cdot \ln OC + \beta \cdot \ln S. \quad (8)$$

V případě, kdy je aktuální výstup nižší než virtuální, potom veřejné domovy pro seniory (resp. soukromé domovy pro seniory) aktuálně produkují méně výstupu, než v případě, kdy by využily plně své vstupy a „produkční technologii“ soukromých domovů pro seniory (resp. veřejných domovů pro seniory).

V případě, kdy je aktuální výstup vyšší než virtuální, potom veřejné domovy pro seniory (resp. soukromé domovy pro seniory) aktuálně produkují více výstupu, než v případě, kdy by využily plně své vstupy a „produkční technologii“ soukromých domovů pro seniory (resp. veřejných domovů pro seniory).

Data

Možnost aplikovat výše uvedené přístupy je dána zejména dostupností údajů, jejich validitou a velikostí datového souboru. V oblasti sociálních služeb však dochází k problémům se statistickým sledováním údajů. Nejen, že se některé údaje nesledují vůbec, ale existuje vedle sebe několik zdrojů, které nejsou mezi sebou snadno propojitelné a vyskytují se v nich nevalidní hodnoty.

Datová základna pro tento příspěvek vznikla sloučením interních údajů MPSV ze systému OK-Poskytovatel a údajů dotačního řízení vztahujících se k roku 2009. Sledované údaje se vztahují nejen

k oblasti nákladovosti jednotlivých zařízení, ale jsou zde i údaje o kapacitě, počtu zaměstnanců a vybavenosti. Pouze tímto sloučením bylo možné získat kompletní údaje za jednotlivé organizace.

Sloučením výše uvedených dvou datových zdrojů došlo k redukci rozsahu souboru z původních 466 domovů pro seniory na 443. Pro možnost porovnání efektivnosti jednotlivých zařízení podle typu poskytovatele byly utvořeny dvě hlavní skupiny, a to „veřejné domovy pro seniory“, jež poskytují příspěvkové organizace obcí a krajů a organizační složky obcí, a „soukromé domovy pro seniory“, kam patří nestátní neziskové organizace a jiné právnické (např. s. r. o., a. s. či fyzické osoby).

V rámci prvního aplikačního přístupu pak byly navíc utvořeny kategorie podle velikosti domovů pro seniory. Malá zařízení obsahují nejvýše 20 lůžek, středně velká zařízení 21 až 60 lůžek a velká zařízení 61 lůžek a více. V kategorii malých zařízení se nacházelo 40 domovů pro seniory, v kategorii středně velkých zařízení 166 domovů pro seniory a v kategorii velkých zařízení 237 domovů pro seniory. V celkovém souboru se pak vyskytovalo 342 veřejných domovů pro seniory a 101 soukromých domovů pro seniory.

I když soukromé a veřejné domovy pro seniory poskytují obdobné služby, existují mezi nimi odlišnosti nejen v počtu lůžek, ale i ve stáří objektu, celkových nákladech i ve výši pracovních úvazků. Základní popisné charakteristiky výše uvedených kategorií jsou uvedeny v tabulce 1, 2 a 3.

Nutno zdůraznit, že výsledky aplikovaných přístupů musí být brány obezřetně, neboť nelze statisticky podchytit, že vyšší náklady jsou spojeny s vyšší kvalitou poskytnuté péče. Rovněž zde musíme zohlednit i skutečnost, že u zařízení, která mají vyšší míru klientů s vyšší potřebou péče, budou logicky vyšší i celkové náklady. Ani tyto informace však z dostupných údajů získat nelze.

Tabulka 3: Srovnání veřejných a soukromých domovů pro seniory, malá zařízení
Zdroj: interní údaje MPSV, vlastní výpočty

Ukazatel	Malá zařízení					
	PO obcí a krajů, obce s rozšířenou působností			NNO a komerce		
	Průměr	Medián	Směrodatná odchylka	Průměr	Medián	Směrodatná odchylka
Lůžka	15,5	18,00	5,3	15,4	15,00	3,7
Stáří objektu	114,5	68,9	122,4	76	60,00	98,4
Celkové náklady (mil. Kč)	6,8	6,3	6,6	4,0	3,9	1,8
Pracovní úvazky	10,5	9,6	5,6	10,2	10,40	4,2

Tabulka 2: Srovnání veřejných a soukromých domovů pro seniory, středně velká zařízení
Zdroj: interní údaje MPSV, vlastní výpočty

Ukazatel	Středně velká zařízení					
	PO obcí a krajů, obce s rozšířenou působností			NNO a komerce		
	Průměr	Medián	Směrodatná odchylka	Průměr	Medián	Směrodatná odchylka
Lůžka	44,2	45,0	11,1	36,6	36,0	10,0
Stáří objektu	81,9	50,0	105,5	71,3	47,5	70,8
Celkové náklady (mil. Kč)	12,5	12,7	4,1	10,0	9,0	5,4
Pracovní úvazky	25,4	26,8	8,0	22,7	20,4	9,2

Tabulka 3: Srovnání veřejných a soukromých domovů pro seniory, velká zařízení
Zdroj: interní údaje MPSV, vlastní výpočty

Ukazatel	Velká zařízení					
	PO obcí a krajů, obce s rozšířenou působností			NNO a komerce		
	Průměr	Medián	Směrodatná odchylka	Průměr	Medián	Směrodatná odchylka
Lůžka	122,8	104,0	55,2	94,2	82,5	46,0
Stáří objektu	53,4	32,5	71,3	98,1	75,5	163,9
Celkové náklady (mil. Kč)	30,4	27,2	14,1	24,0	20,4	12,8
Pracovní úvazky	63,7	56,5	28,3	55,1	48,9	21,0

Výsledky

Poměrové ukazatele – ukazatele výkonnosti

Ukazatele výkonnosti byly propočítány pro jednotlivé velikostní kategorie veřejných a soukromých domovů pro seniory. V tabulkách 4, 5 a 6 jsou u každé kategorie (malá, středně velká, velká zařízení) a typu (veřejné, soukromé) domovu pro seniory uvedeny vždy dva domovy pro seniory, a to se shodně nejnižším a nejvyšším počtem lůžek (vyjma třetí kategorie).

Tabulka 4: Srovnání veřejných a soukromých domovů pro seniory, malá zařízení
Zdroj: interní údaje MPSV, vlastní výpočty

Právní forma	Kraj, počet lůžek	TC/B	W/B	S/B	B/TC (mil. Kč)	B/S
Veřejné DS	Středočeský, 10	168 571	99 440	0,81	5,93	1,23
	Jihomoravský, 20	1 347 023	620 179	0,60	0,74	1,67
Soukromé DS	Moravskoslezský, 10	220 144	108 590	0,76	4,54	1,32
	Liberecký, 20	278 152	129 707	0,57	3,67	1,77

Tabulka 5: Srovnání veřejných a soukromých domovů pro seniory, středně velká zařízení
Zdroj: interní údaje MPSV, vlastní výpočty

Právní forma	Kraj, počet lůžek	TC/B	W/B	S/B	B/TC (mil. Kč)	B/S
Veřejné DS	hl. m. Praha, 23	303 085	142 763	0,73	3,30	1,37
	Jihočeský, 58	290 210	128 547	0,62	3,45	1,61
Soukromé DS	Královéhradecký, 23	254 574	117 369	0,52	3,93	1,94
	hl. m. Praha, 58	201 282	93 676	0,47	4,97	2,14

Tabulka 6: Srovnání veřejných a soukromých domovů pro seniory, velká zařízení
Zdroj: interní údaje MPSV, vlastní výpočty

Právní forma	Kraj, počet lůžek	TC/B	W/B	S/B	B/TC (mil. Kč)	B/S
Veřejné DS	Olomoucký, 61	433 595	177 695	0,59	2,31	1,69
	Jihomoravský, 343	250 951	116 316	0,5	3,98	1,99
Soukromé DS	Moravskoslezský, 62	302 224	106 584	0,74	3,31	1,35
	hl. m. Praha, 258	257 137	113 195	0,48	3,89	2,08

Pokud srovnáme jednotlivá zařízení o stejné kapacitě, je zřejmé, že mezi veřejnými a soukromými domovy pro seniory existují na základě vypočtených ukazatelů efektivnosti rozdíly. Například v kategorii malých zařízení byl vybrán veřejný domov pro seniory s 10 lůžky ze Středočeského kraje a soukromý domov pro seniory s 10 lůžky z Moravskoslezského kraje. Zatímco veřejný domov

pro seniory dosahoval celkových nákladů na jedno lůžko cca 169 tisíc Kč, v případě soukromého domova pro seniory to bylo 220 tisíc Kč na jedno lůžko. U soukromého domova pro seniory připadaly na jedno lůžko i vyšší mzdové náklady. Veřejný domov pro seniory lze tudíž z tohoto hlediska považovat za efektivnější. S rostoucím počtem lůžek se však jako efektivnější jeví soukromé domovy pro seniory. Z hlediska přepočteného počtu zaměstnanců na jedno lůžko pak dosahovaly vyšších hodnot veřejné domovy pro seniory, kromě soukromých domovů pro seniory s kapacitou vyšší než 61 lůžek.

V průměru je za každou velikostní kategorii vyšší celková a mzdová nákladovost na jednotku výstupu u veřejných domovů pro seniory (kromě velkých zařízení, kde vyšší celkové nákladovosti na jedno lůžko dosahují soukromé domovy pro seniory). Celkové postavení soukromých a veřejných domovů pro seniory je možno zhodnotit na základě indexů, které vycházejí z poměru průměrných ukazatelů za soukromé a veřejné domovy pro seniory (tabulka 7). Například index pro mzdové náklady na jednotku výstupu ($I_{W/B}$) vychází z tvaru:

$I_{W/B}$=průměrná hodnota „W/B“ za soukromé domovy pro seniory / průměrná hodnota „WB“ za veřejné domovy pro seniory

Tabulka 7: Indexy ukazatelů výkonnosti podle velikosti domovů pro seniory
Zdroj: vlastní výpočty

Kategorie	$I_{TC/B}$	$I_{W/B}$	$I_{S/B}$	$I_{B/TC}$	$I_{B/S}$
Malá zařízení	0,63	0,64	1,00	1,26	1,08
Středně velká zařízení	0,94	0,89	1,07	1,21	0,87
Velká zařízení	1,02	0,96	1,13	1,02	0,85

Výsledné hodnoty například indikují, že:

- jednotkové náklady jsou o 37 % (resp. 36 % v případě mzdových nákladů) vyšší u malých veřejných domovů pro seniory než u soukromých zařízení;
- středně velké soukromé domovy pro seniory produkují 121 % výstupu veřejných domovů pro seniory na jednotku nákladů (nákladová jednotka = mil. Kč) a 87 % výstupu veřejných domovů pro seniory na jednoho zaměstnance (přepočtené počty);
- u velkých soukromých domovů pro seniory jsou celkové náklady na jednotku výstupu o 2 % vyšší než v případě veřejných domovů pro seniory;
- velké soukromé domovy pro seniory mají o 13 % vyšší počet zaměstnanců na jedno lůžko (přepočtený) než velké veřejné domovy pro seniory.

Podle takto získaných výsledků lze tudíž shrnout, že v úhrnu dosahují nižší nákladovosti na jednotku výstupu soukromé domovy pro seniory, tzn. jsou považovány za efektivnější než veřejné domovy pro seniory. Současně pak v průměru dosahují i vyššího počtu zaměstnanců na jedno lůžko, než je tomu u veřejných domovů pro seniory.

Srovnání podle těchto ukazatelů však může být zavádějící, neboť nemůžeme z dostupných údajů zjistit, zda vyšší náklady u veřejných domovů pro seniory jsou dány tím, že zaměstnávají lépe kvalifikovaný personál, či zda poskytují kvalitnější péči, resp. jsou typické tím, že mají určitou skupinu klientů, kteří potřebují větší rozsah péče a nákladnější vybavení. Může proto nastat dvojí situace - domov pro seniory dosáhne uspokojivé míry efektivnosti, ale bude mít nižší kvalitu služeb, a na druhé straně se může vyskytnout zařízení s vysokými náklady na jednotku výstupu, které mohou znamenat vysokou kvalitu poskytnuté péče a spokojenost klienta.

Statistické odhady nákladových funkcí

Efektivitu daného zařízení lze rovněž vyjádřit jako poměr skutečných nákladů k nákladům očekávaným, vypočteným z odhadnutého modelu nákladové funkce, jež vyjadřuje vztah mezi náklady a výstupem, tj. počtem lůžek. V případě, kdy skutečné náklady jsou vyšší než očekávané, považuje se dané zařízení za neefektivní. Pokud jsou skutečné náklady nižší než očekávané, jde o efektivní zařízení.

Nákladové funkce byly odhadnuty jednotlivě pro veřejné (9, 10) a soukromé domovy pro seniory (11, 12). Výsledky odhadnutých modelů a jejich kvality jsou uvedeny v tabulce 8.

$$OC = e^{13,026} \cdot B^{0,871}, \quad (9) \qquad W = e^{12,410} \cdot B^{0,826}, \quad (10)$$

$$OC = e^{12,384} \cdot B^{1,011}, \quad (11) \qquad W = e^{11,621} \cdot B^{0,987}, \quad (12)$$

Tabulka 8: Výsledné hodnoty odhadnuté nákladové funkce (provozní náklady a mzdové náklady)

Koeficient	Odhadnutá funkce provozních nákladů		Odhadnutá funkce mzdových nákladů	
	Veřejné DS	Soukromé DS	Veřejné DS	Soukromé DS
$\ln\beta_0$	13,02614*	12,38410*	12,40962*	11,62124*
	-0,09332	-0,20556	-0,10218	-0,25495
β_1	0,87070*	1,01146*	0,82562*	0,98656*
	-0,02128	-0,0579	-0,0233	-0,07181
R^2	0,8312	0,7551	0,787	0,6559
R^2_{adj}	0,8301	0,7526	0,7863	0,6525
N	342	101	342	101

Pozn. směrodatná chyba odhadnutého parametru je uvedena v závorce
* statistická významnost odhadnutého parametru na 5% i 1% hladině významnosti

Výsledné poměry skutečných a predikovaných celkových nákladů jsou uvedeny v tabulce 9. Ilustrativně jsou opět uvedeny pro vybrané domovy pro seniory, které již byly popisovány v rámci prvního přístupu. Pohledem na tyto poměry zjišťujeme, že veřejný domov pro seniory s kapacitou 108 lůžek má o 28 % vyšší skutečné provozní náklady než očekávané, tzn. jedná se o neefektivní jednotku. Soukromý domov pro seniory s kapacitou 108 lůžek má však o 51 % nižší skutečné než očekávané provozní náklady, jedná se tedy o velmi efektivní jednotku. Oproti tomu veřejný domov pro seniory s kapacitou 20 lůžek v Jihomoravském kraji má o 337 % vyšší provozní náklady, než by se dalo očekávat. Na druhé straně má pak veřejný domov pro seniory s kapacitou 10 lůžek provozní náklady nižší než predikované, a to o 50 %. Pro existenci těchto „outliers“, tj. bodů extrémně vzdálených od regresní funkce, jsou možná dvě vysvětlení: a) opravdu se jedná o velmi efektivní nebo velmi neefektivní jednotky, nebo b) celkové náklady nelze vysvětlit tak jednoduchým modelem bez dalších proměnných, které zachycují činnost domovů pro seniory.

Tabulka 9: Poměr skutečných a očekávaných nákladů

Právní forma	Kraj, počet lůžek	Poměr provozních nákladů	Poměr mzdových nákladů
Veřejné domovy pro seniory	Středočeský, 10	**0,5**	0,61
	Jihomoravský, 20	**4,37**	**4,27**
Soukromé domovy pro seniory	Moravskoslezský, 10	0,9	1,01
	Liberecký, 20	1,12	1,21
Veřejné domovy pro seniory	hl. m. Praha, 23	1,01	1,01
	Jihočeský, 58	1,08	1,06
Soukromé domovy pro seniory	Královéhradecký, 23	1,03	1,10
	hl. m. Praha, 58	0,8	0,89
Veřejné domovy pro seniory	Olomoucký, 61	**1,62**	**1,48**
	Jihomoravský, 343	1,18	1,31
	Moravskoslezský, 108	1,28	1,55
Soukromé domovy pro seniory	Moravskoslezský, 62	1,21	1,01
	hl. m. Praha, 258	1,01	1,09
	Ústecký, 108	**0,49**	**0,46**

Kalkulace tzv. „produkčních mezer“

Produkční funkce podle výpočtového vztahu (7) byly odhadnuty jak pro veřejné, tak i pro soukromé domovy pro seniory. Pro veřejné domovy pro seniory byla odhadnuta ve tvaru (13) a pro soukromé domovy pro seniory ve tvaru (14). Výsledky odhadnutých modelů a jejich kvality jsou uvedeny v tabulce 10.

$$Y = e^{-5,4732} \cdot OC^{0,4724} \cdot S^{0,5046}, \qquad (13)$$

$$Y = e^{-2,1552} \cdot OC^{0,2407} \cdot S^{0,6077}, \qquad (14)$$

Tabulka 10: Výsledné hodnoty odhadnuté produkční funkce

Koeficient	Veřejné domovy pro seniory	Soukromé domovy pro seniory
lna	-5,4732* (0,73503)	-2,15522* (1,01143)
α	0,4724* (0,05421)	0,24068** (0,07705)
β	0,50455* (0,05242)	0,60774* (0,08277)
R^2	0,868	0,842
R^2_{adj}	0,867	0,839
N	342	101

Pozn. směrodatná chyba odhadnutého parametru je uvedena v závorce
* statistická významnost odhadnutého parametru na 5% i 1% hladině významnosti
** statistická významnost odhadnutého parametru na 5% hladině významnosti

„Produkční mezera“ byla vypočtena jako rozdíl aktuálního výstupu (tj. počet lůžek) a „virtuálního“ výstupu, který byl propočten na základě odhadnutých produkčních funkcí. Výsledné hodnoty jsou ilustrovány pouze na vybraných domovech pro seniory, které již byly uvedeny v rámci aplikace ukazatelů výkonnosti, a jsou uvedeny v tabulce 11.

Všechny vybrané domovy pro seniory (veřejné i soukromé) s kapacitou do 20 lůžek dosahují záporné hodnoty „produkční mezery“. To znamená, že „produkují“ méně výstupu, než by „produkovaly“ při využití svých vstupů a „produkční technologie“ soukromých, resp. veřejných domovů pro seniory. V případě středně velkých zařízení, tj. s počtem lůžek v intervalu 21 až 60, dochází k záporné „produkční mezeře“ pouze u veřejných domovů pro seniory, u soukromých pak dosahuje kladných hodnot. V praxi to tak znamená, že veřejné domovy pro seniory by dosáhly vyššího výstupu, pokud by při svých dosavadních zdrojích byly stejně produktivní jako v případě soukromého sektoru.

I když jsou tyto výsledky pouze ilustrativní, lze shrnout, že soukromé domovy pro seniory se zdají být více produktivní při využití svých produkčních faktorů, než veřejné domovy pro seniory. Tyto výsledky jsou do jisté míry konzistentní s hodnocením efektivnosti podle poměrových ukazatelů i nákladových funkcí.

Tabulka 11: „Produkční mezera“ veřejných a soukromých domovů pro seniory

Právní forma	Kraj, počet lůžek	"Produkční mezera"
Veřejné domovy pro seniory	Středočeský, 10	-1
	Jihomoravský, 20	-28
Soukromé domovy pro seniory	Moravskoslezský, 10	-3
	Liberecký, 20	-1
Veřejné domovy pro seniory	hl. m. Praha, 23	-7
	Jihočeský, 58	-8
Soukromé domovy pro seniory	Královéhradecký, 23	**1**
	hl. m. Praha, 58	**15**
Veřejné domovy pro seniory	Olomoucký, 61	-21
	Jihomoravský, 343	**27**
	Moravskoslezský, 108	-34
Soukromé domovy pro seniory	Moravskoslezský, 62	-5
	hl. m. Praha, 258	**92**
	Ústecký, 108	**35**

Závěr

Cílem toho příspěvku bylo nastínit možné přístupy měření efektivnosti v oblasti sociálních služeb. Oblast sociálních služeb nelze posuzovat jako integritu jednotlivých segmentů. Je nezbytné uplatnit jiné pohledy na hodnocení efektivnosti služeb sociální péče, sociální prevence a sociálního poradenství, neboť tyto oblasti jsou ovlivněny celou řadou kritérií a faktorů, jako například strukturou uživatelů, kritérii pro stanovení míry závislosti, aj.

Tento příspěvek byl zaměřen na zhodnocení efektivnosti domovů pro seniory podle typu poskytovatele. Pro možnost statistických analýz byly utvořeny dvě hlavní skupiny, a to veřejné domovy pro seniory (příspěvkové organizace obcí a krajů a organizační složky obcí) a soukromé domovy pro seniory (nestátní neziskové organizace a komerční sektor). Celkem byly aplikovány tři přístupy – poměrové ukazatele, statistické odhady nákladových funkcí a kalkulace tzv. „produkčních mezer". V úhrnu se pak jevily jako efektivnější soukromé domovy pro seniory. Tyto přístupy jsou však pouze ilustrativní, neboť oblast sociálních služeb disponuje roztříštěnou datovou základnou. Nelze tak stanovit výstupní proměnnou, která by zohledňovala kvalitu poskytnuté péče a spokojenost klienta v zařízení, resp. počet provedených úkonů sociální a zdravotní péče.

Všechny tři výše uvedené přístupy vycházely z hodnocení nákladovosti, přičemž zařízení, která dosahovala nižších nákladů, byla považována za efektivnější. Nicméně je důležité si uvědomit, že vyšší náklady mohou vyjadřovat vyšší ceny vstupů, např. vytápění, potraviny, mzdy pro kvalifikovaný personál, moderní vybavení či lepší péči o klienty. Proto je lepší při měření efektivnosti v oblasti sociálních služeb využít sofistikovanější postupy, které by do modelu zahrnuly více proměnných, jež souvisí s efektivním chodem daného zařízení. V dalším výzkumu tak budou aplikovány ekonometrické modely, zejména analýza obalových dat.

Literatura

ČERVENKOVÁ, A., BRUTHANSOVÁ, D., JEŘÁBKOVÁ, V., 2009. Ošetřovatelská a rehabilitační péče v pobytových zařízeních sociálních služeb. *Zdravotnictví v České republice, roč. 12, č. 3*. Praha. ISSN: 1213-6050.

DLOUHÝ, M., STRNAD, L., 1999. Nemocnice – kvalita, efektivita, finance. *IGA MZ ČR.* Praha.

JABLONSKÝ, J., DLOUHÝ, M., 2004. Modely hodnocení efektivnosti produkčních jednotek. *Professional publishing*, Praha. ISBN: 80-86419-49-5.

SLINTÁKOVÁ, B., 2003. Comparative Analysis of Efficiency of Public and Private Provision of Social Services. *E+M Ekonomie a management. 1/2003, VI ročník, str. 6 – 11*. ISSN:1212-3609.

JEL Classification: C12, C43, D24, H55

Summary

Estimating the efficiency of the social services' system

It is very important to ensure the effeciency of the public sector and public expanditures. Effeciency model's creating is a dynamic process that requires regular evaluation of the socioeconomic conditions. We must be able to measure consumption and quality of the provided services. But also we must consider the impact of other criteria and factors in this area.

This contribution is focused on estimating the effeciency of the most common residantial social care services – homes for the elderly. Effeciency will be examined in terms of type of registered providers. There will be applied three approaches: (1) ratio indicators, (2) statistical estimates of cost functions, and (3) calculation of „productivity gap".

Keywords: social services, efficiency, cost function, efficiency ratios, Cobb-Douglas production function.

Porovnanie sekvenčného a dávkového učenia pri metódach samo-organizujúcich sa máp

Lukáš Pastorek
lukas.pastorek@vse.cz
Doktorand oboru statistika
Školitel: prof. Ing. Hana Řezanková, CSc., (hana.rezankova@vse.cz)

Abstrakt: Táto práca sa zaoberá dvoma prístupmi k učeniu výpočtového modelu prostredníctvom adaptácie váh jednotlivých referenčných vektorov, ktoré tvoria váhovú konfiguráciu modelu v metódach strojového učenia. Špeciálna pozornosť je venovaná metódam učenia bez učiteľa (metóda k-priemerov, neurónový plyn, Kohonenove samo-organizujúce sa mapy). Tieto metódy nemajú žiadneho „učiteľa“, o ktorého by sa mohli oprieť a korigovať správanie v závislosti od vývoja hodnotiacej funkcie. Pri trénovaní nemajú žiadnu informáciu o rozdelení, zatriedení vektorov, alebo počte skupín. Klasická metóda k-priemerov je porovnávaná s výsledkami získanými dvoma metódami spadajúcimi do oblasti umelých neurónových výpočtových modelov. Každý z týchto modelov je natrénovaný s využitím dvoch učiacich algoritmov – sekvenčného a dávkového učenia. Sekvenčný učiaci algoritmus vyberie v každom iteratívnom kroku náhodne jeden vstupný vektor z celkového trénovacieho súboru. Dávkový algoritmus využíva možnosť adaptovať polohu vektorov až potom, ako sú váhovým vektorom predstavené všetky vstupné vektory prostredníctvom priradenia vstupného vektoru k svojmu najbližšiemu modelovému váhovému vektoru. Dopady voľby alternatívneho dávkového nastavenia učenia oproti sekvenčnému algoritmu u vyššie spomenutých metód sú hodnotené prostredníctvom priemernej štvorcovej odchýlky a výpočtového času. Metódy boli otestované na dátovom súbore pozostávajúcom z mapy vonkajšieho prostredia, ktorá je charakterizovaná relatívne vysokou dimenziou a vysokým počtom zhlukov.

Kľúčová slová: sekvenčný tréning, dávkový tréning, metóda *k*-priemerov, neurónový plyn, Kohonenova samo-organizujúca sa mapa

Úvod

Strojové učenie je jeden z technických prostriedkov umelej inteligencie, ktorý slúži k iteračnému natrénovaniu modelov, ktorých podstata stojí v metaforickom kopírovaní evolučne overených alebo prírodou vyvinutých procesov. Učenie ako základný princíp, ktorý spája všetky tieto metódy, je založený na spoznaní predkladanej trénovacej dátovej predlohy a nastavení váhovej konfigurácie modelu tak, aby, bol model po predložení podobných dátových signálov (vektorov) schopný reagovať podobne a zatriedil signály do rovnakých skupín napriek ich viacrozmernému charakteru. My sa zameriame na metódy umelej inteligencie, ktoré spadajú do oblasti *učenia bez učiteľa.*

Tieto metódy nemajú žiadneho „učiteľa“, o ktorého by sa mohli oprieť a korigovať správanie v závislosti od vývoja hodnotiacej funkcie. Pri trénovaní nemajú žiadnu informáciu o rozdelení, zatriedení vektorov, alebo počte skupín. Sú schopné učiť sa samostatne, pričom kopírujú pôvodné rozloženie dát vo vektorovom priestore v závislosti od zvoleného počtu skupín (zhlukov). Pri trénovaní modelu existujú dva prístupy učenia – *sekvenčný* a *dávkový*.

Sekvenčný (inkrementálny) učiaci algoritmus je založený na iteratívnom opakovaní rovnaké postupu, pri ktorom je v každom kroku náhodne vybraný jeden vstupný vektor z celkového trénovacieho súboru. V následnom kroku je počítaná vzdialenosť medzi vstupným vektorom a všetkými váhovými vektormi modelu. Ten váhový vektor, ktorý sa ukáže ako najbližší na základe zvolenej metrickej vzdialenosti, získa najvyššiu možnosť adaptovať svoju polohu. Ostatné vektory upravujú svoju polohu (váhy) podľa určitých pravidiel špecifických pre každú metódu.

Dávkový algoritmus (angl. *batch* algorithm) využíva možnosť adaptovať polohu vektorov až potom, ako sú váhovým vektorom predstavené *všetky* vstupné vektory prostredníctvom priradenia vstupného vektoru k svojmu najbližšiemu modelovému váhovému vektoru. Jediný krok, ktorý prebieha sekvenčne je práve toto rozdelenie priestoru vstupných vektorov do tzv. *Voronoivých regiónov*. Toto pokrytie je známe aj pod označením *Dirichletove pokrytie*. S každým vektorom sa bude asociovať práve jeden región, zodpovedajúci najbližšiemu modelovému vektoru ako „centru“ tohto regiónu. Po rozde-

lení celého priestoru dochádza k adaptácií váhových vektorov cez vážený aritmetický priemer váh okolia a vstupných vektorov.

Metódy

Implemetácia *sekvenčného* a *dávkového* učenia bude prezentovaná na jednom bežne používanom zhlukovacom algoritme (metóda *k-priemerov*) a dvoch významných zástupcoch samo-organizujúcich sa máp (*Kohonenove samoorganizujúce sa mapy* a algoritmus *neurónového plynu*).

Metóda k-priemerov

V prípade metódy *k-priemerov* (angl. *k-means*) so *sekvenčným* učením (MacQueen, 1967), vyberáme v každom iteračnom kroku náhodne jeden zo vstupných vektorov $\mathbf{x}_i$ z dátového súboru $\mathbf{X} = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathbf{R}^D, i = \{1,...,n\}\}$, kde n je dĺžka trénovacieho súboru a $\mathbf{R}^d$ je d-rozmerný vektorový priestor. Tento vstupný vektor následne predložíme populácií váhových vektorov modelu $\mathbf{C} = \{\mathbf{c}_j \mid \mathbf{c}_j \in \mathbf{R}^D, j = \{1,...,k\}\}$, kde c_j je váhový vektor modelu, ktorý označujeme ako *centroid j-teho zhluku* a k je zvolený *počet* zhlukov, do ktorých chceme dáta rozdeliť. *Centroidy* súťažia o najbližšiu pozíciu k predloženému vstupnému vektoru z hľadiska euklidovskej vzdialenosti

$$\|\mathbf{x} - \mathbf{c}_v\| = \arg\min_j \|\mathbf{x}_i - \mathbf{c}_j\|,$$

kde $\mathbf{c}_v$ je *víťazný centroid* (referenčný váhový vektor), ktorý reprezentuje zhluk, do ktorého bude vstupný vektor prvotne patriť. Z celkového množstva *centroidov* svoju polohu adaptuje len tento *víťazný centroid* na základe vzorca:

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + \alpha(t)[\mathbf{x}(t) - \mathbf{c}_j(t)],$$

pričom t reprezentuje *čas* (iteračný krok) a $\alpha(t) \in [0,1]$ *rýchlosť učenia*. Celkový čas (celkový počet iterácií T) sa odvíja od počtu cyklov cez celú trénovaciu množinu (dĺžka trénovacieho súboru násobená počet cyklov). Ako funkciu *rýchlosti učenia* môžeme zvoliť akúkoľvek klesajúcu funkciu. Po adaptácií pozície *centroidu* je *centroidom* predložený nový vstupný vektor a proces sa opakuje až do momentu, kým sa nevyčerpá celá trénovacia množina. Po vyprázdnení množiny sa spustí odznovu nový cyklus.

Dávkový algoritmus učenia v prvom rade priradí vstupné vektory k najbližším centroidom (rezprezentujúcim zhluky) a až potom, ako je jasné, ktoré vektory prináležia ktorému zhluku, sa naraz adaptujú váhy všetkých centroidov podľa vzorca:

$$\mathbf{c}_j(t+1) = \frac{\sum_{i=1}^{n} w_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} w_{ij}}, \forall j = 1,...,k.$$

Váhy w_{ij} definujú, či vstupný vektor danému zhluku patrí alebo nepatrí, podľa predpisu

$$\forall w_{ij} : w_{iv} = \begin{cases} 1 & \|\mathbf{x}_i - \mathbf{c}_v\| = \arg\min_j \|\mathbf{x}_i - \mathbf{c}_j\| \\ 0 & \text{inak} \end{cases}.$$

Kohonenove samo-organizujúce sa mapy

Kohonenove samoorganizujúce sa mapy, známe skôr pod skratkou *SOM* (Kohonen, 2001) využívame k transformácií nelineárnych vzťahov v trénovacej množine dát prostredníctvom sady uzlov (v prípade neurónových sietí sa využíva výraz *neurón*) na menej rozmerný (1- alebo 2- rozmerný) priestor $\mathbf{R}^d \rightarrow \mathbf{R}^2$. S každým neurónom v mriežke je asociovaný práve jeden váhový vektor modelu. S využitím iteračného trénovania a jednoduchých geometrických vzťahov medzi neurónmi na fixnej topolo-

gickej mriežke *SOM* zachováva rozloženie pôvodných viacrozmerných dát a i vzťahy medzi nimi. Podobné signály (vstupné vektory) zo vstupného priestoru evokujú tiež podobnú odozvu v blízkych neurónoch na mriežke vo výstupnom priestore.

Kohonenove mapy pozostávajú z dvoch vrstiev (vstupná a výstupná vrstva). Vstupnú vrstvu považujeme za distribučnú, kde sú vstupné vektory prezentované neurónom vo výstupnom vrstve. Výstupná vrstva je tvorená spomínanými váhovými vektormi.

Zachovávame rovnaké označenie váhových vektorov neurónov a vstupných vektorov ako pri *k*-priemeroch.

Sekvenčný učiaci algoritmus postupuje identicky ako v prípade metódy *k*-priemerov s tým rozdielom, že svoju váhu neupravuje len víťazný neurón ale i jeho topologický susedia v mriežke. Adaptácia sa uskutočňuje podľa vzorca:

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + \alpha(t) h_{v_j}(t) \left[\mathbf{x}(t) - \mathbf{c}_j(t)\right],$$

kde *t* predstavuje čas, $\alpha(t)$ rýchlosť učenia v čase *t*, $h_{v_i}(t)$ funkciu okolia (susedstva) okolo víťazného neurónu c_v v čase t a $\mathbf{x}(t)$ je už vyššie spomínaný vstupný vektor v čase *t*. Funkcia okolia, ktorá hovorí o sile (váhe) adaptácie víťazného neurónu a susedných neurónov, môže mať tvar gaussovskej krivky $h_{v_i}(t) = e^{-d_{vj}^2 / 2\sigma_t^2}$, kde d_{vj} je vzdialenosť neurónov od víťazného neurónu v priestore mriežky $d_{vj} = \left\| r_v - r_j \right\|$, kde *r* sú súradnice neurónov v 2-rozmernej mriežke a σ_t je rádius okolia, ktorý definuje vzdialenosť alebo polomer dosahu na ostatné neuróny na mriežke v okolí víťazných neurónov, ktoré môžu adaptovať svoju polohu.

V prípade *dávkového* učenia sa váhové vektory neurónov upravia podľa vzorca, ktorý je priemerom súčinov váh hodnôt funkcie okolia a hodnôt vstupných vektorov. Každý neurón je adaptovaný podielom sumy súčinov všetkých vstupných vektorov a váh okolia, ktoré boli asociované s daným neurónom pri danom vstupnom vektore a celkovej sumy váh okolia spomínaného neurónu.

$$\mathbf{c}_j(t+1) = \frac{\sum_{i=1}^{n} h_{v_j}(t) \mathbf{x}_i}{\sum_{i=1}^{n} h_{v_j}(t)}, \forall j = 1,...,k.$$

Algoritmus neurónového plynu

Na rozdiel od *SOM*-u algoritmus *neurónového plynu* (angl. *Neural Gas*; Martinetz a Schulten,1991) neobsahuje fixnú topológiu siete, ale obsahuje zoznam tzv. „*poradie susedov*“ (angl. ”neighborhood ranking”).

Počas *sekvenčného* tréningu sú váhové vektory modelu radené podľa euklidovskej vzdialenosti od vstupného vektora $\mathbf{x}_i, (\mathbf{c}_{j_0}, \mathbf{c}_{j_1}, ..., \mathbf{c}_{j_{k-1}})$, kde $\mathbf{c}_{j_0}$ je najbližší váhový vektor k vstupnému vektoru $\mathbf{x}_i$, $\mathbf{c}_{j_1}$ je druhý najbližší váhový vektor a $\mathbf{c}_{j_m}$, $m = 0,...,k-1$ je váhový vektor modelu, pre ktorý existuje *m* vektorov $\mathbf{c}_j$ takých, že

$$\left\| \mathbf{x} - \mathbf{c}_j \right\| < \left\| \mathbf{x} - \mathbf{c}_{j_m} \right\|.$$

Index *m*, ktorý je spojený s každým vektorom $\mathbf{c}_j$, je určený funkciou $m(\mathbf{x}_i, \mathbf{c}_j)$ závislej na $\mathbf{x}_i$ a $\mathbf{C} = \{\mathbf{c}_j \mid \mathbf{c}_j \in \mathbf{R}^D, j = \{1,...,k\}\}$. Adaptačný krok je vyjadrený vzorcom

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + \alpha(t) h_\lambda\left(m(\mathbf{x}_i, \mathbf{c}_j)\right) \left[\mathbf{x}(t) - \mathbf{c}_j(t)\right],$$

kde $\alpha(t)$ je rýchlosť učenia a $h_\lambda\left(m\left(\mathbf{x}_i,\mathbf{c}_j\right)\right)=e^{-m/\lambda(t)}$ je exponenciálna forma funkcie okolia. Parameter $\lambda(t)$ určuje počet váhových vektorov, ktoré upravujú svoje váhy a je definovaný diskrétnou klesajúcou funkciou: $\lambda(t)=\left(\lambda_f/\lambda_i\right)^{t/T}$.

Pri *dávkovom* učení sú váhové vektory upravené ako vážený priemer príslušných váh poradí susedov priradených ku každému vstupnému vektoru

$$\mathbf{c}_j(t+1)=\frac{\sum_{i=1}^{n}h_\lambda\left(m\left(\mathbf{x}_i,\mathbf{c}_j\right)\right)\mathbf{x}_i}{\sum_{i=1}^{n}h_\lambda\left(m\left(\mathbf{x}_i,\mathbf{c}_j\right)\right)},\forall j=1,\ldots,k$$

Experiment

Uskutočnili sme porovnanie všetkých spomenutých metód pri použití oboch spôsobov učenia. Metódy sme otestovali v prostredí *Matlab* a s využitím *SOM toolboxu*, (Vesanto, 2000) na dátovom súbore, ktorý obsahuje 39883 štatistických jednotiek. Jednotka je v našom prípade jeden viacrozmerný obrazový signál tzv. *SURF deskriptor*, ktorý vo svojich 64 rozmeroch nesie lokálnu informáciu o bode na digitálnej fotografií. Tieto fotografie boli nasnímané kamerou robota pri prechode vonkajším prostredím. Algoritmus robota bol nastavený tak, aby si všímal zaujímavé body na každom obraze a vlastnosti bodu spolu s informáciami o okolí registrovaného bodu uložil do jednotlivých zložiek 64-rozmerného vektora – *SURF deskriptora*. Takto bol zostrojený celý dátový súbor, ktorý predstavuje nasnímanú mapu prostredia, ktorú robot pri svojej ceste videl a ktorá mu bude slúžiť ako podpora pri navigácií pri budúcej ceste uskutočnenej bez učiteľa.

Všetky metódy sme inicializovali s rovnakým nastavením váhových vektorov modelu, rovnakým počtom cyklov (10) a identickým počtom zhlukov. Našou snahou bude zhlukovať *SURF deskriptory* naprieč obrazmi tak, aby sme dokázali do zhluku zaradiť rovnaké objekty. Teda, aby zhluky obsahovali všetky *deskriptory* popisujúce rovnaké miesto na každej fotografii.

Tabuľka 1: Porovnanie priemerných štvorcových odchýlok a výpočtového času pri 64-rozmerných vektoroch Zdroj: vlastné výpočty

Metóda	Učiaci algoritmus	Zvolený počet zhlukov	Priemerná štvorcová odchýlka	Výpočtový čas (sekundy)
***k*-priemerov**	*sekvenčný*	9	0,4371	4,21
		99	0,2843	12,73
		1008	0,2282	102,51
	dávkový	9	0,5222	0,81
		99	0,3066	2,74
		1008	0,2178	20,38
Kohonenove samo-organizujúce sa mapy	*sekvenčný*	9	0,5341	22,63
		99	0,3614	40,43
		1008	0,3461	247,84
	dávkový	9	0,5382	1,04
		99	0,3556	1,62
		1008	0,3070	14,34
neurónový plyn	*sekvenčný*	9	0,4361	16,99
		99	0,2799	51,38
		1008	0,1696	428,29
	dávkový	9	0,4386	3,19
		99	0,2557	26,42
		1008	0,1706	208,13

Tabuľka 2: Porovnanie priemerných štvorcových odchýlok a výpočtového času pri 6-rozmerných vektoroch Zdroj: vlastné výpočty

Metóda	Učiaci algoritmus	Zvolený počet zhlukov	Priemerná štvorcová odchýlka	Výpočtový čas (sekundy)
***k*-priemerov**	*sekvenčný*	9	0,0454	3,03
		99	0,0239	4,33
		1008	0,0131	19,03
	dávkový	9	0,0445	0,23
		99	0,0254	1,99
		1008	0,0164	17,87
Kohonenove samo-organizujúce sa mapy	*sekvenčný*	9	0,0531	20,20
		99	0,0340	22,03
		1008	0,0322	54,35
	dávkový	9	0,0540	0,56
		99	0,0317	0,76
		1008	0,0275	9,39
neurónový plyn	*sekvenčný*	9	0,0442	13,53
		99	0,0236	19,58
		1008	0,0129	84,04
	dávkový	9	0,0452	1,53
		99	0,0240	7,01
		1008	0,0127	84,14

Skúsili sme vziať do úvahy rôzne nastavenie počtu zhlukov (9;99;1008) a na porovnanie veľkosti chýb sme zhlukovali i identické dáta ale s redukovaným počtom rozmerom (prvých 6 dimenzií). Počty zhlukov sme volili v závislosti od požiadaviek metódy *SOM*, ktorej topológia mriežky diktuje usporiadanie neurónov a pomer strán. Aby boli zachované rovnaké podmienky, tieto počty sme využili aj pri iných metódach, i keď v prípade *k-priemerov* a *neurónového plynu* by bolo možné použiť zaokrúhlené počty (10;100;1000).

Z tabuľky 1 môžete vyčítať, že z hľadiska najmenšej chybovosti najlepšie obstála metóda *neurónového plynu* nasledovaná metódou *k-priemerov*. Naopak najhoršie, s výrazným desatinným odstupom, sa umiestnila metóda *Kohonenových samo-organizujúcich sa máp*. V prípade *neurónového plynu* sa *dávkový* algoritmus úspešnosťou vyrovnáva *sekvenčnému* algoritmu. U metódy *k-priemerov* sa lepšie umiestnil *sekvenčný* algoritmus pri nižšom počte zhlukov, naopak pri najväčšom počte zhlukov slabo zaostáva za *dávkovým* algoritmom. *Kohonenove samo-organizujúce sa mapy* vykazujú lepší výkon pri *dávkovom* algoritme, ktorý je však stále ďaleko za výkonnosťou zvyšných dvoch metód.

Neurónový plyn napriek úspešnosti z hľadiska priemernej štvorcovej odchýlky skončil s priepastným rozdielmi na poslednom mieste v rýchlosti výpočtu. V prípade veľkého počtu zhlukov výrazne zaostal nielen pri *sekvenčnom* ale dokonca i pri *dávkovom* algoritme.

Pri porovnaní výsledkov z tabuľky 1 a 2 vidíme, že pri nízkej dimenzií nie je výrazný rozdiel medzi metódami, i keď ako najslabšia metóda v tabuľke 1 sa ukázali byť opäť *Kohonenove samo-organizujúce sa mapy*. Rozdiely sú však veľmi malé a je zložité jednoznačne určiť víťaza.

Záver

V práci sme sa snažili poukázať na dva prístupy k učeniu samo-organizujúcich sa výpočtových modelov. Výsledky, ktoré sme získali, nám odkrývajú slabosti, ktoré sa prejavujú s rastúcim počtom zhlukov a dimenziou, pričom sa nemusia nevyhnutne prejaviť pri riešení úloh s malým počtom skupín, či rozmerov. Musíme však vziať do úvahy, že výsledky sú ovplyvnené aj voľbou kritéria, ktoré si zvolíme. Preto by sme sa mali do budúcna obrátiť aj na iné ako sú štandardné ukazovatele odchýlok, aby sme sa vyhli predčasným a mylným záverom o vhodnosti metódy. Dávkový algoritmus, napriek svojej relatívne jednoduchej teoretickej konštrukcií nebol aplikovaný na všetky existujúce modely z dôvodu zložitej alebo nemožnej implementácie. Tieto modely ako sú napr. rastúci neurónový plyn

trpia práve pomalou rýchlosťou výpočtu, preto by som sa rád v budúcej práci zameral na implementáciu tohto učenia do tohto algoritmu.

Literatura

KOHONEN, T., 2001. *Self-organizing maps*. (3rd ed.). Berlin: Springer.

MACQUEEN, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations. In: LUCIEN, M. a J. NEYMAN, *eds. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, s. 281-297.

MARTINETZ, T. M. a K. J. Schulten., 1991. A „neural-gas" network learns topologies. In: KOHONEN, T., K. MAKISARA a K. SIMULA, *eds. Artificial Neural Networks*, North Holland, Amsterdam, s. 397-434

PASTOREK, L., 2011. Využitie samoorganizujúcich sa máp pri navigácií autonómnych robotov. *Forum Statisticum Slovacum*. Bratislava: SŠDŠ, roč. 7, č.7/2011, s. 163-167.

SOBÍŠEK, L. et al., 2011. Aplikácia shlukové analýzy na data z pojišťoven. *Forum Statisticum Slovacum*. Bratislava: SŠDŠ, roč. 7, č.5/2011, s. 145-151.

VESANTO, J. et al., 2000. *SOM Tool-box for Matlab 5*: Manual and tutorial [online]. poslední aktualizace 17. 3. 2005, Dostupné z http://www.cis.hut.fi/somtoolbox/.

JEL Classification: C45.

Dedikace: Táto práca bola spracovaná s finančnou podporou grantu IGA VSE F4/5/2011.

Summary

The Comparison of the Sequential and Batch Learning Methods for Self-Organizing Maps

This paper concerns with two different approaches to computational model learning through adaptation of the reference vector features, which constitutes model weights in the machine learning methods. Special attention is given to unsupervised learning methods (K-means method, Neural Gas and Kohonen's Self-Organizing Maps). These methods do not rely on any "teacher", which would lead dynamic evolutionary process to the correct adaptation with regards to evaluation function. Computational models have no prior information on distribution, correct classification, or the number of groups.

The performance of the classic method, such as k-means, is compared with the results obtained by two methods belonging to the artificial neural computing models. Each of these methods is tested using two learning algorithms – sequential and batch learning algorithms.

In the sequential learning algorithm random input vector is iteratively chosen from the dataset. Batch algorithm adapts position of the model vectors according to the "best matching unit" rule" (closest reference model weight vector adapts position) after all input vectors are presented to the weight vectors to compete for the closest position to the input vector. Impacts of the alternative batch learning against sequential algorithm was evaluated for all above mentioned methods from the point of the mean square error and computation time. Methods were tested on the dataset consisting of the map of the outdoor environment, which is characterized by relatively high dimension and high number of clusters.

Keywords: sequential training, batch training, *K*-means method, Neural Gas, Kohonen´s Self-Organizing Map.

AutoClass – klasifikační metoda založená na bayesovském teorému

Lukáš Sobíšek
Xsobl06@vse.cz
Doktorand oboru statistika
Školitel: prof. Ing. Hana Řezanková, CSc., (rezanka@vse.cz)

Abstrakt: Příspěvek popisuje bayesovskou klasifikační metodu implementovanou v programu AutoClass, který byl vytvořen výzkumníky v NASA. Jedná se o klasifikační algoritmus vycházející z naivního bayesovského klasifikátoru.

Na rozdíl od metod, které jednoznačně přiřazují objekty do tříd, bayesovský modelový přístup hledá v datech latentní třídy, které nejvýstižněji charakterizují jednotlivé skupiny pozorovaných objektů. Tyto třídy popisuje pomocí modelů latentních tříd. AutoClass nabízí čtyři modely latentních tříd. Tyto modely vycházejí z klasického modelu směsí rozdělení, který předpokládá podmíněnou nezávislost. Při odhadu rozdělení pravděpodobnosti parametrů modelu podmíněné empirickými daty se vychází z Bayesovy věty. Parametry modelu odhaduje jednoduše přímo z dat metodou maximální věrohodnosti v případě učení s učitelem. V situaci učení bez učitele využívá iterativní EM algoritmus pro odhad parametrů.

AutoClass byl aplikován a testován na mnoha velkých datových souborech nejen výzkumníky v NASA, ale i statistiky působícími v průmyslu a akademiky. Od novějších verzí AutoClass umožňuje kombinovat vstupní proměnné spojité a nespojité. Poradí si s chybějícími hodnotami. Algoritmus umožňuje klasifikovat pomocí proměnných, které jsou podmíněně nezávislé při platnosti hypotézy o modelu. Automaticky volí počet tříd a komplikovanost popisu tříd.

Klíčová slova: AutoClass, bayesovská klasifikační metoda, naivní bayesovský klasifikátor, EM algoritmus, model směsí rozdělení

Úvod

Prvky libovolného souboru objektů, které je možno popsat, lze začlenit do dílčích skupin. Klasifikační úloha si klade za cíl nalézt v datovém souboru vnitřně homogenní a navenek heterogenní skupiny pozorovaných objektů. Rovněž lze seskupovat sledované vlastnosti objektů, tj. proměnné. Identifikace dílčích skupin pomocí klasifikačních metod se řadí do explorativní analýzy dat a je používána v různorodých vědních oborech, např. biochemii, astrofyzice, robotice, marketingovém výzkumu aj.

AutoClass je softwarový systém vyvinutý pracovníky NASA v Ames Research Center (AMC), který obsahuje robustní klasifikační metodu odvozenou z teoretického základu bayesovské statistiky. Volně dostupný program AutoClass C naprogramovala v jazyce C doktorka Diane Cook a Joseph Potts z University of Texas v Arlingtonu. AutoClass C je volně k dispozici na stránkách NASA věnovaným programu AutoClass.

Cílem klasifikační úlohy z pohledu bayesovské klasifikační teorie je nalézt množinu latentních tříd (LC), které reprezentují data a nalezené třídy popsat. K tomu využívá několik statistických modelů LC. V kontextu s bayesovskou teorií je skupina v dalším textu zmiňována pod pojmem *třída*.

Tento příspěvek se zaměřuje na popis klasifikace pozorovaných objektů s využitím bayesovké statistiky. Následující sekce definuje základní pojmy a princip bayesovké klasifikace. Další část popisuje postup řešení klasifikační úlohy v programu AutoClass, základní model směsí rozdělení (*Finite Mixture Model*) a specifikace modelů LC dostupných v programu. Poslední část ukazuje aplikace a vlastnosti programu AutoClass.

Bayesovská klasifikační teorie

Oproti klasickým metodám, které přiřazují jednotlivé objekty jednoznačně do tříd (jako jsou metody shlukové analýzy pro pevné shlukování), bayesovský modelový přístup hledá v datech takové třídy,

které co „nejlépe“ charakterizují náležící objekty. Místo jednoznačného přiřazení je pro objekty odhadnuta pravděpodobnost příslušnosti k jednotlivým třídám. Součet těchto pravděpodobností se pro každý objekt rovná hodnotě jedna. Pokud každý objekt náleží do jedné ze tříd s 99 % a vyšší pravděpodobností, třídy jsou navzájem heterogenní. Naopak v případě, kdy má několik objektů pravděpodobnost příslušnosti stejnou, třídy mají významné průniky a klasifikaci nelze považovat za relevantní. Jedná se tedy o fuzzy přístup.

Pravděpodobnostní model Gelman et al. (2004) popisují jako pravděpodobnostní funkci nebo hustotu pravděpodobnosti, z níž mohla empirická data vzniknout. Při odhadu rozdělení pravděpodobnosti parametrů modelu podmíněné empirickými daty X[1] se vychází z *Bayesovy věty*

$$P(\theta|X) = \frac{P(\theta)\cdot P(X|\theta)}{P(X)}. \tag{1}$$

Aposteriorní znalost o parametrech modelu $P(\theta|X)$ je úměrná součinu apriorní znalosti o parametrech a věrohodnosti. Aposteriorní znalost je obsažena v posteriorním pravděpodobnostním rozdělením. Apriorní znalost o parametrech (o zastoupení jednotlivých tříd modelu) je vyjádřena pravděpodobnostním rozdělením parametrů $P(\theta)$. Věrohodnost $P(X|\theta)$ obsahuje informace o parametrech, které jsou obsaženy v proměnné X, jedná se tedy o pravděpodobnost, že daná data vzniknou, podmíněná hodnotami parametrů modelu. Marginální věrohodnost $P(X)$ se nachází ve jmenovateli.

Pro dané X existuje T potenciálních modelů $t = 1, 2, \ldots, T$ s parametry θ_t. Při odhadu nejpravděpodobnějšího modelu, tedy modelu s největší aposteriorní pravděpodobností (MAP – maximum posteriory) θ_{MAP} lze ze vzorce (1) vypustit jmenovatel. V případě, že není důvod favorizovat žádný z modelů, ze vzorce se vyřadí $P(\theta)$ a $\theta_t = \theta_{MAP}$ je model s maximální věrohodností

$$P(X|\theta) = \max_t P(X|\theta_t). \tag{2}$$

Berka (2003) popisuje dva možné přístupy, jak stanovit aposteriorní pravděpodobnost pro více proměnných $P(\theta|X_1, X_2, ..., X_K)$. Prvním je *naivní bayesovský klasifikátor*, který vychází z podmíněné nezávislosti jednotlivých proměnných. Druhý přístup se dokáže vypořádat s neplatnou podmínkou nezávislosti proměnných pomocí *bayesovské sítě*.

Klasifikační úloha v programu AutoClass C

Robustní algoritmus v AutoClass C modeluje data jako směs podmíněných nezávislých latentních tříd (dále jen tříd). Základním modelem je klasický model směsí rozdělení (FMM – *Finite Mixture Model*). AutoClass využívá normální rozdělení pro spojité proměnné X a Bernoulliho, Poissonovo rozdělení pro diskrétní proměnné X. Klasifikace je definována jako množina tříd a pravděpodobností každé třídy. Pod pojmem třída se skrývá parametrizované pravděpodobnostní rozdělení.

Algoritmus hledá pro každou třídu C_j, $j = 1, 2, \ldots, J$ nejpravděpodobnější vektor hodnot parametrů $\vec{\theta}_j$ a její model, tedy pravděpodobnostní rozdělení nebo hustotu pravděpodobnosti $P(X|\theta_j)$. Hodnotu parametru lze interpretovat jako oblast hodnot proměnné, ve které dominuje pravděpodobnost náležení do dané třídy.

Axiomem FMM je *podmíněná nezávislost*[2] jednotlivých proměnných $X_1, X_2, \ldots, X_K$. Podmíněnou nezávislost proměnných X_k, $k = 1, 2, \ldots, K$, lze vyjádřit vztahem

$$P(\vec{\mathbf{X}}_i | X_i \in C_j) = \prod_{k=1}^{K} P(X_{ik} | X_i \in C_j), \tag{3}$$

[1] Teorie je vysvětlována pro data s jednou proměnnou.

[2] Jestliže je známa náležitost všech objektů do jednotlivých tříd C_j modelu θ_t, tyto proměnné jsou navzájem nezávislé. Proměnné mohou být selektivně korelovány v rámci jednotlivých tříd.

kde $\vec{\mathbf{X}}_i = \{X_{i1},...,X_{iK}\}$ je vektor hodnot proměnných pro *i*-tý objekt (*i* = 1, 2, …, *n*). Kombinací vzorce podmíněné nezávislosti (3) a vzorce (1) je získán FMM, viz (Cheeseman a Stutz, 1996)

$$P(\theta|\mathbf{X}) = \frac{P(\theta)}{P(\mathbf{X})} \cdot \prod_{k=1}^{K} P(X_k|\theta), \tag{4}$$

kde **X** je datová matice zahrnující všech *n* objektů. MAP odhad θ a $P(X|\theta)$, potřebné pro specifikaci modelu, je mnohem jednodušší získat v případě učení s učitelem. Jmenovatel je v případě hledání MAP odhadů vynechán. Algoritmus spočítá $P(X|\theta)$ a $P(\theta)$ zvlášť v každé třídě na základě četnosti výskytů hodnot metodou maximální věrohodnosti. V případě učení bez učitele se pro hledání $P(X|\theta)$ ve vícerozměrném prostoru hodnot $\mathbf{X}=\{x_{11}, \ldots, x_{ik}, \ldots, x_{IK}\}$ postupuje pomocí *Expectation Maximization (EM) algoritmu*, viz (Dempster et al., 1997). V inicializační fázi se náhodně volí θ_j. V cyklu se na základě zvolených θ počítají $P(X|\theta)$. Z takto spočítaných hodnot se určí θ_{MAP}. Postupuje se iterativně. Na závěr jsou vypočteny pravděpodobnosti náležení objektu do třídy $P(\theta|X)$. Výstupem je nalezená množina tříd a vektor náležitosti do jednotlivých tříd.

Modely LC v programu AutoClass

Rozlišují se dva typy parametry modelů:

- Diskrétní parametry *T*, které popisují model obecně. Většinou specifikují věrohodnostní funkci. Parametry *T* mohou například definovat, zda jsou dvě proměnné korelovány nebo určují počet tříd v klasifikaci.
- Spojité volné parametry *V*, jako jsou například míra korelace nebo relativní velikost tříd.

Software odhaduje na základě věrohodnosti, zda se vyplatí přidat parametr, tedy zda má parametr signifikantní přínos pro vysvětlení dat na úkor zvýšení počtu parametrů.

Z důvodu výpočetních obtíží (součty a integrály ve vzorci 4) program AutoClassu nabízí čtyři relativně jednoduché robustní statistické modely.

Jednoduchý multinomický m. modeluje jednu věrohodnostní funkci, která se řídí multinomickým rozdělením. Parametry *V* jsou vektor pravděpodobností každé hodnoty X_{ik}. Model je určen pro diskrétní podmíněně nezávislé proměnné, které mohou zahrnovat chybějící hodnoty.

Jednoduchý normální CN m. modeluje věrohodnostní funkci s normálním rozdělením s parametry střední hodnota μ a rozptyl σ, kde $P(\mu)$ ~ normální rozdělení a $P(\sigma)$ ~ logaritmicko rovnoměrné rozdělení. Připouští chybu měření, ale předpokládá se, že je konstantní a relativně malá ve srovnání s variabilitou. Model je aplikovatelný pro spojité proměnné bez chybějících hodnot. Nepřímo ho lze použít pro zdola omezené spojité proměnné po provedení logaritmické transformace a pro zdola i shora omezené spojité proměnné po log-odds transformaci.

Jednoduchý normální CM m. – Model lze použít pro stejný typ proměnných jako CN model. Oproti předchozímu modelu se liší v tom, že umožňuje modelovat chybějící pozorování. Parametry modelu jsou binární pravděpodobnost výskytu hodnoty, μ a σ.

Vícerozměrný normální m. lze využít pro spojité lokálně nezávislé proměnné s chybějícími hodnotami. Oproti předchozím metodám modeluje množinu spojitých proměnných jako věrohodnost. Parametry věrohodnosti jsou vektor středníchhodnot **μ** a kovariační matice **C**.

Modely blíže definují[3] Hanson et al.(1991).

[3] Parametry T, V, normovanou věrohodnost a apriorní znalost o parametrech.

Význam programu AutoClass

AutoClass byl aplikován a testován na mnoha velkých datových souborech nejen výzkumníky v NASA, ale i statistiky působícími v průmyslu a akademiky. Díky tomuto programu byly objeveny nové, nezávisle ověřené jevy. Mezi publikované případové studie patří objevení nových tříd infračervených hvězd ve spektrálním katalogu s nízkým rozlišením. Data byla pořízena pomocí družice IRAS. Provedla se nová klasifikace letišť na základě databáze všech letišť v USA. Byly objeveny nové třídy proteinů, intronů a dalších vzorů v DNA/proteinové sekvenci. Kvalita shlukování byla úspěšně otestována také na rozsáhlých datech extrahovaných ze satelitní mapy regionu v Kansasu o rozloze 30 km^2. Tyto studie blíže popisuje Cheeseman a Stutz (1996). V rozsáhlých testech byly prokázány obecně lepší výsledky klasifikace pomocí programu AutoClass ve srovnání s klasickými metodami. O klasických metodách více pojednává Řezanková et al. (2009).

Od novějších verzí AutoClass umožňuje kombinovat vstupní proměnné spojité a nespojité. Poradí si s chybějícími hodnotami a automaticky vybírá počet tříd. Algoritmus umožňuje klasifikovat pomocí proměnných, které jsou podmíněně nezávislé při platnosti hypotézy o modelu. Automaticky volí počet tříd a komplikovanost popisu tříd. Bližší popis a matematické vyjádření vlastností algoritmu viz (Hanson et al., 1991).

Závěr

Ve srovnání s jinými vícerozměrnými metodami shlukování disponuje bayesovská klasifikace řadou výhod. Jedná se o robustní přístup, který není zatížen předpoklady o vstupních datech a umožňuje provádět exploratorní analýzu dat s proměnnými různých typů. V dřívějších desetiletích se z důvodu výpočetní náročnosti o tomto přístupu spíše teoretizovalo. Výpočetní technologie dnešní doby již umožňuje aplikaci na reálná, libovolně velká data. AutoClass je k dispozici zdarma na internetu. Z těchto důvodů systém nabývá na popularitě a nalézá uplatnění v různorodých vědeckých a komerčních oblastech analýzy dat. Publikované výzkumy demonstrují výrazně lepší výsledky v porovnání s jinými metodami.

Literatura

AUTOCLASS HOMEPAGE [online]. http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/.

BERKA, P. 2003. *Dobývání znalostí z databází*. Vyd. Praha: Academia, ISBN 80-200-1062-9.

CHEESEMAN, P. a J. STUTZ 1996. *Bayesian Classification (AutoClass): Theory and Results*. Advances in Knowledge Discovery and Data Mining, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, & Ramasamy Uthurusamy, Eds. AAAI Press/MIT Press.

DEMPSTER, A. P., N. V. LAIRD a D. B. RUBIN 1997. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B, 39(1), 1-38.

GELMAN, A. et al. 2004. *Bayesian Data Analysis*. Boca Raton : Chapman & Hall / CRC, ISBN: 1-58488-388-X.

HANSON, R., J. STUTZ a P. CHEESEMAN 1991. *Bayesian Classification Theory*, Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch.

ŘEZANKOVÁ, H., D. HÚSEK a V. SNÁŠEL 2009. *Shluková analýza dat.* 2. vyd. Praha: Professional Publishing, ISBN 978-80-86946-26-9.

JEL Classification: C38.

Dedikace: Tento článek byl vypracován za podpory projektu IGA VSE F4/5/2011.

Summary

AutoClass – Classification Method Based on Bayesian Theorem

The main goal of this paper is to present a Bayesian classification method implemented in AutoClass software. This software was created by researchers from NASA. It is a classification tool based on naive Bayesian classifier.

The Bayesian approach to classification differs to classification methods which assign objects into classes uniquely. It is looking for latent classes that characterize each group of objects most accurately. These classes are described by using latent class models. AutoClass offers four latent class models. These models are based on the classical finite mixture model, which assumes conditional independence of variables. Estimation of the probability distribution of model parameters conditional on empirical data is based on the Bayes theorem. Model parameters are estimated simply from data by maximum likelihood method in the event of supervised learning. On the contrary, unsupervised learning uses iterative EM algorithm.

AutoClass was applied and tested in many large data sets not only by researchers at NASA, but also by statistics operating in industry and by academics. The newer versions of AutoClass enables to combine continuous and discrete variables. It handles with missing values. The algorithm allows to classify by variables that are conditionally independent on condition that the model hypothesis is valid. Automatically chooses the number of classes and complexity of the class description.

Keywords: AutoClass, Bayesian classification method, naive Bayesian classifier, EM algorithm, finite mixture model.

Odhady naděje dožití u osob se specifickým onemocněním

Ondřej Šimpach
ondrej.simpach@vse.cz
Doktorand oboru statistika
Školitel: doc. Ing. Jitka Langhamrová, CSc., (langhamj@vse.cz)

Abstrakt: Naděje dožití, jinak také známa pod pojmem střední délka života, se v minulosti výrazně odvíjela od vyspělosti zdravotnictví, získaných lékařských poznatků a od znalostí příslušných léčebných procesů. V případě osob nemocných celiakií, což je onemocnění nesnášenlivosti lepku, se naděje dožití od padesátých do osmdesátých let minulého století prakticky neměnila. Tito lidé umírali ve velmi podobném nízkém věku, což bylo způsobeno neznalostí diagnostiky jejich nemoci. Od doby pádu komunistického režimu v Československu však nastal výrazný zlom a vývoj naděje dožití těchto osob se prudce změnil. Tato změna nastala proto, že do svobodné země začaly pronikat lékařské poznatky ze zahraničí, a díky tomu se na dosud neznámá onemocnění nalezly postupy léčby, popřípadě informace ke zmírnění následků těchto onemocnění. Práce poskytne náhled do odhadu vývoje naděje dožití osob nemocných celiakií včetně grafické interpretace výsledků. Dále provede srovnání s vývojem naděje dožití u celkového obyvatelstva. Zároveň bude provedena kointegrační analýza mezi časovými řadami nedějí dožití u celkového obyvatelstva a odhadnutých nadějí dožití u obyvatelstva onemocněného celiakií. Práce se pokusí vyvrátit hypotézu tvrdící, že trend časových řad má přibližně stejný směr. V závěru tato studie poskytne srovnání pravděpodobnosti úmrtí osob onemocněných celiakií a pravděpodobnosti úmrtí osob onemocněných některou další nemocí, jejíž diagnóza byla ve světě lékařství v minulosti již dlouhá léta známa. Toto srovnání bude provedeno za pomoci logitových a probitových modelů, což jsou v současnosti velmi populární modely, aplikovatelné v různých oblastech vědy.

Klíčová slova: naděje dožití, pravděpodobnost úmrtí, celiakie, kointegrace, logit, probit, diskrétní vysvětlovaná proměnná.

Úvod

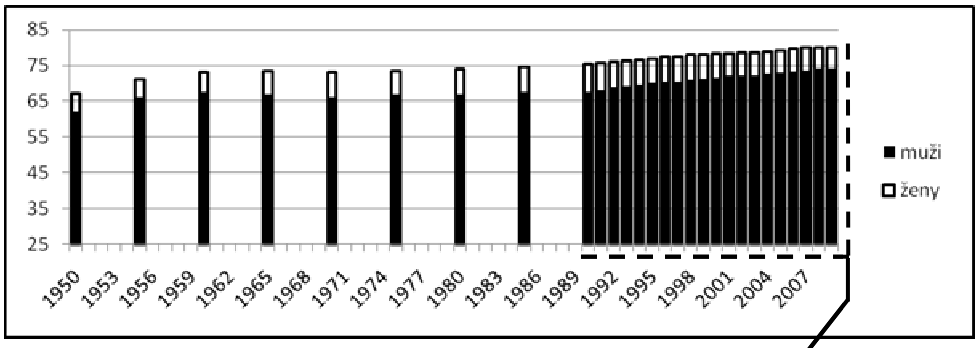
V dobách minulého režimu bylo naprosto běžné, že velké množství chorob a nemocí bylo buď obyvatelstvu země utajeno, nebo se o příznacích, průběhu a dopadech těchto chorob z nějakých důvodů nevědělo. Nevědělo se o nich buď proto, že věda a technika nebyla natolik vyspělá, aby tato onemocnění byla schopná odhalit, nebo z důvodu, že země byla vůči ostatním zemím izolována a neměla informace a znalosti, které již jinde byly známy a několik let již zachraňovaly lidské životy. V jiných zemích existovaly již dávno léky a léčebné procesy na mnohá onemocnění a v tehdejším Československu se o nich buď nevědělo, nebo nebylo možné tyto léky a „informace“ dovézt. Jedním z onemocnění, které v minulosti zkracovalo život mnoha lidem, je i tzv. „nesnášenlivost lepku“, odborně známa pod názvem celiakie, což je dle Společnosti pro bezlepkovou dietu chronické onemocnění sliznice tenkého střeva, způsobené přecitlivělostí na lepek. Práce srovná mezi sebou vývoj naděje dožití celé populace s odhadem vývoje naděje dožití osob nemocných na přecitlivělost na lepek. Lidé nemocní touto chorobou neměli dostatečné informace o tom, jak by se měli stravovat a tím eliminovat důsledky, které jim běžná strava způsobovala. Důležité je upozornit, že samostatné statistiky pro zemřelé osoby na toto onemocnění se nevedou, v současnosti jsou zahrnuty v souhrnné skupině *XI – Nemoci trávicí soustavy*, publikované ČSÚ. Počet nemocných v České republice je v současnosti Společností pro bezlepkovou dietu odhadován asi na 40 000–50 000 (četnost 1:200–250), z nichž je ale asi pouze 10–15 % vedeno v péči lékařů. U naprosté většiny nemocných tato nemoc stále není diagnostikována.

Ve své odborné činnosti jsem se zpočátku zaměřil zejména na experimentální úlohy menších rozsahů, ze kterých bych chtěl postupem času vytvořit souvislou sbírku publikovaných výsledků, pojednávajících převážně o soudobém vývoji demografických ukazatelů, mezi které patří například i naděje dožití obyvatel onemocněných specifickým onemocněním, neboť tito obyvatelé, mezi něž patří zejména mladí lidé, mohou svým počtem, který pro každou konkrétní nemoc nabývá hodnot i několika desítek tisíc jedinců, významně ovlivnit výkonové ukazatele národního hospodářství České republiky. Na základě myšlenek z první části této studie, pojednávající o odhadech naděje dožití osob onemocněných celiakií, budou aplikovány moderní přístupy nelineární regrese, které zažívají v současnosti rozmach

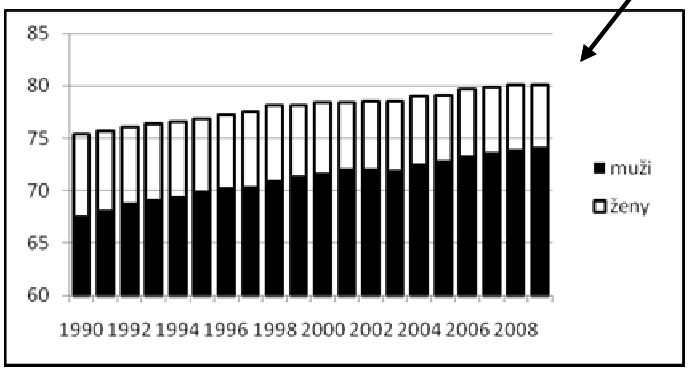
zejména v oblasti průzkumu trhu, průzkumu veřejného mínění, v bankovních aplikacích a v dalších analýzách, kde jde o odhady nastoupení pravděpodobnosti určitého jevu. V práci bude tedy dále provedena pravděpodobnostní analýza nastoupení jevu „úmrtí“ určitých osob, která bude ovlivňována dalšími proměnnými. Půjde o nezávislé potvrzení výsledku odhadu naděje dožití osob onemocněných celiakií, který bude vypočten v první části této studie.

Naděje dožití populace jako celku a odhady naděje dožití u osob onemocnělých celiakií

Z dat publikovaných Českým statistickým úřadem je možno graficky znázornit vývoj naděje dožití mužů a žen od roku 1950 do roku 2009. Statistický úřad má zveřejněn vývoj naděje dožití obyvatel od roku 1950 podrobně na svém webu. Tato analýza do minulosti pro nás ovšem nebude ani tak podstatná. Důležitý vývoj je zachycen na obrázku 1 u celkového obyvatelstva od roku 1990, kdy naděje dožití začala všeobecně růst. Detail tohoto nárůstu zachycuje podrobně obrázek 2. Nebudeme zde rozebírat důvody zvyšování naděje dožití u celkového obyvatelstva podrobněji (zlepšení kvality života lidí, zlepšení dostupnosti léků a nové přístroje a postupy v medicíně, což v našich podmínkách má za následek prodlužovaní naděje dožití v důsledku snižování úmrtnosti ve středním a vyšším věku), neboť primárně se příspěvek zaměřuje na odhad vývoje naděje dožití u osob onemocnělých celiakií.



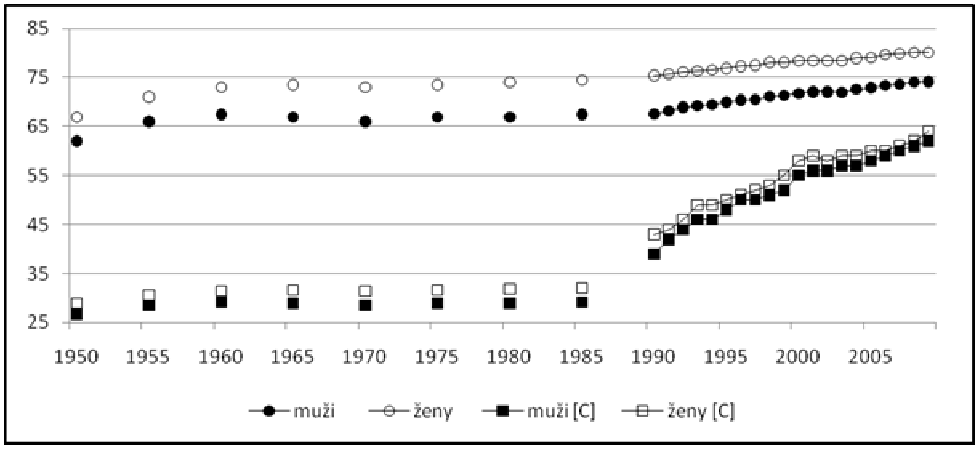
Obrázek 1: Naděje dožití u obyvatelstva celkového v letech 1950–2009 Zdroj dat: ČSÚ, 2011



Obrázek 2: Naděje dožití u obyvatelstva celkového v letech 1990–2009 (detail) Zdroj dat: ČSÚ, 2011

Z vlastních dat, od kterých se nedá předpokládat, že by byly vhodné k usouzení na celou populaci osob onemocněných celiakií, bylo provedeno jednoduché vymodelování průběhu naděje dožití těchto osob do minulosti. Ze zpráv a osobních poznámek lékařů, jejich životních zkušeností, na které v minulosti nebyly brány ohledy, a dalších informací poskytnutých mi odborníky z praxe, byl odhadnut průběh naděje dožití osob onemocněných celiakií od roku 1950 do roku 1990. Na obrázku 3 je tento průběh znázorněn. Časová řada označená „muži“ a „ženy“ zobrazuje skutečné hodnoty naděje dožití populace celkem, které jsou publikovány Českým statistickým úřadem. Časová řada „muži [c]“ a „ženy [c]“ zobrazuje odhadnuté hodnoty naděje dožití osob nemocných celiakií. Je patrné, že naděje dožití těchto osob se pohybovala okolo ustálené hodnoty 30 let. U žen se tato ustálená hodnota pohybovala v přibližném rozsahu 29–32 let v období 1950–1985 a byla nepatrně vyšší než ustálená hodnota u mužů. V případě mužů se naděje dožití pohybovala v rozsahu 27–29 let. Jen málokterý muž přežil s tímto onemocněním hranici 30 let. Kromě všeobecně známých případů mužské nadúmrtnosti, které

zde neuvádíme, jsou pro muže častější také v mladším věku úmrtí na úrazy a otravy. Všeobecně se ví, že muži zkonzumují mnohem více potravy než ženy. Jelikož celiakie je nemoc nesnášenlivosti lepku, který způsobuje onemocnění sliznice tenkého střeva [7], a lidé v dobách minulých nevěděli, že lepek nesmí konzumovat, tak větší konzumace potravin u mužů a tím i potravin obsahující lepek mohla způsobit, že jejich tenké střevo bylo ničeno mnohem silněji a rychleji než u žen. Menší množství konzumované potravy, a tím i menší množství konzumovaného lepku, by mohlo být vysvětlením, proč byla i v minulosti u mužů s tímto onemocněním naděje dožití nižší. Ženy si tak svou střídmostí k potravě mohly prodloužit život i o 3 roky.



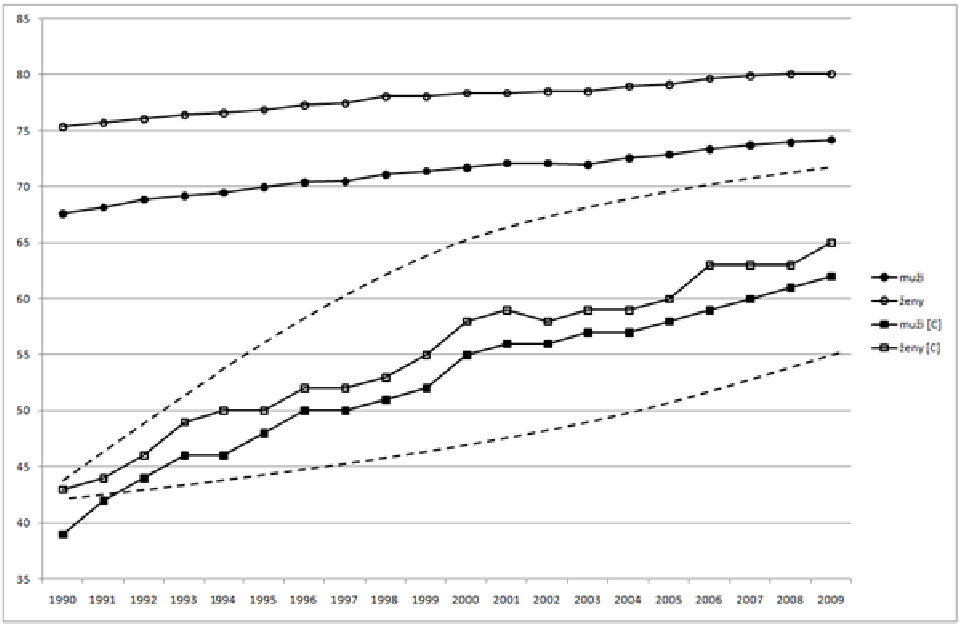
Obrázek 3: Odhad vývoje naděje dožití u mužů a žen nemocných celiakií v období 1950–2009
Zdroj dat: ČSÚ, 2011 a dále vlastní propočty

V další části obrázku 3 je patrný průběh odhadu naděje dožití osob nemocných celiakií od roku 1990 do roku 2009. Je vidět, že zlomový okamžik mohl nastat v období, kdy klesala síla minulého režimu, a poměry izolovanosti informací začaly pomíjet. Přístup nových znalostí a poznatků, nejen získaných doma díky rozvoji domácího lékařství, ale i ze zahraničí, zejména důsledkem otevření hranic v roce 1989, způsobilo to, že naděje dožití těchto osob mohla růst.

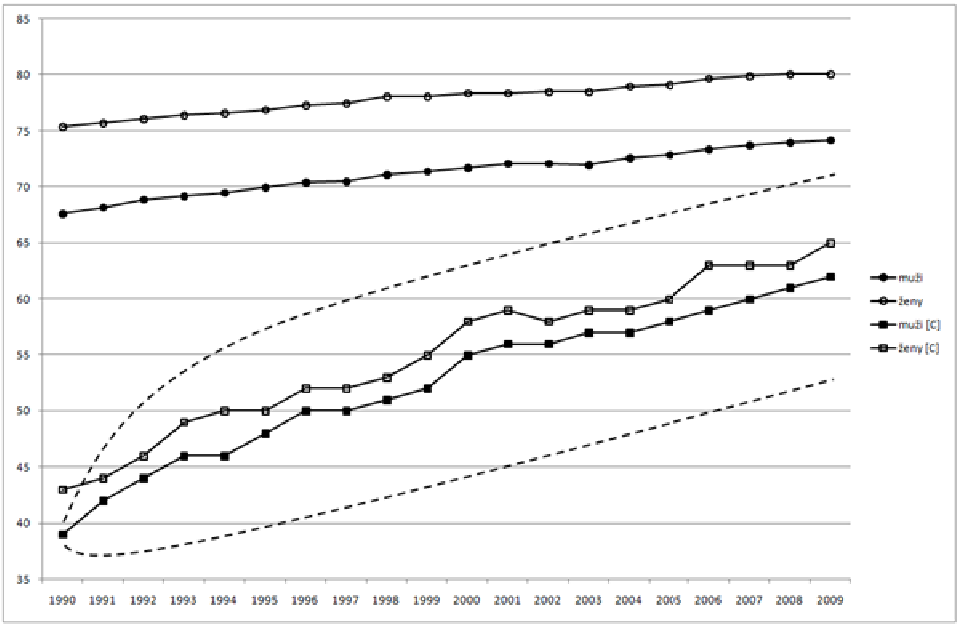
Začaly se vyrábět potraviny neobsahující lepek a alternativní možnosti stravování pro tyto osoby začaly pronikat do České republiky ze zahraničí. Celiakie je v současnosti nevyléčitelná choroba. Jak tvrdí např. Klener (2006), „při dodržování bezlepkové diety však vymizí všechny její příznaky“. Vzhledem k tomu, že se po revoluci Česká republika otevřela světu nejen po ekonomické stránce, ale i po stránce informací, vědy, techniky a vzdělání, dá se předpokládat, že vývoj naděje dožití u osob nemocných touto chorobou bude limitovat k vývoji naděje dožití celkového obyvatelstva. Detailní průběh odhadu vývoje naděje dožití osob nemocných celiakií je uveden na obrázku 4 a 5. Obrázek 4 je doplněn navíc o 95% intervaly stability pro ženy trpící celiakií. V 95% všech případů by se měla střední hodnota odhadu nacházet v tomto intervalu, ovšem vzorek, ze kterého byly provedeny odhady, není reprezentativní. Pro lepší analýzu chybí data, která není možné bez nějakého dlouhodobějšího výzkumu pořídit. Obrázek 5 je opět doplněn o 95% intervaly stability pro muže trpící celiakií. Zrovna tak jako u žen, ani tento vzorek, ze kterého byly pořízeny údaje, není reprezentativní. Na exaktní odhady chybí data a není možné je bez nákladů pořídit.

Analýza vztahů mezi vývojem jednotlivých nadějí dožití

Můžeme provést analýzu, zdali mezi skutečnými hodnotami vývoje naděje dožití u celkového obyvatelstva a odhadnutých hodnot naděje dožití u obyvatel trpících celiakií, existuje kointegrační vztah, neboli zda by řady mohly mít stejný trend. Z postupu uvedeného např. Dickiem a Fullerem (1979) nebo Engelem a Grangerem (1987) můžeme prozkoumat, zda mají řady jednotkový kořen, tj. jsou nestacionární.



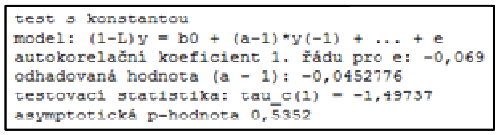
Obrázek 4: Odhad vývoje naděje dožití u mužů a žen nemocných celiakií v období 1990–2009 (detail + 95% intervaly stability odhadu pro ženy) Zdroj dat: ČSÚ, 2011 a dále vlastní propočty

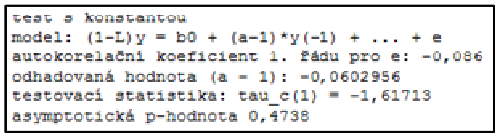


Obrázek 5: Odhad vývoje naděje dožití u mužů a žen nemocných celiakií v období 1990–2009 (detail + 95% intervaly stability odhadu pro muže) Zdroj dat: ČSÚ, 2011 a dále vlastní propočty

Provedeme test jednotkového kořene pro řadu „muži“, tj. vývoj naděje dožití mužů v populaci a pro řadu „muži [c]“, tj. vývoj odhadu naděje dožití mužů trpících celiakií. Vyslovená hypotéza zní:
H_0: časové řady jsou nestacionární, jsou typu I(1)
H_1: časové řady jsou stacionární, jsou typu I(0)

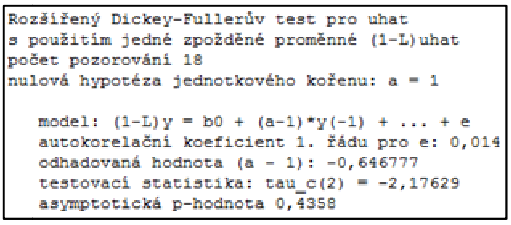
Výstupy ze systému GRETL jsou uvedeny pro řadu „muži“ na obrázku 6 a pro řadu „muži [c]“na obrázku 7. Je vidět, že na 5% hladině významnosti nezamítáme hypotézu H_0, že řady jsou nestacionární.





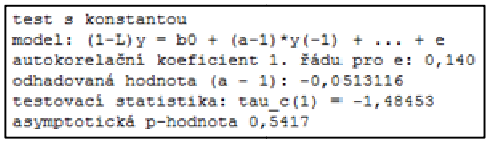
Obrázek 6: Výsledek ADF pro proměn. „muži“ Obrázek 7: Výsledek ADF pro proměn. „muži [c]“
Zdroj: systém GRETL

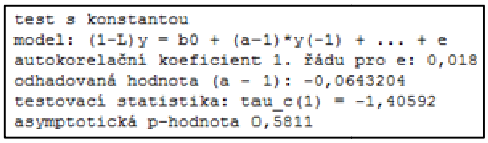
Aby existoval kointegrační vztah, musí být nezamítnuty testované nulové hypotézy pro proměnné a následně musí být zamítnuta testovaná hypotéza pro rezuidua z kointegrační regrese. Kointegrační regrese, uvedená např. Arltem a Arltovou (2007) byla vypočtena a byl proveden rozšířený Dickey-Fullerův test pro proměnnou . Výsledek tohoto testu je uveden na obrázku 8.



Obrázek 8: Výsledek ADF testu pro rezidua kointegrační regrese Zdroj: systém GRETL

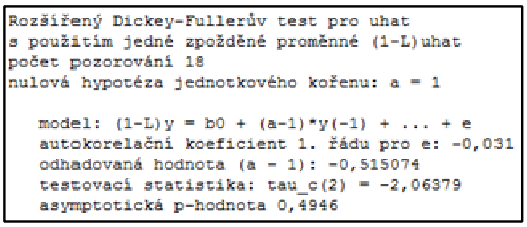
Z výstupu je patrné, že kointegrační vztah nemůže existovat. Způsobeno to může být buď prostou neexistencí vztahu, nebo nereprezentativními odhady, které by mohly být jiné, než jaké byly vypočteny. Případná neexistence vztahu je však pozitivní zjištění, neboť bychom rádi tvrdili, že vývoj naděje dožití mužů onemocněných celiakií je jiný, než vývoj naděje dožití mužů celkem. Obdobně bude proveden test jednotkového kořene pro řadu „ženy“, tj. vývoj naděje dožití žen v populaci a pro řadu „ženy [c]“, tj. vývoj odhadu naděje dožití žen trpících celiakií. Na obrázku 9 jsou opět uvedeny výstupy ze systému GRETL pro řadu „ženy“ a na obrázku 10 výstupy pro řadu „ženy [c]“. Na 5% hladině významnosti nezamítáme testovanou hypotézu H_0 o přítomnosti jednotkového kořene v obou řadách.





Obrázek 9: Výsledek ADF pro proměn. „ženy“ Obrázek 10: Výsledek ADF pro proměn. „ženy [c]“
Zdroj: systém GRETL

Na základě existence jednotkových kořenů v řadách „ženy“ a „ženy [c]“ byla vypočtena kointegrační regrese dle příslušného postupu a následně proveden rozšířený Dickey-Fullerův test pro proměnnou . Výsledek tohoto testu je uveden na obrázku 11.



Obrázek 11: Výsledek ADF testu pro rezidua kointegrační regrese Zdroj: systém GRETL

Ani v tomto případě není možné tvrdit, že řady „ženy“ a „ženy [c]“ jsou spolu kointegrovány. Kointegrační vztah mezi nimi neexistuje, není tedy možné použít např. predikce pomocí VAR modelů pro vývoj naděje dožití osob s celiakií do budoucna.

Modelování pravděpodobnosti úmrtí osob onemocněných celiakií

V následné části studie bude prozkoumána pravděpodobnost úmrtí osoby do určitého časového okamžiku v případě, že má diagnostikovánu určitou chorobu. Z obrázku 3 bude pro tuto studii rok 1990,

kdy se naplno začaly aplikovat nové poznatky z medicíny, které způsobily prodlužování naděje dožití osob onemocněných celiakií. Pro experiment nelineární regrese, aplikovaný v této studii, jsem získal 40 pozorování, sestávající ze dvou vzorků. Nejedná se o reprezentativní výběr a jsou z větší části porušeny základní předpoklady pro aplikaci metod matematické statistiky. Jde však o jediná data, která se podařilo opatřit, proto pouze v dlouhém období a po více experimentech bude moci být mé tvrzení potvrzeno nebo vyvráceno. Vzorek obsahuje 20 pozorování osob onemocněných celiakií a 20 osob onemocněných některou jinou nemocí nebo trpící jinými zdravotními komplikacemi, které jsou však v medicíně již dlouhou dobu známé. Jedná se o nemoci a komplikace:

- cukrovka,
- vysoký obsah cholesterolu v krvi (více než 5 mmol/l),
- vysoký krevní tlak (více než 150 SYS a 95 DIA),
- astma bronchiale (onemocnění dýchacích cest),
- případně jiná onemocnění.

Výběr 40 pozorování je vyčerpávající, opatřený od jednoho lékaře, který se ve své praxi zabýval dospělými osobami s onemocněním trávicí soustavy a některými dalšími onemocněními. Výběr tvoří osoby, které byly pozvány v roce 1990 na celkovou zdravotní prohlídku, a jejich další vývoj zdravotního stavu byl do budoucna sledován. Autoři Spector a Mazzeo (1980) ve své studii sestavili příklad, kde odhadovali, s jakou pravděpodobností uspěje student u zkoušky, s použitím dalších vysvětlujících proměnných, některých i diskrétních. Na základě tohoto příkladu byly rozvinuty pravděpodobnostní logitové a probitové modely, které v současnosti využívá řada autorů ve svých výpočtech a publikacích, jako například Hoyos et al. (2010) pro marketingové účely. Pokusil jsem se sestavit vlastní logitový a probitový model, na jehož základě bych chtěl odhadnout pravděpodobnost nastoupení jevu „úmrtí *x*-leté osoby do pěti let od okamžiku zdravotní prohlídky“ v případě, že osoba má diagnostikovánu nemoc trávicí soustavy (dnes již známá celiakie) a nastoupení tohoto jevu v případě, že má diagnostikovánu některou z dalších výše uvedených nemocí.

Vysvětlovaná proměnná je diskrétní. Pravděpodobnost nastoupení jevu „úmrtí *x*-leté osoby do pěti let od okamžiku zdravotní prohlídky“ je rovno 1, pak osoba zemře. Nebo pravděpodobnost nastoupení jevu „úmrtí *x*-leté osoby do pěti let od okamžiku zdravotní prohlídky“ je rovno 0, pak osoba přežije. Aby bylo možno určit i hodnoty pravděpodobnosti nastoupení tohoto jevu mezi oběma extrémy, bude aplikován logitový a probitový model diskrétní volby, kdy vysvětlovaná proměnná nabývá hodnot z intervalu <0 ; 1>. V pořízených datech vystupují proměnné:

- **VEK** – je věk osoby pozvané ke zdravotní prohlídce,
- **KZRU** – je *Konstanta Zvyšující Riziko Úmrtí*, která byla zkonstruována především pro potřeby tohoto příkladu a jejíž výpočet udává vzorec (1). Konstanta nabývá hodnot z intervalu <5 ; 35>,
- **CEL** – je binární proměnná, nabývající hodnot „0“ = osoba nemá celiakii, avšak trpí některou z dalších výše zmíněných nemocí nebo má zdravotní komplikace, nebo „1“ = osoba má celiakii,
- **UMRTI_5** – je binární proměnná, nabývající hodnot „0“ = osoba do 5 let od zdravotní prohlídky nezemřela v důsledku svého onemocnění, nebo „1“ = osoba do pěti let od zdravotní prohlídky zemřela v důsledku svého onemocnění.

Jak bylo řečeno výše, *konstanta zvyšující riziko úmrtí* byla zkonstruována pouze pro potřeby výpočtů tohoto příkladu a její výpočet vychází z tabulky 1.

Tabulka 1: Odpovědi pacientů na otázky lékaře během celkové prohlídky Zdroj: vlastní sestavení

	$Ш_1$	$Ш_2$	$Ш_3$
Kuřák	ne	příležitostně	pravidelně
Černá káva	ne	příležitostně	pravidelně
Alkohol	ne	příležitostně	pravidelně
Spánek	pravidelný	nepravidelný	špatný
Stravování se	pravidelné	nepravidelné	špatné

Lékař při celkové prohlídce pokládá pacientovi několik jednoduchých otázek. Na základě odpovědí byla vytvořena tato tabulka, kde místo slovně uvedených odpovědí byly zaznamenány $ш_{i,j}$, nabývající hodnot „0“ a „1“, kde

- „0“ = odpověď pacienta se neshoduje se slovem, uvedeném v příslušné buňce,
- „1“ = odpověď pacienta se shoduje se slovem, uvedeném v příslušné buňce.

Po zaznamenání odpovědí do tabulky 2 v „0/1“ formátu

Tabulka 2: Odpovědi pac. na otázky lékaře během prohlídky ve formátu „0/1“ Zdroj: vlastní sestavení

	$Ш_1$	$Ш_2$	$Ш_3$
Kuřák	$ш_{i,j}$ *= 0/1*	…	…
Černá káva	…		
Alkohol	…		
Spánek	…		
Stravování se	…		

Byl proveden výpočet

$$KZRU = \left(1 * \sum_{i=1}^{5} ш_{i,1}\right) + \left(3{,}5 * \sum_{i=1}^{5} ш_{i,2}\right) + \left(7 * \sum_{i=1}^{5} ш_{i,3}\right) . (1)$$

Tato konstanta může nabývat hodnot z intervalu <5; 35>, kde krajní hodnota 5 znamená, že pacient minimálně zvyšuje riziko úmrtí svým špatným životním stylem, a krajní hodnota 35 znamená, že pacient zvyšuje riziko úmrtí tím nejhorším možným způsobem.

Logitový model

Z předpokladů uvedených například Christensenem (1990) je pravděpodobnostní funkce

$$P_i = E(Y = 1|\boldsymbol{X}_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i)}} , (2)$$

která upravená pro potřeby této úlohy má podobu

$$P_i = E(Y = 1|\boldsymbol{X}_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 VEK_i + \beta_2 KZRU_i + \beta_3 CEL_i)}} . (3)$$

Označme

$$Z_i = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i , (4)$$

která je pro potřeby úlohy v podobě

$$Z_i = \beta_0 + \beta_0 + \beta_1 VEK_i + \beta_2 KZRU_i + \beta_3 CEL_i . (5)$$

Pak platí

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}} = F(Z_i) , (6)$$

což je distribuční funkce logistického rozdělení. Pravděpodobnost osoby *x*-leté nezemřít do pěti let od okamžiku zdravotní prohlídky je

$$1 - P_i = \frac{1}{1 + e^{Z_i}} , (7)$$

a proto

$$\frac{P_i}{1 - P_i} = e^{Z_i} . (8)$$

Za pomoci logaritmu obdržíme logit

$$\ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i \,, (9)$$

který má v této úloze podobu

$$\ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_0 + \beta_1 VEK_i + \beta_2 KZRU_i + \beta_3 CEL_i \,. (10)$$

Pro odhad parametrů loktového modelu se nedá využít klasické metody nejmenších čtverců, s dobrým software lze však využít metodu maximální věrohodnosti. Z obecného zápisu logaritmu věrohodnostní funkce, uvedeném Christensenem (1990) jako

$$\ln L(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{N}\left[Y_i \ln\left(\frac{e^{Z_i}}{1+e^{Z_i}}\right) + (1-Y_i)\ln\left(1-\frac{e^{Z_i}}{1+e^{Z_i}}\right)\right] , (11)$$

vychází po dosazení

$$\ln L(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{N}\left[Y_i \ln\left(\frac{e^{\beta_0+\boldsymbol{\beta}'\boldsymbol{X}_i}}{1+e^{\beta_0+\boldsymbol{\beta}'\boldsymbol{X}_i}}\right) + (1-Y_i)\ln\left(1-\frac{e^{\beta_0+\boldsymbol{\beta}'\boldsymbol{X}_i}}{1+e^{\beta_0+\boldsymbol{\beta}'\boldsymbol{X}_i}}\right)\right] , (12)$$

a pro tuto konkrétní úlohu pak

$$\begin{aligned}\ln L(\beta_0, \beta_1, \beta_2, \beta_3) &= \sum_{i=1}^{N}\left[UMRTI_5_i \ln\left(\frac{e^{\beta_0+\beta_1 VEK_i+\beta_2 KZRU_i+\beta_3 CEL_i}}{1+e^{\beta_0+\beta_1 VEK_i+\beta_2 KZRU_i+\beta_3 CEL_i}}\right)\right. \\ &\left. + (1-UMRTI_5_i)\ln\left(1-\frac{e^{\beta_0+\beta_1 VEK_i+\beta_2 KZRU_i+\beta_3 CEL_i}}{1+e^{\beta_0+\beta_1 VEK_i+\beta_2 KZRU_i+\beta_3 CEL_i}}\right)\right].\end{aligned} \quad (13)$$

Probitový model

V logitovém modelu bylo využito logistického rozdělení, v případě probitového modelu musíme využít rozdělení normálního. Jestliže proměnná X má normální rozdělení se střední hodnotou μ a rozptylem σ^2, její hustota pravděpodobnosti je

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (14)$$

a její příslušná distribuční funkce je

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \,. (15)$$

Pro odhad parametrů probitového modelu taktéž nemůžeme využít klasické metody nejmenších čtverců, ale univerzální metodu maximální věrohodnosti. Z obecného zápisu logaritmu věrohodnostní funkce, uvedeném např. Christensenem (1990)

$$\ln L(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{N}[Y_i \ln(F(Z_i)) + (1-Y_i)\ln(1-F(Z_i))] \,, (16)$$

vychází po dosazení

$$\ln L(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{N}[Y_i \ln(F(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i)) + (1-Y_i)\ln(1-F(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i))] \,, (17)$$

což je pro tuto konkrétní úlohu

$$\begin{aligned}\ln L(\beta_0, \beta_1, \beta_2, \beta_3) &= \sum_{i=1}^{N}\big[UMRTI_5_i \ln\big(F(\beta_0 + \beta_1 VEK_i + \beta_2 KZRU_i + \beta_3 CEL_i)\big) \\ &+ (1-UMRTI_5_i)\ln\big(1-F(\beta_0 + \beta_1 VEK_i + \beta_2 KZRU_i + \beta_3 CEL_i)\big)\big]\end{aligned} \quad (18)$$

a distribuční funkce

$$F(\beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\beta_0+\boldsymbol{\beta}'\boldsymbol{X}_i)} e^{\frac{-z^2}{2}} dz \ , (19)$$

po dosazení pro tuto konkrétní úlohu

$$F(\beta_0 + \beta_1 VEK_i + \beta_2 KZRU_i + \beta_3 CEL_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_0+\beta_1 VEK_i+\beta_2 KZRU_i+\beta_3 CEL_i} e^{\frac{-z^2}{2}} dz \ . (20)$$

Odhadování neznámých parametrů modelu bylo v minulosti velmi složité a ani s příchodem programových systémů se pohodlí uživatele při odhadech příliš nezlepšilo. Odhady byl software schopen spočítat až po zadání konkrétního matematického zdrojového kódu, který byl pro každou konkrétní úlohu jedinečný. Navíc bylo třeba zadat počáteční, tzv. startovací hodnoty odhadů, které software iteračně vylepšoval. Byl velký prostor pro vznik chyb. Času potřebného k odhadům a jejich následnému zlepšení bylo třeba velké množství. Téměř všechny systémy, které vydaly své aktualizace a nové verze v roce 2007 a později již bezpečně umí automatické dosazování startovacích hodnot a automatický výpočet iterací. V práci byl využit analytický systém GRETL, s jehož pomocí byly spočteny odhady neznámých parametrů a testy modelu.

Odhady parametrů – logitový model

Po výpočtu odhadů neznámých parametrů metodou maximální věrohodnosti systém zobrazil tabulku, uvedenou na obrázku 12 s následujícími výsledky:

Model 4: Logit, za použití pozorování 1-40
Závisle proměnná: UMRTI_5

	koeficient	směr. chyba	t-podíl	p-hodnota	
const	-23,6617	9,35780	-2,529	0,0115	**
VEK	0,337744	0,152168	2,220	0,0265	**
KZRU	0,365498	0,150880	2,422	0,0154	**
CEL	8,32785	3,36660	2,474	0,0134	**

Obrázek 12: odhady parametrů logitového modelu Zdroj: systém GRETL

Je vidět, že přestože bylo pořízeno jen velmi malé množství pozorování, všechny odhady neznámých parametrů jsou na 5% hladině významnosti statisticky významné. Můžeme vypsat odhadnuté parametry do pravděpodobnostní funkce (2):

$$P_i = E(Y = 1|\boldsymbol{X}_i) = \frac{1}{1 + e^{-(-23{,}6617+0{,}337744*VEK_i+0{,}365498*KZRU_i+8{,}32785*CEL_i)}}$$

a zároveň vyjádřit podobu logitu (10):

$$\ln\left(\frac{P_i}{1-P_i}\right) = Z_i = -23{,}6617 + 0{,}337744 * VEK_i + 0{,}365498 * KZRU_i + 8{,}32785 * CEL_i$$

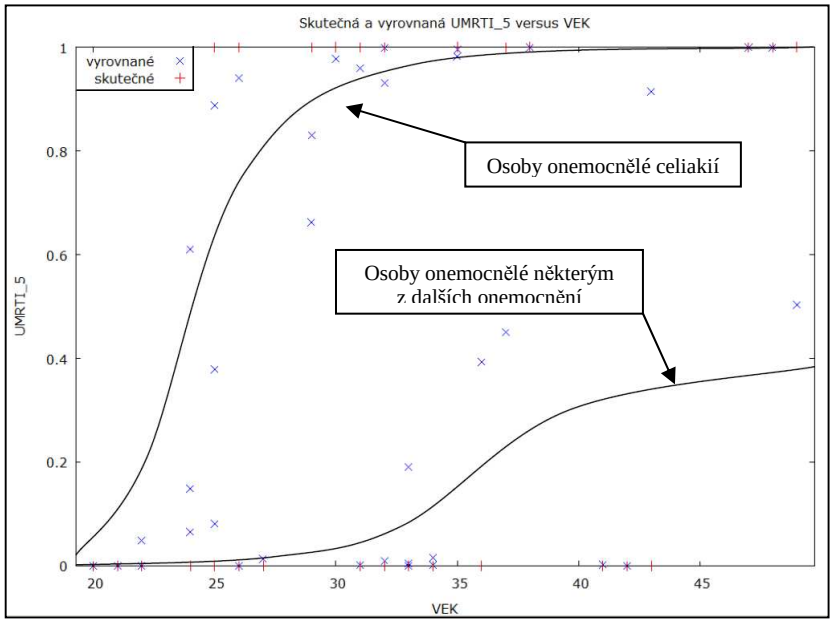
Odhadnutá hodnota funkce logaritmu věrohodnosti (13) je

$$\ln L(\beta_0, \beta_1, \beta_2, \beta_3) = -7{,}92245785380$$

a její dosažení bylo získáno po 8 iteracích.

Nejdůležitější grafický výstup této části práce se nachází na obrázku 13. Jde o grafické zobrazení závislosti pravděpodobnosti úmrtí osoby *x*-leté do pěti let od okamžiku lékařské prohlídky v roce 1990 na věku konkrétní osoby. Horní křivka odpovídá příslušným hodnotám pravděpodobnosti úmrtí osob onemocnělých celiakií. Dolní křivka odpovídá příslušným hodnotám pravděpodobnosti úmrtí osob onemocnělých některou z dalších, dříve zmíněných nemocí. Je patrné, že pravděpodobnost, že osoba mezi lety 1990 a 1995 zemře, byla u osob onemocnělých celiakií mnohem vyšší než u osob, jejichž nemoc byla ve světě medicíny již dlouhou dobu známá. Pravděpodobnost úmrtí 35leté osoby do 5 let od návštěvy lékaře je již téměř jednotková a tudíž lze říci, že odhady nadějí dožití, které byly uvedeny na obrázku 3 v úvodu této práce, mohou být správné nebo se od reálných hodnot liší jen velmi málo. Tato pravděpodobnost je tak vysoká zejména proto, že lidé, kteří 10 nebo 20 let ještě žili v minulém

režimu, ve kterém onemocnění celiakie nebylo v medicíně diagnostikováno, ničili nevědomky svůj organismus potravou, kterou nesměli konzumovat. Když se na jejich onemocnění přišlo, bylo již bohužel pozdě pokoušet se něco zachránit.



Obrázek 13: Závislost pravděpodobnosti úmrtí osoby *x*-leté do pěti let od okamžiku lékařské prohlídky v roce 1990 na věku konkrétní osoby, logitový model Zdroj: vlastní výpočty

Odhady parametrů – probitový model

Obdobným způsobem jako v případě logitového modelu byl i zde proveden v systému GRETL odhad neznámých parametrů metodou maximální věrohodnosti. Systém zobrazil tabulku s výsledky, uvedenými v obrázku 14:

```
Model 5: Probit, za použití pozorování 1-40
Závisle proměnná: UMRTI_5

             koeficient   směr. chyba   t-podíl   p-hodnota
  ------------------------------------------------------------
  const      -13,5867      5,09631     -2,666     0,0077   ***
  VEK          0,191213    0,0830420    2,303     0,0213   **
  KZRU         0,208414    0,0810669    2,571     0,0101   **
  CEL          4,88747     1,91153      2,557     0,0106   **
```

Obrázek 14: odhady parametrů probitového modelu Zdroj: vlastní výpočty

Směrodatné chyby odhadů parametrů modelu jsou dostatečně malé a na 5% hladině významnosti můžeme s jistotou zamítnout testovanou hypotézu o jejich statistické nevýznamnosti. Je tedy možno vypsat odhadnuté parametry do logaritmu věrohodnostní funkce (18):

$$\ln L(\beta_0,\beta_1,\beta_2,\beta_3) = \sum_{i=1}^{N}\left[UMRTI_5_i \ln\left(F(-13{,}59+0{,}19*VEK_i+0{,}21*KZRU_i+4{,}89*CEL_i)\right)\right.$$
$$\left.+(1-UMRTI_5_i)\ln(1-F(-13{,}59+0{,}19*VEK_i+0{,}21*KZRU_i+4{,}89*CEL_i))\right]$$

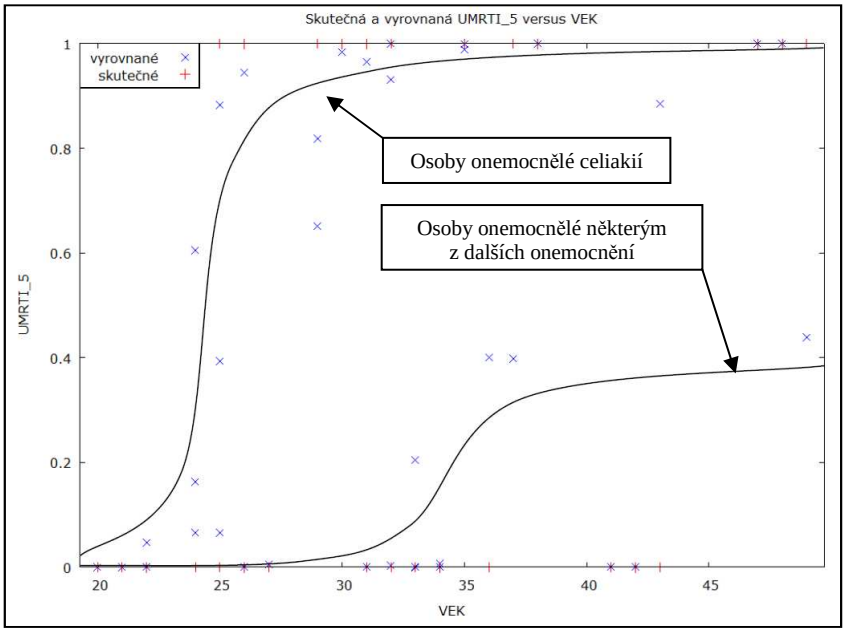
a do distribuční funkce (20) pro tuto úlohu:

$$F(\beta_0+\beta_1 VEK_i+\beta_2 KZRU_i+\beta_3 CEL_i)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(-13{,}59+0{,}19*VEK_i+0{,}21*KZRU_i+4{,}89*CEL_i)} e^{\frac{-z^2}{2}}dz$$

Samotná odhadnutá hodnota funkce logaritmu věrohodnosti (18), které bylo dosaženo iterační metodou po 10 iterací, je:

$$\ln L(\beta_0,\beta_1,\beta_2,\beta_3) = -7{,}83512785597\,.$$

Nejdůležitější grafický výstup je z této části na obrázku 15. Jde o grafické zobrazení závislosti pravděpodobnosti úmrtí osoby *x*-leté do pěti let od okamžiku lékařské prohlídky v roce 1990 na věku konkrétní osoby. Horní křivka odpovídá příslušným hodnotám pravděpodobnosti úmrtí osob onemocnělých celiakií. Dolní křivka odpovídá příslušným hodnotám pravděpodobnosti úmrtí osob onemocnělých některou z dalších, dříve zmíněných nemocí. Graf je obdobný jako v případě logitového modelu. Odhady jsou správné a oba modely se dají použít.



Obrázek 15: Závislost pravděpodobnosti úmrtí osoby *x*-leté do pěti let od okamžiku lékařské prohlídky v roce 1990 na věku konkrétní osoby, probitový model Zdroj: vlastní výpočty

Závěr

Vzhledem k tomu, že vzorek pro odhad naděje dožití není statisticky reprezentativní, není možné bezpečně tvrdit, že vypočtená naděje dožití je správná. Taktéž není možné tvrdit, že žádný kointegrační vztah mezi nadějí dožití u celkového obyvatelstva a u obyvatel trpících celiakií nemůže existovat. Nicméně jde o jediný pokus, představit podobu tohoto vývoje veřejnosti. V případě, že by se opatřila vhodnější a reprezentativnější data, mohly by být odhady zpětně přepočteny, modifikovány a znovu publikovány jako odhady přesnější. Zároveň by na základě těchto přesnějších odhadů mohly být provedeny statistické a demografické analýzy, jak by se mohla naděje dožití osob trpících celiakií do budoucna vyvíjet. Můžeme však předpokládat, že i když se nenajde lékařská metoda, jak vyléčit tyto osoby trvale, vzhledem k již dostatečně objasněným příčinám a důsledkům tohoto onemocnění se dá předpokládat, že do budoucna bude naděje dožití těchto osob limitovat k naději dožití celkového obyvatelstva. Z výsledků, které byly uvedeny v části logitových a probitových modelů (byť splnění předpokladů pro jejich vznik není silné), je možno potvrdit odhadnutou hodnotu naděje dožití osob onemocněných celiakií z roku 1990. Během vývoje již samostatné České republiky naděje dožití těchto osob rostla, pravděpodobnost úmrtí osob *x*-letých tudíž určitě začala klesat. S největší pravděpodobností se podaří do budoucna opatřit další obdobná data o těchto pacientech, ať už od lékaře stejného, nebo jiného, ale posunutá v čase o několik let dál. Doufám, že tato práce vznikne a poskytne odhady pravděpodobnosti úmrtí nižší, než poskytla tato práce. Tak by došlo k dalšímu potvrzení předpokládaných výsledků. Z dalších studií, kterých bych chtěl do budoucna dosáhnout, uspořádám co nejpřesnější odhady naděje dožití osob onemocněných různými nemocemi, společně s ekonomickou analýzou významu těchto osob pro produkt této republiky, neboť jak jsem dříve uvedl, osob nemocných každou konkrétní nemocí, která může jejich život předčasně ukončit, je několik desítek tisíc. Případná předčasná ztráta potenciálu, který spočívá v těchto lidech, je krom morálních hodnot také ztrátou ekonomickou.

Literatura

ARLT, J., ARLTOVÁ, M.: „*Ekonomické časové řady*“, Grada Publishing, 2007.

DICKEY, D.A., FULLER, W.A.: „Distribution of the Estimators for Autoregressive Time Series with a Unit Root“, *Journal of the American Statistical Association*, 74, 1979, str. 427–431.

ENGLE, R.F., GRANGER, C.W.J.: „Co-integration and error correction: Representation, estimation and testing“, *Econometrica*, 55 (2), 1987, str. 251-276.

HOYOS, D., MARIEL, P., MEYERHOFF, J.: „*Comparing the performance of different approaches to deal with attribute non-attendance in discrete choice experiments: a simulation experiment*„, BILTOKI, Universidad del País Vasco - Departamento de Economía Aplicada III (Econometría y Estadística), 2010.

CHRISTENSEN, R.: „*Log-Linear Models*“ Springer-Verlag, New York, 1990.

KLENER, P. et al.: *Vnitřní lékařství*. Praha: Galén a Karolinum 2006, str. 591.

SPECTOR, L.C., MAZZEO, M.: „*Probit Analysis and Economic Education*“, Journal of Economic Education. Spring, 11, 1980, pp. 37–44.

ČESKÝ STATISTICKÝ ÚŘAD. „*Naděje dožití při narození v letech 1950 – 2010*“ Zdroj: http://www.czso.cz/csu/redakce.nsf/i/nadeje_doziti_pri_narozeni_v_letech_1950_2010

SPOLEČNOST PRO BEZLEPKOVOU DIETU O.S.. „*Alergie na lepek*“Zdroj: http://celiak.cz/o-nemoci/alergie-na-lepek

JEL Classification: C40.

Summary

Estimates of life expectancy for people with specific diseases

Because of that sample to estimate the life expectancy is not statistically representative, it can't safely say, that the calculated life expectancy is correct. Also, it is not possible to say, that no co-integration between life expectancy for the total population and the population suffering from celiac disease can't exist. However, it is the only attempt to introduce this form of development to the public. In the event, that would seek better and more representative data, estimates could be retroactively restated, modified and republished as more accurate estimates. At the same time, based on these more accurate estimates could be made statistical and demographic analysis, how could life expectancy of people suffering from celiac disease develop in the future. Even if we will find a medical method to permanently cure these people, due to the already known causes and consequences of this disease, we can say, that the life expectancy of these people will limit to the life expectancy of the total population in the future. The results, which were placed in the LOGIT and PROBIT models, it is possible to confirm the value of the estimated life expectancy of people suffering from celiac disease in 1990. During development of the independent Czech Republic could the life expectancy of these people grow, the probability of death x-year-old certainly began to decline. Most likely in the future will find further information about these patients, whether by the same doctor or another, but shifted in time a few years away. I hope that this work will be, and provide estimates of the probability of death lower than that provided by this work. Then there was another confirmation of the expected results. In other studies, which I would like to achieve in the future, arrange the most accurate estimates of life expectancy of people sick in different diseases, together with the economic analysis of the importance of these people to the product of the Republic, because as I said earlier, there are tens of thousands of people sick each these diseases, which can prematurely terminate human life. Potential loss of these people is in addition to moral values, a loss of economic value.

Keywords: Life Expectancy, Probability of Death, Celiac Disease, Co-integration, LOGIT, PROBIT, Discrete Dependent Variable.

Analýza CHAID jako nástroj vyhodnocování marketingových kampaní

Jiří Vild
jiri.vild@vse.cz
Doktorand oboru statistika
Školitel: doc. Ing. Iva Pecáková, CSc., (pecakova@vse.cz)

Abstrakt: Článek se zabývá analýzou CHAID jako nástrojem vyhodnocování marketingových kampaní. CHAID je metoda konstrukce rozhodovacích stromů pro kategoriální vysvětlované proměnné a kategoriální prediktory. Algoritmus metody probíhá ve třech krocích, přičemž využívá Pearsonova chí-kvadrát testu o nezávislosti v kontingenční tabulce.

V článku je jako praktický příklad uvedeno vyhodnocení marketingové kampaně nejmenované finanční instituce. Během kampaně byli osloveni současní i bývalí zákazníci, kterým byla předložena zvýhodněná nabídka nového produktu. Pomocí analýzy CHAID se podařilo odhalit segment respondentů, u kterých byla větší tendence na nabídku „kývnout“, i zákazníky, kteří naopak na nabídku reagovali minimálně. CHAID rovněž pomohl odhalit interakce mezi prediktory, které byly na začátku analýzy skryty.

Kromě toho, že CHAID je jednoduchým nástrojem pro odhalování zajímavých segmentů oslovených zákazníků, může díky schopnosti odhalovat interakce mezi proměnnými sloužit i jako zdroj informací pro jiné statistické metody jako jsou logistická regrese, diskriminační analýza či kovariační analýza.

Klíčová slova: Rozhodovací stromy, CHAID, kategoriální data, marketingová kampaň, velké datové soubory, závislost proměnných.

Úvod

Pokud chce firma provést retenci stávajících nebo akvizicí nových zákazníků, spustí marketingovou kampaň (telefonickou, televizní, e-mailovou,…), při které nabídne svůj produkt za zvýhodněných podmínek. Po proběhnutí kampaně je pak užitečné prozkoumat, kteří zákazníci na kampaň reagovali a zda je možné najít určitou skupinu, segment zákazníků, kteří reagovali výrazně jinak než ostatní. Reakce zákazníka je obvykle binární proměnná, která nabývá hodnoty 1 v případě pozitivní reakce a 0 v případě negativní reakce. Na hodnoty této proměnné mohou mít vliv další proměnné, které máme u každého zákazníka k dispozici, ať už jde o proměnné demografického charakteru (věk, pohlaví, bydliště,…) či obchodního charakteru (zda jde o stálého zákazníka, konkurenčního zákazníka, údaje o platební morálce zákazníka,…). Sledují se tedy charakteristiky oslovených zákazníků a vliv těchto charakteristik na reakci zákazníka. Mezi často používané techniky v této oblasti patří diskriminační analýza, vícenásobná regrese nebo logistická regrese, ale v poslední době jsou stále populárnější tzv. rozhodovací stromy, což jsou techniky na rozhraní statistických a data miningových metod. Analýza CHAID je technika konstrukce rozhodovacího stromu určená primárně pro kategoriální data.

Rozhodovací stromy

Rozhodovací stromy jsou techniky zobrazující vztahy mezi veličinami pomocí hierarchického uspořádání, které má formu obráceného stromu. Rozhodovací stromy mají dvě základní roviny využití. První je rozdělení heterogenního souboru (např. zákaznický kmen) do menších homogenních skupin (skupiny zákazníků, které si jsou podobné), druhá je hledání závislosti hodnot námi sledované vysvětlované proměnné (např. průměrná měsíční útrata zákazníka, reakce zákazníka na nabídku) na hodnotách jiných veličin, prediktorů, které máme k dispozici (věk, rodinný stav, příjem, historické údaje o zákaznícovi…).

Základními stavebními prvky rozhodovacího stromu jsou *uzly* a *větve*. Uzly vznikají dělením souboru podle hodnot některého z prediktorů na menší podskupiny (např. podle pohlaví na muže a ženy) a obsahují hodnoty vysvětlované proměnné v příslušné podskupině (např. podíl mužů, kteří reagovali na kampaň, podíl žen, které reagovaly na kampaň). Větve spojují uzly a tedy i jednotlivé hierarchické

úrovně stromu. Nejvýše postavený uzel se nazývá *kořen stromu*. *Koncové uzly* jsou uzly, ze kterých již nevycházejí žádné větve. Rozhodovací stromy, u kterých z každého uzlu, který není koncový, vychází právě dvě větve, se nazývají *binární stromy*, jinak jde o tzv. *vícenásobné rozhodovací stromy*.

Analýza CHAID

Analýza CHAID (Chi-square Automated Interaction Detection) byla představena v roce 1980 v článku G. V. Kasse. Vede ke konstrukci vícenásobných rozhodovacích stromů. Podle (Řezanková & Húsek, 2000) je možné CHAID zařadit do klasifikačních stromů, které využívají tzv. „učitele“. Je určena primárně pro kategoriální proměnné - jsou-li prediktor nebo vysvětlovaná proměnná spojitého typu, je možné je převést na kategoriální ordinálního typu rozdělením oboru jejich hodnot do vhodných intervalů. Algoritmus probíhá ve třech krocích. V prvním kroku testuje u každého prediktoru, zda jsou všechny jeho kategorie statisticky významně odlišné. Pokud ne, pak tyto kategorie sloučí. Ve druhém kroku hledá prediktor, který nejlépe diferencuje hodnoty vysvětlované proměnné a podle kterého následně větví hodnoty vysvětlované proměnné. V prvním i druhém kroku se využívá Pearsonova chí-kvadrát testu o nezávislosti v kontingenční tabulce. Třetí krok algoritmu stanovuje pravidla, při kterých analýza končí.

Algoritmus CHAID – 1. krok

Předpokládáme, že kategoriální vysvětlovaná proměnná má d kategorií a sledovaný ordinální nebo nominální prediktor c kategorií. Uspořádáním dat do kontingenční tabulky získáme tabulku o rozměru c x d. V prvním kroku jsou nejprve pro každý prediktor spojeny ty kategorie, které statisticky významně *neodlišují* hodnoty vysvětlované proměnné.

Pro každou dvojici kategorií testujeme (pomocí chí-kvadrát testu o nezávislosti v kontingenční tabulce), zda jsou tyto kategorie z hlediska vysvětlované proměnné statisticky významně odlišné. U ordinálního prediktoru testujeme pouze každé dvě po sobě jdoucí kategorie, u nominálního prediktoru všechny možné dvojice kategorií. Pokud test neprokáže na zvolené hladině významnosti, že sledované kategorie jsou z hlediska vysvětlované proměnné statisticky významně odlišné, je možné tyto dvě kategorie sloučit do nové, tzv. *redukované kategorie*. Pro všechny z s možných dvojic kategorií vypočteme p-hodnotu χ^2 testu o nezávislosti v odpovídající kontingenční tabulce o rozměru 2 x d jako

$$p^{(i)} = P(\chi^{2(i)} < G^{(i)}[d-1]) \qquad i = 1, 2, \ldots s \qquad (1)$$

a dvojici kategorií s nejvyšší p-hodnotou testu sloučíme.

Iterativně pokračujeme do té doby, dokud nejsou všechny kategorie na zvolené hladině významnosti statisticky významně odlišné. Dostaneme tak nový prediktor s redukovaným počtem kategorií ($r \le c$). Původní kontingenční tabulka o rozměru c x d se tedy nejdříve „rozpadne“ na s kontingenčních tabulek o rozměru 2 x d, které se poté sloučí do *redukované kontingenční tabulky* o rozměru r x d (viz obrázek 1). Tento postup aplikujeme na každý prediktor.

Algoritmus CHAID – 2. krok

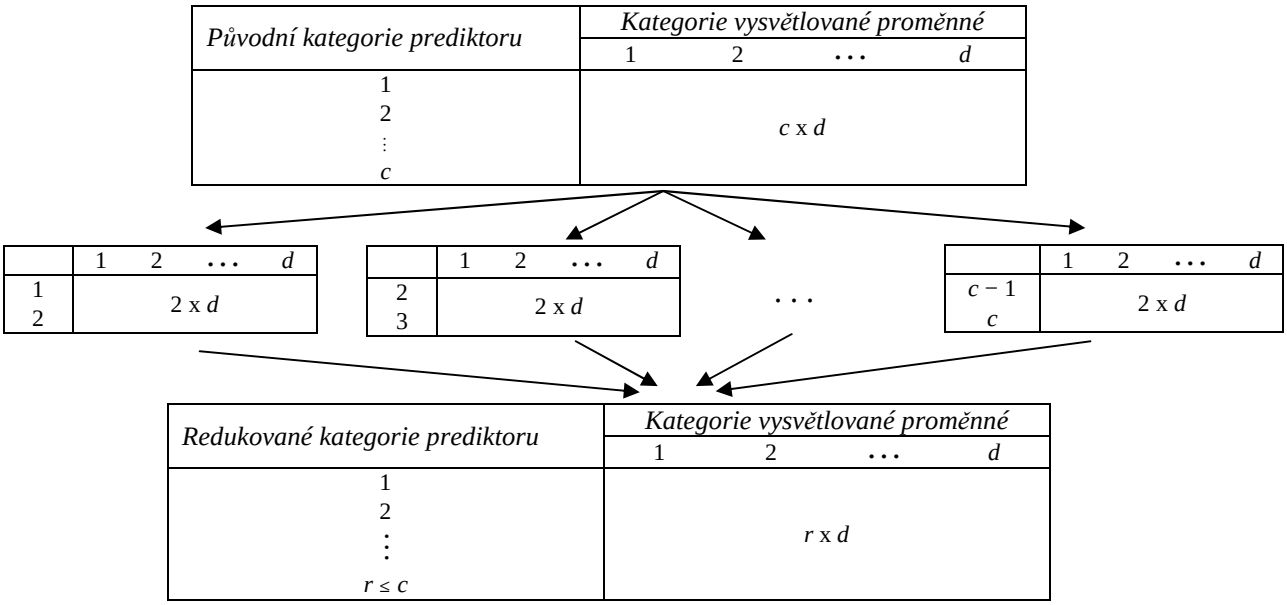
Ve druhém kroku je pro větvení zvolen ten prediktor, u něhož bylo zjištěno, že nejlépe diferencuje hodnoty vysvětlované proměnné. Kritériem pro výběr je *upravená p-hodnota* (Kass, 1980) chí-kvadrát testu o nezávislosti v redukované kontingenční tabulce r x d, která (míněno upravená p-hodnota) zohledňuje počet kategorií, které nový prediktor má (Bonferroniho úprava). Upravenou p-hodnotu, která slouží pro srovnání diskriminační schopnosti jednotlivých prediktorů, získáme vynásobením původní p-hodnoty tzv. *Bonferroniho multiplikátorem* (Kass, 1980). Jeho výpočet se liší podle typu prediktoru. Bonferroniho multiplikátor má tvar

$$B_{ord} = \binom{c-1}{r-1} \qquad (2)$$

pro ordinální prediktor a

$$B_{nom} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!} \qquad (3)$$

pro nominální prediktor. (SPSS, 2008) Prediktor s nejnižší upravenou p-hodnotou se stává první větvící proměnnou a každá z jeho kategorií tvoří jeden uzel rozhodovacího stromu, ve kterém opět sledujeme četnosti kategorií vysvětlované proměnné.



Obrázek 1: Schéma procesu slučování podobných kategorií prediktoru

Algoritmus CHAID – 3. krok

Ve třetím kroku se pro každý vzniklý uzel opakuje první a druhý krok až do té doby, dokud není splněno některé z pravidel pro zastavení algoritmu. Pravidla pro zastavení algoritmu CHAID (Kass, 1980) jsou:

- Pro žádný uzel již neexistuje další prediktor, jehož upravená p-hodnota, by byla nižší než zvolená hladina významnosti.
- Byl dosažen maximální počet úrovní stromu.
- V žádném novém uzlu nelze splnit podmínku pro minimální počet pozorování.

Hladina významnosti, maximální počet úrovní stromu i minimální počet pozorování pro uzel jsou parametry, které se volí před analýzou – pro jejich volbu nejsou pevně stanovená kritéria, ale volí se taková kombinace parametrů, aby výsledný rozhodovací strom měl dostatečnou vypovídací hodnotu.

Výhody a nevýhody analýzy CHAID

Zásadní výhodou analýzy CHAID je, že neklade téměř žádné podmínky na data. Podmínkou chí-kvadrát testu o nezávislosti v kontingenční tabulce je dostatečný počet pozorování a společně s ohledem na dostatečnou vypovídací hodnotu výsledného stromu je prakticky jedinou podmínkou analýzy CHAID dostatečně velký datový soubor. Další výhodou je, že CHAID dává výsledky v přehledné grafické formě, rozhodovací strom je dostatečně dobře čitelný i pro laického uživatele. Kromě toho, že stromová struktura umožňuje odhalit zajímavé shluky pozorování (např. zákaznické segmenty), lze z ní také vyčíst interakci mezi sledovanými prediktory, takže může dobře posloužit např. i pro kovariační analýzu.

Nevýhodou je, že je primárně určen pro kategoriální proměnné a pokud máme v souboru proměnné numerické, pak jejich převedením na kategoriální proměnné přicházíme o část informace obsažené v datech. CHAID navíc, na rozdíl např. od logistické regrese, vícenásobné regrese nebo diskriminační analýzy, neposkytuje výsledek ve formě matematického modelu, kterým by bylo možné závislosti a interakce mezi proměnnými explicitně vyjádřit. Jde tedy o nástroj, který je vhodné použít především

na datové soubory obsahující primárně kategoriální proměnné, což je typický případ dat používaných v marketingových průzkumech a analýzách. Lze jej využít nejen k odhalování vztahů v datech, ale také jako nástroj exploratorní a konfirmatorní analýzy dat a východisko pro další statistické metody. Další výhody a nevýhody jsou uvedeny v (Nisbet, et al., 2009).

Aplikace – retenční kampaň finanční instituce

Nejmenovaná finanční instituce spustila retenční kampaň, při které nabízela svým současným i bývalým zákazníkům nový produkt za zvýhodněných podmínek. Osloveno bylo 12 432 zákazníků, přičemž na nabídku jich pozitivně zareagovalo 267 (2,1 %). Firmu zajímá, jestli je možné najít skupiny zákazníků, které na nabídku reagovaly výrazně jinak než zbytek oslovené zákaznické báze. U zákazníků byly sledovány hodnoty následujících proměnných, v závorce je vždy uveden typ proměnné:

- *Responze* (binární) - vysvětlovaná proměnná, kde hodnota 1 představuje pozitivní reakci na nabídku hodnota 0 negativní reakci na nabídku
- *Pohlaví* (binární)
- Doba ve dnech, která uplynula od konce smlouvy se sledovanou institucí (numerická), resp. *Doba od konce smlouvy*
- *Typ právního subjektu* (kategoriální nominální)

Provedeme analýzu CHAID se zvolenou hladinou významnosti 5 %, maximálním počtem úrovní 3 a minimálním počtem 200 pozorování pro větvící uzel a 100 pozorování pro koncový uzel.

1. krok – redukce kategorií

V prvním kroku testujeme pro každý prediktor, zda jsou všechny jeho kategorie z hlediska vysvětlované proměnné statisticky významně odlišné. Vzhledem k tomu, že proměnná Pohlaví je binární, tak u ní tento krok nemá smysl. Zaměříme se tedy na proměnné Doba od konce smlouvy a Typ právního subjektu.

Doba od konce smlouvy je numerická proměnná, jejíž záporné hodnoty vyjadřují, že zákazník měl k rozhodnému datu stále aktivní smlouvu a kladné hodnoty vyjadřují počet dní, které k rozhodnému datu uplynuly od konce smlouvy se sledovanou finanční institucí. Proměnnou převedeme na kategoriální rozdělením oboru hodnot do intervalů, přičemž záporné hodnoty spojíme do kategorie „Současný zákazník“ a kladné hodnoty spojíme do kategorií, které představuje časový interval poloviny roku, tzn. „Do 0,5 roku“; „Od 0,5 roku do 1 roku“; „Od 1 roku do 1,5 roku“; „Od 1,5 roku do 2 let“ a „Více než 2 roky“ od konce smlouvy. Rozdělení do kategorií a četnosti hodnot vysvětlované proměnné v jednotlivých kategoriích jsou uvedeny v tabulce 1.

Tabulka 1: Původní kategorie prediktoru Doba od konce smlouvy
Zdroj: Data nejmenované finanční instituce, vlastní výpočty

Index kategorie	*Doba od konce smlouvy*	*Responze*		
		NE	ANO	*Celkem*
1	Současný zákazník	2 384	78	2 462
2	Do 0,5 roku	1 973	56	2 029
3	Od 0,5 roku do 1 roku	1 931	43	1 974
4	Od 1 roku do 1,5 roku	2 013	37	2 050
5	Od 1,5 roku do 2 let	1 989	26	2 015
6	Více než 2 roky	1 875	27	1 902
Celkem		12 165	267	12 432

Protože jde o kategoriální proměnnou ordinálního typu, budeme testovat odlišnost pouze těch dvojic kategorií, které za sebou následují (viz tabulka 2). V první iteraci tedy provedeme celkem pět testů nezávislosti v kontingenční tabulce 2 x 2. Výsledky testů jsou uvedeny v tabulce 2. V první iteraci vyšly všechny dvojice kategorií statisticky nevýznamné, ale nejvyšší p-hodnotu má dvojice kategorií „Od 1,5 roku do 2 let“ a „Více než dva roky“. Kategorie tedy sloučíme do nové redukované kategorie a po-

kračujeme 2. iterací. Ve druhé iteraci se jako nejméně odlišné jeví kategorie „Současný zákazník“ a „Do 0,5 roku“. Ve třetí iteraci sloučíme kategorie „Od 0,5 roku do 1 roku“ a „Od 1 roku do 1,5 roku“. Ve čtvrté iteraci máme k dispozici již jen tři kategorie, které na hladině významnosti 5 % vycházejí odlišné z hlediska vysvětlované proměnné Response. Výsledné redukované kategorie prediktoru Doba od konce smlouvy jsou uvedeny v tabulce 3.

Tabulka 2: Iterativní sloučení kategorií prediktoru Doba od konce smlouvy

Indexy testované dvojice kategorií	*p-hodnota chí-kvadrát testu o nezávislosti v kontingenční tabulce 2 x d*
1. iterace	
[1] – [2]	0,42
[2] – [3]	0,24
[3] – [4]	0,40
[4] – [5]	0,18
[5] – [6]	**0,73**
2. iterace	
[1] – [2]	**0,42**
[2] – [3]	0,24
[3] – [4]	0,40
[4] – [5,6]	0,17
3. iterace	
[1,2] – [3]	0,07
[3] – [4]	**0,40**
[4] – [5,6]	0,17
4. iterace	
[1,2] – [3,4]	0,00
[3,4] – [5,6]	0,03

Tabulka 3: Redukované kategorie prediktoru Doba od konce smlouvy

Doba od konce smlouvy	*Responze*		
	NE	ANO	*Celkem*
Současný a do 0,5 roku od konce smlouvy	4 357	134	4 491
Od 0,5 roku do 1,5 roku	3 944	80	4 024
Více než 1,5 roku	3 864	53	3 917
Celkem	12 165	267	12 432

Typ právního subjektu je kategoriální proměnná nominálního typu. V souboru se vyskytují tři kategorie zákazníků – Fyzické osoby nepodnikající (FON), Fyzické osoby podnikající (FOP) a Právnické osoby (PO). Četnosti hodnot vysvětlované proměnné v jednotlivých kategoriích jsou uvedeny v tabulce 4.

Při zkoumání odlišnosti jednotlivých dvojic kategorií budeme postupovat stejně jako v předchozím případě s tím rozdílem, že tentokrát budeme testovat všechny možné dvojice kategorií (viz tabulka 5). V první iteraci tedy provedeme celkem tři testy, na jejichž základě spojíme kategorie FON a PO, ve druhé iteraci pouze potvrdíme, že vůbec existuje závislost response zákazníka na typu právního subjektu zákazníka. V tabulce 6 je uvedena výsledná podoba prediktoru.

Tabulka 4: Původní kategorie prediktoru Typ právního subjektu
Zdroj: Data nejmenované finanční instituce, vlastní výpočty

Typ právního subjektu	*Responze*		
	NE	ANO	Celkem
Fyzická osoba nepodnikající (FON)	8 364	163	8 527
Fyzická osoba podnikající (FOP)	2 486	72	2 558
Právnická osoba (PO)	1 315	32	1 347
Celkem	12 165	267	12 432

Tabulka 5: Iterativní sloučení kategorií prediktoru Typ právního subjektu

Testované dvojice kategorií	*p-hodnota χ^2 testu o nezávislosti v kontingenční tabulce 2 x d*
1. iterace	
[FON] – [FOP]	0,01
[FOP] – [PO]	0,01
[FON] – [PO]	**0,14**
2. iterace	
[FON; PO] – [FOP]	0,01

Tabulka 6: Redukované kategorie prediktoru Typ právního subjektu

Typ právního subjektu	*Response*		
	NE	ANO	*Celkem*
Fyzická osoba nepodnikající (FON) a Právnická osoba (PO)	9 679	195	9 874
Fyzická osoba podnikající (FOP)	2 486	72	2 558
Celkem	12 165	267	12 432

2. krok – volba větvící proměnné

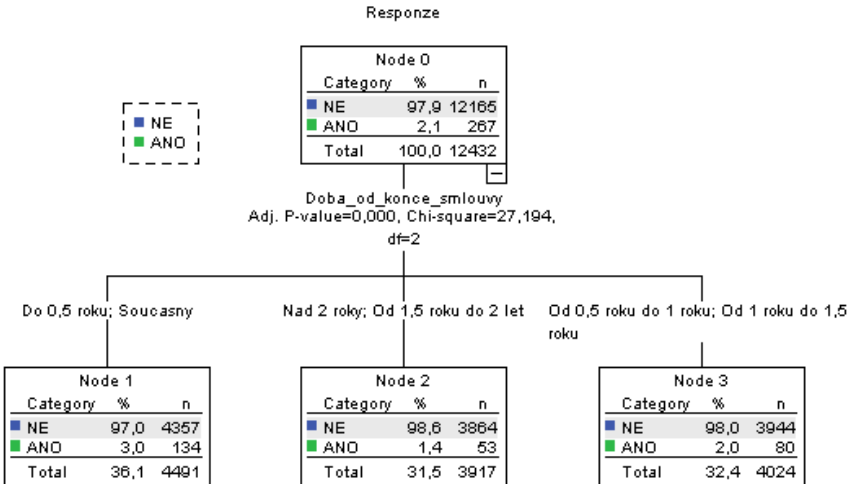
Ve druhém kroku vybíráme větvící proměnnou, tj. prediktor s nejnižší upravenou p-hodnotou χ^2 testu o nezávislosti v redukované kontingenční tabulce. V tabulce 7 jsou uvedeny výsledky testů pro jednotlivé prediktory. Statisticky významně vychází na hladině významnosti 5 % závislost response zákazníka na době uplynuté od konce smlouvy a na typu právního subjektu. Silnější tato závislost vychází pro dobu od konce smlouvy, a tak na první úrovni budeme rozhodovací strom větvit podle redukovaných kategorií prediktoru Doba od konce smlouvy (viz obrázek 1).

Tabulka 7: Výběr první větvící proměnné

Prediktor	*p-hodnota*	*Bonferroniho multiplikátor*	*upravená p-hodnota*
Doba od konce smlouvy	0,00	10	**0,00**
Typ právního subjektu	0,01	3	0,03
Pohlaví	0,40	1	0,40

Z grafu je patrné, že čím kratší doba uplynula od konce smlouvy zákazníka s institucí, tím vyšší byla response. Zatímco ze skupiny současných zákazníků a zákazníků, kterým skončila smlouva nejdéle půl roku před rozhodným datem, kývlo na nabídku 3 % respondentů, u skupiny zákazníků, kteří nejsou ve smluvním vztahu s institucí více než 1,5 roku přijalo o nabídku o 1,4 procentního bodu respondentů méně.

Nyní budeme opakovat kroky 1 a 2 pro každý z nově vzniklých uzlů do té doby, dokud nebude splněno některé z pravidel pro zastavení algoritmu.



Obrázek 2: Rozhodovací strom po prvním větvení

3. krok – zastavení algoritmu a výsledný rozhodovací strom

Průběh druhého a třetího větvení je stručně popsán v tabulce 8. U každého z uzlů je výčet všech možných větvících proměnných včetně upravených p-hodnot. Je patrné, že u zákazníků, kterým smlouva skončila před více než půlrokem (uzly č. 2 a 3), nezávisela jejich response na pohlaví ani na tom, jaký právnický subjekt představují. Tyto uzly proto již není možné dále větvit. Naproti tomu současní zákazníci a zákazníci, kterým skončila smlouva nejvýše před půl rokem, reagovali různě, a to v závislosti na tom, zda šlo o fyzické osoby nepodnikající, nebo právnické či fyzické osoby podnikající (uzel č. 1). Protože uzly 2 a 3 není možné dále větvit, všímáme si již pouze uzlů 4 a 5. V těchto uzlech je navíc možné sledovat již jen závislost na pohlaví. Zatímco fyzické osoby nepodnikající (uzel č. 4) reagovaly nezávisle na pohlaví, u skupiny fyzických osob podnikajících a právnických osob (uzel č. 5) test prokázal závislost response na pohlaví. Algoritmus v tomto momentě končí, protože pro žádný uzel již neexistuje další prediktor, jehož upravená p-hodnota by byla nižší než zvolená hladina významnosti 5 %.

Tabulka 8: Průběh druhého a třetího větvení

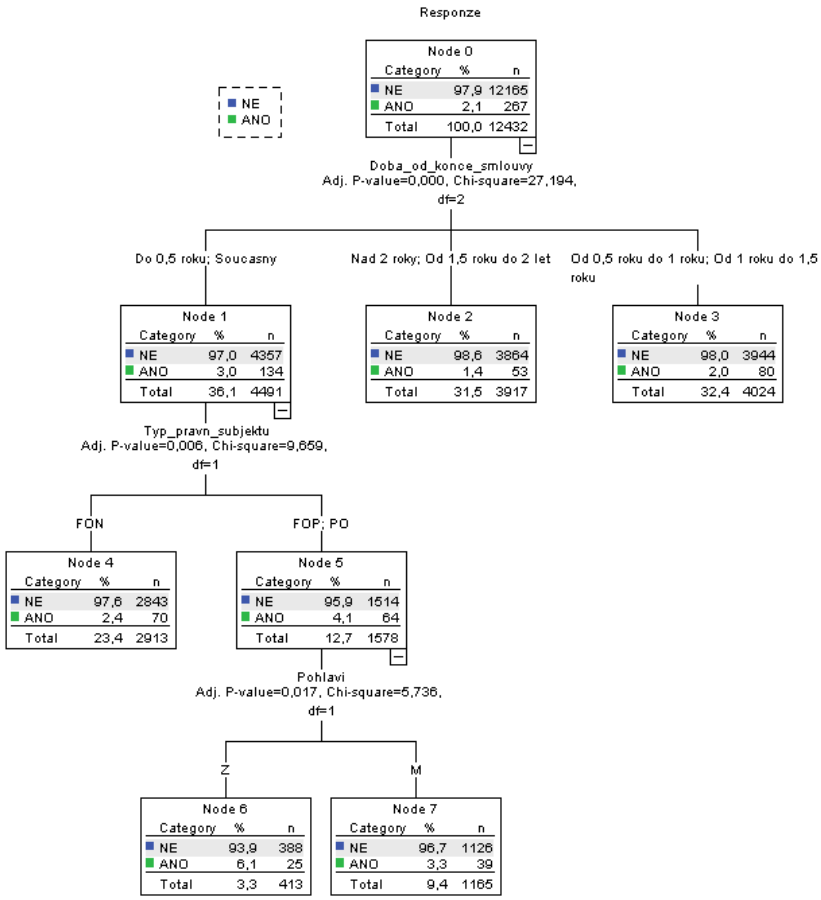
2. větvení			*3. větvení*		
Uzel	Redukované kategorie	Prediktory a upravené p-hodnoty	Uzel	Redukované kategorie	Prediktory a upravené p-hodnoty
1	Současný; Do 0,5 roku	*Typ právnického subjektu* ***0,006***	4	FON	Pohlaví 0,869
		Pohlaví 0,209	5	FOP; PO	*Pohlaví* ***0,017***
3	0d 0,5 roku do 1 roku; Od 1 roku do 1,5 roku	Typ právnického subjektu 0,405			
		Pohlaví 0,209			
2	0d 1,5 roku do 2 let; Nad 2 roky	Typ právnického subjektu 0,345			
		Pohlaví 0,605			

Na obrázku 3 je výsledný rozhodovací strom (výstup je z programového softwarového systému SPSS Statistics 17.0). Na druhé úrovni proběhlo větvení pouze u skupiny současných a nedávných zákazníků (uzel č. 1) podle typu právnického subjektu. V tomto případě byla evidentně vyšší response u redukované kategorie fyzických osob podnikajících a právnických osob (uzel č. 5), tedy obecně u podnikajících subjektů (4,1 %). V dalším opakování algoritmu se v této skupině projevila rozdílná

response podle pohlaví. Zatímco muži reagovali pozitivně ve 3,3 % případů, u oslovených žen dosáhla response výše 6,1 %.

Při pohledu na výsledný rozhodovací strom na obrázku 3 je možné vyčíst některé zajímavé segmenty respondentů. Nejslaběji na nabídku reagoval segment bývalých zákazníků (uzel č. 2), kteří s firmou nejsou ve smluvním vztahu více než 1,5 roku. Tento segment reagoval pouze v 1,4 % případů. Naopak nejsilnějším segmentem z hlediska response byly podnikající ženy, které buď stále jsou zákaznicemi sledované instituce, nebo jim skončila smlouva nejvýše před půl rokem. Ze 413 oslovených respondentů jich nabídku přijalo 25, což představuje responsi ve výši 6,1 %. Na základě této skutečnosti by v budoucnu měla firma při zavedení podobné kampaně cílit zejména na tuto skupinu, zatímco nižší náklady by měly vynaloženy na oslovení zákazníků, kteří nejsou s firmou ve vztahu více než 1,5 roku.

Další informace, kterou lze z rozhodovacího stromu vyčíst, jsou interakce mezi jednotlivými prediktory. Zatímco při výběru první větvící proměnné se response zdála zcela nezávislá na pohlaví, při dalším zkoumání se u skupiny současných a nedávných zákazníků, kteří podnikají, závislost na pohlaví prokázala. CHAID tedy v tomto případě odhalil skrytý vztah mezi responsí klienta na nabídku a pohlavím klienta. Analýza CHAID může tedy velmi dobře posloužit jako zdroj informací pro jiné analýzy jako je logistická regrese, kovariační analýza apod., protože již dopředu podává informaci o možných skrytých interakcích mezi proměnnými.



Obrázek 3: Výsledný rozhodovací strom

Závěr

Analýza CHAID je technika konstrukce rozhodovacích stromů na rozmezí data miningových a statistických metod. Je založena na sérii testů nezávislosti v kontingenční tabulce a je tedy určena primárně pro kategoriální data – s těmi se hojně pracuje především v oblasti marketingu a marketingových výzkumů.

Na praktickém příkladu, který se zabýval proběhnutou marketingovou kampaní nejmenované finanční instituce, se ukázaly největší přednosti této techniky. Přehledný grafický výstup ukázal, které segmenty zákazníků na kampaň reagovaly pozitivně a kteří respondenti naopak na kampaň nereagovali, a proto by měli být z dalších podobných kampaní v rámci zvýšení efektivity vynecháni. Dále analýza pomohla odhalit skryté vztahy mezi proměnnými, které lze využít v dalších analýzách. CHAID je možné použít jako samostatnou techniku vyhodnocení marketingové kampaně, ale také jako techniku exploratorní analýzy dat, která poskytne užitečné informace využitelné v jiných statistických metodách, nebo jako nástroj konfirmatorní analýzy dat k ověření výsledků např. v analýze latentních tříd.

Literatura

KASS, G. V., 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics,* Vol. 29, s. 119-127.

NISBET, R., J. ELDER a G. MINER, 2009. *Handbook of Statistical Analysis and Data Mining Applications.* Boston: Elsevier Academic Press Inc..

ŘEZANKOVÁ, H. a D. HÚSEK, 2000. Klasifikace v programových systémech pro analýzu dat. In: *ROBUST 2000*. Nečtiny: JČMF, s. 257-266.

SPSS, Inc, 2008. *SPSS Statistics 17.0 Algorithms.* Chicago: SPSS.

JEL Classification: C49.

Summary

CHAID Analysis as a Marketing Campaign Evaluation Tool

This paper deals with CHAID analysis as a marketing campaign evaluation tool. CHAID is a method of building decision trees designed for categorical data mainly. CHAID is a simple tool of revealing interesting segments of respondents, but it can also be used as a source of input information in other statistical methods such as logistic regression or discrimination analysis and also as a confirmatory analysis tool in latent class analysis.

In the paper a practical example of using CHAID was presented. An unnamed financial institution launched campaign in which it addressed current and former customers with an advantageous offer. By using CHAID analysis we were able to identify which segments of customers were more attracted by the offer and on the others side which customers didn't react positively. We also found interactions among predictor variables which were hidden at the beginning of the analysis.

Keywords: Decision Trees, CHAID, Categorical Data, Marketing Campaign, Large Datasets, Dependent Variables.

Analýza heterogenních preferencí pomocí hierarchických modelů

Ondřej Vilikus

ondrej.vilikus@vse.cz

Doktorand oboru statistika

Školitel: prof. Ing. Petr Hebák CSc., (hebak@vse.cz)

Abstrakt: V příspěvku je srovnáván hierarchický bayesovský model s modelem latentní proměnné a dvěma referenčními přístupy k modelování preferencí v analýze conjoint. Přístupy jsou srovnávány na základě Monte Carlo studie, ve které byly simulovány datové soubory s různou mírou a povahou heterogenity. Srovnání je založeno na hodnocení modelů z hlediska jejich schopnosti reprodukovat simulované hodnoty parametrů a pomocí vlastního kritéria i z hlediska schopnosti odhalit podobnosti preferencí jednotlivých respondentů. Výsledky studie naznačují, že hierarchický model lépe vystihuje heterogenitu populace a je úspěšnější při odhadu preferencí i v situacích, kdy jsou podmínky nepříznivé z hlediska povahy heterogenity preferencí nebo množství dostupných dat.

Klíčová slova: conjoint analýza, heterogenní populace, hierarchický model, model latentní proměnné.

Úvod

Metoda založená na výběru je v současnosti nejpoužívanějším přístupem v analýze conjoint (Sawtooth, 2011). Za svůj úspěch tato metoda vděčí mimojiné širšímu využití hierarchických bayesovských modelů v posledních letech. Ty umožnily analyzovat preference na úrovni jednotlivců bez neúměrných požadavků na množství dat získaných od každého respondenta. Přesto je zřejmé, že využití hierarchických modelů není všemocné a že v některých případech může být dosaženo lepších výsledků použitím jiného přístupu.

Cílem tohoto článku je zmapovat přínos hierarchických modelů v různých obtížných situacích na základě srovnání s alternativními postupy na simulovaných datech. Pro srovnání jsou použity umělé datové soubory, které se liší v míře heterogenity preferencí (homogenní vs. heterogenní populace) i její povahy (unimodální vs. segmentovaná populace) a v dalších parametrech.

Metody analýzy conjoint

První metody analýzy conjoint používané od 70. let 20. století, mezi které patří *tradiční conjoint* a *metoda částečného profilu*, byly založeny na postupech, kdy respondenti porovnávají, řadí nebo hodnotí na škále předkládané popisy všech nebo některých vlastností výrobků či služeb. Takto získaná data byla používána k pochopení preferencí spotřebitelů a často i k modelování poptávky po existujících nebo uvažovaných produktech.

Žádný z těchto přístupů však věrně nereprezentuje reálné rozhodování spotřebitelů, kteří si skutečně vybírají výrobek nebo službu. Spotřebitelé netráví čas srovnáváním výrobků, z nichž by reálně neuvažovali ani o jednom, ani nehodnotí jejich atraktivitu nebo pravděpodobnost, že si je koupí. Naopak hledají výrobek, který ze všech nabízených nejlépe splňuje jejich požadavky. *Metoda založená na výběru*, která vychází z modelu diskrétní volby, využívá simulace tohoto rozhodování pro analýzu preferencí spotřebitelů.

Modely diskrétní volby byly při analýze preferencí používány již v 70. letech McFaddenem (1974). Šlo ale o analýzu dat popisujících reálná rozhodnutí spotřebitelů. Teoretické základy metody založené na výběru byly položeny až Louveierem a Woodworthem (1983). Jejich přístup kombinoval základní principy analýzy conjoint s modely diskrétní volby. Respondenti při použití této metody opakovaně vybírali ze sady několika profilů popisujících zkoumané vlastnosti výrobku. Předkládané profily byly vytvářeny podobně jako u tradičního conjointu na základě ortogonálního návrhu. Kromě nabízených profilů mohl respondent zvolit také možnost, že žádná z nabízených alternativ pro něj není přijatelná (*nulová alternativa*). Zařazení nulové varianty umožnilo modelovat minimální užitek, který musí výrobek spotřebiteli přinést, aby si ho byl ochoten koupit.

V situaci, kdy *n* respondentům předložíme každému *T* úloh, v nichž mají za úkol vybrat jeden z *J* profilů, odhadujeme pro *t*-tou úlohu *i*-tého respondenta parametry multinomického logitového modelu

$$P\left(y_{ij't}=1\right)=\frac{\exp\left(\sum_{k=1}^{K}x_{j'kt}\beta_{ik}\right)}{\sum_{j=1}^{J}\exp\left(\sum_{k=1}^{K}x_{jkt}\beta_{ik}\right)},$$

kde β_{ik} představují preference *i*-tého respondenta ohledně *k*-tého atributu, x_{jkt} popis *j*-tého profilu v dané úloze a $y_{ij't}$ informaci o tom, zda respondent v dané úloze *j'*-tou variantu vybral (1), nebo nevybral (0). (Pro zjednodušení značení budu v celém textu pro každý atribut uvádět pouze jeden parametr, přestože v praxi je jejich počet závislý na počtu úrovní tohoto atributu.)

Výběr z nabízených profilů je sice intuitivním, ale z hlediska množství získaných informací neefektivním způsobem zjišťování preferencí respondenta. V každé úloze musí respondent prostudovat několik profilů, z nichž vybere jen ten nejatraktivnější. Výběrem získáme informaci o tom, že daný profil je preferován před ostatními, ale pořadí dalších profilů neznáme. V případě, že respondent vybere nulovou alternativu, dokonce získáme pouze informaci o tom, že žádný ze zobrazených profilů nepřináší respondentovi větší užitek než nulová varianta.

Již u tradičního conjointu platilo, že při větším počtu uvažovaných atributů a jejich úrovní je obtížné nebo dokonce nemožné dosáhnout spolehlivých odhadů preferencí každého jednotlivého respondenta. V případě metody založené na výběru to platí dvojnásob. Nižší efektivita získávání informací, která je cenou za jeho vyšší realističnost, vedla k tomu, že u metody založené na výběru bylo zpravidla možné modelovat preference spotřebitelů pouze na úrovni agregovaných dat.

Použití agregovaných modelů je v situacích, kdy mají spotřebitelé rozdílné preference, příliš zjednodušující a vede k chybným závěrům. Pro řešení tohoto problému, který zpočátku brzdil praktické využití metody založené na výběru, bylo navrženo několik možných přístupů. Ty můžeme rozdělit na přístupy snažící se o efektivnější využití dostupných informací modelováním heterogenity preferencí a přístupy, které se naopak snaží o zvýšení efektivity samotného dotazování. Cílem tohoto příspěvku je na simulovaných datech porovnat úspěšnost přístupů, které patří do první skupiny.

Model latentní proměnné

Model latentní proměnné (*latent class model*) (DeSarbo et al, 1995) je zobecněním pravděpodobnostního přístupu k segmentaci pro případy, kdy nedostatek dat neumožňuje odhad parametrů na úrovni každého respondenta. Vychází z předpokladu existence *S* homogenních latentních (skrytých) segmentů tvořených respondenty se shodnými preferencemi. Velikost těchto segmentů je dána parametry α_s, které představují podíl respondentů, náležících do *s*-tého segmentu, a pro které platí:

$$0\le\alpha_s\le 1,\ \sum_{s=1}^{S}\alpha_s=1.$$

Pro respondenta patřícího do *s*-tého segmentu, je pravděpodobnost volby *j'*-té možnosti z *J* nabízených dána jako

$$P\left(y_{sj'}=1\right)=\frac{\exp\left(\sum_{k=1}^{K}x_{j'k}\beta_{ks}\right)}{\sum_{j=1}^{J}\exp\left(\sum_{k=1}^{K}x_{jk}\beta_{ks}\right)},$$

kde β_{ks} představují parametry, vyjadřující preference *s*-tého segmentu ohledně *k*-tého atributu. Skutečné zařazení respondentů do jednotlivých segmentů neznáme, a tak ho odhadujeme spolu s parametry β_{ks} iterativním postupem maximalizujícím hodnotu věrohodnostní funkce

$$L = \prod_{i=1}^{n} \sum_{s=1}^{S} \alpha_s \prod_{t=1}^{T} \prod_{j=1}^{J} \left[\frac{\exp\left(\sum_{k=1}^{K} x_{j'k} \beta_{ks} \right)}{\sum_{j=1}^{J} \exp\left(\sum_{k=1}^{K} x_{jk} \beta_{ks} \right)} \right]^{y_{ijt}} .$$

V první iteraci jsou parametry β_{ks} a α_s zpravidla určeny náhodně. V dalších iteracích střídavě odhadujeme pravděpodobnosti zařazení respondentů do jednotlivých segmentů jako

$$p_{is'} = \frac{\alpha_{s'} \prod_{t=1}^{T} \prod_{j=1}^{J} \left[P(y_{s'j} = 1) \right]^{y_{ijt}}}{\sum_{s=1}^{S} \alpha_s \prod_{t=1}^{T} \prod_{j=1}^{J} \left[P(y_{sj} = 1) \right]^{y_{ijt}}}$$

a následně po přepočítání parametrů α_s jako

$$\alpha_s = \frac{\sum_{i=1}^{n} p_{is}}{n}$$

i samotné parametry β_{ks} pro nové rozřazení respondentů do segmentů. Jelikož tento postup nezaručuje nalezení globálního maxima věrohodnostní funkce, opakujeme obvykle tyto kroky pro několik výchozích stavů a volíme řešení, kterým jsme dosáhli nejvyšší hodnotu věrohodnostní funkce.

Hierarchický bayesovský model

Zatímco model latentní proměnné představuje diskrétní přístup k reprezentaci heterogenity, kdy předpokládáme, že rozdíly v preferencích lze vysvětlit zařazením respondentů do několika segmentů, hierarchický bayesovský model reprezentuje spojitý přístup k modelování heterogenity. Vychází z předpokladu, že každý respondent má preference odlišné od ostatních, ale že odchylky preferencí jednotlivců od „průměrného spotřebitele" můžeme popsat pomocí vícerozměrného pravděpodobnostního rozdělení.

Tento typ modelu je formulován ve formě hierarchie podmíněných rozdělení, ve kterých jsou parametry jedné úrovně vysvětlovány v dalších úrovních. V případě modelu diskrétní volby je jedna úroveň tvořena samotným modelem náhodného užitku

$$P\left(y_{ij} \middle| \boldsymbol{\beta}_i \right) = P\left(y_{ij'} = 1 \right) = \frac{\exp\left(\mathbf{x}_{j'} \boldsymbol{\beta}_i \right)}{\sum_{j=1}^{J} \exp\left(\mathbf{x}_j \boldsymbol{\beta}_i \right)},$$

kde vektor $\boldsymbol{\beta}_i$ reprezentuje preference *i*-tého respondenta a $\mathbf{x}_j$ popis *j*-tého profilu, a druhá úroveň specifikuje rozdělení, ze kterého pocházejí vektory $\boldsymbol{\beta}_i$

$$P\left(\boldsymbol{\beta}_i \middle| \overline{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \right) = N\left(\overline{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \right).$$

Hyperparametry $\overline{\boldsymbol{\beta}}$ a $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, které popisují rozdělení parametrů $\boldsymbol{\beta}_i$ v populaci, neznáme, a proto informaci o těchto parametrech specifikujeme ve formě neinformativního apriorního rozdělení. Odhad parametrů $\boldsymbol{\beta}_i$ je následně založen na odhadu pravděpodobnostního rozdělení $P\left(\boldsymbol{\beta}_i, \overline{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \middle| \mathbf{y}_i \right)$, které reprezentuje naši informaci o odhadovaných parametrech, založenou na získaných datech $\mathbf{y}_i$. Vzhledem k tomu, že toto pravděpodobnostní rozdělení není možné vyjádřit analyticky, odhadujeme jej po-

mocí Monte Carlo simulací založených na markovských řetězcích. Vzhledem k omezené délce příspěvku najde čtenář bližší vysvětlení tohoto přístupu například v (Rossi et al, 2005).

Návrh experimentu

Generování umělých dat

Srovnáváme-li alternativní přístupy k modelování preferencí nezávisle na způsobu sběru dat, máme neocenitelnou možnost využít k tomu simulovaných dat. Ta jsou snadno získatelná a oproti reálným datům máme všechny jejich charakteristiky plně pod kontrolou. Navíc známe v případě simulovaných dat hodnoty parametrů, na základě kterých byla vytvořena a které tak můžeme využít pro posouzení úspěšnosti srovnávaných postupů. Základní myšlenka Monte Carlo simulační studie spočívá ve vygenerování „pravdy“ (parametrů, které jsou středem našeho zájmu) a následně jejich odhadu srovnávanými přístupy na základě simulovaných dat.

Použitý algoritmus generování umělých dat volně vychází z postupů publikovaných v jiných studiích srovnávajících metody modelování preferencí spotřebitelů (Vriens et al 1996; Wirth 2010). Vzhledem ke specifickým cílům mého příspěvku jsem však v některých krocích navrhnul některá níže uvedená zobecnění a úpravy.

Proces generování dat probíhá ve třech fázích:

- generování „skutečných“ hodnot parametrů vyjadřujících preference hypotetických respondentů,
- vytvoření návrhu experimentu pro simulované dotazování,
- simulace pořízení dat nahrazující dotazování skutečných respondentů.

Cílem první fáze je vytvoření matice **B** která je složena z *n* vektorů $\boldsymbol{\beta}_i$ reprezentujících „skutečné“ preference simulovaných respondentů. Prvky této matice β_{ik} bývají zpravidla generovány z nezávislých normálních rozdělení se stejnými (Vriens et al 1996) nebo různými (Wirth 2010) rozptyly. Jelikož je však v praxi nezávislost preferencí jednotlivých úrovní atributů spíše výjimečná, navrhnul jsem obecnější postup, který vede k parametrům, které jsou korelované, přičemž míru závislosti lze snadno měnit.

Postup je založen na následujících krocích:

- Vygenerujeme matici ***Z*** o rozměru $n \times K$, jejíž prvky z_{ik} pocházejí z normovaného normálního rozdělení.
- Vygenerujeme matici ***F*** o rozměru $K \times K$, jejíž prvky f_{ik} pocházejí z rozdělení $N\left(0; q^{k-1}\right)$, kde $0 \le q \le 1$ je parametr ovlivňující míru závislosti generovaných parametrů (pro $q = 1$ získáváme nekorelované parametry).
- Součinem těchto matic získáme matici $\mathbf{B}_0 = \mathbf{FZ}$.

Jelikož jedním z hlavních cílů simulační studie bylo srovnání přístupů v situacích různé povahy heterogenity populace, zařadil jsem do postupu ještě jeden krok, ovlivňující strukturu získané populace z hlediska povahy heterogenity.

Kromě matice $\mathbf{B}_0$ vygenerujeme podobným způsobem ještě vektor $\boldsymbol{\beta}_c = \mathbf{F}z_c$, reprezentující „společné“ preference celé populace a *S* vektorů $\boldsymbol{\beta}_s = \mathbf{F}z_s$, reprezentujících preference charakteristické pro jednotlivé segmenty. Prvky vektorů z_c a z_s jsou opět generovány z normovaného normálního rozdělení. Konečné hodnoty parametrů získáme jako vážený součet

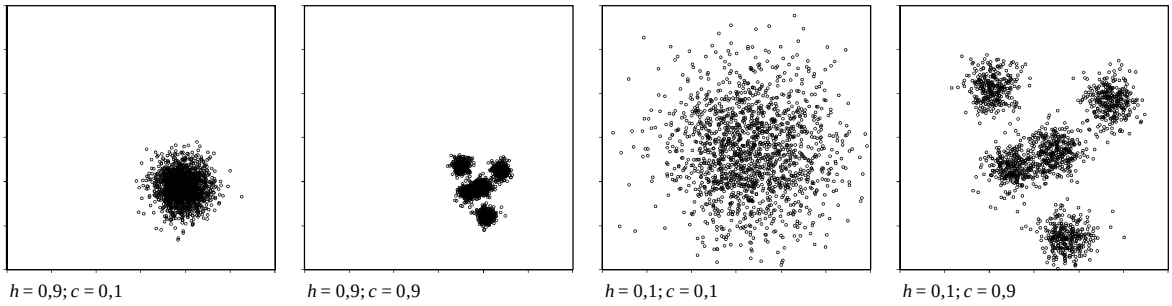
$$\beta_{iks} = w_c\beta_{c,k} + w_s\beta_{s,k} + w_0\beta_{0,ik},$$

kde parametry w_c, w_s a w_0 představují váhy ovlivňující výslednou podobu heterogenity. Tyto váhy získáme jako $w_c = h/d$, $w_s = (1-h)c/d$ a $w_0 = (1-h)(1-c)/d$, kde parametr *h* ovlivňuje míru

heterogenity (pro $h = 1$ jsou preference všech respondentů shodné, při $h = 0$ jsou maximálně odlišné), parametr c určuje podíl variability preferencí, který lze vysvětlit příslušností k jednomu ze segmentů, a d představuje výraz, který zajišťuje, aby se změnou parametrů h a c neměnil rozptyl generovaných parametrů β_{iks} a pro který platí

$$d = \sqrt{h^2 + (1-h)^2(2c^2 - 2c + 1)}\,.$$

Efekt volby různých hodnot parametrů h a c je ilustrován na obrázku 1. Z obrázku je patrná i další výhoda popsaného postupu. Volbou různých hodnot parametru h a c můžeme vytvořit datové soubory, které se při zachování matic F a Z a vektorů z_c a z_s budou lišit pouze v povaze heterogenity, ale jinak budou maximálně srovnatelné.



Obrázek 1: Efekt volby parametrů *h* a *c* (zobrazeny jsou dvě hlavní komponenty matice *B*)

Návrhy experimentů jsem vytvořil pro zvolený počet hypotetických respondentů pomocí procedury *balanced overlap* v programu CBC/Web 7.0. Jde o využití pseudonáhodného přístupu, který zajišťuje ortogonalitu návrhu, ale umožňuje, aby se v rámci jedné úlohy opakovaly úrovně atributů napříč zobrazenými profily. Více informací čtenář najde v (Sawtooth, 2008).

Pro simulaci dotazování jsem použil stejný postup založený na modelu náhodného užitku, jako Wirth (2010). V každé úloze simulovaný respondent vybírá profil, který mu přináší nejvyšší užitek získaný jako

$$U_j = \mathbf{x}_j \boldsymbol{\beta}_i + \varepsilon_{ij}\,,$$

kde náhodná složka ε_{ij} má Gumbelovo rozdělení se střední hodnotou 0 a rozptylem $\pi^2/6$.

Hodnotící kritéria

Při srovnání popsaných přístupů jsem použil některá běžně užívaná kritéria úspěšnosti modelu a vlastní kritérium pro hodnocení schopnosti modelu vystihnout strukturu respondentů z hlediska podobnosti jejich preferencí. Jako hlavní kritérium úspěšnosti modelu jsem použil ukazatel *cor*, představující průměrnou hodnotu korelačního koeficientu mezi simulovanými parametry každého respondenta β_{ik} a jejich odhadnutými protějšky $\hat{\beta}_{ik}$.

Jako vedlejší kritéria úspěšnosti modelu jsem použil ukazatel *hit* udávající průměrný podíl správně odhadnutých rozhodnutí respondentů ve 100 kontrolních úlohách shodných pro všechny respondenty, které nebyly použity pro odhad parametrů, a *pct* představující průměrnou predikovanou pravděpodobnost vybrání správné varianty v těchto kontrolních úlohách.

Vlastní kritérium hodnotící schopnost zachytit strukturu souboru z hlediska podobnosti je rovněž založeno na kontrolních úlohách. Kritérium c_U vyjadřuje míru schopnosti modelu odhalit, kteří respondenti hodnotí profily podobně a kteří odlišně. Definujeme-li v souladu s návrhem Greena (1975) podobnost dvou respondentů $\delta_{(2)}(\boldsymbol{\beta}_i, \boldsymbol{\beta}_{i'})$ jako korelační koeficient mezi jejich užitky spočítanými pro všechny myslitelné profily, které můžeme na základě testovaných atributů vytvořit, získáme hodnotu kritéria c_U jako korelační koeficient mezi takto získanými podobnostmi pro všechny dvojice respondentů založe-

ných na skutečných preferencích a stejným způsobem spočítanými podobnostmi $\hat{\delta}_{(2)}\left(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_{i'}\right)$ založených na odhadech preferencí. Je zřejmé, výpočet přesné hodnoty tohoto ukazatele by byl výpočetně příliš komplikovaný, a tak jsem použil pouze jeho odhad na základě 100 použitých kontrolních úloh a 1 200 náhodně vybraných dvojic respondentů.

Referenční modely

Jak model latentní proměnné, tak hierarchický model kombinují při odhadu preferencí informace získané od více respondentů. Zatímco model latentní proměnné odhaduje preference skrytých segmentů reprezentujících určitou část populace, v případě hierarchického modelu probíhá „vypůjčování" informace od jiných respondentů prostřednictvím předpokladu o rozdělení individuálních parametrů, díky kterému jsou jejich odhady stahovány ke středu (v anglické literatuře bývá tento efekt nazýván *bayesian shrinkage*).

V obou případech bychom měli získat lepší výsledky než použitím modelu, který s heterogenitou preferencí nepočítá vůbec (*model založený na agregovaných datech*), nebo modelem, který používá čistě informace získané od jednotlivých respondentů (*individuální model*). Z tohoto důvodu jsem do srovnání zařadil i výsledky pro tyto dva modely. Jelikož parametry individuálního modelu zpravidla nelze určit běžným způsobem (počet odhadovaných parametrů pro každého respondenta je výrazně vyšší než počet pozorování), jsou jejich odhady získány s využitím hřebenové regrese.

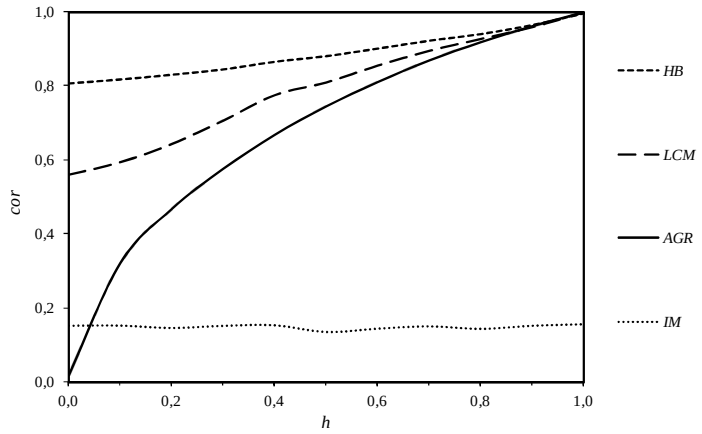
Výsledky

Níže uvedené výsledky jsou založené na simulovaných datech, které se vždy liší pouze v jednom uvažovaném parametru. Pokud není dále uvedeno jinak, všechny použité datové soubory obsahovaly simulované odpovědi $n = 400$ respondentů, z nichž každý v $T = 10$ úlohách, volil jeden z $J = 4$ profilů, které byly popsány $K = 6$ atributy, každý o $L = 6$ úrovních. Preference respondentů tedy byly reprezentovány $K(L-1) = 30$ parametry. Generované parametry jsou lehce korelované díky volbě parametru $q = 0{,}9$. Pro datové soubory, ve kterých tvoří respondenti přirozené segmenty, je jejich počet vždy $S = 5$.

Parametry hierarchického modelu jsem odhadl pomocí programu *CBC/HB* 5.2.8 a parametry ostatních modelů programem *Latent Class* v4.0.8. Pro lepší srovnatelnost výsledků s jinými studiemi jsem ponechal volitelné parametry procedur v původním nastavení. Při použití modelu latentní proměnné jsem odhadoval pouze model se „správným" počtem segmentů. Díky tomu odpadla v praxi nutná fáze hledání optimálního počtu segmentů.

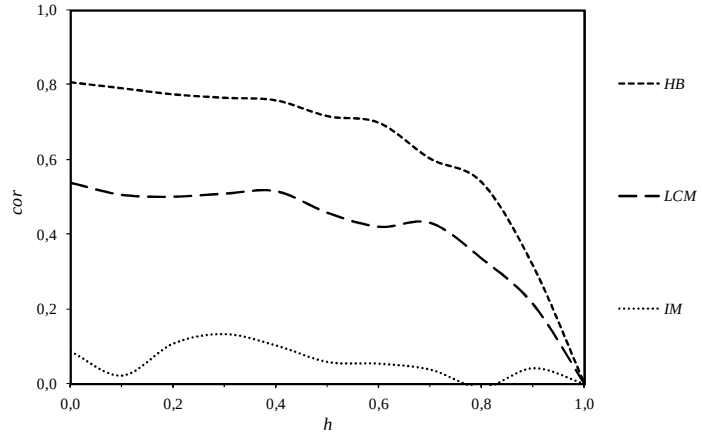
Citlivost na strukturu populace

Schopnost modelu úspěšně predikovat rozhodnutí heterogenní populace jsem otestoval pomocí 11 souborů, u nichž byl parametr *h* volen od 0 do 1. Parametr *c* byl ponechán na hodnotě 0. Šlo tedy ve všech případech o unimodální rozdělení. Dosažené hodnoty ukazatele *cor* pro srovnávané modely jsou zachyceny na obrázku 2. Je zřejmé, že zatímco schopnost agregátního modelu zachytit heterogenní preference s klesající hodnotou *h* (a tedy menší podobností respondentů) rychle klesá, u modelu latentní proměnné a hierarchického modelu je pokles mnohem pomalejší. Úspěšnost individuálního modelu je na míře heterogenity nezávislá, ale ve všech případech velmi nízká.



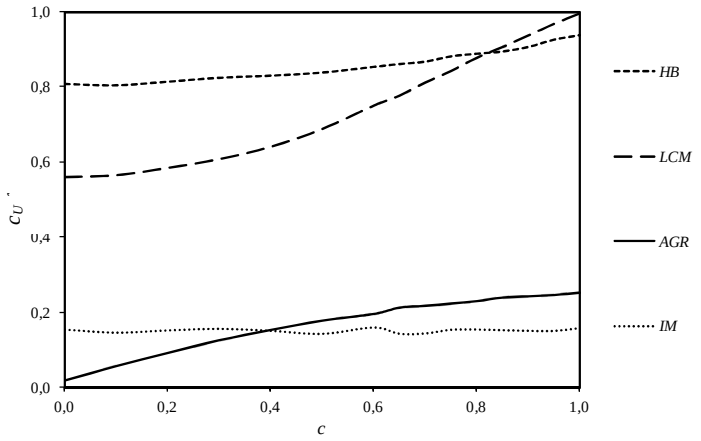
Obrázek 2: Úspěšnost srovnávaných modelů v závislosti na volbě parametru *h*
HB – hierarchický model, LCM – model latentní proměnné, AGR – agregátní, IM – individuální

Rozdíly v úspěšnosti srovnávaných modelů dobře vysvětluje ukazatel c_U, jehož dosažené hodnoty jsou znázorněny na obrázku 3. Hierarchický model dobře zachycuje rozdíly v preferencích i v případě poměrně homogenní populace a vždy lépe než model latentní proměnné. Schopnost individuálního modelu odhalit odlišnosti v preferencích respondentů je díky nestabilitě odhadů velmi omezená. Výsledky agregátního modelu nejsou zobrazeny, jelikož ten heterogenitu vůbec neuvažuje.



Obrázek 3: Schopnost modelů vystihnout heterogenní preference v závislosti na volbě parametru *h*
HB – hierarchický model, LCM – model latentní proměnné, IM – individuální

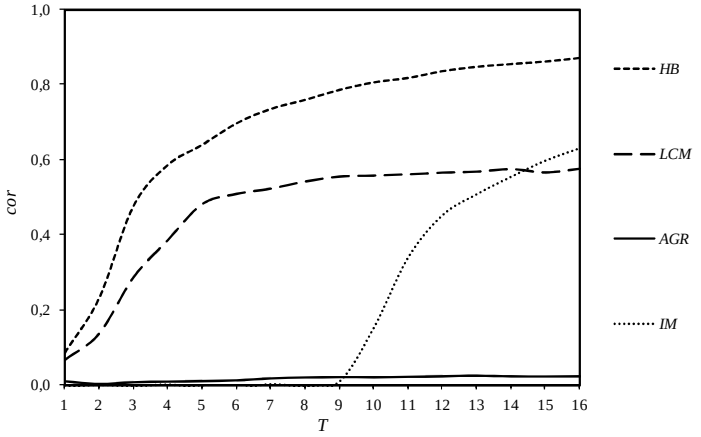
Kromě míry heterogenity populace jsem dále zkoumal vliv její struktury dané volbou parametru *c*. Zatímco parametr *h* byl tentokrát ponechán na nulové hodnotě (maximální míra heterogenity), parametr *c* jsem měnil v 15 souborech od 0 do 1. Struktura populace v podobě izolovaných segmentů při větších hodnotách parametru *c* staví metodu latentní proměnné do role jasného favorita (hierarchický model předpokládá unimodální rozdělení individuálních parametrů). Hodnoty ukazatele *cor*, zachycené na obrázku 4 však tento předpoklad zcela nepotvrzují. Model latentní proměnné dosahuje s rostoucí hodnotou parametru *c* sice relativně lepších výsledků, úspěšnost hierarchického modelu však překonává až pro hodnotu $c = 0{,}85$, která je u reálných dat jen těžko představitelná.



Obrázek 4: Úspěšnost srovnávaných modelů v závislosti na volbě parametru *c*
HB – hierarchický model, LCM – model latentní proměnné, AGR – agregátní, IM – individuální

Citlivost na dostatek dat

Primární motivací jak modelu latentní proměnné, tak hierarchického modelu byla potřeba modelovat heterogenní preference jednotlivých respondentů i v situaci, kdy jsme omezeni množstvím dat, které jsme od nich schopní získat. Pokud bychom analyzovali preference neúnavných robotů, od kterých bychom mohli získat tolik dat, kolik bychom potřebovali, tyto přístupy (a v podstatě ani analýzu conjoint) bychom nepotřebovali. Cílem posledního provedeného experimentu tedy bylo zjistit, jak nízko můžeme stanovit naše požadavky na množství získaných dat, abychom ještě dokázali zachytit odlišné preference respondentů.



Obrázek 5: Úspěšnost srovnávaných modelů v závislosti na počtu předložených úloh *T*
HB – hierarchický model, LCM – model latentní proměnné, AGR – agregátní, IM – individuální

Obrázek 5 zachycuje hodnoty ukazatele *cor* v situaci, kdy jsou parametry *h* a *c* rovny nule a kdy měníme počet úloh *T* předložených každému respondentovi od 1 do 16. Na obrázku je dobře vidět, že zatímco hierarchický model v tomto případě dává už pro tři až čtyři úlohy slušné výsledky a s každou další se schopnost vystihnout preference jednotlivých respondentů dále zlepšuje, v případě modelu latentní proměnné jsou přijatelné výsledky až v případě pěti úloh a dále už nedochází k podstatnému zlepšení.

V případě individuálního modelu jsou požadavky na množství dat nejvyšší. Potřebujeme alespoň deset úloh, abychom získali výsledky, které budou lepší než čistě náhodné určení odhadovaných parametrů. Pro čtrnáct úloh však už úspěšnost individuálního modelu překonává výsledky dosažené modelem latentní proměnné a pro další zvyšování počtu úloh můžeme očekávat, že se bude úspěšnost přibližovat úspěšnosti hierarchického modelu.

Agregátní model je v případě heterogenní populace nevhodný bez ohledu na počet úloh zobrazených jednotlivým respondentům.

Závěr

Výsledky experimentu potvrdily, že hierarchický model je pro analýzu heterogenních preferencí nejlepší volbou. Zatímco v případě homogenní populace z hlediska preferencí je úspěšnost použitých modelů srovnatelná, s rostoucími rozdíly v preferencích respondentů je hierarchický model díky schopnosti lépe odhadnout preference na individuální úrovni výrazně lepší než model založený na agregovaných datech, ale i než model latentní proměnné. Srovnání s individuálním modelem ukazuje, že přínos hierarchického modelu není jen v odhadu individuálních preferencí, ale i v zajištění stability těchto odhadů.

Zajímavým zjištěním je, že hierarchický model funguje dobře i v situaci, kdy je populace tvořena izolovanými segmenty, což je v rozporu s představou, ze které model vychází. Kromě případu výjimečně izolovaných segmentů hierarchický model lépe reprodukuje podobnosti respondentů z hlediska jejich preferencí než model latentní proměnné, což nasvědčuje tomu, že použitím hierarchického modelu pro odhad individuálních preferencí a následně některé z metod shlukové analýzy na hodnotách získaných parametrů můžeme získat lepší rozdělení populace do homogenních segmentů, než samotným použitím modelu latentní proměnné.

Přestože se robustnost hierarchického modelu prokázala i v případě velmi nízkého počtu úloh zodpovězených respondenty, stále považuji tvrzení, že hierarchický model je preferovanou volbou pro úplně všechny situace za předčasné. Další výzkum hodlám věnovat detailnější analýze vlivu korelovanosti odhadovaných parametrů a dalším vlivům, jako jsou počet atributů či jejich úrovní nebo relativní důležitost jednotlivých atributů v rozhodování spotřebitele.

Literatura

DESARBO, W. S. et al., 1995. Market Segmentation with Choice-based Conjoint Analysis. *Marketing Letters*, Vol. 6 (2), s. 137–147.

GREEN, P. E. and KRIEGER, A. M., 1985. Buyer Similarity Measures in Conjoint Analysis: Some Alternative Proposals. *Journal of Classification,* Vol. 1, s. 41–61.

LOUVIERE, J. J. and WOODWORTH, G., 1983. Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data. *Journal of Marketing Research,* Vol. 20, 350–67.

MCFADDEN, D., 1974. Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics*, Vol. 1, (2), s. 105–142.

ROSSI, P. E., et al., 2005. *Bayesian Statistics and Marketing*. Hoboken: Wiley. ISBN: 0-470-86367-6.

SAWTOOTH SOFTWARE INC., 2008. CBC v6.0 Technical Paper. Sequim, WA: *Sawtooth Software Technical Paper Series*. Dostupné z: http://www.sawtoothsoftware.com/download/techpap/-cbctech.pdf.

SAWTOOTH SOFTWARE INC, 2011. Report on Conjoint Usage. *Sawtooth Solutions Newsletter*, Summer 2011. Dostupné z: http://www.sawtoothsoftware.com/education/ssolutions.shtml#news.

VRIENS, M et al, 1996. Metric Conjoint Segmentation Methods: A Monte Carlo Comparison. *Journal of Marketing Research*, Vol. 33 (1), s. 73–85.

WIRTH, R., 2010. HB-CBC, HB-Best-worst-CBC or no HB at all? Sawtooth Software Inc. (ed.): *2010 Sawtooth Software Conference Proceedings*, s. 321–356. Dostupné z: http://www.sawtoothsoftware.com/download/techpap/2010Proceedings.pdf.

JEL Classification: C11, C18, C25

Summary

Analyzing heterogeneous preferences with hierarchical models

Comparison of hierarchical Bayesian model with latent class model and two reference approaches to preference estimation in conjoint analysis is given. The approaches are compared empirically based on a Monte Carlo study in which simulated datasets with different amount and type of heterogeneity are used. The comparison pertains to measures that assess simulated parameter values recovery, goodness of fit and own measure of model ability to capture structure of the sample with respect to similarity of individuals in preferences. Hierarchical model is found to be best capturing heterogeneity of the sample and outperforms other approaches in parameter recovery even under adverse conditions of strongly segmented sample or low amount of data available for each respondent.

Keywords: conjoint analysis, heterogeneous population, hierarchical model, latent class model.

Neostré shlukování

Tomáš Vintr
tomas.vintr@vse.cz
Doktorand oboru statistika
Školitel: prof. Ing. Hana Řezanková, CSc., (rezanka@vse.cz)

Abstrakt: Použití metod rozkladu má několik známých a často citovaných omezení. V článku jsou prováděny testy robustnosti vybraných shlukovacích algoritmů (k-průměry, FCM, PCA, UPFC) při různých počtech objektů ve shluku, počtech shluků, dimenzích, mírách přidaného šumu a překryvech shluků. Výsledky jsou zobrazeny v grafech a komentovány.

Existují tři obecně známé problémy metod rozkladu. První z nich spočívá v tom, že pokud jsou jednotlivé shluky výrazně odlišné svou velikostí (počtem objektů patřících do jednotlivých shluků), početnější shluky přitahují vzorové objekty k sobě a tyto vzorové objekty pak buď zůstanou v oblasti mezi shluky, nebo jsou přitaženy přímo do velkého shluku a ten pak rozdělí na několik částí. Druhý problém spočívá v určení počtu shluků, což je parametr, který je nutno zadat algoritmu před spuštěním. Počet shluků je buď určován na základě expertního odhadu výzkumníka, případně existuje velké množství indexů pro odhad tohoto parametru. Od toho odvozeným třetím problémem je inicializace, tedy umístění příslušného počtu vzorových objektů do prostoru objektů. Inicializace je nejčastěji prováděna náhodně, ať už náhodným výběrem objektů z datového souboru a jejich prohlášení za vzorové objekty, nebo vytvořením nových objektů náhodně rozmístěných v prostoru objektů.

V tomto článku jsou testovány algoritmy při přesně známém počtu shluků s identickým (normálním) rozdělením a počtem objektů ve shluku s cílem odhalit, zda algoritmy trpí i jinými neduhy. Pro testování se používají tři typy inicializací: vzorové objekty jsou umístěny na pozice středních hodnot rozdělení, které shluky vygenerovalo, na pozice náhodně vybraných objektů a na náhodně vybrané pozice v prostoru objektů.

Klíčová slova: k-průměry, FCM, PCM, PCA, UPFC, metody rozkladu.

Úvod

Shluková analýza je soubor vícerozměrných statistických metod sloužících ke klasifikaci objektů. Metody shlukové analýzy patří do třídy metod, kterým se říká *učení bez učitele*. Jejich cílem je vytvořit podmnožiny datového souboru takové, aby podobnost objektů z jedné podmnožiny a rozdílnost objektů z různých podmnožin byla co nejvyšší. Takovýmto podmnožinám se pak říká *shluky*. V tomto článku se budu zabývat shlukováním založeným na *metodách rozkladu*, kde jsou jednotlivé shluky reprezentovány vektory, kterým se říká *vzorové objekty*.

Existují tři obecně známé problémy této třídy algoritmů. První z nich spočívá v tom, že pokud jsou jednotlivé shluky výrazně odlišné svou velikostí (počtem objektů patřících do jednotlivých shluků), početnější shluky přitahují vzorové objekty k sobě a tyto vzorové objekty pak buď zůstanou v oblasti mezi shluky, nebo jsou přitaženy přímo do velkého shluku a ten pak rozdělí na několik částí (Kaymak and Setnes, 2000). Druhý problém spočívá v určení počtu shluků, což je parametr, který je nutno zadat algoritmu před spuštěním. Počet shluků je buď určován na základě expertního odhadu výzkumníka, případně existuje velké množství indexů pro odhad tohoto parametru (Wang et al., 2009). Od toho odvozeným třetím problémem je *inicializace*, tedy umístění příslušného počtu vzorových objektů do *prostoru objektů*. Inicializace je nejčastěji prováděna náhodně, ať už náhodným výběrem objektů z datového souboru a jejich prohlášení za vzorové objekty, nebo vytvořením nových objektů náhodně rozmístěných v prostoru objektů.

V tomto článku budu testovat algoritmy při přesně známém počtu shluků s identickým (normálním) rozdělením a počtem objektů ve shluku s cílem odhalit, zda algoritmy trpí i jinými neduhy. Pro testování budu používat tři typy inicializací: vzorové objekty umístím na pozice středních hodnot rozdělení, které shluky vygenerovalo, na pozice náhodně vybraných objektů a na náhodně vybrané pozice v prostoru objektů.

Metody rozkladu

Myšlenka shlukování metodou rozkladu je poměrně jednoduchá, její realizace však s sebou nese významná úskalí. Předpokládejme, že známe polohy středních hodnot jednotlivých shluků a že známe přiřazení každého objektu ke svému shluku. Každý shluk chápejme jako výstup měření nějaké náhodné veličiny. Položme vzorové objekty do středních hodnot shluků. Potom součet rozptylů jednotlivých shluků bude číslo J, dané vztahem

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d_{ij}^2,$$

kde c je počet shluků, n je počet objektů v datovém souboru, u_{ij} je hodnota přiřazení j-tého objektu i-tému shluku s podmínkou

$$\sum_{j=1}^{n} u_{ij} = 1, \forall i \in \{1,..., c\},$$

a $d_{ij} = ||\mathbf{x}_j - \mathbf{c}_i||$ je zvolená metrika, kde $\mathbf{c}_i$ je střední hodnota i-tého shluku a $\mathbf{x}_j$ je j-tý objekt.

Jednoduchou úvahou lze dojít k závěru, že pokud bychom přesunuli některý vzorový objekt mimo střední hodnotu, případně nějaký objekt přiřadili špatnému shluku, součet rozptylů shluků by se zvýšil.

Z obdobné úvahy vychází základní algoritmus známý pod názvem *k-průměry* (MacQueen, J. B., 1967). V tomto článku jej budu označovat zkratkou HCM (Hard c-Means). Princip algoritmu spočívá v iterativním hledání minima účelové funkce

$$J_{HCM}(\mathbf{X}; \mathbf{U}_{HCM}; \mathbf{C}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} d_{ij}^2,$$

kde $\mathbf{X} = \{x_1, ..., x_n\}$, $\mathbf{X} \subset \mathbf{R}^{\dim}$, je množina objektů a $\mathbf{C} = \{C_1, ..., C_c\}$ je množina shluků, $\mathbf{U}_{HCM} = (u_{ij})$ označuje *matici příslušností* s podmínkami

$$u_{ij} = \{0,1\}, \forall i \in \{1,...,c\}, \forall j \in \{1,...,n\},$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1,...,n\},$$

a $d_{ij} = ||\mathbf{x}_j - \mathbf{c}_i||$ je zvolená metrika, kde $\mathbf{c}_i$ je vzorový objekt z C_i.

Vzhledem k tomu, že jsou všechny objekty datového souboru přiřazeny nějakému shluku, je HCM velmi citlivý na přítomnost odlehlých pozorování. Hranice jednotlivých shluků se nachází uprostřed mezi dvěma sousedními vzorovými objekty.

Pro setření „ostré“ hranice mezi shluky byl vytvořen přístup neostrého shlukování (Fuzzy Clustering). Základním algoritmem neostrého shlukování je algoritmus označovaný běžně zkratkou FCM (Fuzzy c-Means) (Bezdek, J. C., 1981). Princip algoritmu spočívá v iterativním hledání minima účelové funkce

$$J_{FCM}(\mathbf{X}; \mathbf{U}_{FCM}; \mathbf{C}) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2,$$

kde $\mathbf{X} = \{x_1, ..., x_n\}$, $\mathbf{X} \subset \mathbf{R}^{\dim}$, je množina objektů a $\mathbf{C} = \{C_1, ..., C_c\}$ je množina shluků, $\mathbf{U}_{FCM} = (u_{ij})$ označuje matici příslušností s podmínkami

$$\sum_{j=1}^{n} u_{ij} > 0, \forall i \in \{1,...,c\},$$

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j \in \{1,...,n\},$$

$d_{ij} = ||\mathbf{x}_j - \mathbf{c}_i||$ je zvolená metrika, kde $\mathbf{c}_i$ je vzorový objekt C_i, a $m > 1$ je *mlžitel*.

Důležitým parametrem účelové funkce je mlžitel. Pokud by jeho hodnota byla položena rovna 1, algoritmus FCM by se redukoval na HCM (Miyamoto et al., 2008). Ani FCM však neřeší problém vlivu odlehlých pozorování, neboť součet měr příslušnosti každého objektu datového souboru je 1.

Algoritmus, jehož cílem bylo vypořádat se odlehlými pozorováními se v literatuře označuje zkratkou PCM (Possibilistic c-Means) (Krishnapura and Keller, 1993). V tomto algoritmu byla upravena podmínka pro míru příslušnosti v tom smyslu, že součet měr příslušnosti každého objektu datového souboru je číslo mezi 0 a 1, s cílem, aby odlehlá pozorování měla součet měr příslušnosti co nejnižší. Aby se zamezilo triviálnímu řešení, kdy jsou všechny míry příslušnosti rovny 0, musela být přeformulována účelová funkce, ve které se objevuje sporný parametr. Buď se rozhodneme za tento parametr dosadit nějakou konstantu, k čemuž potřebujeme významnou znalost datového souboru, nebo použijeme autory algoritmu doporučovaný výpočet. V tomto případě je však tento parametr funkcí příslušností objektů ke shlukům, čímž dochází k problémům s derivací účelové funkce a k nemožnosti vygenerovat matici příslušností. I přesto, že matici příslušností lze vygenerovat tím, že provedeme jednu či několik iterací algoritmu FCM, je algoritmus PCM výpočetně nestabilní (Krishnapuram, R. and Keller, J. M., 1993). Kvalitativním problémem PCM je pak velká náchylnost ke splývání shluků. Pokud je inicializace dvou vzorových objektů realizována v prostoru jednoho shluku, je velmi pravděpodobné, že dojde k nalezení minima obou vzorových objektů ve středu tohoto shluku.

Výše uvedené algoritmy tvoří myšlenkový základ algoritmů založených na rozkladu datového souboru na pevně stanovený počet podmnožin. Jistě stojí za zmínku, že užití mlžitele není jediný způsob, jak dosáhnout neostrosti hranic mezi shluky (Miyamoto et al., 2008). Množství úprav účelové funkce je však velmi rozsáhlé, a jistě přesahuje rámec jednoho článku. Pro účely svého článku jsem vybral pouze úpravy přímo vycházející z výše uvedených algoritmů, které mají za cíl je spíše poupravit než přetvořit.

Jedním z takových počinů je algoritmus známý pod zkratkou PCA (Possibilistic Clustering Algorithm) (Yang and Wu, 2006). Tento algoritmus zachovává vlastnosti PCM, ale zbavuje algoritmus výpočetních nestabilit. Účelová funkce PCA je

$$J_{PCA}(\mathbf{X};\mathbf{U}_{PCA};\mathbf{C}) = \sum_{i=1}^{c}\sum_{j=1}^{l} u_{ij}^{m} d_{ij}^{2} + \frac{\beta}{m^2\sqrt{c}} \sum_{i=1}^{c}\sum_{j=1}^{l} \left(u_{ij}^{m} \log u_{ij}^{m} - u_{ij}^{m}\right),$$

kde $\mathbf{X} = \{x_1, ..., x_n\}$, $\mathbf{X} \subset \mathbf{R}^{\dim}$, je množina objektů a $\mathbf{C} = \{C_1, ..., C_c\}$ je množina shluků, $\mathbf{U}_{PCA} = (u_{ij})$ označuje matici příslušností s podmínkami

$$\sum_{j=1}^{n} u_{ij} > 0, \forall i \in \{1,...,c\},$$

$$0 < \sum_{i=1}^{c} u_{ij} \le c, \forall j \in \{1,...,n\},$$

$d_{ij} = \|\mathbf{x}_j - \mathbf{c}_i\|$ je zvolená metrika, kde $\mathbf{c}_i$ je vzorový objekt C_i, $m > 1$ je mlžitel a β je kladný parametr, který autoři definují jako kovarianci datového souboru

$$\beta = \frac{1}{n}\sum_{j=1}^{n} \left\| x_j - \bar{x} \right\|^2,$$

kde $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$.

Mezitím autoři FCM a PCM pracovali na algoritmu, který by v sobě udržel dobré vlastnosti obou algoritmů a zbavil se těch nežádoucích (Pal et al., 1997, Pal et al., 2005). Požadovaných vlastností bylo dosaženo sečtením obou účelových funkcí v jednu. Tento princip pak byl okopírován a použit při tvorbě algoritmu označovaného zkratkou UPFC (Unsupervised Possibilistic Fuzzy Clustering) (Wu et al., 2010). Jeho účelová funkce je součtem účelových funkcí FCM a PCA

$$J_{UPFC}(\mathbf{X};\mathbf{U}_{FCM};\mathbf{U}_{PCA};\mathbf{C})=\sum_{i=1}^{c}\sum_{j=1}^{n}\left(au_{ij,FCM}^{m}+bu_{ij,PCA}^{h}\right)d_{ij}^{2}+\frac{\beta}{n^{2}\sqrt{c}}\sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij,PCA}^{h}\log u_{ij,PCA}^{h}-u_{ij,PCA}^{h}\right),$$

kde $\mathbf{X} = \{x_1, ..., x_n\}$, $\mathbf{X} \subset \mathbf{R}^{\dim}$, je množina objektů a $\mathbf{C} = \{C_1, ..., C_c\}$ je množina shluků, $\mathbf{U}_{FCM} = (u_{ij,FCM})$ označuje matici příslušností s podmínkami

$$\sum_{j=1}^{n}u_{ij,FCM}>0,\forall i\in\{1,...,c\},$$

$$\sum_{i=1}^{c}u_{ij,FCM}=1,\forall j\in\{1,...,n\},$$

$\mathbf{U}_{PCA} = (u_{ij,PCA})$ označuje matici příslušností s podmínkami

$$\sum_{j=1}^{n}u_{ij,PCA}>0,\forall i\in\{1,...,c\},$$

$$0<\sum_{i=1}^{c}u_{ij,PCA}\le c,\forall j\in\{1,...,n\},$$

$d_{ij} = \|\mathbf{x}_j - \mathbf{c}_i\|$ je zvolená metrika, kde $\mathbf{c}_i$ je vzorový objekt C_i, $m,h > 1$ jsou mlžitele a β je kladný parametr, který autoři definují jako kovarianci datového souboru

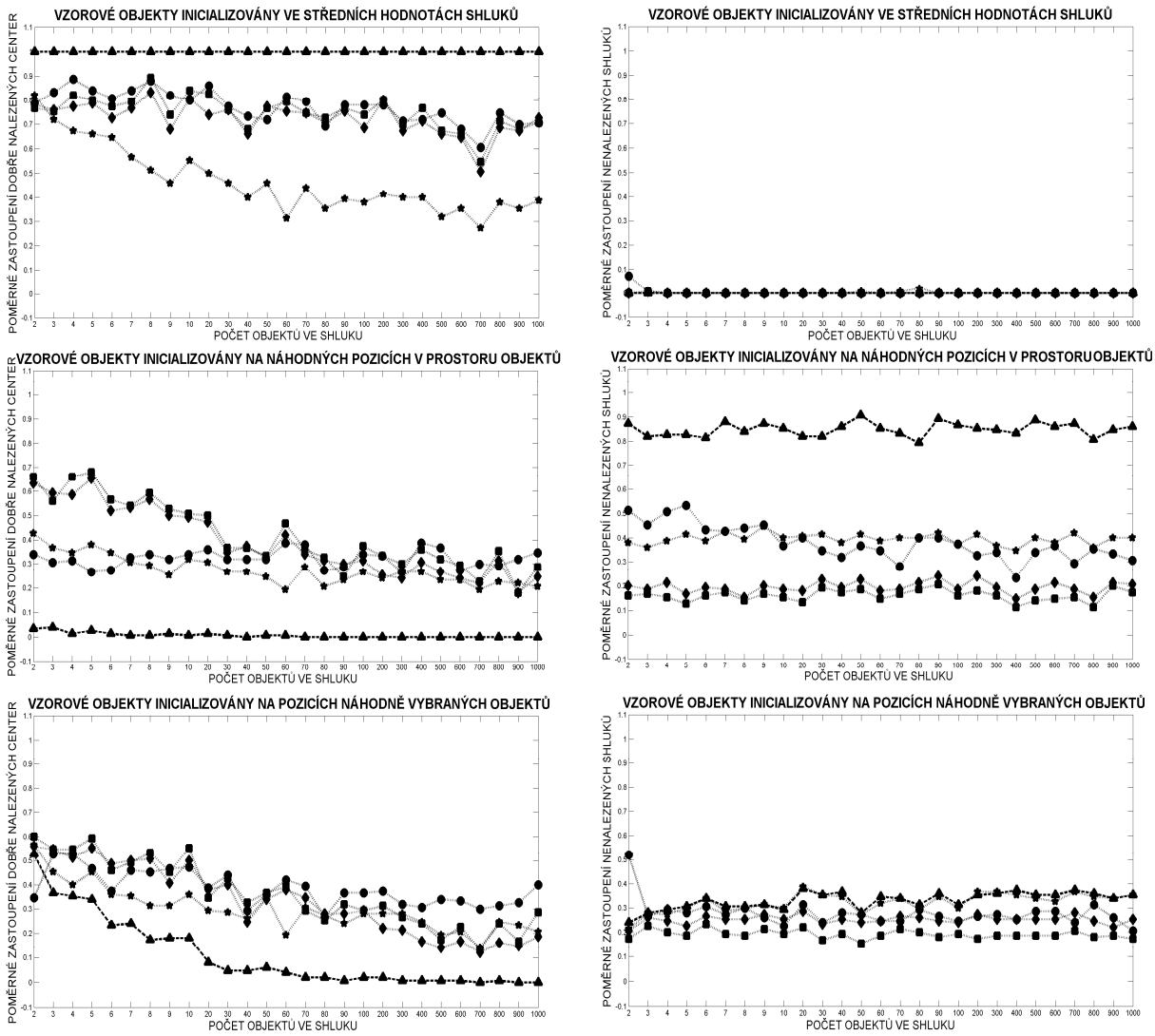
$$\beta=\frac{1}{n}\sum_{j=1}^{n}\left\|x_j-\bar{x}\right\|^{2},$$

kde $\bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j$.

Algoritmus by měl mít vlastnosti FCM i PCM a výpočetní stabilitu PCA.

Experimenty

Experimenty byly prováděny na generovaných datových souborech. Každý objekt datového souboru je reprezentován číselným vektorem příslušné dimenze. Datový soubor je generován tak, že jsou nejprve v oblasti (hyper)krychle o straně 1 se středem v počátku soustavy souřadnic vygenerovány rovnoměrným rozdělením vektory, které jsou dále chápány jako střední hodnoty normálního rozdělení, které následně vygeneruje zadaný počet bodů v okolí každé střední hodnoty, čímž vytvoří shluky. Střední hodnotu normálního rozdělení, které tímto způsobem vygenerovalo shluk v datovém souboru, budu pro jednoduchost označovat *střední hodnota shluku*. V základním nastavení pro generování datového souboru je $dim = 2$ (počet proměnných každého objektu), počet shluků $c = 150$, směrodatná odchylka každého shluku $\sigma = 0{,}00\overline{6}$, poměr přidaného šumu vzhledem k počtu objektů v datovém souboru 0 % a počet bodů v každém shluku 150. V každém testu je jeden parametr nastavení položen jako proměnná, pro každou hodnotu proměnné je provedeno shlukování algoritmy HCM, FCM, PCA a UPFC. Každý z těchto algoritmů provádí shlukování do dosažení ukončovací podmínky, minimálně však musí provést 11 iterací, a pokud nedosáhne ukončovací podmínky do 100. iterace, je ukončen. Mlžitele u algoritmů, které je obsahují, jsou nastaveny na hodnotu 2, parametry *a* a *b* algoritmu UPFC jsou položeny rovny jedné.

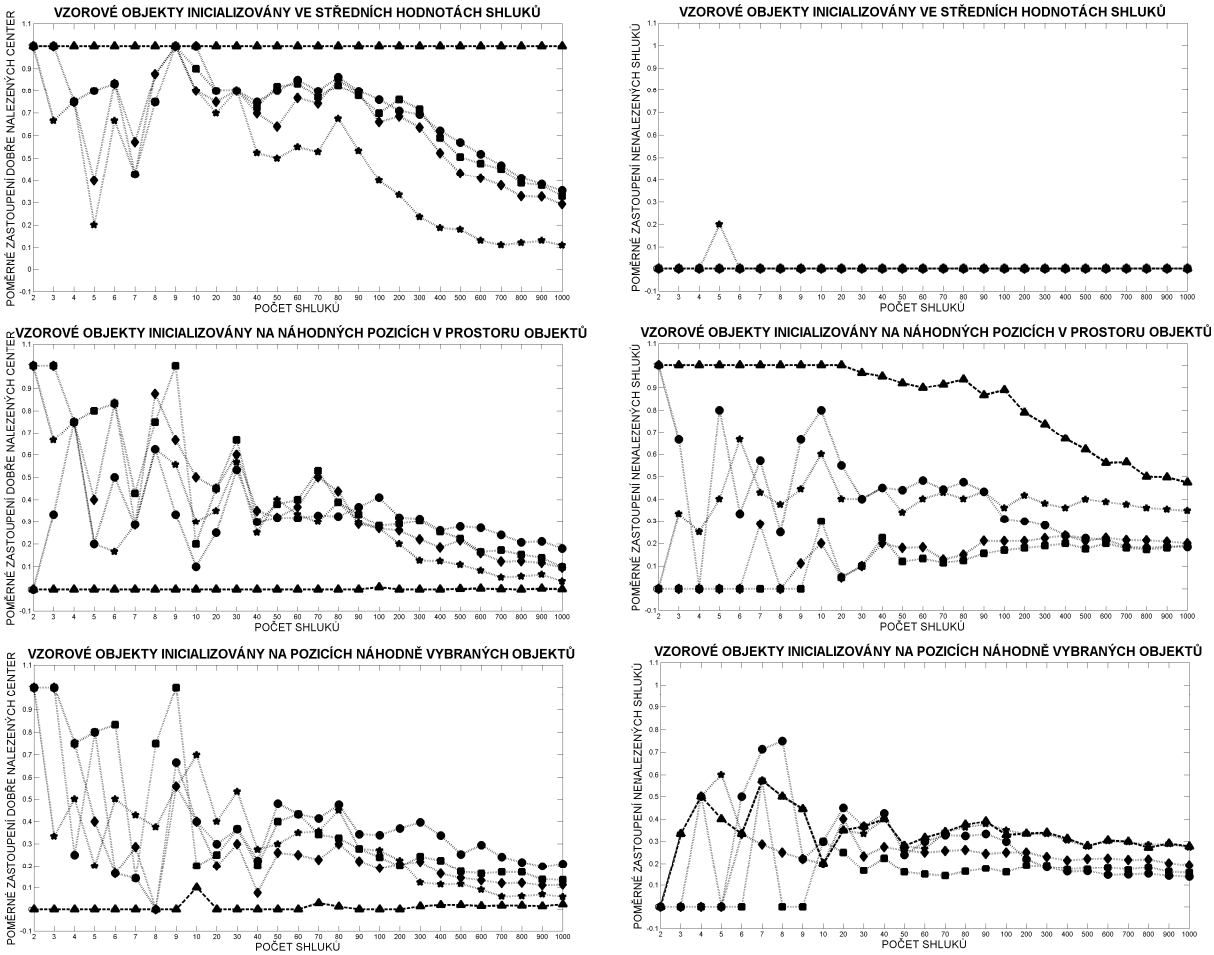


Obrázek 1: Test dopadu počtu objektů ve shluku na kvalitu shlukování. Levý sloupec grafů vyjadřuje závislost poměrného zastoupení dobře nalezených center na počtu bodů ve shluku. Pravý sloupec grafů vyjadřuje závislost poměrného zastoupení nenalezených shluků na počtu bodů ve shluku. První dvojice grafů shora jsou výsledky při inicializaci ve středních hodnotách shluků, druhá dvojice jsou výsledky při inicializaci v náhodných pozicích datového souboru a třetí dvojice je inicializace na pozicích náhodně vybraných objektů. Trojúhelníky značí hodnoty inicializací, kruhy výsledky HCM, čtverce výsledky FCM, pěticípé hvězdy výsledky PCA a kosočtverce výsledky UPFC.

Ve všech algoritmech se používá pro výpočet vzdálenosti mezi objekty euklidovská metrika. Každý algoritmus je při jednom nastavení datového souboru spuštěn třikrát, pokaždé s principiálně odlišnou inicializací. Tyto inicializace jsou shodně nastaveny pro všechny testované algoritmy při daném nastavení parametrů datového souboru. První inicializace vzorových objektů je na pozicích středních hodnot shluků. Druhá inicializace je na náhodných pozicích uvnitř krychle, ve které jsou rozmístěny střední hodnoty shluků, a třetí inicializace je na pozicích náhodně vybraných objektů datového souboru (pojem *pozice objektu* užívám ve smyslu pozice bodu v euklidovském prostoru se souřadnicemi odpovídajícími hodnotám atributů příslušného objektu vzhledem k počátku kartézské soustavy souřadnic).

Kvalita shlukování je vyhodnocována pomocí dvou kritérií. Pokud střední hodnota shluku leží v 95% intervalu spolehlivosti vzorového objektu, který má po ukončení algoritmu reprezentovat průměrnou hodnotu objektů ve shluku, je tento vzorový objekt označen jako *dobře nalezené centrum*. Shluky, u nichž se nevyskytuje vzorový objekt, který nebyl přiřazen jinému shluku, ve vzdálenosti menší než 3σ, je označen jako *nenalezený shluk*. Výsledky experimentů jsou pak zaznamenány v sériích grafů.

V prvním experimentu je proměnnou počet bodů ve shluku. Z výsledků (obrázek 1) lze vyčíst, že počet bodů ve shluku nemá vliv na kvalitu shlukování.



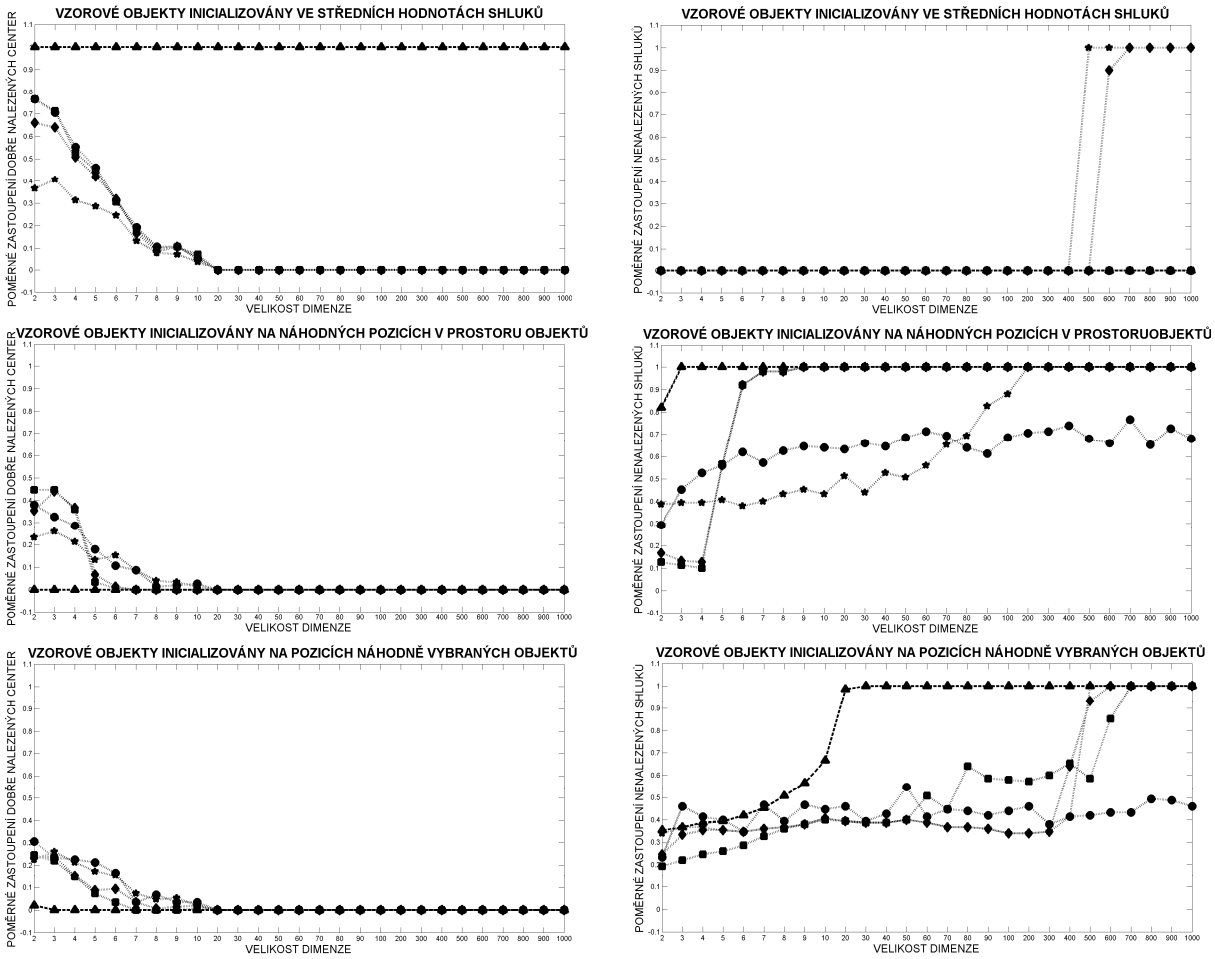
Obrázek 2: Test dopadu počtu shluků na kvalitu shlukování. Levý sloupec grafů vyjadřuje závislost poměrného zastoupení dobře nalezených center na počtu shluků. Pravý sloupec grafů vyjadřuje závislost poměrného zastoupení nenalezených shluků na počtu shluků. První dvojice grafů shora jsou výsledky při inicializaci ve středních hodnotách shluků, druhá dvojice jsou výsledky při inicializaci v náhodných pozicích datového souboru a třetí dvojice je inicializace na pozicích náhodně vybraných objektů. Trojúhelníky značí hodnoty inicializací, kruhy výsledky HCM, čtverce výsledky FCM, pěticípé hvězdy výsledky PCA a kosočtverce výsledky UPFC.

V druhém experimentu je proměnnou počet shluků. Shledáváme zde (obrázek 2), že výsledek shlukování je při počtu shluků v řádu jednotek velmi nepředvídatelný. V řádu stovek (tisíců) shluků pak dochází k poklesu kvality shlukování. Domnívám se, že je tento jev způsobený působením „masy datového souboru“, kdy velké množství velmi malých vah významně ovlivňuje vážený průměr,

kterým se u neostrého shlukování odhaduje pozice centra shluku. Nicméně je z tohoto pohledu překvapivé, že HCM v tomto testu nevyniklo výrazněji.

Ve třetím experimentu je proměnnou dimenze objektů (obrázek 3). Testované algoritmy jsou zřetelně funkční při velmi nízké dimenzi úlohy. Nečekaným výsledkem je pak to, že při dimenzi nad 500 nejsou algoritmy neostrého shlukování schopny detekovat shluk. Výjimkou mezi testovanými je HCM. Zde bych si dovolil poznámku, že hodnocení shlukování při dimenzi 600 může být i zcestné vzhledem k tomu, že je sporné, zda jsou „počty“ vůbec stavěné na takový prostor, nebo spíš zda interpretace vzdáleností v takovém prostoru nepostrádá smysl.

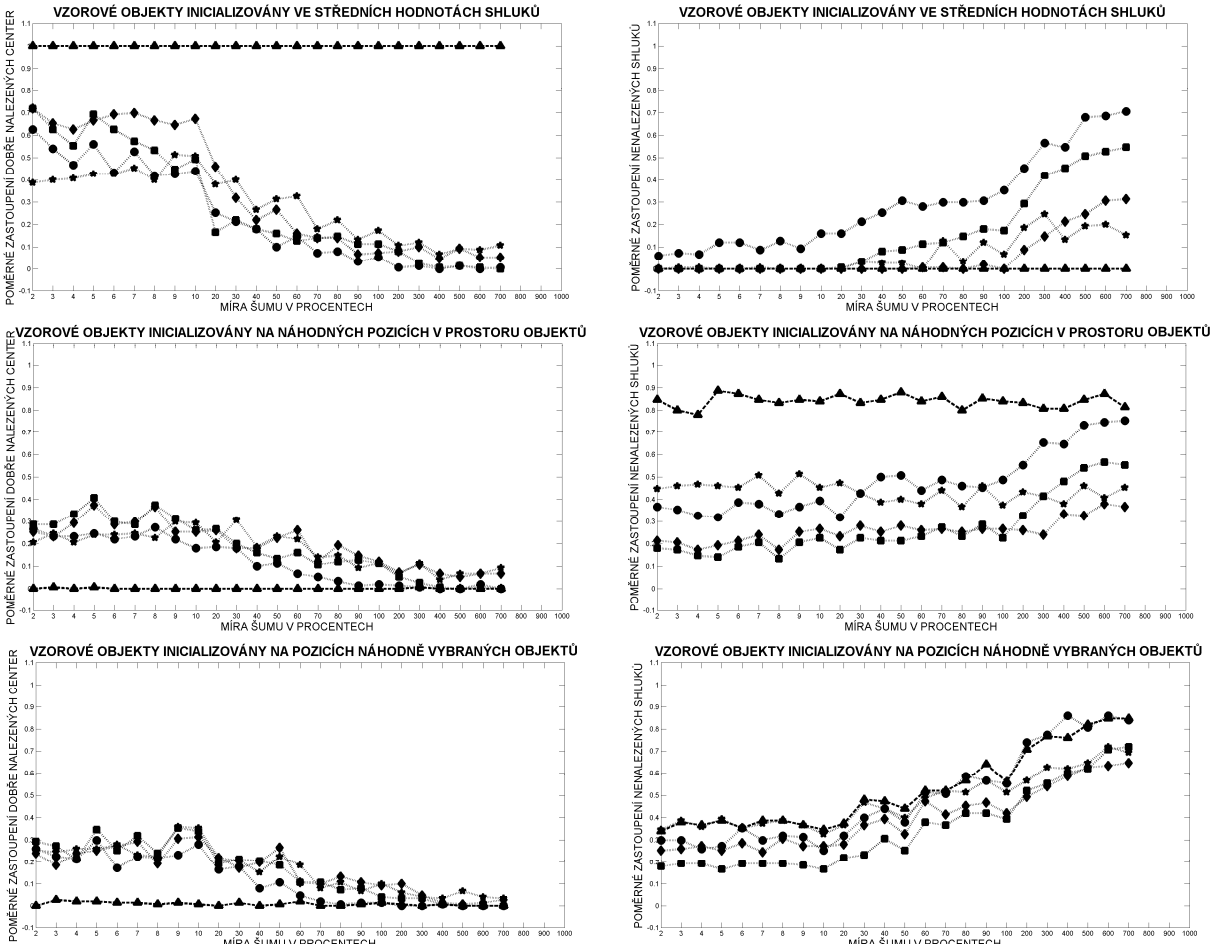
Ve čtvrtém experimentu (obrázek 4) je proměnnou poměrné zastoupení přidaného šumu (s rovnoměrným rozdělením). Algoritmy jsou funkční přibližně do deseti procent přidaného šumu. S vyšším zastoupením šumu pak kvalita shlukování pozvolně klesá.



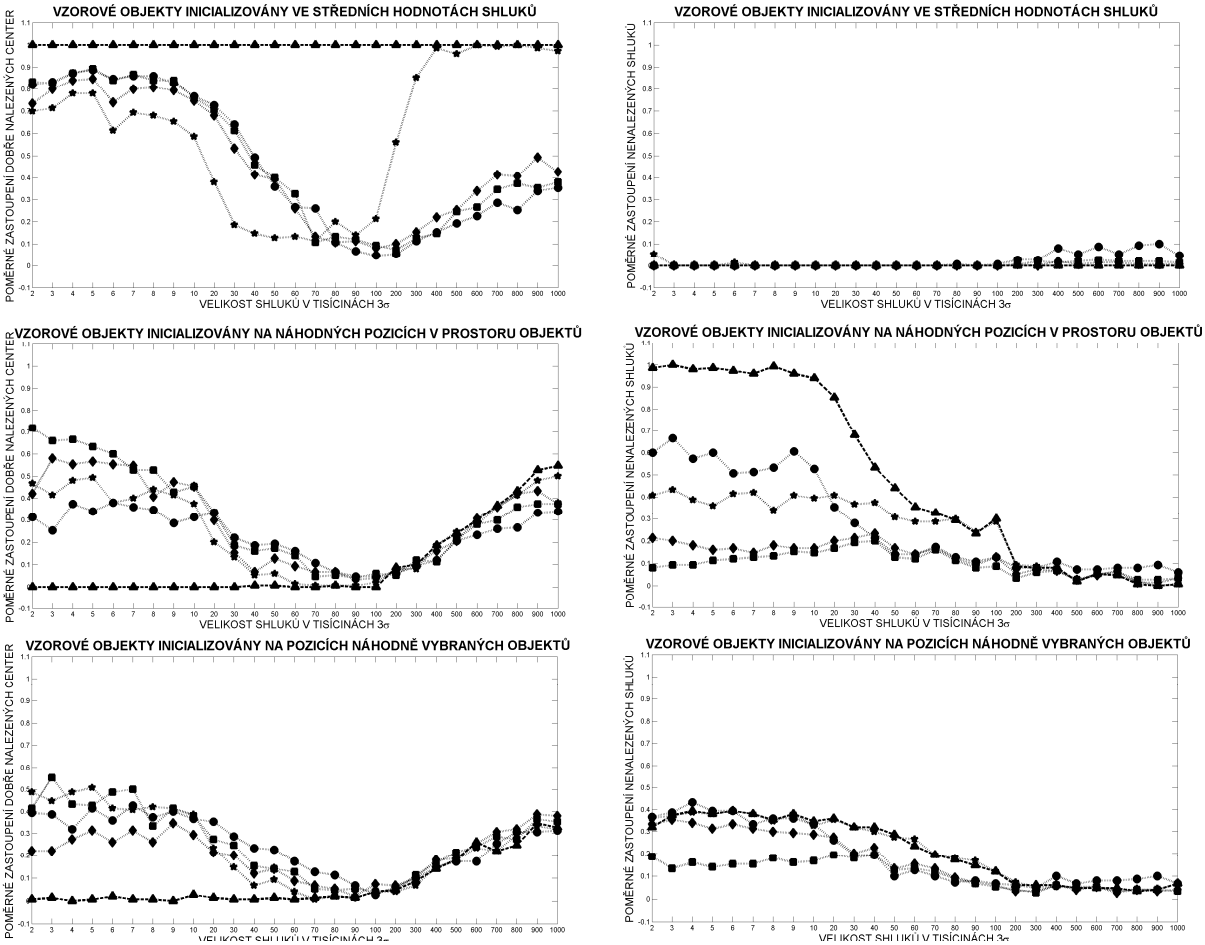
Obrázek 3: Test dopadu velikosti dimenze na kvalitu shlukování. Levý sloupec grafů vyjadřuje závislost poměrného zastoupení dobře nalezených center na velikosti dimenze. Pravý sloupec grafů vyjadřuje závislost poměrného zastoupení nenalezených shluků na velikosti dimenze. První dvojice grafů shora jsou výsledky při inicializaci ve středních hodnotách shluků, druhá dvojice jsou výsledky při inicializaci v náhodných pozicích datového souboru a třetí dvojice je inicializace na pozicích náhodně vybraných objektů. Trojúhelníky značí hodnoty inicializací, kruhy výsledky HCM, čtverce výsledky FCM, pěticípé hvězdy výsledky PCA a kosočtverce výsledky UPFC.

V posledním experimentu se postupně zvětšuje směrodatná odchylka rozdělení, které generuje shluky (obrázek 5). Kvalita shlukování je zřetelně závislá na míře překryvu mezi shluky. Zajímavé je ale chování PCA, které se výrazně lépe, je-li dobře inicializováno, dokáže udržet v blízkosti střední hodnoty shluku.

Srovnám-li jednotlivé algoritmy, musím konstatovat, že se od sebe kvalitou shlukování při mnou zvoleném nastavení příliš nelišily. Jako nejhorší se jeví PCA, které však bylo nejlépe schopné odhalit střední hodnoty shluků při výrazných překryvech. HCM zase bylo schopno jako jediné detekovat shluky při dimenzi nad 600. Na druhou stranu, s přidaným šumem se vypořádalo nejhůř. Zbylé dva algoritmy nevykazovaly v nastavených podmínkách extrémní chování v porovnání s ostatními.



Obrázek 4: Test dopadu množství šumu na kvalitu shlukování. Levý sloupec grafů vyjadřuje závislost poměrného zastoupení dobře nalezených center na množství šumu. Pravý sloupec grafů vyjadřuje závislost poměrného zastoupení nenalezených shluků na množství šumu. První dvojice grafů shora jsou výsledky při inicializaci ve středních hodnotách shluků, druhá dvojice jsou výsledky při inicializaci v náhodných pozicích datového souboru a třetí dvojice je inicializace na pozicích náhodně vybraných objektů. Trojúhelníky značí hodnoty inicializací, kruhy výsledky HCM, čtverce výsledky FCM, pěticípé hvězdy výsledky PCA a kosočtverce výsledky UPFC. Hodnoty horizontální osy jsou v procentech zastoupení šumu vzhledem k počtu objektů v nezašumělém datovém souboru.



Obrázek 5: Test dopadu velikosti shluků (a od toho odvozené množství překryvů shluků) na kvalitu shlukování. Levý sloupec grafů vyjadřuje závislost poměrného zastoupení dobře nalezených center na velikosti shluků. Pravý sloupec grafů vyjadřuje závislost poměrného zastoupení nenalezených shluků na velikosti shluků. První dvojice grafů shora jsou výsledky při inicializaci ve středních hodnotách shluků, druhá dvojice jsou výsledky při inicializaci v náhodných pozicích datového souboru a třetí dvojice je inicializace na pozicích náhodně vybraných objektů. Trojúhelníky značí hodnoty inicializací, kruhy výsledky HCM, čtverce výsledky FCM, pěticípé hvězdy výsledky PCA a kosočtverce výsledky UPFC. Hodnoty horizontální osy jsou tisíciny trojnásobků směrodatných odchylek rozdělení, která vygenerovala shluky.

Závěr

V tomto článku jsem si dal za cíl nalézt a popsat méně známé či neznámé vlastnosti shlukovacích algoritmů založených na metodě rozkladu. Pro testování jsem zvolil tři zástupce této třídy algoritmů, kteří zastupují hlavní myšlenkové proudy a jednoho, který je propojuje.

Výsledky testů ukazují, že kvalita shlukování této třídy algoritmů je významně nezávislá na počtu bodů ve shlucích. To nelze říci o počtu shluků. Při počtu shluků v řádu jednotek je kvalita shlukování nepředvídatelná, v řádu stovek shluků pak významně klesá. Významným problémem pro algoritmy je počet proměnných zkoumaných objektů. Kvalita shlukování už při poměrně nízké dimenzi prostoru objektů významně klesá, při dimenzích v řádu stovek algoritmy nedokáží rozeznat shluky ani v případě, že se inicializují v centrech shluků. V testech, kdy se do datového souboru přidával šum, se projevilo, že kvalita shlukování není ovlivněna do přibližně 10 % přidaného šumu. Zajímavé je, že se algoritmy neostrého shlukování, zvláště pak UPFC, dokázaly vypořádat i s šumem několikanásobně převyšujícím počet objektů datového souboru bez přidaného šumu. Výsledek testu zaměřeného na překryv shluků pak ukazuje zajímavou skutečnost, že se algoritmy dokáží vypořádat s datovým souborem, kde jsou shluky výrazně odděleny stejně, jako když mezi nimi dochází k významným překryvům.

Doufám, že výsledky těchto experimentů přispějí k porozumění chování shlukovacích algoritmů a k jejich dalšímu rozvoji. Tímto směrem bych rád směřoval svůj další výzkum.

Literatura

BEZDEK, J. C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers. ISBN: 0306406713.

KAYMAK, U. and M. SETNES, 2000. Extended fuzzy clustering algorithms. *ERIM report series Research in Management*. No. ERS-2000-51-LIS

KRISHNAPURAM, R. and J. M. KELLER, 1993. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*. Vol. 1, iss. 2, p. 98-110. ISSN: 1063-6706.

MACQUEEN, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. University of California Press, vol. 1, p. 281-297.

Miyamoto, S. et al., 2008. *Algorithms for fuzzy clustering*. Springer.

PAL, N. R. et al., 1997. A mixed c-means clustering model. In: *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems*, 1997. IEEE, vol. 1, p. 11-21. ISBN: 0-7803-3796-4.

PAL, N. R. et al., 2005. A possibilistic fuzzy c-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*. Vol. 13, iss. 4, p. 517-530. ISSN: 1063-6706.

WANG, Y. et al., 2009. A Selection Model for Optimal Fuzzy Clustering Algorithm and Number of Clusters Based on Competitive Comprehensive Fuzzy Evaluation. *IEEE Transactions on Fuzzy Systems.* Vol. 17, iss. 3, p. 568-577. ISSN: 1063-6706.

WU, X. et al., 2010. Unsupervised Possibilistic Fuzzy Clustering⋆. *Journal of Information & Computational Science*. Vol. 7, iss. 5, p, 1075-1080.

YANG, M. S. and K. L. WU, 2006. Unsupervised possibilistic clustering. *Pattern Recognition*. Vol. 39, iss. 1, p. 5-21.

JEL Classification: C38

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu IGA VSE F4/5/2011.

Poděkování: Autor děkuje své manželce Vandě Vintrové za mnohou pomoc s prostředím Microsoft Word a za nadstandardní konzultační činnost týkající se normy ČSN ISO 690 (01 0197) a své školitelce Haně Řezankové za rady v oblasti slohové a gramatické.

Summary

Fuzzy Clustering

The usage of the methods of the partitioning clustering has several known and often quoted limits. The article presents the tests of the robustness of the selected clustering algorithms (k-means, FCM, PCA, UPFC) for different number of the objects in a cluster, of the clusters, of the dimensions, for the different portion of the added noise and the different coincidence of the clusters.

Keywords: k-means, FCM, PCM, PCA, UPFC, partitioning clustering.

Approaches to Outlier Detection

Vanda Vintrová

xvilv03@vse.cz

Doktorandka oboru statistika

Školitelka: prof. Ing. Hana Řezanková, CSc., (rezanka@vse.cz)

Abstract: The aim of this article is to give a global overview on contemporary approaches to outlier detection and to describe their main advantages and shortcomings. The article also presents some selected algorithms as examples.

The article introduces main approaches to outlier detection and presents a possible classification of algorithms for outlier detection. We can distinguish labeling and scoring algorithms according to their output, whether the algorithm labels an object as an outlier or as a normal object or whether an outlier assigns a score to an object representing a degree of outlierness. Most labeling algorithms can be transformed on scoring algorithms. It is also possible to classify algorithms as local or global depending on whether they search an outlier relative to the whole dataset or only relative to a subset of objects. The problem of global approaches is that outliers influence the calculations what can falsify the results. On the other hand, the problem of local approaches is to choose a reference set correctly. Lastly this article offers a classification according to whether an algorithm makes an assumption about a model of a generating mechanism or if it takes into account only spatial proximities of objects. The latter group includes distance-based and density-based approaches that have been widely used recently.

The distance-based algorithms are very popular for their computational simplicity as they only calculate distances between objects. It captures global outliers, therefore it is applicable only on datasets which contain clusters with similar density, what is not a case for many real world datasets which have more complex structure. The density-based algorithms overcome this shortcoming. They compare a local density of an object with local densities of its neighbors and produce outlier scores.

Key words: outliers, distance-based algorithms, density-based algorithms, LOF.

Introduction

In real data files, outliers, as objects considerably inconsistent with other objects from the same dataset, can be present. Some statistical and data mining methods regard these outliers as a noise that should be identified and eliminated as they falsify the analysis. However, outliers can also contain useful information and therefore it is important to investigate outliers in detail.

The definitions of outliers usually differ according to a method of outlier identification. There are three following general definitions which are very often cited. An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs (Grubbs, 1969). An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism (Hawkins, 1980). A definition of Hawkins is intuitive. We suppose that normal data objects follow the same generating mechanism and therefore objects that deviate so much that are suspicious to be generated by another mechanism are labeled as outliers. An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data (Barnett and Lewis, 1994). The problem is that different approaches define „inconsistent" differently.

Outliers can occur because of human or instrument error, fraudulent behavior, deviations, changes or faults in systems. Outlier detection is a very important task in many fields, it can indicate abnormal conditions from which performance degradation may result, such as an aircraft rotation defect. Outlier identification is also very useful instrument for detection of fraudulent usage, which could be signified by sudden change in the usage pattern. Purchasing behavior of a credit card owner usually changes when a credit card is stolen. Abnormal buying patterns can signify a credit card abuse. It can also detect suspicious trades in the equity markets. Outlier detection is also helpful in medicine, where some symptoms or test results can signify potential health problems of a patient. Outlier identification

is also very important in databases, where abnormal values may indicate a measurement error. Removing such errors may be important for other data mining tasks. But we cannot forget that one person's noise could be another person's signal, and that an abnormal value can signify not only an error or also a fraudulent case or a system deviation.

Below I will describe main approaches to the outlier detection. My point is not to give a full list of existing algorithms but rather to give a global view on approaches to the outlier detection and to classify these algorithms into several groups. Different classifications of outlier algorithms are possible.

Approaches to outlier identification

We can distinguish three main classifications of approaches to outlier detection:

- global versus local outliers,
- labeling versus scoring algorithms, and
- model based versus proximity based approaches.

Global versus local outliers

The difference between global and local outliers is weather the outlierness of an object is determined with respect to the whole database (global outliers) or only with respect to some subset of objects (local outliers). A global approach to outliers assumes only one generating mechanism. A problem of algorithms of finding global outliers is that they take as a reference set all the objects of a database, including outliers, what can falsify a result. On the other hand local approaches do not make an assumption on a number of generating mechanisms, there can be several, as it takes as a reference set only a small subset of objects. A problem of this approach is to determine a correct subset of objects as a reference set.

Labeling versus scoring algorithms

A labeling algorithm strictly states if an object is an outlier or not, whereas a scoring algorithm assigns to an object a degree of being an outlier, an outlier score. The objects can be sorted by their scores and therefore scoring algorithms can be usually transformed on labeling algorithms by defining a threshold for labeling objects as outliers. Many scoring algorithms find top n outliers. A problem is to set n correctly, because if there are only few outliers in a dataset, an algorithm could label normal objects as outliers and oppositely. For most algorithms, we can say that global outlier detection algorithms produce binary labels and local outlier detection algorithms usually assign scores to objects.

Model based versus proximity based approaches

Model-based algorithms make an assumption about a model of a generating mechanism. If an object does not fit to this model, it is considered as an outlier. Model-based algorithms are statistical algorithms, depth-based algorithms and deviation-based algorithms. Proximity-based algorithms examine a spatial proximity of every object. Outliers are objects which proximity deviates from proximities of other objects. According to a way of spatial proximity calculation we distinguish distance-based algorithms, density-based algorithms and angle-based algorithms. The last one is a specific case, where angles between a considered object and all other objects are calculated.

Algorithms for outlier detection

Statistical algorithms

Statistical algorithms were the earliest approaches used for outlier detection. They are based on an assumption about a data distribution. Most of the earliest algorithms were applicable only for single dimensional datasets.

There are several statistical tests, which one to choose depends on the data distribution, whether the parameters of the distribution are known, type or number of outliers. However, these parameters generally cannot be easily determined, which makes these algorithms difficult to apply.

Statistical algorithms either assume a known underlying distribution of the dataset or at least they are based on statistical estimates of unknown distribution parameters. These algorithms label as outlier those objects that considerably deviate from the model assumptions. Statistical algorithms are usually labeling algorithms that generate global outliers.

The problem of these algorithms is that they are not suitable for high-dimensional datasets and for real-world data without prior knowledge of the underlying data distribution. Another problem is that statistical algorithms are not robust and therefore are very sensitive to outliers that entre the computations and consequently falsify the results

Depth-based algorithms

The second category is a depth-based approach (Ruts and Rousseeuw, 1996, Tukey, 1977), which is inspired by computer graphics, is based on a principal that outliers lie at the border of the data space. Every object is represented as a point in a *k*-dimensional space and is assigned a depth value. Objects with smaller depth value are detected as outliers. Such methods do not have the distribution fitting problem. However, for multidimensional spaces they may not be applicable. Depth-based algorithms are inefficient for large datasets for $k \geq 4$. Depth-based algorithms generate global labeled outliers, but can also generate scoring outliers for example by taking a depth as a scoring value.

Deviation-based algorithms

The idea of deviation-based algorithms (Arning et al., 1996) is similar to statistical approaches, but it is independent on the chosen distribution. Outliers are points which deviate from the general characteristic of the set of objects (either local group or the whole dataset). Deviation-based algorithms generate global labeled outliers.

Distance-based algorithms

Distance-based algorithms (DB algorithms) are closely related to statistical approach to outliers. The advantage of the DB approach in comparison with the statistical approach is that DB algorithms are often non-parametric, therefore do not assume an underlying generating model for the data. Distance-based algorithms compare distances of objects to their neighbors. Normal objects are supposed to lie in a dense region whereas outliers are objects in a sparse region, it means that an outlier is an object far away from its neighbor objects. It generalizes the notion of outliers provided by several discordance tests developed in statistics and enjoys better computational complexity than the depth-based approaches for larger values of *k*. It is probably an origin of developing outlier detection algorithms in the context of databases.

Density-based

Density-based algorithms compare a local density of an object with local densities of its neighbors. Density-based algorithms assume that a density around a normal object is similar to a density around its neighbor objects, whereas an outlier has a density different form its neighbor objects. Density-based algorithms compare a relative density around an object to a density of its neighbor points and consequently compute an outlier score.

Angle-based approaches

This approach was proposed by Kriegel et al. (2008). It uses distances between points in a vector space and also it compares the angles between pairs of distance vectors to other points and outliers. Unlike the previous methods, it is supposed to work well also for high-dimensional data.

Development of distance-based and density-based approaches

In recent years mainly distance-based and density-based approaches have been very popular, therefore I shall describe a development in these areas in more details.

Distance-based approach

Distance-based algorithms are very simple and therefore popular and widely used algorithms, it only calculates distances between objects.

The earliest approach introduced by (Knorr and Ng, 1998) uses two parameters, d and p, and does not rank the outliers.

Definition 1:

An object o from a dataset $\boldsymbol{T}$ is a DB(p; d)-outlier, if at least a fraction p of the objects in $\boldsymbol{T}$ lies greater than distance d from o.

The problem of this approach is finding optimal d and p so that outliers would be correctly detected with only a few false detections. The parameters d and p have to be defined in advance by user.

Some variants of distance-based algorithm were introduced. The optimized algorithm proposed by (Ramaswamy et al., 2000) is based on a distance of an object to its kth nearest neighbor, denoted d_k.

Definition 2:

Given n and k, an object o is an outlier if no more than $n - 1$ objects in a dataset $\boldsymbol{T}$ have a d_k higher than an object o.

Outliers are top n objects which distances to the kth nearest neighbor d_k are greatest. It is supposed that objects with higher d_k values lie in a sparse region and therefore are outliers whereas normal objects lie in a dense region and therefore have lower d_k values. The objects are ranked according to their distance to their kth nearest neighbor and then the top n objects are identified as outliers. Parameters k and n are defined by user in advance. A problem is that the user has to know how many outliers are in a dataset.

Another algorithm has been proposed by Angiulli and Pizzuti (2002).

Definition 3:

Given parameters k and n, an object o in a dataset $\boldsymbol{T}$ is the nth outlier with respect to k if there are exactly $n - 1$ objects p in a dataset $\boldsymbol{T}$ such that $w_k(p) > w_k(o)$.

It computes an outlier score $w_k(o)$ of an object o as a sum of distances from this object to its k nearest neighbors.

A problem of this approach is that it can only capture global outliers as it compares distances within the whole database. Therefore distance-based outliers are applicable only on datasets which contain clusters with similar density. However, this is not a case for many real world datasets which have more complex structure.

Density-based approaches

To overcome the shortcomings of a distance-based approach, a density-based approach has been proposed. Density-based algorithms compare a local density of an object with local densities of its neighbors. Together with distance-based these are the most established approaches to outlier detection in databases. Breunig et al. (2000) is the first to introduce the concept of local density outliers and a local outlier factor (LOF). It compares a local density of an object with local densities of its k-nearest neighbors. The local density is estimated by a specific distance at which a point can be reached from its neighbors, called reachability distance, what produces more stable results within clusters. LOF value of approximately 1 indicates that the point is located in a region of homogenous density. Higher LOF values signify an outlier, LOF score is a degree of being an outlier, but the scaling is different for different datasets and even for different objects within one dataset the same score can signify a different degree of being an outlier, depending on different local data distribution.

An advantage of the LOF algorithm is that it can detect outliers even if the clusters of a dataset have different density and different size. This algorithm depends only on one parameter *k*; however this parameter strongly affects the outlierness of an object. If the parameter *k* is set too low, LOF does not detect outliers which are close to a dense cluster if the parameter is set too high, small clusters are regarded as outlier clusters.

The LOF algorithm was modified several times especially with an aim to speed up the algorithm. An improvement of LOF known as Connectivity Outlier Factor (COF) was proposed by Tang et al. (2002) to overcome an ineffectiveness of LOF in detecting outliers in sparse datasets.

Another modification of the LOF algorithm was proposed by Chiu and Fu (2003). They introduce LOF´, LOF´´ and GridLOF. LOF´ simplifies the formula of LOF for ease of understanding, LOF´´ distinguishes between a neighborhood for computing the density of an object and a neighborhood for comparing the densities of the neighbors of an object. The GridLOF utilizes grid-based method to prune objects that are not outliers and then compute LOF score.

Recently, a local distance-based outlier factor called LDOF was proposed by Zhang et al. (2009). LDOF is a ratio of *k* nearest neighbors' average distance to *k* nearest neighbors inner distance. The *k* nearest neighbors' inner distance is the average mutual distance of all pairs of the *k* nearest neighbors of an object *o*. A LDOF value higher than 1 signifies an outlier. This algorithm was proposed to overcome a problem of the LOF algorithm, which generates high LOF scores for cluster points, if the parameter *k* is set too high (higher than the number of cluster points), and therefore misses genuine outlier points. However, if an object *o* is between two dense clusters, it will not be classified as an outlier.

Another density-based algorithm was proposed by Papadimitriou et al. (2003) named Local Outlier Integral (LOCI) based on the idea of a multi-granularity deviation factor (MDEF). The difference between LOF and LOCI is that LOCI uses neighborhood instead of *k* nearest neighbors. LOCI is less sensitive to input parameters than LOF.

The outlier scores provided by various outlier algorithms differ widely in their scale, range and meaning. For most methods the outlier scores are not comparable from dataset to dataset, for many methods the outlier scores are not comparable even within one dataset. The same outlier score for one object means this object is an outlier and for another object (even within the same dataset, but from a different cluster) the same outlier score means that this object is not extraordinary.

To overcome this problem a new method has been proposed by Kriegel et al. (2011) to unify outlier scores provided by different outlier algorithms. They propose two types of operations, regularization and normalization. Regularization means that the score is transformed into a range [0; 1); score equals approximately 0 for inliers, higher values signify outliers. Normalization transforms scores into a range [0; 1]. These transformations do not change the ordering obtained by the original score. The contribution of this approach is not only unification of outlier scores, but also the fact that these operations can increase a contrast between outlier and inlier scores. A transformed outlier score is a rough probability, if an object is an outlier. Transformed outlier scores are therefore easier to understand and to interpret.

Conclusion

In this paper I have introduced several outlier detection algorithms. Different algorithms define an outlier differently, detect an outlier either as local or global, assign to an outlier either label or a score signifying a degree of being an outlier and therefore the results of different algorithms are difficult to compare. A shortcoming of these algorithms is that they are dependent on some parameter defined by a user in advance. There is any universally applicable algorithm. It is difficult to say which one is the best as every algorithm is designed for different dataset, to solve different problem of outlier detection. All the algorithms are supposed to produce fine results, but it is necessary to choose an algorithm carefully so it is appropriate for the specific dataset. A choice mainly depends on a fact whether a user knows a number of outliers, a size of clusters and whether the clusters are of the same size and density.

References

ANGIULLI, F. and C. PIZZUTI, 2002. Fast outlier detection in high dimensional spaces. In: *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*. London: Springer-Verlag, p. 15-26. ISBN: 3-540-44037-2.

ARNING, A., R. Agrawal and P. Raghavan, 1996. A linear method for deviation detection in large databases. In: *Proc. Int. Conf. On Knowledge Discovery and Data Mining* (KDD). Portland: AAAI, p. 164-169.

BARNETT, V. and T. LEWIS, 1994. *Outliers in statistical data*. 3rd edition. John Wiley & Sons. ISBN 0-471-93094-6.

BREUNIG, M. M., H. P. Kriegel, R. T. Ng and J. Sander, 2000. LOF: identifying density-based local outliers. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. New York: ACM, p. 93-104. ISBN: 1-58113-217-4

GRUBBS, F. E., 1969. Procedures for detecting outlying observations in samples. *Technometrics*. Vol. 11, iss. 1, p. 1-21.

HAWKINS, D., 1980. *Identification of outliers*. London, New York: Chapman and Hall.

CHIU, A. L. and A. W. FU, 2003. Enhancements on local outlier detection. In: *Proceedings of seventh International Database Engineering and Applications Symposium*. IEEE, p. 298-307. ISBN: 0-7695-1981-4.

KNORR, E. M. and R. T. NG, 1997. A unified approach for mining outliers. In: *Proceedings of the 1997 conference of the Centre for Advanced Studies on Collaborative research*. IBM Press, p. 11.

KRIEGEL, H. P., M. Schubert and A. Zimek, 2008. Angle-based outlier detection in high-dimensional data. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM: New York, p. 444-452. ISBN: 978-1-60558-193-4.

KRIEGEL, H. P., P. Kroger, S. Schubert and A. Zimek. *Interpreting and unifying outlier scores* [online]. [quoted 2012-01-25]. http://www.dbs.ifi.lmu.de/~zimek/publications/SDM2011/SDM11-outlier-preprint.pdf

PAPADIMITRIOU, S., H. Kitagawa, P. B. Gibbons and C. Faloutsos, 2003. LOCI: Fast outlier detection using the local correlation integral. In: *Proceedings of the 19th International Conference on Data Engineering.* IEEE, p. 315-326. ISBN: 0-7803-7665-X.

RAMASWAMY, S., R. Rastogi and K. Shim, 2000. Efficient algorithms for mining outliers from large data sets. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. New York: ACM, p. 427-438. ISBN: 1-58113-217-4.

RUTS, I. and P. J. ROUSSEEUW, 1996. Computing depth contours of bivariate point clouds. *Computational Statistics and Data Analysis*. Elsevier, vol. 23, p. 153-168.

TANG, J., Z. Chen, A. Fu and D. Cheung, 2002. Enhancing effectiveness of outlier detections for low density patterns. In: *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. London: Springer, p. 535-548. ISBN: 3-540-43704-5.

TUKEY, J. W., 1977. *Exploratory data analysis.* Addison-Wesley.

ZHANG, T., M. Hutter and H. Jin, 2009. A new local distance-based outlier detection approach for scattered real-world data. In: *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.* Berlin: Springer, p. 813-822. ISBN: 978-3-642-01306-5.

JEL Classification: C89.

Shrnutí

Přístupy k detekci odlehlých pozorování

Cílem článku je poskytnout celkový přehled přístupů k detekci odlehlých pozorování a popsat jejich hlavní výhody a nevýhody. Článek také představí některé vybrané algoritmy.

Článek představí hlavní přístupy k hledání odlehlých pozorování a také možné třídění algoritmů. Algoritmy můžeme třídit podle výstupu na klasifikační, tedy takové, které objekt označí jednoznačně jako odlehlé nebo normální pozorování, nebo na skórovací, které objektu přiřadí určité skóre označující stupeň odlehlosti objektu. Většina skórovacích algoritmů může být převedena na klasifikační. Dále můžeme algoritmy dělit na globální nebo lokální podle toho, zda hledají odlehlé pozorování vzhledem k celému souboru nebo pouze vzhledem k podmnožině objektů. Problém globálního přístupu je v tom, že odlehlá pozorování vstupují do výpočtů a ovlivňují je. Na druhou stranu, u lokálních odlehlých pozorování je problematické určit správnou referenční podmnožinu, ve vztahu ke které bude posuzována odlehlost objektů. Nakonec článek ukáže také dělení podle toho, zda algoritmus předpokládá určitý model generujícího mechanismu, nebo zda bere v úvahu pouze blízkost objektů v prostoru. Článek popisuje vývoj přístupů založených na vzdálenosti a na hustotě a uvádí základní algoritmy z těchto oblastí, jelikož tyto dva přístupy jsou v posledních desetiletích velmi oblíbené a často aplikované. Poslední jmenovaná skupina zahrnuje dva v poslední době velmi často používané přístupy, a to přístup založený na vzdálenostech a přístup založený na hustotě.

Algoritmy založené na vzdálenostech jsou velmi oblíbené pro svou výpočetní jednoduchost, neboť počítají pouze vzájemné vzdálenosti mezi objekty. Jejich nevýhodou je, že nacházejí pouze globální odlehlá pozorování, jsou tedy využitelné pouze v datasetech, kde mají shluky podobnou hustotu. Tento problém se snaží řešit algoritmy založené na hustotě, které srovnávají hustotu okolí objektu s hustotou okolí jeho sousedů a které přiřazují objektům skóre.

Klíčová slova: odlehlá pozorování, algoritmy založené na vzdálenosti, algoritmy založené na hustotě, LOF.