SBORNÍK

prací účastníků vědecké konference doktorského studia
Fakulta informatiky a statistiky
Vysoké školy ekonomické v Praze

Vědecký seminář se uskutečnil dne 8. února 2018
pod záštitou děkana FIS doc. RNDr. Luboše Marka, CSc.

Sestavení sborníku
prof. Ing. Petr Doucek, CSc.
proděkan pro vědu a výzkum

Nakladatelství Oeconomica – Praha 2018

ISBN 978-80-245-2259-3

OBSAH

Předmluva

Dne 8. února roku 2018 se konal seminář „Den doktorandů“ Fakulty informatiky a statistiky VŠE v Praze, který patří mezi tradiční akce, na nichž se setkávají doktorandi se svými školiteli. Seminář slouží jako přátelská, leč náročná platforma pro presentaci výsledků vědecké a odborné práce studentů všech doktorských oborů fakulty. Pro mnohé z doktorandů je to první vystoupení před odbornou veřejností, na němž získávají cenné zkušenosti a zpětnou vazbu ke své vědecké práci. Zde mají také příležitost si vytříbit své schopnosti formulovat srozumitelně a jasně své názory a hypotézy spolu s uplatňováním argumentů na jejich podporu a obhajobu. Nedílnou součástí jejich vystoupení je presentace závěrů výzkumné práce a diskuse nad jejími výsledky. Vlastní příprava vystoupení umožní doktorandům zvykat si na fakt, že výsledky své často dlouhodobé práce musí stěsnat do relativně krátkého časového úseku presentace, v níž musí jasně, srozumitelně, přehledně a poutavým způsobem přednést to, na čem pracovali.

Do dvacátého třetího ročníku „Dne doktorandů“ se přihlásilo celkem šestnáct účastníků. Z toho na oboru „Aplikovaná informatika“ to byly čtyři příspěvky, na oboru „Ekonometrie a operační výzkum“ také čtyři příspěvky a na oboru „Statistika“ šest příspěvků. Semináře se zúčastnili studenti obou forem studia, jak presenční, tak i kombinované.

Za práci v komisích bych rád poděkoval všem jejím členům, kteří pracovali pod vedením předsedů příslušných komisi. Na oboru „Aplikovaná informatika“, který je součástí i stejnojmenného studijního oboru, byl předsedou jako již v několika uplynulých ročnících doc. Ing. Vojtěch Svátek, Dr. (KIZI) a členové komise byli v tomto roce prof. Ing. Zdeněk Molnár, CSc. (KIT) a doc. Ing. Vlasta Střížová, CSc. (KSA). Pro studijní program „Kvantitativní metody v ekonomice“, studijní obor „Ekonometrie a operační výzkum“ pracoval komise ve složení předseda - prof. Ing. Petr Fiala, CSc. (KEKO), členy jeho komise pak byli prof. Ing. Roman Hušek, CSc. (KEKO) a prof. RNDr. Ing. Michal Černý, Ph.D. (KEKO). Pro obor statistika byla předsedkyní komise paní prof. Ing. Hana Řezanková, CSc. (KSTP), a dalšími členy byly doc. Ing. Iva Pecáková, CSc. a doc. Ing. Diana Bílková, CSc.

Komise pečlivě sledovaly předvedené výkony a stanovily následující pořadí nejlepších:

Studijní program – Aplikovaná informatika
obor Aplikovaná Informatika
1. místo: Ing. Filip Vencovský: E-Service Quality Diagnosis
2. místo: Ing. Daniel Vodňanský: Možnosti vizualizace metrik dat
3. místo: Ing. Václav Zeman: Dolování pravidel z RDF grafů

Studijní program – Kvantitativní metody v ekonomice
obor Ekonometrie a operační výzkum
1. místo: Mgr. Vladimír Holý: Why Use High-Frequency Data in Quantitative Finance?
2. místo: Ing. MSc. Petra Tomanová: Time-varying quantiles and their applications
3. místo: RNDr. Petr Štourač: Vehicle Routing with Three Kinds of Carriers: A Computational Study

obor Statistika
1. místo: RNDr. Samuel Flimmel: Rozptyl odhadu PACF založenom na useknutí časového radu AR(p)
2. místo: Mgr. Michal Gerthofer: Frequency modelling within claim reserving using micro data
3. místo: Ing. Michaela Jirková: Analýza dopadů národních ekvivalenčních škál založených na různých přístupech na české ukazatele

Oceněným studentům blahopřeji a věřím, že získané zkušenosti uplatní ve své další práci, ať už vědecké nebo v praxi. Uznání také patří všem vědeckým a pedagogickým pracovníkům, školitelům doktorandů, kteří se „Dne doktorandů“ zúčastnili a kteří svým vedením a radami byli nápomocni při zpracování a presentaci příspěvků.

Zvláštní poděkování pak patří studijní referentce doktorského studia paní Jitce Krajíčkové, díky níž byl seminář skvěle organizačně zajištěn, dále paní Petře Šarochové za administrativní podporu akce a Mgr. Lee Nedomové za práci při editaci a sestavení tohoto sborníku.

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

Kvalita a interní audit

Josef Muknšnábl

josef.muknsnabl@protonmail.com

Doktorand oboru Aplikovaná informatika

Školitel: Doc. Ing. Vlasta Svatá, CSc., (svata@vse.cz)

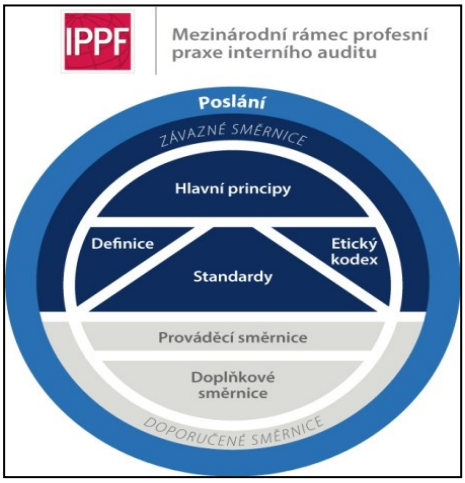
Abstrakt: Článek se zabývá kvalitou v interním auditu, z pohledu Mezinárodního rámce profesní praxe interního auditu a jeho součástí. Popisuje zejména, jak tento rámec pracuje s pojmem kvalita, jaký na kvalitu obecně klade důraz, jak pojem kvalita vykládá, kde všude vidí potencionální aplikace a jaká jsou pro tyto aplikace doporučení. Tato zjištění jsou dále využita při konstrukci jednotného hodnotícího rámce kvality vyvíjeného v rámci doktorské dizertační práce autora. Základem vyvíjeného rámce bude dotazníkové šetření prováděné elektronickým způsobem, postavené na hodnotícím dotazníku, jehož konstrukce probíhá. Článek zmiňuje způsob, jakým jsou zjištění ohledně kvality promítána do konstrukce dotazníku.

Klíčová slova: audit, interní audit, kvalita, kvalita v interním auditu, měření kvality, měření kvality v interním auditu, mezinárodní rámec pro profesní praxi interního auditu, mezinárodní standardy pro profesní praxi interního auditu, etický kodex, definice interního auditu, základní standardy, standardy pro výkon, prováděcí směrnice, doplňkové směrnice, metriky pro měření kvality, výkonnostní ukazatele kvality.

Úvod

S pojmem kvalita se můžeme setkat v běžném životě téměř na každém kroku v nespočtu situací. Přestože je to poměrně běžné slovo, není vždy zdaleka jasné, co se kvalitou přesně myslí a co ne. Je to dáno nejednoznačností samotného chápání pojmu kvalita, ke kterému dochází z různých důvodů. Například ve velké části aplikací platí, že kvalita není zpravidla jednoduchý jev, číslo, myšlenka nebo koncept, ale je to složitější, multidimenzionální pojem. Kvalitu také nebývá často z tohoto důvodu lehké prvoplánově měřit. Dále je důležitá také použitá úroveň abstrakce, tzn. o kvalitě můžeme hovořit jak ve zcela obecném/filozofickém smyslu, tak velmi konkrétně (např. o kvalitě výrobku, konceptu, modelu). I přes všechny tyto obtíže je ale nutné s pojmem kvality pracovat a nějak ji uchopit. Nejinak tomu je v interním auditu, pro který tvoří kvalita základní pilíř činnosti, protože bez kvality práce nemůže být důvěra a interní audit bez důvěry neboli bez vnímané kvality ze strany klíčových zákazníků, ztrácí takřka smysl a jeho činnost je významně znemožněna a zkomplikována. Pojďme se tedy podívat, jak je koncept kvality v činnosti interního auditu uchopen.

Činnost útvarů interního auditu je celosvětově upravena tzv. „Mezinárodním rámcem profesní praxe interního auditu“ (“ČIIA - Český institut interních auditorů z. s.,” n. d.). Vydává jej IIA (The Institute of Internal Auditors) v angličtině a je překládán do jazyků zemí, kde působí lokální pobočky IIA, v ČR je to ČIIA (Český Institut interních auditorů). Dlužno říci, že vzhledem k nedávným změnám v rámci profesní praxe (nová verze platná od 01. 01. 2017) a počtu dokumentů ještě není překlad zdaleka dokončen a rozsáhlé části zůstávají nepřeloženy. Z hlediska struktury se tento rámec dělí na několik částí, viz Obrázek 1.



Obrázek 1 – Mezinárodní rámec profesní praxe interního auditu (“ČIIA - Český institut interních auditorů z. s.,” n. d.)

Z pohledu tématu kvalita jsou nejdůležitější tzv. „Standardy“, „Doplňkové směrnice“ a „Prováděcí směrnice“. Pojďme se podívat blíže, jak se v těchto vybraných částech pracuje s kvalitou.

Kvalita z pohledu Mezinárodních standardů pro profesní praxi interního auditu

Mezinárodní standardy pro profesní praxi interního auditu (dále jen „Standardy“) jsou základní normou pro výkon interního auditu a tvoří jádro samotného profesního rámce. Soulad s těmito „Standardy“ je všeobecně považován ne pouze za jakousi volitelnou, dobrovolnou součást výkonu interního auditu ale za nutnost/povinnost, která je zcela nezbytná pro řádné naplnění odpovědností interního auditu nezávisle na zemi působnosti, sociálním, jazykovém, kulturním a právním prostředí, typu organizace, velikosti, oboru působnosti atd. „Standardy“ mají dvě hlavní části, a sice tzv. „Základní standardy“ a „Standardy pro výkon interního auditu“. Jak „Standardy“ uvádí „Základní standardy“ obsahují tzv. zásadní požadavky, a „Standardy pro výkon“ pak popisují charakter služeb a poskytují kvalitativní kritéria pro hodnocení činnosti útvarů interního auditu. Přestože by se z této definice mohlo zdát, že kvalita je tedy spíše záležitostí „Standardů pro výkon“, není tomu tak a s pojmem kvalita a její kritéria se pracuje běžně i v „Základních standardech“.

Kvalita a „Základní standardy“

Klíčovými pilíři, na nichž „stojí“ výkon interního auditu, jsou nezávislost a objektivita. Bez naplnění těchto dvou podmínek ztrácí výkon interního auditu jakýkoliv smysl. Je zajímavé, že „Základní standardy“ dávají, byť nepřímo, v interpretační části (část 1100 – Nezávislost a objektivita/Interpretace) kvalitu do souvislosti s objektivitou, a definují, že objektivita je „*nezaujatý myšlenkový postoj“, který mimo jiné interním auditorům „zajišťuje důvěru ve výsledek jejich práce a zamezuje přijímání kompromisů ohledně její kvality.*“ (The Institute of Internal Auditors, 2016). Kompromisy ohledně kvality jsou tedy vnímány jako porušení objektivity, nezávisle na tom, jestli se jednalo o úmyslné či neúmyslné jednání.

Přímo a nepřímo (v interpretační části) se o kvalitě hovoří v „Základních standardech“ v celkem 7 částech z 18 (38,88 %). Konkrétně se jedná o části:

- 1300 – Program pro zabezpečení a zvyšování kvality
- 1310 – Požadavky kladené na Program pro zabezpečení a zvyšování kvality
- 1311 – Interní hodnocení
- 1312 – Externí hodnocení
- 1320 – Podávání zpráv o Programu pro zabezpečení a zvyšování kvality
- 1321 – Užívání výrazu „Je v souladu s Mezinárodními standardy pro profesní praxi interního auditu“
- 1322 – Informování týkající se nesouladu

„Základní standardy“ (The Institute of Internal Auditors, 2016) ve vztahu ke kvalitě definují tyto, níže uvedené povinnosti:

„*Vedoucí interního auditu musí vypracovat a pravidelně aktualizovat program pro zabezpečení a zvyšování kvality interního auditu, který zahrnuje všechna hlediska funkce interního auditu.*“ Program musí být navržen tak, aby „*umožnil hodnocení souladu činnosti interního auditu se Standardy a dále umožnil hodnocení, zda se interní auditoři řídí Etickým kodexem. Tento program také hodnotí účinnost a efektivnost činností interního auditu a identifikuje příležitosti ke zlepšení.*“. Oproti minulé verzi standardů platné do 31. 12. 2016 vypadlo hodnocení souladu s „Definicí interního auditu“. Dále se předpokládá se, že takový program bude dozorován ze strany orgánů společnosti (firmy, organizace veřejné správy) a „Základní standardy“ říkají, že „*Vedoucí interního auditu by měl podporovat orgány společnosti v jejich dohledu nad programem pro zabezpečení a zvyšování kvality. Program pro zabezpečení a zvyšování kvality musí zahrnovat jak interní, tak externí hodnocení.*“. Interní hodnocení by mělo být průběžné a mělo by být prováděno osobami „*v rámci společnosti, které mají dostatečné znalosti postupů interního auditu.*“. V rámci termínu „dostatečná znalost postupů interního auditu“ „Základní standardy“ rozumí to, že se „*vyžaduje alespoň porozumění všem prvkům Mezinárodního rámce profesní praxe interního auditu.*“ Externí hodnocení se musí provádět minimálně jednou za pět let, „*odborně způsobilým a nezávislým externím hodnotitelem nebo externím hodnoticím týmem. Vedoucí interního auditu musí s orgány společnosti projednat: formu a četnost*

externích hodnocení, odbornou způsobilost a nezávislost externího hodnotitele nebo hodnoticího týmu, včetně jakéhokoli možného střetu zájmů.“

„*Vedoucí interního auditu musí informovat vedení a orgány společnosti o výsledcích programu pro zabezpečení a zvyšování kvality. Předávané informace by měly zahrnovat: rozsah a frekvenci jak interních, tak externích hodnocení, informace o kvalifikaci a nezávislosti hodnotitele/ů nebo hodnoticího týmu, včetně možných střetů zájmů, závěry učiněné hodnotiteli a plány nápravných opatření.*“. Dle interpretační části, konkrétně forma, obsah a frekvence se stanovují „*po dohodě s vedením a orgány společnosti“, a „výsledky externích hodnocení a výsledky pravidelných interních hodnocení předávány po ukončení příslušného hodnocení. Výsledky průběžného sledování jsou sdělovány nejméně jednou ročně.“.*

„*Uvedení, že činnost interního auditu je v souladu s Mezinárodními standardy pro profesní praxi interního auditu, je možné pouze pokud to výsledky programu pro zabezpečení a zvyšování kvality podporují.*“

Jak je vidět z výše uvedeného povinnosti ve vztahu ke kvalitě jsou v „Základních standardech“ vymezeny spíše rámcově než detailně. Sumárně „pouze“ definují nutnost pro útvary interního auditu mít program pro zabezpečování a zvyšování kvality, zaměřený na všechny funkce, které IA ve firmě vykonává. Program by měl být dozorován ze strany vedení společností, mít jak interní (průběžnou a pravidelnou) a externí část. Interní hodnocení by měla provádět osoba/instituce s dostatečnou znalostí postupů IA bez dalšího vymezení. Výsledky sledování kvality by se měly předávat vedení společnosti a definovat a provádět nápravná opatření. Velmi obecné je také uvedení informace že činnost IA je v souladu s mezinárodními standardy, kdy je zde pouze jedna podmínka, a sice ta, že to výsledky programu pro zabezpečení a zvyšování kvality toto tvrzení podporují, což není dále nikterak specifikováno a je předmětem dalšího výkladu. Bohužel není ani naznačen návod či systém hodnocení případně řešení rozporných situací, které mohou nastat např., že externí hodnocení je výrazně jiné než interní hodnocení atd. Základní standardy tedy přistupují ke kvalitě pouze principiálně, z hlediska klíčových nástrojů a mechanismů bez detailu. Ty jsou potom rozpracovány až v tzv. Prováděcích směrnicích. Pojďme se podívat, jak jsou na tom dále standardy pro výkon.

Kvalita a „Standardy pro výkon“

Standardy pro výkon (The Institute of Internal Auditors, 2016) se zabývají pojmem kvalita celkem ve 4 částech z 32 (12,5 %). Konkrétně se jedná o části:

- 2070 – Externí poskytovatel služeb a organizační odpovědnost za provádění interního auditu
- 2340 – Dohled (supervize) nad prováděním zakázky
- 2420 – Kvalita zpráv
- 2430 – Užívání výrazu „Provedeno v souladu s Mezinárodními standardy pro profesní praxi interního auditu“

O kvalitě se zde dočteme zejména, následující:

„*Nad zakázkami musí být prováděn řádný dohled (supervize) tak, aby bylo zaručeno, že jsou splněny jejich cíle, je zajištěna jejich odpovídající kvalita a že kompetence zaměstnanců jsou dostatečné.*“

„Zprávy musí být přesné, objektivní, jasné, stručné, konstruktivní, úplné a včasné.“ Význam jednotlivých atributů se pak dále rozvádí v interpretační části, kde se doslova píše „*Přesné zprávy neobsahují chyby a zkreslení a věrným způsobem odpovídají zjištěným skutečnostem. Objektivní zprávy jsou nestranné, nezaujaté a nezkreslené a jsou výsledkem spravedlivého a vyváženého ohodnocení všech souvisejících skutečností a okolností. Jasné zprávy jsou snadno pochopitelné a logické, neobsahují nepotřebné technické výrazy a poskytují všechny významné a relevantní informace. Stručné zprávy jdou k podstatě věci a vyhýbají se nepotřebným podrobným popisům, přemíře detailů, nadbytečnosti informací a rozvláčnosti. Konstruktivní zprávy přinášejí klientovi a společnosti prospěch a zdokonalení tam, kde je to potřebné. Úplné zprávy nepostrádají nic, co by bylo nezbytné z hlediska cílové skupiny uživatelů, a obsahují všechny významné a související informace a pozorování nezbytná pro zdůvodnění doporučení a závěrů. Včasné zprávy jsou dobře*

načasované, odpovídajícím způsobem reagující na vznik nenadálých situací a to vzhledem k důležitosti zjištěného problému. Včasné zprávy umožňují vedení přijmout odpovídající nápravné opatření.“

„Označení, že zakázky jsou „Provedeny v souladu s Mezinárodními standardy pro profesní praxi interního auditu”, lze „*použít pouze tehdy, pokud je to podloženo výsledky programu pro zabezpečení a zvyšování kvality.*“

„Standardy pro výkon“ specifikují tu skutečnost, že nad konkrétními auditními zakázkami prováděnými útvarem interního auditu, musí být prováděn dohled. Dále specifikuje požadavky na auditní zprávy a naznačují, jaké mají vypadat výsledné auditní zprávy. Zdůrazněno je zde pět základních atributů včetně času.

Podobně jako v případě „Základních standardů“ tak i u „Standardů pro výkon“ kvalitativní požadavky definují spíše obecné pilíře, na nichž pojem kvalita stojí, než konkrétní, jasné návody včetně způsobu měření a vyhodnocování. Pro další detail je nutné se ponořit do „Doplňkových směrnic“ a „Prováděcích směrnic“.

Kvalita a „Doplňkové směrnice“

„Doplňkové směrnice“ se sestávají celkem z 54 separátních dokumentů věnujících se různým aspektům činnosti oddělení interního auditů a provádění auditů. Věnují se jak zcela obecným tématům, jako jsou například komunikace auditních zpráv auditovaným, interakce s managementem a správní radou, tak specifickým aspektům auditování vybraných témat, jako je například auditování interního kontrolního prostředí, auditování antikorupčních programů, auditování odměn managementu atd. Z pohledu kvality a měření výkonu interního auditu obsahují doplňkové směrnice dva klíčové dokumenty, a sice „Quality Assurance and Improvement Program (česky „Program pro zabezpečení a zvyšování kvality“)“ (The Institute of Internal Auditors, 2012) a „Measuring Internal Audit Effectiveness and Efficiency (česky Měření efektivity a účinnosti interního auditu)“ (The Institute of Internal Auditors, 2010).

Program pro zabezpečení a zvyšování kvality

Program pro zlepšení a zajištění kvality (anglicky Quality Assurance and Improvement Program) specifikuje detailněji, čemu je třeba při zajištění kvality a zlepšování věnovat pozornost. Z pohledu konkrétních kroků jednak rozvádí obecné principy specifikované ve „Standardech“ jednak přidává dílčí aspekty, na které je dobré se zaměřit na jednotlivých úrovních práce (tzn. od nejmenších konkrétních auditních zakázek až po externí hodnocení). V případě auditních zakázek zdůrazňuje (The Institute of Internal Auditors, 2010):

- Nutnost nastavení procesů, které zajistí adekvátní transformaci auditního plánu do konkrétních auditních zakázek.
- Zachování a následování klasického cyklu auditních zakázek tzn. plánování, zjišťování údajů, vytváření zprávy a komunikace včetně řádného sledování plnění odsouhlasených opatření.
- Zjišťování zákaznické zpětné vazby formou dotazníku, sumarizaci zkušeností získaných z jednotlivých auditních zakázek.

V případě úrovně celého oddělení/útvaru interního auditu zdůrazňuje (The Institute of Internal Auditors, 2010):

- Existenci písemných politik, instrukcí pro řízení činnosti celého útvaru a zajištění souladu s „Definicí interního auditu“, „Etického kodexu“ a „Standardy“[1]. Nutnost shody mezi činností útvaru auditu a činností všeobecně vymezenou v chartě interního auditu (Internal Audit Charter)
- Nutnost naplňování očekávání klíčových zákazníků interního auditu a přinášení přidané hodnoty
- Efektivní a účinné využívání zdrojů

[1] zde je nesoulad s novými standardy, které obecně soulad s „Definicí interního auditu“ již nevyžadují.

Z pohledu externího hodnocení kvality mluví (The Institute of Internal Auditors, 2010) o:

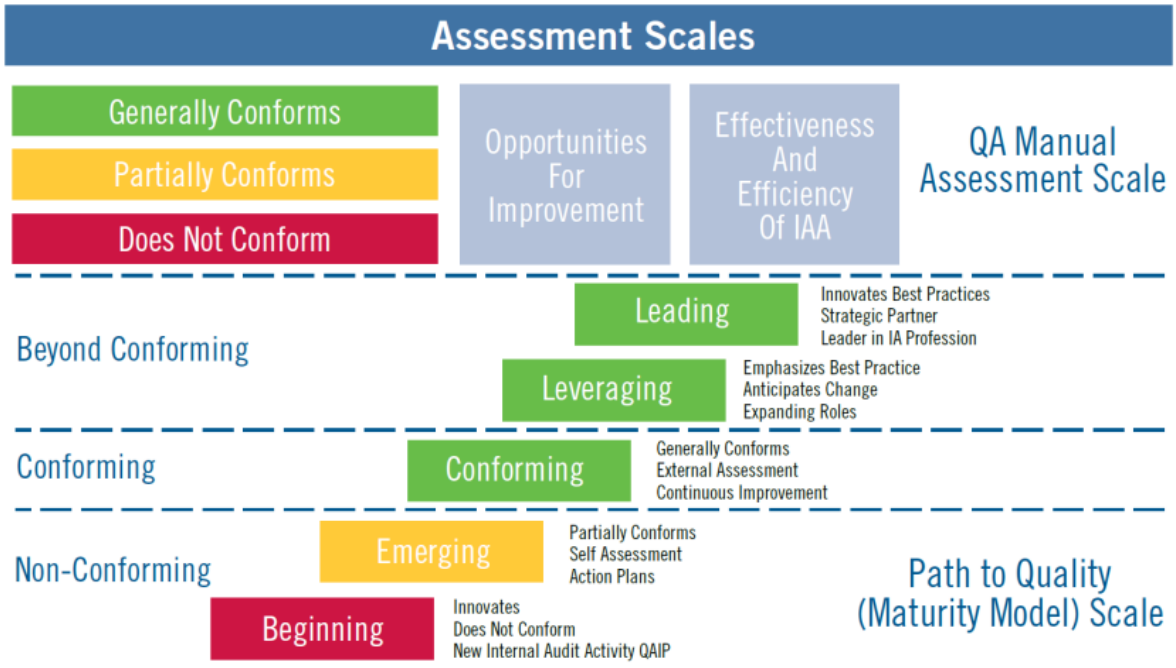
- šíři hodnocení, kdy externí hodnotitel hodnotí nejen auditní, ale i konzultační zakázky včetně hodnocení činnosti útvaru interního auditu ze strany klíčových zákazníků.

U konkrétních auditních zakázek dále hovoří o monitorování těchto zakázek pomocí kontrolních seznamů (např. ohledně požadovaných vyplněných auditních dokumentů), nebo za pomoci auditního software, který zaručuje automatizovaně správný průchod životním cyklem auditu (jehož nasazením se řeší existence a schvalování auditních dokumentů). Zmiňuje se také, že je dobré mít zadefinované metriky pro měření hodnoty poskytované zákazníkům, byť je nijak dále nespecifikuje. Požaduje také periodické revize materiálů, které jsou během auditu vyprodukovány a jejich souladu s Definicí interního auditu (dle starší verze „Standardů"), Etickým kódem, „Standardy". Tyto revize by měly být prováděny jinými zaměstnanci, než těmi, kdo pracovali na konkrétní zakázce. Požaduje také pravidelné provádění hloubkových rozhovorů a průzkumů s klíčovými zákazníky interního auditu a porovnání s nejlepší praxí. U externího hodnocení připouští dvě možné varianty hodnocení kvality, a sice buď prostřednictvím kvalifikovaného nezávislého hodnotitele, nebo formou sebehodnocení s nezávislou (externí) validací výsledků sebehodnocení a odkazuje na konkrétní dokumenty zabývající se touto problematikou.

Oproti „Standardům" ale program pro zlepšení a zajištění kvality již navrhuje (The Institute of Internal Auditors, 2010) použít hodnotící stupnici pro vyjádření míry souladu aktivit útvaru interního auditu se „Standardy". Konkrétně doporučuje čtyři vhodné:

- Třístupňovou stupnici z manuálu IIA pro hodnocení kvality (IIA Quality Assessment Manual Scale) s hodnoceními „Nevyhovuje", Částečně vyhovuje" a „Celkově vyhovuje" a dvěma podstupni v kategorii „Zcela vyhovuje".
- Pětistupňovou hodnotící stupnici IIA (The IIA's Assessment Scale – IIA Path to Quality) s hodnoceními "Úvodní", "Rozvíjející se", "Zavedený","Pokrokový" a „Pokročilý".
- Pětistupňový IIA model vyspělosti pro veřejný sektor (IIA Capability Model for the Public Sector): „Počáteční", „Základní", „Integrovaný", „Řízený", „Optimální"
- DIIR (Guideline for Conducting a Quality Assessment): 3-Uspokojivý/2-Prostor pro zlepšení/1-Je potřeba významné zlepšení / 0-Neaplikovatelný.

Porovnání prvních dvou viz přiložený obrázek číslo 2.



Obrázek 2 – Hodnotící stupnice (The Institute of Internal Auditors, 2012)

Měření efektivity a účinnosti interního auditu

S pojmem kvalita velmi úzce souvisí pojmy efektivita a účinnost. Tyto pojmy jsou v detailu rozpracovány v dokumentu s názvem „Measuring Internal Audit Effectiveness and Efficiency" (česky Měření efektivity a účinnosti interního auditu). Dle tohoto dokumentu by měl útvar interního auditu *„zavést výkonnostní metriky a související kritérium vhodné pro dané prostředí/organizaci pro měření stupně (včetně stupně kvality) dosažení cílů kvůli kterým byl útvar interního auditu zřízen. Efektivita a účinnost útvaru interního auditu by měly být monitorovány a vyhodnocovány pravidelně jakožto součást auditního procesu.*" (The Institute of Internal Auditors, 2010)

Pro zavedení takovýchto metrik je nutné především rozumět Mezinárodnímu rámci pro profesní praxi, zejména pak standardům, a identifikovat klíčové zákazníky auditu a jejich potřeby. Tyto metriky by měly být jak kvantitativní (např. počet realizovaných auditů vs. plán, průměrný počet doporučení na audit včetně závažnosti, procentuální vyjádření realizace nápravných opatření k danému dni, pokrytí auditního prostředí k danému dni atd.), tak kvalitativní (strukturovaná data, slovní vyjádření z dotazníků).

Kvalita a prováděcí směrnice

K již zmíněným částem „Standardů" existují tzv. prováděcí směrnice (Český institut interních auditorů. z. s., 2017), které dále upřesňují požadavky jednotlivých částí. U částí standardů zmíněných v kapitolách Kvalita a „Základní standardy" a Kvalita a „Standardy pro výkon" se jedná o následující vybrané zajímavé části a upřesnění:

- 1300 – Program pro zabezpečení a zvyšování kvality: zde se upozorňuje na to, že ohodnocení kvality neznamená explicitní ohodnocení každé dílčí auditní zakázky, ale že „*Zakázky by měly být spíše provedeny v souladu se zavedenou metodikou, která podporuje kvalitu, a v základním nastavení soulad se Standardy.*". Co se týká prokazování souladu, zmiňuje, že soulad může být „*prokázán prostřednictvím dokumentace dokončených pravidelných hodnocení, včetně rozsahu a plánu prověření, pracovní dokumentace a zpráv ze zakázek.*"
- 1311 – Interní hodnocení" definuje, že „*Průběžné sledování je realizováno v první řadě prostřednictvím průběžných činností, jako je například plánování zakázek a vykonávání dohledu nad nimi,*". Mezi možnými mechanismy průběžného sledování jsou již zmíněny konkrétní ukazatele (KPI) jako například „*poměr certifikovaných a všech interních auditorů v útvaru, počet let zkušeností v oblasti interního auditu, počet hodin průběžného profesního rozvoje získaných v průběhu roku, včasnost zakázek a spokojenost zainteresovaných subjektů.*" Dále zmiňuje projektové, rozpočtové a jiné ukazatele. Pravidelné sebehodnocení může být provedeno na základě vzorku realizovaných auditů. Sledované ukazatele (KPI) mohou zahrnovat např. „*porovnání rozpočtovaných a skutečných hodin strávených na zakázce, procento splnění auditního plánu, počet dní uplynulých mezi ukončením práce v terénu a vydáním zprávy, procento zavedených auditních pozorování a včasnost zavedení nápravných opatření souvisejících s auditními pozorováními*"
- 1312 – Externí hodnocení: může být po dohodě s managementem prováděno i častěji. Relevantní důvody mohou být např. vysoká fluktuace zaměstnanců nebo slučování dvou auditních oddělení.
- 1320 – Podávání zpráv o Programu pro zabezpečení a zvyšování kvality: říká že „*výsledky pravidelných interních hodnocení by měly být předány po ukončení příslušného hodnocení"* a dále že *„by měly být alespoň jednou ročně předány orgánům společnosti"*.
- 2070 – Externí poskytovatel služeb a organizační odpovědnost za provádění interního auditu „*Různé dokumenty mohou prokazovat soulad se Standardem 2070. V první řadě může být důkazem o odpovědnosti organizace týkající se udržování programu pro zabezpečení a zvyšování kvality interního auditu smlouva (tj. pověření k provedení zakázky) uzavřená mezi organizací a externím poskytovatelem služeb. Mezi dva základní výsledky těchto odpovědností patří dokumentovaný program pro zabezpečení a zvyšování kvality interního auditu a výsledky interních a externích hodnocení.*"
- 2340 – Dohled (supervize) nad prováděním zakázky: „*Důkaz souladu se Standardem 2340 může zahrnovat pracovní dokumentaci zakázky, ať už je parafovaná a datovaná osobou pro-*

vádějící dohled nad zakázkou (pokud je dokumentována ručně) nebo elektronicky schválena (pokud je zpracovaná softwarovým systémem). Dodatečné důkazy mohou zahrnovat vyplněný kontrolní seznam prověření pracovní dokumentace zakázky a/nebo sdělení obsahující komentáře k prověřované zakázce.“

Odraz požadavků na kvalitu při konstrukci jednotného hodnotícího rámce kvality

Přestože je vidět, že kvalitě, efektivnosti a účinnosti v interním auditu je v rámci Mezinárodního rámce pro profesní praxi interního auditu věnována nemalá pozornost, neexistuje do současnosti společný, jednotný a sdílený hodnotící rámec, skrze který by bylo možné nějakým nejlépe jednoduchým a ideálně automatizovaným a rychlým způsobem vyhodnotit kvalitu (úroveň) a výkon konkrétního útvaru interního auditu resp. vícero útvarů interních auditů a porovnat je mezi sebou.

Externí hodnocení činnosti útvarů interního auditu samozřejmě v praxi běžně probíhají, nicméně konkrétní postupy, použité metody, způsob vyhodnocení nejsou sjednoceny a liší se podle hodnotitelů, byť všichni vycházejí ze stejného základu. To velmi znesnadňuje vzájemné porovnávání výsledků a činí takové porovnání minimálně pracné, a zejména z hledisek opakovatelnosti a sledování trendů v čase napříč několika organizacemi prakticky nemožné. Zahrnutí velkého počtu organizací např. 30+ je z hlediska pracnosti v podstatě utopií.

Z tohoto důvodu jsem si toto téma vybral jako téma doktorské disertační práce. Hlavním cílem práce je sestavení vlastní metriky/rámce hodnocení kvality IA (vyspělosti, výkonu), která bude co nejjednodušší a ve svém provedení automatizovaná, flexibilní, programovatelná, rychlá, pokud možno univerzální (pro všechny typy organizací) a bude umožňovat kontinuální získávání dat. Téma to není úplně nové, v rámci komunity je i dlouhodobě diskutované, ale konkrétní výstupy z této diskuze nejsou, protože jde o návrh rámce, který by měl zjednodušit komplexní auditní prostředí a současně postihovat všechny regulatorní a další relevantní aspekty týkající se hodnocení kvality. Jde tedy o problém značně teoreticky i prakticky náročný.

Důkazem toho jsou neshody, které se týkají základní podoby takového jednotného hodnotícího rámce. Jak uvádí například (Arena and Azzone, 2009) existuje „*skupina autorů/výzkumníků, kteří vztahují efektivitu zejména ke kvalitě procedur interního auditu, jako je například soulad se standardy, schopnost plánovat, provádět audity a komunikovat auditní zjištění*“, jinými slovy vztahují kvalitu a výkon útvaru IA převážně ke schopnosti provádět audity v souladu s předdefinovanými postupy ze „Standardů“ a neohlíží se až tak moc na potřeby a přání zákazníka, inovace a další věci. Jiní autoři/výzkumníci zase vztahují kvalitu a efektivitu interního auditu jako „*schopnost interního auditu splnit očekávání auditovaných“* a „*zdůrazňují například výsledky zákaznického průzkumu spokojenosti a procento úspěšně implementovaných doporučení“*. Třetí skupina pak vztahuje kvalitu/efektivitu „*k výsledku auditu; např. když audit najde problémy a doporučení k těmto problémům jsou aplikována a problém odstraní*“ nebo „*celkově přispěje k výkonnosti společnosti a ukazatelům jako jsou zisk, cena akcie, růst“*. Poslední skupina pak spojuje kvalitu a výkonnost „*s pozitivním vlivem na kvalitu firemního řízení a monitorování a zlepšování procesů pro řízení rizik a interních kontrolních procesů*“.

Já se při konstrukci hodnotícího rámce nekloním výrazněji ani k jedné z těchto skupin. Domnívám, se že není možné se zaměřit pouze na dílčí aspekty a ostatní zcela ignorovat. Snažím se pojmout celou problematiku z hlediska všech typů aspektů. Tento přístup samozřejmě vede okamžitě k problému rozsahu měřených veličin (desítky, nižší stovky) pokud bych chtěl adresovat detailně všechny aspekty kvality uvedené ve standardech a reálné vyplnitelnosti takového dotazníku. S kolegy z ČIIA, se kterými konstrukci dotazníku konzultuji, máme aktuálně trochu rozpor v tom, co je ještě z hlediska vyplnitelnosti dotazníku, tzn. pracnosti pro auditované, únosné a co již ne.

Základní rozpracovaný přístup k dotazníku je zatím dvouúrovňový tzn. základní sada otázek identifikující klíčové věci ohledně organizace a shody výkonu auditu se standardy např. Jaké je hodnocení činnosti útvaru interního auditu ze strany Správní rady, zda existuje Výbor pro Audit, zda je přítomen Vedoucí auditu na Výboru pro audit, zda se realizuje pravidelný reporting, zda existuje program na zabezpečení a zvyšování kvality, zda jsou nadefinována KPI atd.

Druhá úroveň pak rozpracovává detail ve smyslu vybraných aspektů tak, jak o nich hovoří doplňkové a prováděcí směrnice a zaměřuje se i na konkrétní nejběžnější KPI, jako jsou např. plnění auditního plánu, plnění auditních opatření, počty a poměry auditních a konzultačních zakázek, průměrné zpoždění při implementaci opatření (ve dnech), počet auditorů, počet auditorů na 100 plných pracovních úvazků, počet provedených auditů na auditora za rok atd., hodnocení ze strany auditovaných per audit, zkušenost členů auditního týmu v letech, zkušenost členů auditního týmu per auditor atd.

Smyslem první úrovně je provést základní rozřazení auditních útvarů do základních kvalitativních úrovní ve smyslu objektivně změřené shody a vnímání ze strany organizace, např. kvalitní útvar ve smyslu shody se standardy /vysoké vnímání hodnoty ze strany organizace vs. kvalitní útvar ve smyslu shody se standardy /nízké vnímání hodnoty ze strany organizace atd. Smyslem druhé pak upřesnit pozici v rámci dané úrovně pro vzájemné porovnání s podobně kvalitními/výkonnými útvary. Ideálním výstupem druhé úrovně bude jedno, vypočtené číslo.

Sestavený dotazník bude ověřen a kalibrován na malém vzorku, detailněji ověřen s vybranými institucemi a porovnán s existujícími hodnoceními kvality. Poté vyzkoušen na větším vzorku a ideálně odladěn a nasazen do praxe.

Závěr

Vyhodnocení kvality a výkonnosti útvaru interního auditu není ani přes existence zakotvení konceptu kvality v rámci Mezinárodního rámce pro profesní praxi interního auditu jednoduché a dodnes (leden 2018) neexistuje shoda ani jednotný, společný hodnotící rámec. V rámci komunitních debat existuje několik teoretických přístupů. Práce na takovém rámci představuje výzvu, která se ukazuje být poměrně náročnou, mnohem náročnější než autor tohoto příspěvku předpokládal, nicméně o to více inspirující.

Literatura

ALMATARNEH, G.F., 2011. Factors determining the internal audit quality in banks: Empirical evidence from Jordan. *International Research Journal of Finance and Economics*. **73**, 110–119.

ANON., 2017. *ČIIA - Český institut interních auditorů z.s.* [online] [vid. 2018-01-29]. Dostupné z: http://www.interniaudit.cz/ippf/interaktivni-prehled.php?idKategorie=12

ARENA, Marika a Giovanni AZZONE, 2009. INTERNAL AUDIT EFFECTIVENESS: RELEVANT DRIVERS OF AUDITEES SATISFACTION.

ČESKÝ INSTITUT INTERNÍCH AUDITORŮ. Z.S., 2017. *Prováděcí směrnice* [online]. leden 2017. B.m.: Český institut interních auditorů, z.s. Dostupné z: https://na.theiia.org/translations/MemberDocuments/2017-Implementation-Guides-ALL-Czech.pdf

THE INSTITUTE OF INTERNAL AUDITORS, 2010. *Measuring Internal Audit Effectiveness and Efficiency* [online]. prosinec 2010. B.m.: The Institute of Internal Auditors. Dostupné z: https://na.theiia.org/standards-guidance/recommended-guidance/practice-guides/Pages/Measuring-Internal-Audit-Effectiveness-and-Efficiency-Practice-Guide.aspx

THE INSTITUTE OF INTERNAL AUDITORS, 2012. *Quality Assurance and Improvement Program* [online]. březen 2012. B.m.: The Institute of Internal Auditors. Dostupné z: https://na.theiia.org/standards-guidance/recommended-guidance/practice-guides/Pages/Quality-Assurance-and-Improvement-Program-Practice-Guide.aspx

THE INSTITUTE OF INTERNAL AUDITORS, 2016. *Mezinárodní standardy pro profesní praxi interního auditu (Standardy)* [online]. 2016. B.m.: Český institut interních auditorů, z.s. Dostupné z: http://www.interniaudit.cz/ippf-file/205CZ-77ab-5-revidovane-standardy-s-ucinnosti-od-1-1-2017-s-vyznacenymi-zmenami.pdf

THE INSTITUTE OF INTERNAL AUDITORS AUSTIN CHAPTER, 2009. *Performance Measures for Internal Audit Functions: A Research Project* [online]. únor 2009. B.m.: The Institute of Internal Auditors Austin Chapter. Dostupné z: https://www.google.cz/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwji0-W1w_7YAhVHbVAKHfWoCb8QFggsMAA&url=https%3A%2F%2Fglobal.theiia.org%2Fiiarf%2FPublic%25

20Documents%2FPerformance%2520Measures%2520for%2520Internal%2520Audit%2520Functions%2520-%2520A%2520Research%2520Project%2520-%2520Austin.pdf&usg=AOvVaw0Ke3JpogITdN7H4c16UIfc

THE INSTITUTE OF INTERNAL AUDITORS NETHERLANDS, 2016. *Measuring the Effectiveness of the Internal Audit Function*. červen 2016. B.m.: The Institute of Internal Auditors Netherlands.

WATKINS, A.L., W HILLISON a S.E. MORECROFT, 2005. Audit Quality: A Synthesis of Theory and Empirical Evidence. *Journal of Accounting Literature*. **23**.

JEL Classification: H83, M4, M42, M40, M49, L15

Summary

Název článku v angličtině Quality in Internal Audit

Evaluation of internal audit quality and performance is not despite existence of basic approaches within IPPF easy and there has not been up to these days (Jan 2018) a universal, common, shared evaluation framework in place. There have been debates held within internal audit community and there are numerous approaches for this problem solution but still work on the problem represents an issue that seems to be more demanding that expected but even more inspiring and fascinating.

Keywords: Audit, Internal Audit, quality, quality in internal audit, quality measurement, measurement of quality in internal audit, IPPF, International Professional Practice Framework, standards, implementation guidance, supplemental guidance, code of ethics, quality metrics, quality KPIs.

E-Service Quality Diagnosis

Filip Vencovský

filip.vencovsky@vse.cz

Ph.D. student of applied computer science

Supervisor: Prof. Ing. Jiří Voříšek, CSc., (vorisek@vse.cz)

Abstract: This study focuses on the issue of e-service quality diagnosis. E-services lack face-to-face communication which limits service feedback to on-line communication. On the other hand, it leads to a lot of user generated content. On-line reviews is a type of user generated content that consumers use to express their quality judgements. Although on-line reviews were explored in the literature, their interpretation regarding service quality is insufficient, particularly for full-text review body. The works that explored the review body focus on aggregated level of service quality and used statistical methods for quality diagnosis. This study, in contrast, dives deeper and explores expressed quality on the level of a review. For this purpose, qualitative research according to principles of grounded theory was used. Five research question was proposed regarding review body composition, language patterns, service aspects, sentiment expressions and the relationship between quality rating and review body.

Keywords: E-Service, Service Quality, On-line Review, Qualitative Research

Introduction

As the traditional economy transfers into the digital economy, the importance of e-services rises. The main purpose of *e-services* is to facilitate traditional services (Ladhari 2010). (Parasuraman, Zeithaml, Malhotra 2005) see e-service as the extent to which a web site facilitates efficient and effective shopping, purchasing and delivery.

One of the main tasks of service management is to deliver hi-quality services. *Service quality* is an abstract and elusive construct (Lepmets, Cater-Steel, Gacenga, Ras 2012). It is more than a result of technical attributes; it is a relationship with a service consumer (van Bon, Jong, Kolthof, Pieper, Rozemeijer, Tjassing, van der Veen, Verheijen 2007; Vargo, Lusch 2008). Due to that fact, the paper aims at service quality diagnosis from consumers' point of view.

On-line review is a common instrument for gathering data about e-service quality. It is a form of feedback where individuals express their opinions about products, services or organisations. *"Online reviews are a common form of socialised data representing spontaneously shared opinions by customers on review platforms"* (Mudambi, Schuff 2010).

On-line review usually takes form of a five-point Likert-type scale for quality rating and a textbox for a full-text quality review. However, it was found (Hu, Pavlou, Zhang 2006) that quality rating is an insufficient measure because of its bi-modal character. Moreover, the information that consumers feel weakly or strongly satisfied with a service is not enough to plan a meaningful service delivery and a future improvement of the service.

Traditional structured service quality surveys, such as (Li, Tan, Xie 2002; Rolland, Freeman 2010; El-Bayoumi 2012), are rigid and limit consumers in expressing themselves (Qu, Zhang, Li 2008). On-line reivews are less structured and harder to interpret, but contain more detailed information about quality from consumers' point of view.

Service science literature describes many techniques of service quality diagnosis, but only five papers use on-line reivew as a source of information about service quality (Duan, Cao, Yu, Levy 2013; Song, Lee, Yoon, Park 2015; Palese, Piccoli 2016; James, Calderon, Cook 2017; Vencovsky, Bruckner, Sperkova 2016). All the studies used an computer-aided analysis of full-text content to identify service aspects[1] and related sentiment[2], but they lack the systematic service quality interpretation.

1 A service aspect is a component or attribute of a service that customers see as crucial when assessing the service (Song, Lee, Yoon, Park 2015; Liu 2015).

Figure illustrates levels of consumer point of view – levels on which a service quality can be evaluated. The mentioned studies explored service quality on the aggregated level where all particular opinions were put together and categorised according to service quality aspects.

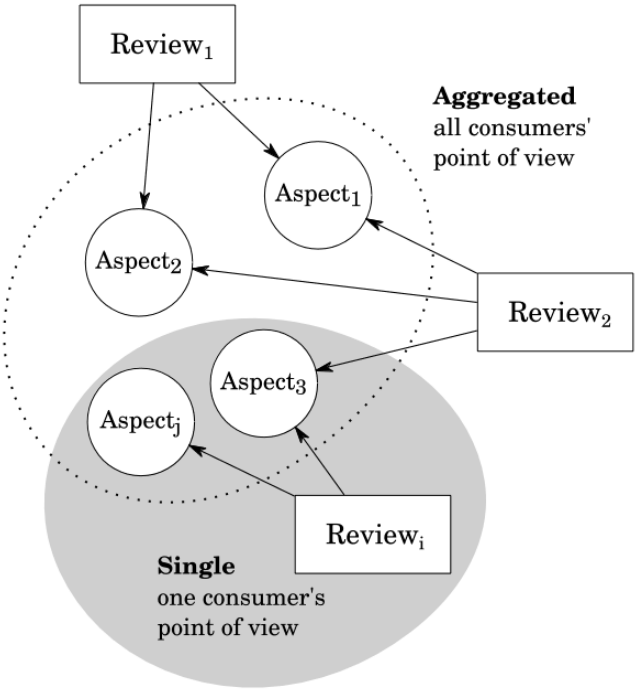


Figure 1: Levels of consumer point of view (author)

Aggregated level act as an indicator of aspects' performance. For example, a service provider needs to know whether communication with a consumer is perceived positively. However, this aggregated information is hard to join into a service quality model.

The most challenging part of the quality diagnosis is setting the right aspects weights and quality dimensions weights. Empirical studies (Gefen 2002; Wolfinbarger, Gilly 2003; Parasuraman, Zeithaml, Malhotra 2005; Yang, Fang 2004) that explored service dimensionality resulted in different quality dimension weights, but, which weights should be used and how to map them to a service if every service has different dimensions? Moreover, if the nature of service is perishable and the provision of service is an individual act, it is always different and dependent on circumstances. Even the same service artefacts may be perceived differently by the same consumer over time.

Because service is an individual act, it is necessary to explore quality on the level of a single quality review. Only this could lead to the right assumptions about quality aggregation. A reviewed literature reflects this effort as well. (Duan, Cao, Yu, Levy 2013) observed that different quality attributes collected from text have different influence weights on an overall rating. On the other hand, (Song, Lee, Yoon, Park 2015) used overall rating for setting individual weights of aspects for calculation of service quality.

Hence, it is essential to understand how quality is expressed in a review, what is the relationship between quality rating and sentiment of addressed service aspects and whether it is possible to identify aspect weights.

Because quality on the level of a single review needs to be explored, unlike the previously presented studies, the analysis in this study is not based on big data and statistical approaches. It focuses on discovery and description of related phenomena, rather than its quantification, and thus qualitative research was chosen.

---

2 Sentiment is the underlying positive or negative feeling implied by opinion (Liu 2015)□.

Method

Before addressing the issues, research questions should be set and suitable data have to be gathered and analysed. University information system was selected as a focal service in this case. Participants were also consumers of focal service. Data were gathered through a questionnaire in the common form of on-line quality review.

On-line service' quality of university information system was surveyed using two questions. The survey was captioned as *"UIS (University Information System) User Review"*. The first question was a five-point rating scale labelled *"Overall perceived quality"*. The first point was labelled *"Low"*; the fifth point was labelled *"High"*. No other description was provided because an implicit interpretation of online review with the same structure was suitable for the research. The scale was followed by an open-ended question labelled "*Full-text review"*. A description was included in the question, *"Please, write at least five sentences long review that reflects your subjective experience with the university information system."* Research data were then analysed using principles of grounded theory (Strauss, Corbin 1990) research. In specific terms, it means that:

- no hypothesis was set;
- a theory was built on data only;
- open coding of research data was used;
- research memos were written.

Five different categories of code were collected reflecting these five research questions:

(1) What type of statements do consumers use to build a review?
(2) What kind of language patterns do reviewers use?
(3) How do reviewers address service aspects?
(4) How do reviewers express sentiment?
(5) How does overall review rating relate to review body?

Because there are few perspectives that need to be considered, more than one open coding process was undertaken. Analyses considered seven particular coding steps: (1) sentence coding, (2) language pattern coding, (3) parsed dependencies coding, (4) aspect coding, (5) sentiment pattern coding, (6) sentiment coding, (7) comparison with an overall quality rating.

To enable the coding, each review needed to be divided into single clauses and classified by parser into word dependencies and parts-of-speech.

Results

The survey resulted in thirty responses collected from students of University of Economics, Prague. They had been asked to follow the description cited in the method section of this study. The students filled the survey online, but also were present in a classroom during the data collection phase. Because the quality survey was optional, not all of attending students submitted it.

The research results are limited given the fact that the survey was in English and the respondents were not native English speakers. Notably, the question on language patterns is difficult to generalise. On the other hand, service customers are not always native speakers, and natural language analysis must adapt to a specific vocabulary and language patterns used by customers.

In total, seventy-eight clauses were coded. Each clause was evaluated according to eight different perspectives. For example, the sentence *"I can really fast find some informations about my courses."* was processed as one clause as follows:

The sentence was coded (1) as a *quality statement*.

Language pattern of the statement was coded (2) as an *ability to do + activity aspect pattern + sentiment + intensity modifier*.

Moreover, every single parsed dependency was coded (3) separately in order to identify patterns of quality judgements[3]:

- VB_1-nsubj→PRP_1 (*actor*);
- VB_1-aux→MD (*ability*);
- VB_1-dobj→NN_1 (*core/aspect*);
- NN_1-nmod→NNS (*aspect*);
- VB_1-advmod→RB_1 (*sentiment/aspect*);
- RB_1-advmod→RB_2 (*sentiment intensity modifier*).

There were three aspects expressed in the single example statement coded (4) as follows:

- "find information", *2-word VB/NN Activity aspect*;
- "courses", *NNS Component aspect*;
- "fast", *RB Attribute aspect*.

The sentiment pattern was coded (5) as *RB over adverb modifier dependency plus RB as sentiment intensity modifier*.

The sentiment of the statement was coded (6) as *positive*, although the general English sentiment model from Stanford classified (7) the statement as *neutral*. Code *positive* was chosen because the attribute in the expression *"finding information"* is "*fast*", which positively influences a consumer regarding the purpose of the selected service.

Besides the presented codes, the expressed sentiment and addressed aspects were compared (8) with an overall quality rating on the five-point scale on the level of the whole quality review of a service consumer.

To enable coding, each review needed to be divided into single clauses and classified by a parser into word dependencies and parts-of-speech. A sentiment of clauses was also classified for comparison purpose.

Statement types

Open coding of statements resulted in twelve statement types. **The statement with the highest potential for quality diagnosis was *quality statement* and *emotion expression statement.*** The difference between these two statements was that no particular service attributes was mentioned, but only emotions was expressed towards service aspects. Another type that relates to quality was *quality comparison*. Close to quality comparison, but only stating that one aspect is more important than another, was *preference statement.* All quality statements were supported by *explanation statements*. Reviewers also described their *experience* with a service or *observation* of reality that surrounds a service. They also made *suggestions* which makes a great potential for service improvement. Also, *benefits* and *costs* of using service were stated. Reviews were also *summarised*, which offer an important view on consumer' perception of final quality judgements.

Regarding weights settings, *preference* and *structure statements* and *summary* are useful. Unfortunately, from no review in the sample was possible to collect enough information to set meaningful weighs.

Language patterns

The second question, *what kind of language patterns do reviewers use*, was answered by parsing sentences with Stanford English parser[4], that outputted word dependencies, and open coding of each dependency. They were coded according to its role in expressing quality. Only quality and emotion statements were analysed because they contain information about service aspects and quality. It was found that **aspects are often expressed with a sentiment in one dependency**. The dependency was,

3 All syntactic relations' names used in this thesis follow the taxonomy of Universal Dependencies v2. For more detailed information, see http://universaldependencies.org/u/dep/.

4 Neural netrwork parser (Chen, Manning 2014) was used with a standard Eglish model: edu/stanford/nlp/models/parser/nndep/english_UD.gz

regarding service quality, called *core dependency*. If aspect or sentiment was expressed in multiple words, dependency was coded as *aspect* or *sentiment dependency*. Sentiment was modified, in accordance with opinion mining literature, using two kinds of dependencies – *sentiment intensity modifier, sentiment orientation modifier*. Other dependencies that extended the core were *ability, actor, condition,* and *subjectivity modifier*.

Service aspects

Third question, *how do reviewers address service aspects*, was answered by coding expressed aspects and aspect expression patterns. It was observed that aspects could be seen, in keeping with opinion mining literature (Liu 2015), as *component* or *attribute*. Interesting was that aspects were also expressed as an *activity*. Activity aspects were mapped mostly to sub-services. The last observation was that aspects are often *implicit*. Implicit aspects were, in contrast to Liu's point of view, only pronouns. Language patterns that appeared in aspect expressions were identified.

Surprising was that clauses were usually coded with more than one aspect because **aspects appeared in multiple dimensions and were also often expressed together one expression**. For example, the phrase *"hard to find"* in the statement *"sometimes its hard to find information what I need"* refers to *Navigation* aspect. It also seems to be part of *Ease of use* aspect, because the sentence contains the word *"hard to"*, which is a constraint in using a system. Last part of the sentence refers to an information need of a consumer and can be seen as *Information support* aspect.

Aspects are often expressed together with sentiment in one expression. For example, in the statement *"system is chaotic"*, *system* is one clear opinion target and *chaotic* behaves as sentiment word, but it is also an expression of *User interface* aspect because it refers to an orderliness of a system to a user.

Without specific knowledge of a service background, such as service documentation, **an example model of service aspects was built from aspects expressed in reviews**. The modelling was done according to the same conditions consumers have when they write a quality review; it was based on the act of using a service, voice of customer and a certain level of IT knowledge and knowledge of reality in which the service is consumed. The model presented in figure Figure is subjective and cannot be generalised to e-service, nor this particular service. The model illustrates, how a consumer's mental model of service composition may be formalised.

The model uses a simplified ER notation. It contains entities and relationships. There is one central entity – *service*. All relationships are oriented in the direction to the central point. The direction implies how the quality of all components is composed.

Squares represent aspects expressed in reviews and coded as aspects. The aspects are at the lowest possible level of generalisation. Most of the aspects were left in the same form as in the reviews; attribute aspects that were converted into the noun form and generalised are the only exception. For example, *responsive* was modelled as *responsiveness,* but also *clumsy* was modelled as *responsiveness* because the reviewer referred to the ability of a website to accommodate to a mobile screen. **All attribute aspect expressions were modelled in an *attribute of* relationship.**

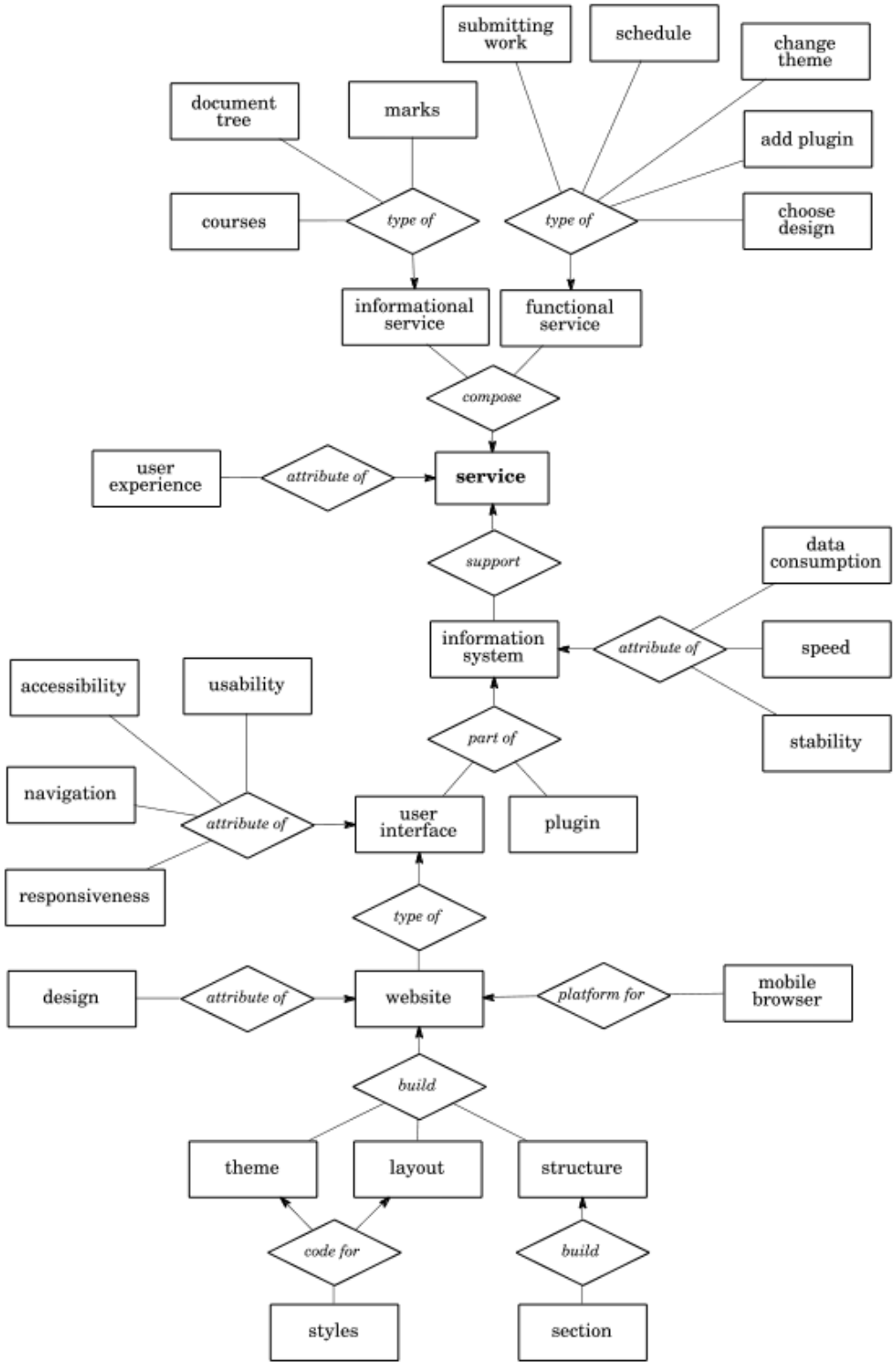


Figure 2: Model of a service aspects (author)

All of the activity aspect expressions were modelled as sub-services of two types: *informational services* and *functional services.* These two generalisations were not explicitly stated in reviews, but correspond to general statements about the ability to *work* or *do something* for functional services and *finding information* for informational services. The two types of sub-service are in line with two other types of service from the theoretical division of IT services: information, application, infrastructure and support (Voříšek, Basl 2008). Only the object aspect expressions were modelled in all relations.

Sentiment

The fourth question, *how do reviewers express sentiment,* was answered by coding expressed sentiment and sentiment expression patterns. As it was already mentioned, sentiment often appeared together with aspects in one expression. Sentiment was modified by adverbial modifiers in intensity and negation modifiers in orientation.

As already discussed, beside aspect, sentiment appears as the other part of a core of a quality statement, but can also be expressed in more words. Adjectives were the dominant part of speech for words coded as sentiment words. Nonetheless, adverbs, nouns and verb were also present. Following list contains all expressed parts of speech with affiliated dependency.

JJ/nsubj, JJ/ccomp, JJ/xcomp, JJ/csubj, JJ/dep, JJ/nsub, MD/advcl, NN/nmod, NN/nsubj, RB/advmod, RB/neg (attribute negation), VBG/nsubj, VBN/nsubjpass, VBN/csubjpass

The sentiment expressed by reviewers is strongly dependent on a specific context. For example, in the statement *"buttons are small"*, the sentiment was coded as negative. It is hard to guess what the reviewer expected to be the right buttons size. If the statement was followed by the explanatory statement *"it is hard to hit them", s*entiment of small would be clear. An adverb used as sentiment intensifier might be another possible clue. In the statement *"buttons are too small"*, it is obvious that the reviewer expected bigger buttons. For some aspects, readers can understand sentiment, regardless of the fact whether the word is accompanied by other words or not. For example, *"small screen"* would be probably perceived negatively, whereas *"small price"* might be seen positively.

Sentiment expressed in emotion statements is easier to analyse than in quality statements because it is less context dependent. Sentiment words coded in emotion expressions were: *satisfied, disappointed, like.* In contrast, sentiment words in quality statements were, for example: *complicated, unstable, small.*

Review rating and review body

The last question, *how overall review rating does relate to review body*, was answered by evaluation of the relationship between quality rating and count, sentiment and categories of aspects. Special attention was paid to quality expressed in summary statements. Based on the small sample of review, it was found that there is no clear relationship between quality rating and content of review body. Only summary statements were similar to quality rating, but there was also small contradictions.

Reviewers tend to use review rating as an overall quality metric. It is supported by observation of sentimental expressions in summary sentences such as *"Overall, I would have rated it as a weak average"* together with rating of two points from five, *"To summarize, I am perfectly satisfied with the services the university information system can provide us"* with full five stars rating or *"it is fine"* with three stars out of five. However, not every review contains such overall quality judgements.

According to the expectation disconfirmation theory (Parasuraman, Zeithaml, Berry 1988; Brown, Swartz 1989; Boulding, Kalra, Staelin, Zeithaml 1993; Gotlieb, Grewal, Brown 1994; Cronin Jr, Taylor 1994; Brown, Venkatesh, Goyal 2014; Brady, Cronin Jr 2001) , high overall rating means that a service quality exceeded expectation or met the expectation of high-quality service. It is an interesting assumption, but it can not be validated easily. In the analysed reviews **different strategies for scale rating was observed**. One reviewer rated service quality with five points and used the expression *"perfectly satisfied"* in a summary statement, whereas other reviewer used the expression *"I like"* with the same five-point rating. Possible explanations of the second case are: (a) The expression mirrors actual meeting of expectations of high-quality service. (b) The reviewer tends to use moderate language. (c) The reviewer tends to rate quality higher. However, these hypotheses can not be confirmed in this research.

The most remarkable observation was made regardin non-expressed aspects. Figure Figure illustrates a possible structure of the relationship between individual aspect groups. The figure is based on a review sample. Aspects mentioned in the review was generalised into more abstract groups. A possible influence of perceived quality is modelled between groups using arrows. The figure was

created for illustration despite it is obvious that these kinds of relationships cannot be certainly captured.

The example review was rated with four points from five possible. The review contained two summary statements: one emotion expression *"personally satisfied"* and one quality statement *"work as it should"*, which clearly refer to the meeting of expectations. Emotion statement supports the fact that the rating is higher than average.

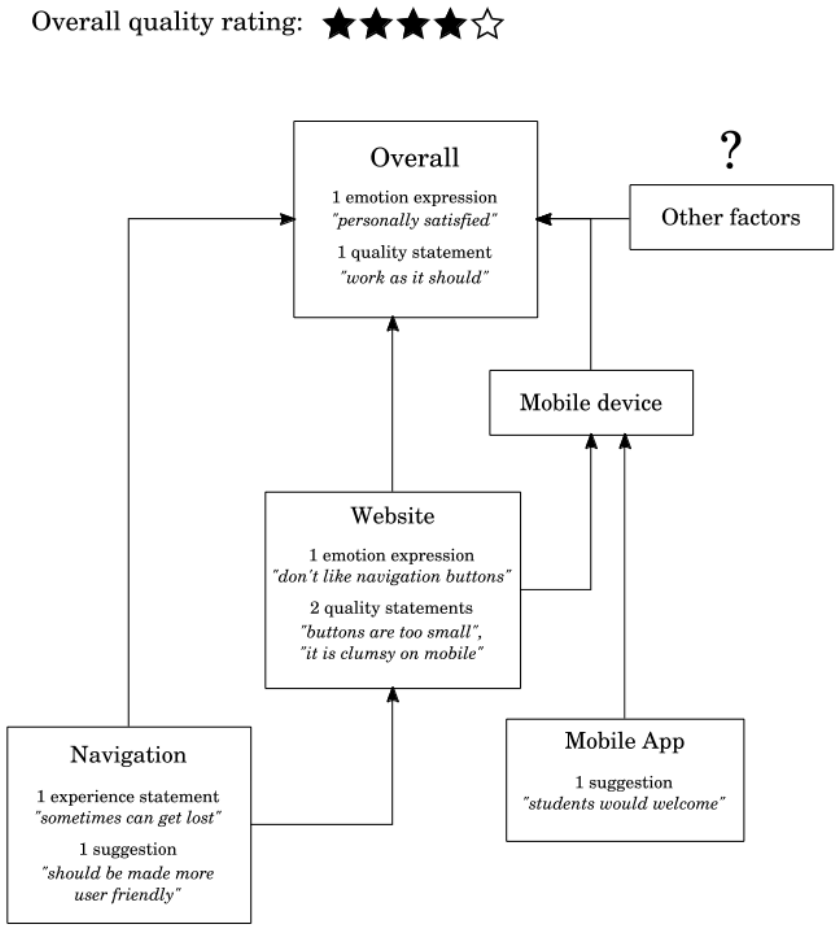


Figure 3: Example review content (author)

What does not correspond is the relationship of overall rating or statement to aspect groups. All mentioned aspects are linked with negative sentiment. From this example, it is clear **overall quality judgement does not always result from mentioned aspects only**. Perceived weight of aspects mentioned in the example must be on a low level, and other factors with positive sentiment must be considered.

In cases where explicit weights are not expressed in a review, like in case of the sample review, weights cannot be satisfactorily set. The only way is an approximative setting, but this setting is limited to one situation only. The situation was described in (Song, Lee, Yoon, Park 2015) and is based on the following assumption. **If only two aspects are expressed, one with negative sentiment and second with positive, and rating is high, relative importance of the positive aspect is higher than the relative importance of the negative aspect** – the positive aspect overweighted the negative one. Unfortunately, this situation happens rarely, because reviewers usually mention more than two aspects. If more aspects appear in a review, it is right to assume that only a group of positive or negative aspects overweight the other group.

Another review from analysed sample illustrates the findings from the literature. Respondent wrote mainly about bad *user experience* caused by *navigation* problems. On the other hand, respondent stated that *information support* is on the expected level. No other service aspects were mentioned. The

limitation of review to these two groups of aspects was supported by a summary statement that mentioned both navigation and information support. As per the overall quality rating of *two points*, *navigation* aspect was perceived probably as more important than *information support*.

However, the assumption of aspects weighting has another limitation. It was observed that **overall rating is influenced by other unexpressed aspects.** The question is how strong that influence might be? According to the literature review, consumers mention only the aspects that did not meet or exceed their expectations (Zhang, Yu, Li, Lin 2016), which is in contradiction to the example review where non-present aspects changed quality judgement completely.

In conclusion, the main finding of this study is that aspect sentiment cannot explain overall quality rating by itself. According to **observations made in this study, it is clear that aspects have different weights, which are hard to extract from a text**. Moreover, **overall quality is clearly influenced by non-expressed aspects**.

Discussion

Service quality literature proposed aspect hierarchy tree (Hu, Liu 2004; Liu 2012, 2015), but this concept should be modelled on the individual level. Findings of this study call for a different mechanism for quality aggregation and diagnosis.

How does the optimal service quality diagnosis look like? On account of perceived service quality as an individual concept, it is necessary to diagnose service quality on the same level – an individual's level. Each review should be transformed into a quality model – more specifically to a tree of quality attributes. All individual's quality trees should be placed in a database to enable their querying.

A quality tree reflects how a consumer understands a service; it is a mental model of service regarding its structure, quality and emotions, a result of implicit quality expectation confirmation. A tree is composed of weighted relations between service quality attributes and corresponding emotions including sentiment and its orientation and intensity.

Qualitative research in this study found that a quality tree cannot be composed based on a single consumer review because consumers mention only a few quality attributes, whereas full quality structure, influence relationships and weights remain hidden.

Service quality research should focus on finding a way how to capture such quality trees. It could also be used for perceived quality prediction in case of planning a change of production quality. Unfortunately, certain constraints still exist. One of the main issues is that no consumer review uncovers a complete quality tree. There are always non-expressed attributes and hidden perceived service quality composition.

There are at least two possible ways how to complete a quality tree: (1) survey consumers with additional questions, (2) start from the assumption that non-expressed quality attributes are unimportant or on the expected level of quality, and thus their influence is negligible.

Conclusions

This study focused on the issue of e-service quality diagnosis on the level of a reivew. For this purpose, qualitative research according to principles of grounded theory was used. Five research question was proposed regarding review body composition, language patterns, service aspects, sentiment expressions and the relationship between quality rating and review body.

As an intrument for data gathering, the usuall form of on-line quality review was used. The survey resulted in thirty responses collected from students of University of Economics, Prague. They had been asked to follow the description cited in the method section of this study. Students filled the survey online, but they were present in a classroom during the data collection phase. Because the quality survey was optional, not all of attending students submitted it.

To enable coding, each review needed to be divided into single clauses and classified by parser into word dependencies and parts-of-speech. Because there were more perspectives that needed to be

considered, not only one open coding process was undertaken. Analyses considered seven particular coding steps: (1) sentence coding, (2) language pattern coding, (3) parsed dependencies coding, (4) aspect coding, (5) sentiment pattern coding, (6) sentiment coding, (7) comparison with an overall quality rating. In total, seventy-eight clauses were coded.

Twelve statement types were that build reviews found. As the most useful, regarding the service quality diagnosis, were identified quality, emotion expression and summary statements. Language structure of these statements was explored.

According to analysis of sentiment and aspect expressions, it was found that aspects are multidimensional, more aspects can be found in one expression, and that sentiment may be expressed together with aspects in one word.

It was found that aspects have different weights, which are hard to extract from a text, and moreover, that overall quality is clearly influenced by non-expressed aspects.

Research results are limited given the fact that the survey was in English and respondents are not native English speakers. Notably, the question on language patterns is difficult to generalise. On the other hand, service customers are not always native speakers, and natural language analysis must adapt to a specific vocabulary and language pattern used by customers.

The optimal service quality diagnoses was proposed based on the results of the study in the discussion chapter. It took into account individuality of quality perception and impossibility to aggregate these perceptions on a higher level and recommend to build individual quality trees.

References

DUAN, Wenjing, CAO, Qing, YU, Yang and LEVY, Stuart, 2013. Mining Online User-Generated Content: Using Sentiment Analysis Technique to Study Hotel Service Quality. In: *2013 46th Hawaii International Conference on System Sciences*. 2013. p. 3119–3128. ISBN 978-1-4673-5933-7.

EL-BAYOUMI, Janice G, 2012. Evaluating IT service quality using SERVQUAL. In: *Proceedings of the ACM SIGUCCS 40th annual conference on Special interest group on university and college computing services - SIGUCCS '12*. New York, New York, USA: ACM Press. 2012. p. 15. SIGUCCS '12. ISBN 9781450314947.

GEFEN, David, 2002. Customer loyalty in e-commerce. *Journal of the association for information systems*. 2002. Vol. 3, no. 1, p. 2.

HU, Minqing and LIU, Bing, 2004. Mining and summarizing customer reviews. In: *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*. 2004. p. 168–177. ISBN 1581138889.

HU, Nan, PAVLOU, Paul A and ZHANG, Jennifer, 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. In: *Proceedings of the 7th ACM conference on Electronic commerce*. ACM. 2006. p. 324–330. ISBN 1595932364.

CHEN, Danqi and MANNING, Christopher D, 2014. A Fast and Accurate Dependency Parser using Neural Networks. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* [online]. 2014. p. 740–750. Available from: https://cs.stanford.edu/~danqi/papers/emnlp2014.pdf

JAMES, Tabitha L., CALDERON, Eduardo D. Villacis and COOK, Deborah F., 2017. Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*. 2017. Vol. 71, p. 479–492. DOI 10.1016/j.eswa.2016.11.004.

LADHARI, Riadh, 2010. Developing e-service quality scales: A literature review. *Journal of Retailing and Consumer Services*. 2010. Vol. 17, no. 6, p. 464–477. DOI 10.1016/j.jretconser.2010.06.003.

LEPMETS, Marion, CATER-STEEL, Aileen, GACENGA, Francis and RAS, Eric, 2012. Extending the IT service quality measurement framework through a systematic literature review. *Journal of Service Science Research*. June 2012. Vol. 4, no. 1, p. 7–47. DOI http://dx.doi.org/10.1007/s12927-012-0001-6.

LI, Y N, TAN, K C and XIE, M, 2002. Measuring web-based service quality. *Total Quality Management*. August 2002. Vol. 13, no. 5, p. 685–700. DOI 10.1080/0954412022000002072.

LIU, Bing, 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. 2012. Vol. 5, no. 1, p. 1–167.

LIU, Bing, 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. First edit. New York, NY, USA: Cambridge University Press. ISBN 978-1-316-29832-9.

MUDAMBI, Susan M and SCHUFF, David, 2010. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*. 2010. Vol. 34, no. 1, p. 185–200. DOI Article.

PALESE, Biagio and PICCOLI, Gabriele, 2016. Online Reviews as a Measure of Service Quality. In: *2016 Pre-ICIS SIGDSA/IFIP WG8.3 Symposium, Dublin 2016*. 2016.

PARASURAMAN, A., ZEITHAML, Valarie A. and MALHOTRA, Arvind, 2005. E-S-QUAL A Multiple-Item Scale for Assessing Electronic Service Quality. *Journal of Service Research*. 2005. Vol. 7, no. X, p. 1–21. DOI 10.1177/1094670504271156.

QU, Zhe, ZHANG, Han and LI, Haizheng, 2008. Determinants of online merchant rating: Content analysis of consumer comments about Yahoo merchants. *Decision Support Systems*. 2008. Vol. 46, no. 1, p. 440–449. DOI 10.1016/j.dss.2008.08.004.

ROLLAND, S. and FREEMAN, I., 2010. A new measure of e-service quality in France. *International Journal of Retail and Distribution Management*. 2010. Vol. 38, no. 7, p. 497–517. DOI 10.1108/09590551011052106.

SONG, Bomi, LEE, Changyong, YOON, Byungun and PARK, Yongtae, 2015. Diagnosing service quality using customer reviews: an index approach based on sentiment and gap analyses. *Service Business*. 2015. DOI 10.1007/s11628-015-0290-1.

STRAUSS, A L and CORBIN, J M, 1990. *Basics of qualitative research: grounded theory procedures and techniques*. Sage Publications. ISBN 9780803932500.

VAN BON, Jan, JONG, Arjen de, KOLTHOF, Axel, PIEPER, Mike, ROZEMEIJER, Eric, TJASSING, Ruby, VAN DER VEEN, Annelies and VERHEIJEN, Tieneke, 2007. *IT service management: an introduction*. First edit. Van Haren Publishing, Zaltbommel. ISBN 978-90-8753-051-8.

VARGO, StephenL. and LUSCH, RobertF., 2008. Service-dominant logic: continuing the evolution. *Journal of the Academy of Marketing Science*. 2008. Vol. 36, no. 1, p. 1–10. DOI 10.1007/s11747-007-0069-6.

VENCOVSKY, Filip, BRUCKNER, Tomas and SPERKOVA, Lucie, 2016. Customer Feedback Analysis: Case of E-banking Service. In: *10th European Conference on Information Systems Management: ECISM 2016*. 2016. p. 404.

VOŘÍŠEK, J and BASL, J, 2008. *Principy a modely řízení podnikové informatiky*. Oeconomica. ISBN 9788024514406.

WOLFINBARGER, Mary and GILLY, Mary C., 2003. eTailQ: Dimensionalizing, measuring and predicting etail quality. *Journal of Retailing*. 2003. Vol. 79, no. 3, p. 183–198. DOI 10.1016/S0022-4359(03)00034-4.

YANG, Zhilin and FANG, Xiang, 2004. Online service quality dimensions and their relationships with satisfaction: A content analysis of customer reviews of securities brokerage services. *International Journal of Service Industry Management*. 2004. Vol. 15, no. 3, p. 302–326. DOI 10.1108/09564230410540953.

ZHANG, X.a, YU, Y.b, LI, H.c and LIN, Z.d, 2016. Sentimental interplay between structured and unstructured user-generated contents. *Online Information Review*. 2016. Vol. 40, no. 1, p. 119–145. DOI 10.1108/OIR-04-2015-0101.

JEL Classification: M31, L86

Shrnutí

Diagnostika kvality e-služeb

Studie se zabývá problémem diagnostiky kvality e-služeb. Ze své podstaty, e-služby neumožňují osobní komunikaci zákazníka s poskytovatelem. Online prostředí ale na druhou stranu přináší velké množství uživatelsky generovaného obsahu. Jedním typem tohoto obsahu jsou online recenze, do kterých zákazníci vnáší svá hodnocení kvality. I když byly online recenze předmětem výzkumu, interpretace kvality služeb z nich je nedostatečně zpracovaná, a to platí zejména pro jejich textovou část. Práce, které se zabývají touto částí, se pohybují výhradně na agregované úrovni kvality. Oproti tomu, tato studie zkoumá kvalitu hlouběji, na úrovni jedné recenze. Pro tento účel byla použita metoda kvalitativního výzkumu s principy zakotvené teorie. Bylo stanoveno pět výzkumných otázek týkajících se složení recenze, jazykových vzorců, aspektů služby, výrazů sentimentu a vztahu mezi bodovým hodnocením kvality a textovou recenzí.

Klíčová slova: E-služby, Kvalita služeb, Online recenze, Kvalitativní výzkum.

Možnosti vizualizace metrik dat

Daniel Vodňanský

xvodd00@vse.cz

Doktorand oboru Aplikovaná informatika

Školitel: prof. RNDr. Jiří Ivánek, CSc., jiri.ivanek@vse.cz

Abstrakt: Práce uvádí tři metriky, umožňující rozčlenění dat na základě jejich typu struktury i vnitřní charakteristiky – míru informace, hierarchičnost a strukturovanost. Následně je představena trojrozměrná vizualizační metoda, umožňující vizualizace hodnot pro konkrétní data, uspořádaná do datových objektů skládajících se z datových jednotek.

Klíčová slova: Trojrozměrná vizualizace, redundance, míra informace, hierarchičnost, strukturovanost, metrika, dimenze, datový objekt, datová jednotka, relační databáze, XML, JSON, RDF, semistrukturovaná data.

Úvod

Motivace

Jednou z motivací výběru tohoto tématu je zájem o problém heterogenity hlavních ve světě používaných datových struktur a stále stejných opakujících se problémů v jejich vzájemné kompatibilitě (nejtypičtěji se jedná o relační databáze a formát XML, což studuje například (Krishnamurthy et al., 2001)). Tato práce staví na názoru, že tato nekompatibilita není způsobena pouhou nekompatibilitou technologií, nýbrž často nekompatibilitou samotné informace a ontologie, kterou data reprezentují. Tuto kompatibilitu se práce snaží studovat prostřednictvím sady metrik, které jsou multidimenzionálně porovnávány.

Ačkoli ve světě vznikla a stále vzniká celá řadů výzkumů na téma měření informačních a datových charakteristik, v naprosté většině případů cílí buď na jednotlivé datové technologie (zejména relační databáze) a ignorují podstatu jimi reprezentované informace, aniž by přistupovaly k heterogenní datové základně jako k celku. Tato práce usiluje a propojení dotyčný vzájemného nesouladu jak v přístupu k rovině dat a informace, tak i v přístupu k různým datovým technologiím.

Cíl

Cílem této práce je návrh vizualizační metody aplikovatelné na stanovený výčet metrik a datových struktur, která přehledně diverzifikuje tyto struktury navzájem a umožní vizuální zařazení konkrétní množiny dat. Součástí metody je i její praktická aplikace.

Datový objekt

Datovým objektem je myšlena jednoznačně ohraničená množina vzájemně souvisejících datových jednotek. Takovýto objekt zpravidla reprezentuje jistou podobně ohraničenou ontologii reálného světa spojenou vazbou na konkrétní entitu – nejtypičtěji podnik, organizaci nebo jejich stejně vymezenou podmnožinu. Datovým objektem je myšlena jednoznačně ohraničená množina vzájemně souvisejících datových jednotek. Takovýto objekt zpravidla reprezentuje jistou podobně ohraničenou ontologii reálného světa spojenou vazbou na konkrétní entitu – nejtypičtěji podnik, organizaci nebo jejich stejně vymezenou podmnožinu.

Pro účely této práce se jedná o prostou množinu jednotek o hodnotách všech zkoumaných metrik, které pro ni jsou jistým způsobem agregovány.

Metrika dat

Tato práce využívá výhradně metriky patřící do kategorie tzv. tvrdých metrik, tedy zcela objektivně a automatizovaně měřitelné a číselně vyjádřitelné ukazatele. Princip jejího měření se v zásadě shoduje a

je i inspirován datovými sklady na bázi OLAP – jedná se o *multidimenzionální datové modely* (jak je nazývá (Chaudhuri a Dayal, 1997). V tomto případě ovšem nejde o zkoumání dat na základě jejich obsahu a informačního významu, nýbrž na základě jejich vnitřní struktury a způsobu, jakým jsou navzájem (případně i vně) propojena.

V souladu s pojetím OLAP má takto vytvořená metrika význam vždy jen v multidimenzionálním kontextu. Určitou odlišností ovšem je, že metrika s dimenzí splývá, resp. její hodnota je využitelná coby metrika i dimenze. Lze tedy například zkoumat hierarchičnost na různých úrovních strukturovanosti právě tak jako zkoumat strukturovanost na různých úrovních hierarchičnosti, což v případě klasických datových skladů pravděpodobně není běžné.

Dále je vhodné podotknout, že existuje ještě čtvrtá metrika, která pro svou jednoznačnost není dále zkoumána, a sice prostá bitová velikost.

Strukturovanost

Strukturovaností je dle (Boehm, B. W. et al., 1973) myšlen „stupeň, do jaké míry systém či jeho komponenta disponuje jistým vzorem svých vzájemně závislých částí“.

Na úrovni informace je relativně sporné, nakolik je adekvátní ptát se na strukturovanost informace samotné a zda tato otázka nemíří ve skutečnosti na strukturovanost konkrétního informačního sdělení lze chápat spíše jako datovou reprezentaci informace nežli informaci samotnou. Pokud navíc informace (abstrahujeme-li od jakékoli vágnosti či neurčitosti) je normalizována do tvaru subjekt – predikát – objekt (A-Box), činí ji to samu o sobě zcela strukturovanou. Totéž platí pro terminologický popis ontologie (T-box).

Např. (Florescu, 2005) představuje obecnou úvahu nad problémem méně strukturovaných informací (a tedy i dat) mimo jiné coby jednoho z hlavních omezení RDBS ve svém klasickém pojetí. (Cardelli, 2001) se zabývá tvorbou deskripčního jazyka upraveného pro kontext semistrukturovaných, a navíc hierarchických dat[1]. Strukturovanost tak zpravidla není viděna coby spojitá metrika, nýbrž coby tři obecně uznávané úrovně strukturovanosti, jaké popisuje např. (Bartmann et al., 2011) – strukturovaná, semistrukturovaná a nestrukturovaná.

Proto je vhodné stanovit následující interní definici: informace je strukturovaná právě do takové míry, jako je strukturovaná její nejstrukturovanější možná přesná datová reprezentace. Strukturovanost dat je míra, do níž je možné při automatizované (strojové, bez použití metod strojového učení a umělé inteligence) transformaci určité datové struktury do struktury maximálně odlišné zachovat množství reprezentované informace.

Strukturovanost tedy v kontextu této práce nerozlišuje pouze tři zmíněné úrovně, nýbrž je agregována na úrovni zkoumaného datového objektu a jedná se tak o spojitou hodnotu. Tři úrovně jsou rozlišovány pouze na úrovni datové jednotky a sice hodnota 1 pro zcela strukturovanou jednotku (zpravidla obsahující samostatnou atomickou hodnotu), 0,5 pro jednotku obsahující strukturu, která není dále dělitelná a 0 pro jednotku strukturu zcela postrádající (zpravidla text obsahující věty). Spolu s nižší mírou strukturovanosti je míra informace i hierarchičnost hůře měřitelná, přičemž na nestrukturovaných datech není přímo měřitelná vůbec.

Hierarchičnost

Pojem hierarchie je zpravidla chápán jako struktura mající stromový charakter, tedy skládající se z jedné „kořenové“ entity, větvící se do entit dalších až po koncové, někdy také nazývané „listy“. Dle (Musca et al., 2011) je struktura hierarchická, pokud jsou „jednotky na nižší úrovni součástí jednotek na úrovni vyšší“. Je-li tedy možné informační strukturu zanést do orientovaného grafu (což do značné míry vyžaduje vyšší míru strukturovanosti) pak platí, že množina uzlů orientovaného grafu je hierarchická, pakliže každý takovýto uzel odkazuje maximálně na jeden další uzel a sám je odkazován libovolným počtem dalších jiných uzlů. Matematicky je možné toto pravidlo formalizovat následovně:

[1] kde mimo jiné do značné míry chybně mluví o datech, ačkoli deskripční jazyky a tím i jím zkoumaná oblast je na informační úrovni

$$\forall a, b: \big(H(a) \wedge R(a,b)\big) \Rightarrow \big(\neg \exists c: c \neq b \wedge R(a,c)\big),$$

kde a, b a c jsou uzly grafu, H je unární predikát – hierarchická vlastnost uzlu a R je binární relace dvou uzlů, kde první odkazuje na druhý.

Informace hierarchického charakteru jsou patrně dominantní, ne-li fakticky jedinou strukturou, s niž v jeden okamžik pracuje lidský mozek, která mu dodává pocit přehlednosti a je tak snadno zpracovatelná i dále sdělitelná. Je tedy pochopitelné, že na hierarchickou strukturu narážíme při každé strukturované (nejen elektronické) komunikaci – v dokumentech, organizačních schématech, jednotlivých webových stránkách, formulářích, prezentacích i při ostatních zpracováních vjemů reálného světa. To ovšem neznamená, že podstata ontologie reálného světa musí být většinově hierarchická; hierarchie zpravidla reprezentuje jistý úhel pohledu.

Pro datovou jednotku je hierarchičnost čistě binární hodnota – 0 v případě porušení definice hierarchičnosti, tedy stavu, kdy datová jednotka odkazuje na více než jednu další a 1 v případě, že tato definice splněna je.

Míra informace

Použití pojmu „míra“ je na informační úrovni do jisté míry problematické, protože se názvem kryje se Shannonovým pojmem, který ovšem vyjadřuje spíše redundanci (stav, kdy kdy je konkrétní fakt (informace) ukládaný v datech vícenásobně (Halpin 2001, s. 112)) datové reprezentace již dané informace. Je-li měřena informace samotná, cílovou hodnotou je množství sdělení samotného, bez ohledu na metody s nimiž bylo sděleno.

Tato hodnota je tedy zejména kvalitativním vyjádřením, nakolik neredundantně je daty reprezentována informace. Shannonův přístup, zkoumající informační zdroj na základě opakovanosti jednotlivých zpráv známý také jako informační entropie (Shannon, 2001) je nicméně prakticky použitelný pouze na jednorozměrnou a lineární (tedy plně strukturovanou) sekvenci bez ohledu na pořadí.

Zde je jistým řešením algoritmická míra informace (AIT), založená dle (Grünwald a Vitányi, 2008) na (Grünwald a Vitányi, 2008) „minimálním počtu bitů potřebných k popisu pozorování“. Tento maximálně zkrácený bitový zápis dat lze také zjednodušeně označit za datovou kompresi. Míru informace v prakticky libovolných datech je poté možné měřit pomocí prosté velikosti nejvíce zkomprimované, a tedy nejmenší množiny komprimovaných dat, případně pak kompresním poměrem těchto dat vůči tatům původním.

Datová jednotka

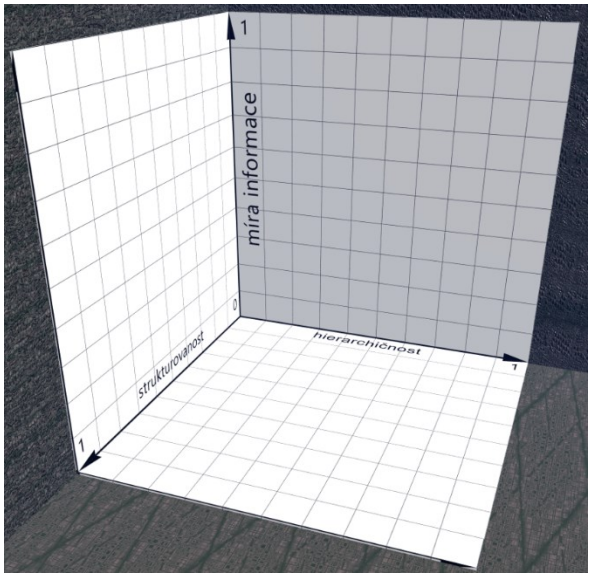
Jednotka je nejmenší možná struktura, pro kterou má smysl výpočet určených metrik a zařazení do dimenzí. Jedná se tedy o určitou obdobu elementárních částic známých z fyziky, jimiž je inspirována.

Po praktické stránce tak jde o prostý řádek v tabulce (například relační databáze v rámci metadatového skladu), pro který jsou měřitelné nejen samotné metriky, ale také bitová velikost a který je odkazatelný na samotná data v rámci datového objektu.

Trojrozměrný datový vizualizační prostor

Vizualizační metody užívané v této práci znázorňují informačně-datový objekt coby geometrický tvar v trojrozměrném prostoru, kde 3 osy (lze použít i jiný počet) reprezentují vybrané metriky.

Hlavní výhodou tohoto přístupu je možnost snadno zkoumat vzájemnou informačně-datovou kompatibilitu více objektů na základě průniku jejich geometrických reprezentací.



Obrázek 1 – Trojrozměrný prostor datových metrik

Tento samotný (pro názornost prázdný) prostor ukazuje Obrázek 1 – Trojrozměrný prostor datových metrik Tato vizualizační metoda umožňuje znázorňovat širokou škálu aspektů teorie zkoumané v této práci.

Tento přístup bezpochyby je aplikovatelný i na metadatový sklad. Aktuálně vyfiltrované hodnoty mohou být místo klasického grafu okamžitě zobrazovány v trojrozměrném prostoru. Datové transformace je dále pak možné znázornit jako animaci přechodu mezi dvěma stavy nebo prosté zobrazení obou stavů.

Jediným hlavním omezením tohoto přístupu jsou tři rozměry, což přirozeně neomezuje kontext této práce, zabývající se třemi metrikami, ovšem může být limitující pro následné rozšiřování a v zásadě i limituje znázornění pomyslné čtvrté metriky, a sice velikosti. Tento potenciální další rozměr lze doplnit například animací, tedy časovou dimenzí, barvou jednotlivých entit (která teoreticky na škále tří základních barev vytváří prostor až pro další tři rozměry) nebo i velikostí těchto entit, což dává smysl především v případě datové velikosti, ovšem zároveň toto z geometrického hlediska staví jistou názornost na úkor přesnosti zobrazení.

Datovou strukturu je možné vizualizovat coby geometrický útvar pokrývající části prostoru, v nichž se mohou vyskytovat struktuře odpovídající datové objekty. Pakliže jsou rozsahy možných hodnot určeny prostými intervaly, jsou datové struktury reprezentovány kvádrem. Nelze ovšem vyloučit, že v případě dalších, prací nezkoumaných struktur se může jednat i o jiné komplexnější geometrické útvary.

V případě datových objektů se jedná o data reprezentující konkrétní fakta, tedy na bázi A-Boxu. V tomto případě se jedná o matici D, jejíž sloupce odpovídají jednotlivým metrikám, a sice s – strukturovanost, h - hierarchičnost, i – míra informace (neredundance) a řádky představují jednotlivé datové jednotky 1 až n. Datový objekt lze potenciálně rozšířit o bitovou velikost jednotlivých jednotek a další identifikační a popisné sloupce.

$$D = \begin{bmatrix} s_1 & h_1 & i_1 \\ \vdots & \vdots & \vdots \\ s_n & h_n & i_n \end{bmatrix}$$

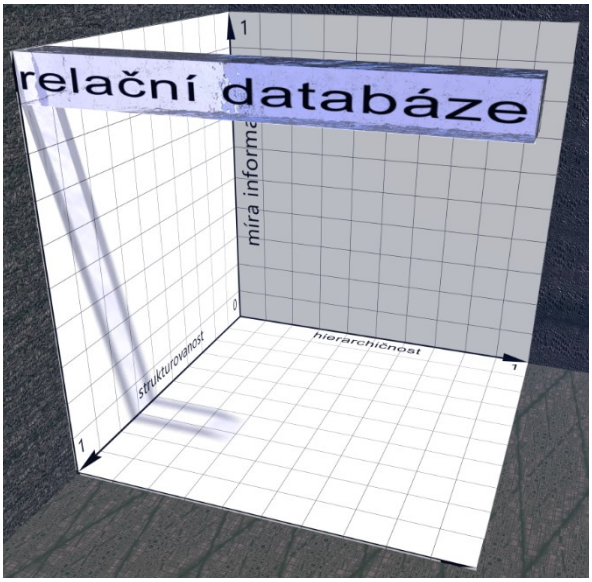
Jak bylo již zmíněno, průniky dvou i více geometrických reprezentací datových objektů nebo celých struktur umožňují analyzovat kompatibilitu dotyčných dat. Jednotky náležící průniku v tomto případě pak značí data přímo kompatibilní, a tedy převoditelná mezi oběma objekty či datovými strukturami bez informační ztráty.

Struktury a intervalové oblasti

Tato část vizualizuje nejběžnější prakticky používané datové struktury za pomoci trojrozměrného prostoru. Pro jistou větší přehlednost jsou hodnoty škálovány do intervalů o velikosti $0{,}1$.

Relační databáze (RDBS) jsou zcela bezpochyby dominantní technologií, především pak v podnikové oblasti, jak potvrzuje (Ramel, 2015). Tento velmi elegantní koncept známý na vědecké úrovni především na základě publikace (Codd, 1990) dodnes v mnoha ohledech není překonán. I přes to se v kontextu této práce potýká s řadou nepružností, které systém metrik pomáhá podrobněji specifikovat.

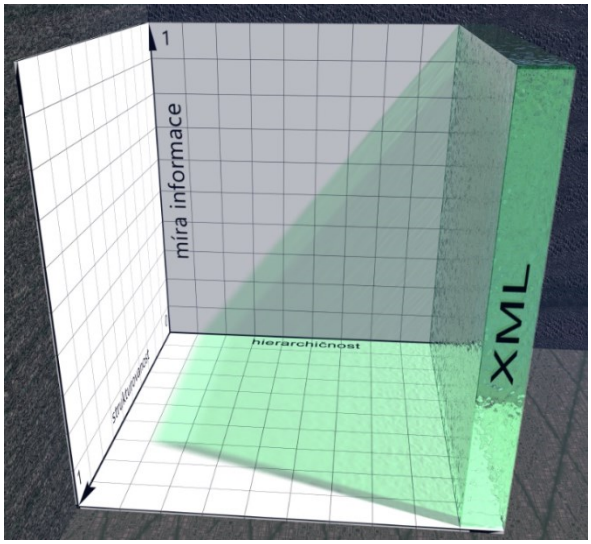
Tímto omezením je zejména zaměření na čistě strukturovaná data – ačkoli prakticky lze v RDBS uložit data o jakékoli strukturovanosti, dochází tím k porušení základních principů této technologie, především principů normalizace dat.



Obrázek 2 – vizualizace datového prostoru optimálně navržené relační databáze

Obrázek 2 znázorňuje prostor, který vymezuje oblast pokrývanou daty na základě zkoumaných metrik vhodně navržené relační databáze, tedy míra informace a strukturovanost rovny jedné, zatímco míra hierarchičnosti je libovolná, tedy $\in \langle 0; 1 \rangle$. Jak je již uvedeno, prakticky je umožněna libovolná redundance i strukturovanost, což je ovšem v rozporu s pojetím relační databáze již od jejího původního konceptu (Codd, 1990) a kvádr v dotyčné vizualizaci by obsáhl celý znázorněný krychlový prostor.

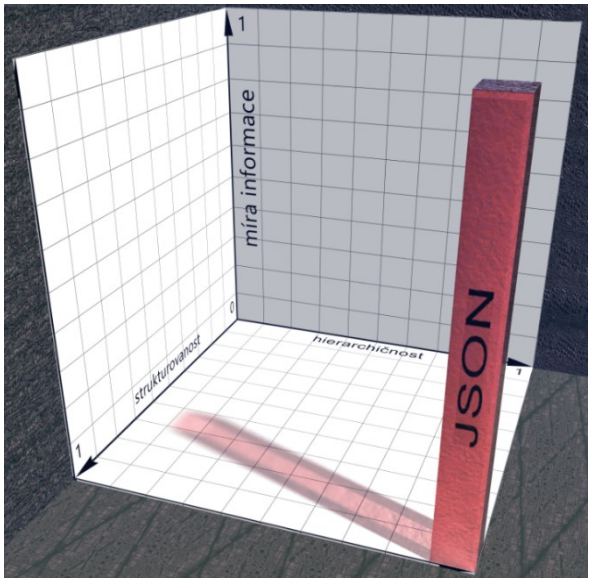
Dalším zkoumaným formátem je XML. XML je hierarchickým formátem pouze ve svém tradičním užití, tj. ve chvíli, kdy relace jsou vyjádřeny pouze na základě vztahů elementů rodiče nebo potomka. Tímto způsobem v praxi nemusí být a často není XML používáno; jazyky schémat XML umožňují různými způsoby užívat obdoby cizích klíčů v relační databázi, které vytváří relace napříč hierarchií. Tento přístup, ačkoli umožňuje kontrolu referenční integrity, je však do jisté míry nekoncepční, neboť reprezentuje relace dvojím způsobem a může tak umožnit modelovat jednu informaci mnoha různými způsoby.



Obrázek 3 – vizualizace datového prostoru formátu XML

Obrázek znázorňuje prostor, který na základě zkoumaných metrik obsáhnou data formátu XML Hierarchičnost je při tradičním užití rovna jedné.

Formát JSON (JavaScript Object Notation) je pro svou úspornost zápisu (byť částečně na úkor přehlednosti) v poslední době populární zejména v AJAXových metodách webových aplikací. S XML se do značné míry překrývá, jak je patrné z vizualizací (Obrázek a Obrázek 4).

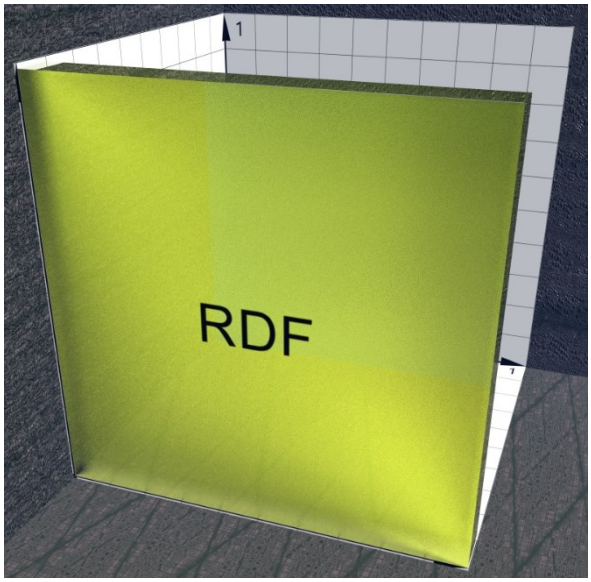


Obrázek 4 - vizualizace datového prostoru formátu JSON

Na rozdíl od XML ovšem nativně nepodporuje semistrukturovaná a nestrukturovaná data (byť jejich samotné uložení opět je možné) a vzhledem k tomu, že vznikl coby zápis objektu programovacího jazyka nemá své schéma.

Ontologické jazyky (RDF, RDFS a OWL) mají při shodné sémantice (představující do jisté míry informační úroveň modelování) řadu tzv. serializací, tedy způsobů textové datové reprezentace. Jednou z těchto hlavních serializací je formát založený na XML, ačkoli z hlediska datových metrik nemá s pojetím XML z předchozí část mnoho společného.

Tato data lze tedy pokládat za zcela strukturovaná, libovolně hierarchická a libovolně redundantní, jak zachycuje Obrázek 5.



Obrázek 5 - vizualizace datového prostoru formátu RDF (a příbuzných souvisejících formátů)

Dále pak existuje celá řada od podstaty nestrukturovaných formátů. Do této skupiny lze jednoznačně zařadit zejména volné texty, obsah multimediálních souborů, ovšem (Beach a Schiefelbein, 2014) říkají, že celkově jde o nejednoznačně vymezený pojem, kam často bývají řazena i data obsahující částečné struktury (fakticky semistrukturovaná), jako jsou e-maily. Vizualizace není pro heterogenitu a spornou měřitelnost hierarchičnosti a míry informace uvedena.

Systém tří metrik umožňuje přehledným způsobem klasifikovat jednotlivé datové struktury v následující tabulce.

Tabulka 1 – srovnání datových formátů

	Strukturovanost	Hierarchičnost	Míra informace
Relační databáze	$\{1\}$	$\langle 0;1\rangle$	$\{1\}$
XML	$\langle 0;1\rangle$	$\{1\}$	$\langle 0;1\rangle$
JSON	$\{1\}$	$\{1\}$	$\langle 0;1\rangle$
RDF	$\{1\}$	$\langle 0;1\rangle$	$\langle 0;1\rangle$
Nestrukturované formáty	$\{0\}$	-	$\langle 0;1\rangle$

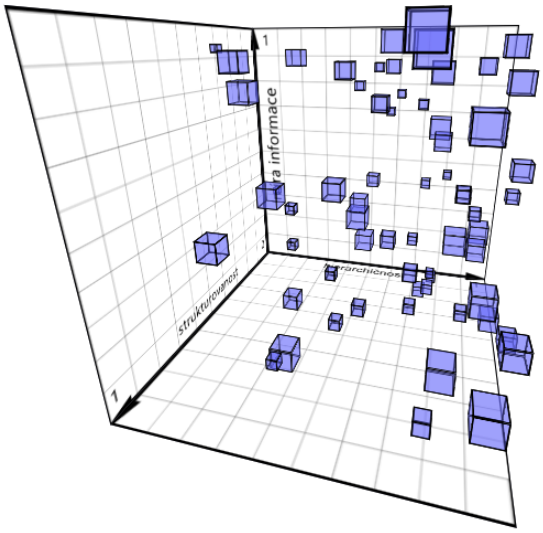
Tabulka 1 shrnuje jednotlivé hodnoty. Míra informace je myšlena coby inverzní pojem k redundanci. Hierarchičnost nestrukturovaných formátů není pro svou neměřitelnost uvedena. Dále je patrné, že metriky dokáží tyto jednotlivé a četně používané struktury plně diverzifikovat, neboť hodnoty se pro žádnou jejich dvojici neopakují.

Popis aplikace

Aplikace 3D-histogram (Vodňanský nedatováno) vyvíjená v rámci disertační práce umožňuje načíst datový objekt ve formě seznamu datových jednotek ve formátu CSV. Jednotlivé hodnoty jsou následně znázorněny v trojrozměrném datovém prostoru.

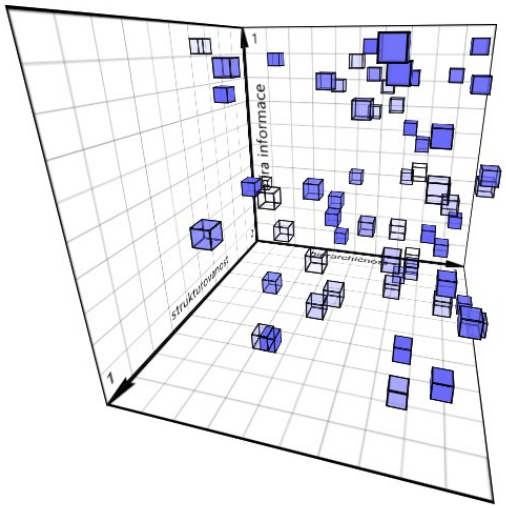
Jako čtvrtý parametr je brána bitová velikost jednotlivých jednotek. Z výše zmíněných důvodů je její zobrazení do jisté míry problematické, proto je nabízeno přepnutí dvou základních režimů vizualizace.

Prvním je prosté zobrazení jednotek coby krychlí na souřadnicích odpovídajících třem základním jednotkám. V tomto případě je velikost znázorněna velikostí hrany dotyčné krychle. Tento přístup lze pokládat za názorný v jistých případech (zpravidla kdy není jednotek příliš mnoho nebo nejsou příliš velké), protože do jisté míry narušuje geometrickou podstatu trojrozměrného prostoru a může vzniknout zcela nepřehledná změť vzájemně se překrývajících krychlí.



Obrázek 6 – první způsob vizualizace datového objektu

Druhý režim rozděluje datový prostor po třech osách na stanovené intervaly (mající také tvar krychle) jejíž barva určuje počet a velikost v něm se nacházejících datových jednotek. Tento přístup odbourává nevýhody předchozího a je tak použitelný zejména pro rozsáhlejší datové objekty o více jednotkách, ovšem s menší velikostí intervalu narůstá výpočetní složitost o třetí mocninu počtu těchto intervalů na jedné ose a znemožňuje zkoumání samostatných jednotek.



Obrázek 7 – druhý (intervalový) způsob vizualizace datového objektu

Závěr

Tato práce představila tři metriky (zaměnitelné s dimenzemi) a jejich aplikaci zejména na čtyři nejběžnější datové struktury. Dále byla prakticky ověřena jejich měřitelnost i na samotných datech prostřednictvím aplikace.

Tento výstup je využitelný především pro analýzu kompatibility dat (členěných do objektů), návrh jejich transformace a hledání problémových oblastí (redundance a nestrukturovaná data v relační databázi apod).

Do budoucna je plánován navazující výzkum možnosti budování tzv. metadatových skladů na bázi OLAP, měřících podobným způsobem tyto vnitřní charakteristiky dat namísto jejich sémantického významu využívající tuto vizualizační metodu. Také lze předpokládat další vylepšování dotyčné aplikace a případnou tvorbu jejího vnějšího API.

Literatura

BARTMANN, Dieter, Freimut BODENDORF, Elmar J. SINZ a Otto K. FERSTL, 2011. *Dienstorientierte IT-Systeme für hochflexible Geschäftsprozesse*. B.m.: University of Bamberg Press. ISBN 978-3-86309-010-4.

BEACH, Christopher S. a William R. SCHIEFELBEIN, 2014. Unstructured Data. *Journal of Accountancy*. **217**(1), 46–51. ISSN 00218448.

BOEHM, B. W. ET AL., 1973. Characteristics of software quality. In: *TRW-SS-73-09*.

CARDELLI, Luca, 2001. Describing semistructured data. *ACM SIGMOD Record*. **30**(4), 80–85.

CODD, E. F., 1990. *The relational model for database management: version 2*. Reading, Mass: Addison-Wesley. ISBN 978-0-201-14192-4.

FLORESCU, Daniela, 2005. Managing Semi-Structured Data. *Queue* [online]. **3**(8), 18–24. ISSN 1542-7730. Dostupné z: doi:10.1145/1103822.1103832

GRÜNWALD, Peter D. a Paul MB VITÁNYI, 2008. Algorithmic information theory. *Handbook of the Philosophy of Information*. 281–320.

HALPIN, Terry, 2001. *Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-672-2.

CHAUDHURI, Surajit a Umeshwar DAYAL, 1997. An overview of data warehousing and OLAP technology. *ACM Sigmod record*. **26**(1), 65–74.

KRISHNAMURTHY, Rajasekar, Jeffrey F. NAUGHTON, Jayavel SHANMUGASUNDARAM a Eugene SHEKITA, 2001. Dealing with (un) structuredness in XML Data and Queries Using Relational Databases. In: *DB seminar at Wise university* [online]. [vid. 2017-10-15]. Dostupné z: https://pdfs.semanticscholar.org/acb6/72e6feea4893192c74fc4cf3dcce31b3ad65.pdf

MUSCA, Serban C., Rodolphe KAMIEJSKI, Armelle NUGIER, Alain MÉOT, Abdelatif ER-RAFIY a Markus BRAUER, 2011. Data with Hierarchical Structure: Impact of Intraclass Correlation and Sample Size on Type-I Error. *Frontiers in Psychology* [online]. **2** [vid. 2016-02-13]. ISSN 1664-1078. Dostupné z: doi:10.3389/fpsyg.2011.00074

RAMEL, David, 2015. Relational Databases Still Reign in Enterprises, Survey Says. *Enterprise Systems* [online]. **2015** [vid. 2017-09-17]. Dostupné z: https://esj.com/articles/2015/04/23/database-survey.aspx

SHANNON, C. E., 2001. A Mathematical Theory of Communication. *SIGMOBILE Mob. Comput. Commun. Rev.* [online]. **5**(1), 3–55. ISSN 1559-1662. Dostupné z: doi:10.1145/584091.584093

VODŇANSKÝ, Daniel, nedatováno. 3D Histogram [online]. Dostupné z: http://3d-histogram.8u.cz/

JEL Classification: D80, D83, L86, Y10, C80

Summary

Data measures vizualisation possibilities

The paper presents three measures, allowing data to be broken down by their type of structure and internal characteristics – measure of information, hierarchicallity and structuredness. Subsequently, a three-dimensional visualization method is introduced, allowing visualization of values for specific data, that are arranged into data objects consisting of data units.

Keywords: Three-dimensional visualization, redundancy, measure of information, hierarchicallity, structuredness, measure, dimension, data object, data unit, relational database, XML, JSON, RDF, semistructured data.

Dolování pravidel z RDF grafů

Václav Zeman
vaclav.zeman@vse.cz
Doktorand oboru aplikovaná informatika
Školitel: doc. Ing. Vojtěch Svátek, Dr., svatek@vse.cz

Abstrakt: Propojená data na webu tvoří obrovskou znalostní bázi, ze které je možné získávat nové dosud neobjevené znalosti a souvislosti pro použití v různých komerčních i vědeckých oblastech. K publikacím strojově čitelných znalostních bází na webu se obecně používá standard RDF, dle kterého lze vytvářet množiny vzájemně propojených faktů. Tento článek se zabývá komplexní analýzou takovýchto datových množin za účelem automatického získávání nových znalostí použitelných pro další aplikace v různých znalostních systémech. Za nově objevenou znalost lze považovat nějaký vzor nebo nějaké pravidlo, které se v datech objevuje s neobvyklou četností. Hledání takovýchto zajímavých a relevantních pravidel z propojených RDF dat vyžaduje víceúrovňovou analýzu faktů, efektivní algoritmus pro rychlý výčet důležitých souvislostí a výběr těch znalostí, které jsou podstatné pro řešený byznys problém. V rámci tohoto článku je představen nástroj *rdfrules*[1], s jehož využitím lze provést všechny fáze k úspěšnému získání znalostí z RDF datasetů v podobě logických pravidel. Nástroj používá ověřené algoritmy pro předzpracování propojených dat, efektivní dolování pravidel a selekci zajímavých znalostí dle shlukové analýzy s velkou nabídkou měr zajímavostí.

Klíčová slova: propojená data, dolování dat, asociační a logická pravidla, dobývání znalostí z databází, RDF

Úvod

Proces získávání pravidel z databází patří mezi základní úlohy z oblasti dolování dat. Pravidlo zapisujeme buď ve tvaru IF-THEN nebo jako implikaci představující podmíněný výrok definovaný v matematické logice jako $A \Rightarrow B$, který se skládá z levé a pravé strany (antecedent a sukcedent) a dá se interpretovat následovně „pokud platí výrok *A*, potom platí i výrok *B*“. Na takovéto pravidlo lze nahlížet jako na způsob reprezentace znalosti, a proto lze úlohu hledání pravidel vyložit i jako proces objevování nových znalostí (Friedman, a další, 2001). Vstupem takovéto úlohy mohou být data v různých formátech. Pro tabulky či množiny transakcí nejčastěji hledáme tzv. asociační pravidla, která se týkají právě jednoho kontextu (Agrawal, a další, 1994). Naopak pro otevřené znalostní databáze tvořené množinami faktů často dolujeme tzv. logická pravidla (Lavrac, a další, 1994) s více proměnnými známá z oblasti induktivního logického programování (dále jen ILP). S rozdílem typů dolovaných pravidel odlišujeme i přístup vyhledávání.

x livesIn "Prague" $\Rightarrow$ x wasBornIn "Czech rep."

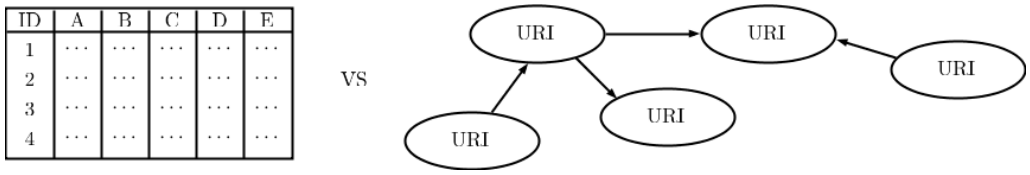
Příklad 1: Asociační pravidlo týkající se pouze jednoho kontextu (jedné proměnné)

x livesIn y $\Rightarrow$ x wasBornIn y

Příklad 2: Logické pravidlo s více proměnnými

Grafové datasety tvořené dle specifikace RDF[2] jsou řazeny mezi tzv. pětihvězdičkové[3] datové zdroje, kde mohou být jednotlivé uzly grafu propojeny s jinými datasety v rámci konceptu otevřených a propojených dat[4] (Yu, 2011). Hlavním rysem RDF specifikace je způsob popisu faktů ve tvaru trojice: **subjekt-predikát-objekt**, kde jsou zdroje a vlastnosti zdrojů popsány jednoznačným URI identifikátorem. RDF dataset je tvořen množinou takovýchto trojic a lze si ho představit jako orientovaný popsaný multigraf obsahující subjekty a objekty jako uzly a predikáty jako hrany (viz obrázek 1).

[1] https://github.com/KIZI/EasyMiner-Rdf
[2] Resource Description Framework
[3] http://5stardata.info/en/
[4] Angl. Linked Open Data



Obrázek 1: Ukázka datové reprezentace v tabulkovém formátu (vlevo) a grafovém RDF formátu (vpravo)

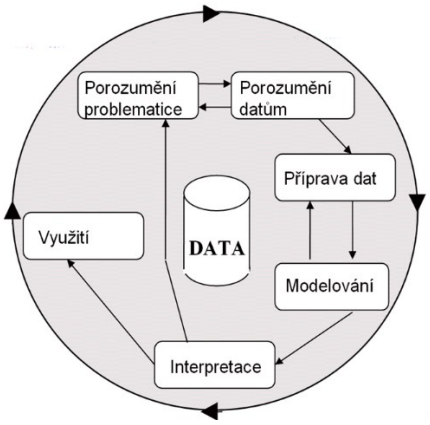
Na rozdíl od tabulkových formátů, kde můžeme objektu (reprezentovanému řádkem tabulky) přiřadit pouze dané vlastnosti (definovaných ve sloupcích tabulky), lze v rámci RDF grafu zachytit mnohem hlubší škálu vazeb mezi objekty a dolovat tak logická pravidla nad více proměnnými.

$$x \text{ hasFather } y \wedge x \text{ hasMother } z \Rightarrow y \text{ marriedTo } z$$

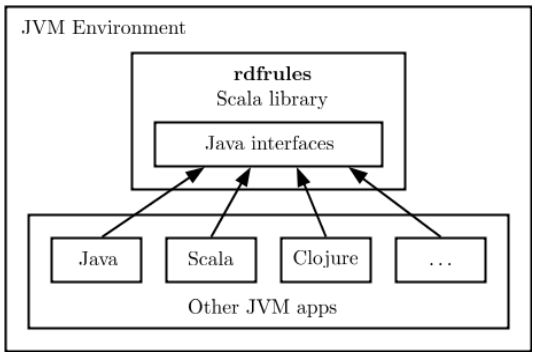
Příklad 3: Logické pravidlo se třemi proměnnými

Pro hledání pravidel z RDF grafů v současnosti existuje několik řešení. Jedním z nich je algoritmus AMIE+, který kombinuje rysy z oblasti hledání asociačních pravidel a ILP (Galarraga, a další, 2015). Podobný přístup je použit v rámci algoritmů RDF2Rules (Wang, a další, 2015) a SWARM (Barati, a další, 2017). Poslední zmíněný navíc do procesu dolování zahrnuje i informace z ontologií (neboli schémat pro daný RDF dataset). Každý z těchto algoritmů dokáže na základě vstupních dat a parametrů efektivně hledat logická pravidla přímo z RDF grafů bez nutnosti degradace na nižší stupeň datové reprezentace kde je již možné použít známé a prozkoumané algoritmy pro hledání asociačních pravidel z tabulek jako jsou apriori (Agrawal, a další, 1994), fp-growth (Han, a další, 2000), eclat (Borgelt, 2003), aj.

Všechny výše zmíněné algoritmy se zabývají pouze problematikou samotného dolování pravidel; nikoliv již neřeší komplexní postup pro získávání relevantních a zajímavých znalostí z RDF dat. Pokud se rozhodneme např. použít algoritmus AMIE+, je nutné si nejprve uvědomit, jaká konkrétní data jsou pro naší úlohu relevantní, jakým způsobem pracovat s numerickými rysy, jak určit zajímavost pravidla a jak správně nastavit parametry pro hledání těch pravidel, která nás skutečně zajímají. Data je tedy potřeba nejprve analyzovat, předzpracovat a následně definovat způsob dolování. Výsledné znalosti je poté nutno filtrovat, určit jejich relevanci a na konci správně interpretovat. Takovýto komplexní postup pro získávání znalostí z databází řeší např. metodiky CRISP-DM (viz obrázek 2) nebo SEMMA (Wirth, a další, 2000).



Obrázek 2: Jednotlivé fáze metodiky CRISP-DM



Obrázek 3: Prostředí nástroje rdfrules

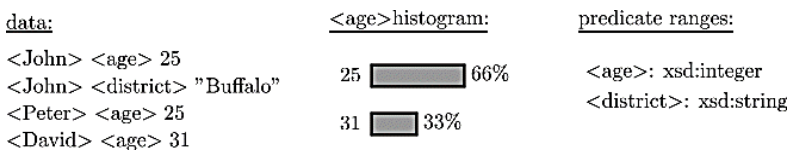
V tomto článku je představen komplexní nástroj **rdfrules** pro dolování pravidel z RDF grafů, ve kterém je implementován algoritmus AMIE+ s řadou rozšířeními a optimalizacemi zefektivňující průběh dolování. Tento nástroj dovoluje uživateli pracovat s množinou RDF grafů a jejich schématy, analyzovat jejich obsah, hledat relevantní rysy, filtrovat, mapovat či předzpracovávat různé výroky a

tím tak lépe porozumět datům pro správné nastavení parametrů dolovacího algoritmu. Mimoto lze zefektivnit průběh dolování velkou škálou nabízených měr zajímavostí a omezeních, čímž lze rychleji vyčíst ta pravidla, která uživatele skutečně zajímají. V poslední fázi nástroj nabízí možnosti filtrovat a řadit nalezená pravidla dle různých měřítek relevance, či shlukovat nebo naopak identifikovat odlehlá pravidla, která vykazují určitou míru podobnosti resp. odlišnosti. Pravidla lze exportovat pro další využití, např. tvorbu business pravidel nebo klasifikačních modelů.

V následujících odstavcích jsou popsány jednotlivé fáze dolování, které je možné spouštět v implementovaném nástroji pro komplexní dolování pravidel z RDF grafů. Tento nástroj je implementován v jazyce Scala a je možné ho použít v jakémkoliv JVM[5] prostředí (viz obrázek 3).

Porozumění datům

Abychom mohli zahájit proces dolování, je nutné nejprve nahrát data do systému a získat o nich souhrnné informace v podobě základních statistik a agregovaných dat (viz obrázek 4). Nástroj by měl tedy nabízet možnosti pro základní analýzu dat a náhled na data z různých dimenzí. Tím pak uživatel může více porozumět vstupním datům a lépe se rozhodnout jaké části datasetu předzpracovat, jak definovat proces dolování a jaká pravidla bude chtít získat na výstupu.



Obrázek 4: Ukázka základní datové analýzy nástroje rdfrules

V rámci RDF specifikace je možné dataset dělit na více pojmenovaných grafů. Každý graf obsahuje určitou množinu výroků (neboli trojic), jejichž domény a rozsahy lze specifikovat RDF schématem nebo ontologií. Jednotlivé grafy a schémata lze zapisovat v různých RDF formátech, např. n-triples, turtle, rdf/xml, json-ld aj. Nástroj pro dolování RDF dat je schopen načíst několik grafů v rámci jednoho datasetu ve všech standardních formátech vč. schémat a ontologií. Základní operace pro práci s pojmenovanými grafy a schématy jsou popsány v tabulce 1 a 2.

Tabulka 1: Popis základních metod třídy Graph

Třída: Graph		
	Metody	**Popis**
Vytváření objektu	fromFile(*file*, *format*, *name*)	Načtení grafu/grafů do systému ze souboru.
	fromInputStream(*is*, *format*, *name*)	Načtení grafu/grafů do systému z proudu dat.
Transformace	map(*func*)	Mapuje trojice na jiné trojice dle funkce *func*.
	filter(*func*)	Filtruje trojice dle funkce *func*.
	take(*n*), drop(*n*), slice(*from*, *until*)	Vybírá určitou podmnožinu dat.
	histogram(*s*, *p*, *o*)	Vytvoří objekt histogramu s agregovanými daty a jejich frekvencemi v grafu na základě vybraných částí výroku.
	types()	Vrátí domény a rozsahy pro všechny predikáty z grafu – identifikuje typy entit buď ze schémat, nebo přímo z dat.
	addSchema(*schema*)	Přiřadí schéma nebo ontologii ke grafu.
Akce	foreach(*func*)	Aplikuje funkci *func* na každou trojici v grafu.
	export(*os*, *format*)	Exportuje graf do výstupního datového proudu ve zvoleném RDF formátu.

[5] Java Virtual Machine

Tabulka 2: Popis základních metod třídy Schema

Třída: Schema		
	Metody	**Popis**
Vytváření objektu	fromFile(*file*, *format*) fromInputStream(*is*, *format*)	Načtení schématu do systému ze souboru nebo datového proudu.
Transformace	map(*func*), filter(*func*)	Mapuje trojice na jiné trojice nebo filtruje dle funkce *func*.
	take(*n*), drop(*n*), slice(*from*, *until*)	Vybírá určitou podmnožinu dat.
Akce	foreach(*func*)	Aplikuje funkci *func* na každou trojici ze schématu.

Tabulka 3: Popis základních metod třídy Dataset

Třída: Dataset		
	Metody	**Popis**
Vytváření objektu	fromGraphs(*graphs*)	Vytvoření datasetu z množiny grafů.
	fromFile(*file*, *format*) fromInputStream(is, *format*)	Vytvoření datasetu z RDF souboru nebo proudu dat.
	fromCache(*is*)	Načtení datasetu z keše.
Transformace	map(*func*), filter(*func*)	Mapuje či filtruje všechny trojice dle funkce *func.*
	take(*n*), drop(*n*), slice(*from*, *until*)	Vybírá určitou podmnožinu dat.
	histogram(*s*, *p*, *o*), types()	Vrací agregované informace o datech stejně jako v třídě grafu.
	addGraph(*graph*)	Přidává graf do datasetu.
	addSchema(*schema*)	Přiřazuje schéma k datasetu.
	addPrefixes(*prefixes*)	Přidává prefixy pro zkrácení popisů zdrojů.
Akce	foreach(*func*)	Aplikuje funkci *func* na každou trojici v datasetu.
	export(*os*, *format*)	Exportuje kompletní dataset do výstupního datového proudu ve zvoleném RDF formátu.
	cache(*type*)	Uložení datasetu do operační paměti nebo na disk.

Z načtených grafů a schémat se vytvářejí datasety, ze kterých se následně provádí předzpracování a dolování napříč kompletní množinou trojic ze všech grafů. Rozdělení na grafy je nutné zachovat pro pozdější fáze, ve kterých je možné filtrovat nalezená pravidla na základě příslušnosti k jednotlivým grafům. V rámci datasetů můžeme definovat prefixy pro redukci dlouhých popisů zdrojů či provádět podobné operace jako se samotnými grafy (viz tabulka 3). Zkompletované datasety je možné exportovat do jednoho souboru v libovolném RDF formátu či kešovat do paměti systému a disku pro pozdější a opakovaná použití. Základní operace pro práci s daty se dělí na transformace a akce.

Transformace

Všechny prováděné operace s daty jsou tzv. immutable – tedy nemění stav aktuálně načteného objektu. Transformační metody jsou líné (neboli lazy) tzn., že zavoláním funkce nedojte k okamžitému provedení dané transformace, ale pouze se deklaruje, že data budou při provádění následných akcí zpracovávána dle nějaké uživatelsky definované funkce. Provedení transformace se tedy provede až při zavolání akčních funkcí.

Podle základních transformačních metod lze načíst filtrovat, nahrazovat a vybírat načtené trojice. S ohledem na uživatelské filtry je možno zobrazovat datové typy predikátů (domény, rozsahy) a histogramy pro jednotlivé subjekty, predikáty či objekty.

Akce

Akční metody zpracovávají načtená data a vytvářejí z nich nějaký konkrétní výstup. Při zavolání akce jsou rovněž aplikovány všechny transformační funkce, které byly definovány v předchozích krocích.

Základní akční metody mají složitost $O(n)$, kde *n* je celkový počet trojic napříč všemi načtenými grafy. Data jsou zpracovávána proudově a nijak nezatěžují operační paměť daného prostředí. Mezi základní operace patří sekvenční výpis všech trojic, export dat na uživatelsky definovaný výstup nebo kešování dat pro pozdější a opakované použití.

Předzpracování dat

Na základě pečlivé analýzy dat, kterou je možné provádět dle výše zmíněných operací, můžeme získat agregované i podrobné informace o všech vlastnostech různých části výroků. Tyto informace lze využít pro následné předzpracování dat.

Při dolování pravidel dochází k prořezávání stavového prostoru na základě minimálních prahů měr zajímavostí vycházející z frekvence rysů, a tudíž lze dopředu odhadnout, jaké frekventované vlastnosti budou, resp. nebudou, zahrnuty ve výsledných pravidlech. Obecně numerické spojité atributy mívají velkou škálu hodnot a frekvence jednotlivých čísel nemusí dosahovat minimální prahové míry pro výběr potenciálně zajímavého pravidla. Mějme atribut věk, jehož rozsahem jsou kladná celá čísla $\langle 18; 80 \rangle$. Tabulka 4 zobrazuje relativní podporu v datech pro hodnoty $\langle 18; 21 \rangle$.

Tabulka 4: Ukázka histogramu

Věk	Podpora
18	1%
19	2%
20	2%
21	6%

Tabulka 5: Popis základních metod pro předzpracování dat třídy Dataset

Třída: Dataset (předzpracování)		
	Metody	**Popis**
Transformace	discretization(*algorithm*)(*func*)	Definuje slučování hodnot dle algoritmu z knihovny EasyMiner-Discretization. Funkce *func* slouží pro mapování a filtraci trojic na příslušný vstup nutný pro zvolený diskretizační algoritmus.

Pokud si zvolíme míru podpory 5 % jako minimální práh pro prořezávání stavového prostoru, potom si na základě zobrazeného histogramu můžeme být jisti, že se vlastnosti věk(18), věk(19) a věk(20) nebudou nikdy vyskytovat v žádných pravidlech, jelikož nedosahují nastavené minimální míry podpory. Pokud si ovšem přejeme i takovéto hodnoty zahrnout do výsledků tak buďto musíme snížit minimální práh podpory, a tím i zvýšit složitost dolování, nebo sloučit „nefrekventované“ hodnoty do intervalu $\langle 18; 20 \rangle$, čímž společně dosáhnou prahu 5 %.

Vedle možností filtrace a mapování trojic může uživatel v rámci načteného datasetu diskretizovat resp. slučovat numerické a nominální hodnoty dle vlastního ručně definovaného nastavení (viz tabulka 5). Ke slučování hodnot je použita knihovna EasyMiner-Discretization[6], vyvíjená speciálně pro potřeby diskretizace numerických hodnot, kterou je možné integrovat do jakékoliv JVM aplikace. Tato knihovna implementuje mnoho diskretizačních algoritmů, které lze použít i v rámci nástroje pro dolování pravidel z RDF grafů.

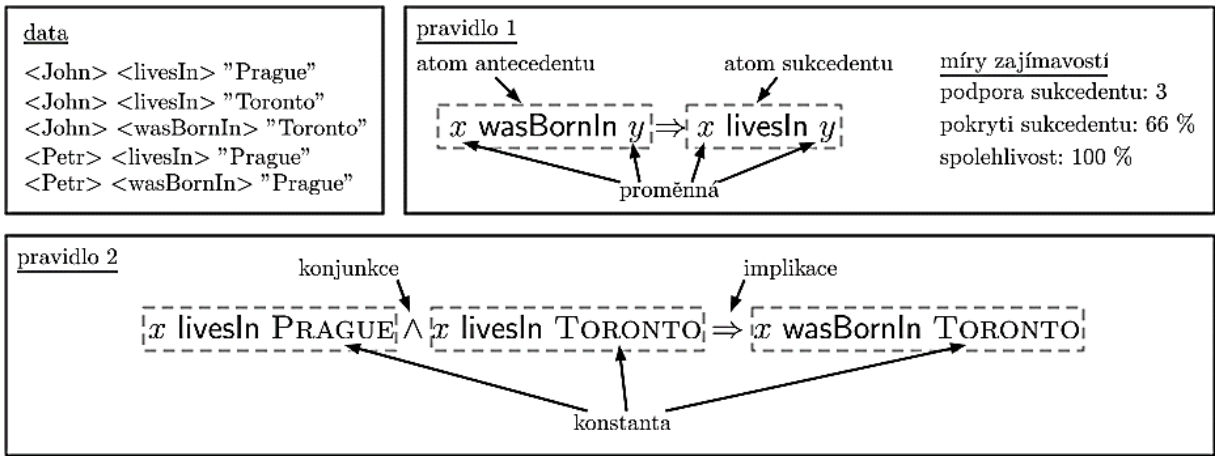
Dolování pravidel

Dolování pravidel z RDF grafů v nástroji *rdfrules* je prováděno na základě algoritmu AMIE+. Základní funkce tohoto algoritmu byly implementovány jako jádro dolovacího nástroje, avšak obsahuje řadu dalších rozšíření přejatých z jiných příbuzných algoritmů.

[6] https://github.com/KIZI/EasyMiner-Discretization

AMIE+

Algoritmus AMIE+ se zaměřuje na dolování Hornových klauzulí, které si lze představit jako pravidla s jedním atomem na pravé straně pravidla a konjunkcí atomů na levé straně pravidla. Atomem nazýváme nedělitelný logický výrok, kterým může být právě trojice subjekt-predikát-objekt. Takovouto trojici lze interpretovat jako binární relaci *x* R *y*, kde *x* = subjekt, *y* = objekt a R = predikát.



Obrázek 5: Ukázka pravidla a jeho částí

Vstupem algoritmu jsou trojice, které jsou převedeny na speciální index. Tento index si lze představit jako sadu hashovacích tabulek SPO, SOP, PSO, POS, OPS, OSP. Hashovací tabulka SPO má jako klíče všechny subjekty ze vstupního RDF datasetu. Hodnotou tabulky pro daný klíč je další hashovací tabulka s klíčem predikátu. Hodnotou vnořené hashovací tabulky je množina objektů. Hashovací tabulka PSO má stejnou strukturu, avšak klíčem tabulky jsou predikáty a klíčem vnořené tabulky subjekty. Takto tedy získáme soubor všech kombinací hashovacích tabulek, které lze vygenerovat z množiny všech trojic. Celkově je tedy nutné data 6x replikovat a uložit do šesti hashovacích tabulek. Díky takovémuto indexu je možné pravidla rychle vyčíst a počítat míry zajímavostí za účelem efektivního prořezávání vyhledávacího prostoru.

Způsob prořezávání je velmi podobný metodám pro dolování asociačních pravidel. Základními prořezávacími prahy jsou míry podpory a pokrytí sukcedentu. Řekněme, že pravidlo je frekventované pokud dosahuje uživatelem nastavené minimální míry podpory a pokrytí sukcedentu. V případě, že pravidlo není frekventované, potom jakákoliv další rozšíření tohoto pravidla nebudou rovněž frekventovaná. Algoritmus postupně generuje všechna možná pravidla, která splňují následující kritéria:

- Pravidlo musí obsahovat alespoň jednu proměnnou.
- Pravidlo obsahuje právě jeden atom na pravé straně pravidla, tzv. sukcedent pravidla.
- Pravidlo musí obsahovat minimálně jeden atom na levé straně pravidla, tzv. antecedent pravidla.
- Všechny atomy v pravidlu musí být tranzitivně propojeny, tzn., že každý atom musí být dosažitelný z jakéhokoliv jiného atomu dle sdílených proměnných.
- Sukcedent musí mít minimální podporu v datech.
- Pravidlo musí mít minimální pokrytí sukcedentu, tj. relativní počet trojic vyhovující sukcedentu, které jsou tranzitivně propojitelné s antecedentem pravidla.
- Pravidlo nesmí mít délku větší než uživatelem definovaná maximální délka pravidla.

Pokud pravidlo nesplňuje tato kritéria, nelze ho dále rozšiřovat; tím je rapidně prořezáván stavový prostor generovaný všemi kombinacemi pravidel, které lze z databáze vyčíst. Výpočet měr zajímavostí, které slouží k prořezávání (podpora a pokrytí sukcedentu), probíhá přímo při výčtu všech možných rozšíření daného pravidla nad vytvořeným indexem. Algoritmus rovněž obsahuje mechanizmy pro eliminaci duplicitních výpočtů.

Při rozšiřování pravidla generujeme takové typy atomů, ze kterých získáme následující typy pravidel:

- **Uzavřené pravidlo**: všechny proměnné v pravidlu se musejí nacházet minimálně ve dvou atomech: $x\ R1\ y \Rightarrow x\ R2\ y$
- **Neúplné pravidlo**: pravidlo obsahuje minimálně jednu a maximálně dvě visící proměnné, které se nacházejí pouze v jednom atomu: $x\ R1\ z \Rightarrow x\ R2\ y$
- **Instancované pravidlo**: pravidlo, které vzniklo nahrazením některé visící proměnné (nacházející se pouze v jednom atomu) konkrétní konstantou (hodnotou). Takovéto pravidlo může být buď neúplné (v případě, že obsahovalo dvě visící proměnné) nebo uzavřené (v případě, že obsahovalo jednu visící proměnnou): $x\ R1\ A \Rightarrow x\ R2\ y$ nebo $x\ R1\ A \Rightarrow x\ R2\ B$

Výstupem algoritmu jsou logická uzavřená pravidla splňující všechna definovaná kritéria.

Rozšíření dolovacího algoritmu

V rámci implementace algoritmu AMIE+ bylo přidáno několik rozšíření, které důsledněji zmenšují vyhledávací prostor na základě nových **omezení**. Jedním z nich je automatická eliminace redundantních a potenciálně nezajímavých pravidel. V dalších omezeních si uživatel může zvolit množinu predikátů, ze kterých chce generovat pravidla, nebo nastavit jaké vlastnosti má vykazovat výstupní pravidlo, např. neobsahuje duplicitní predikáty, neobsahuje žádné konstanty, obsahuje konstanty pouze na pravé straně pravidla apod.

Důležitým rozšířením je možnost definování **vzoru pravidla**, kterému mají vyhovovat výstupní pravidla. Vzor je aplikován přímo při procesu dolování a jeho použitím můžeme rapidně zmenšit vyhledávací prostor a celkově zrychlit běh výpočtu. Tento proces je přirozeně mnohem rychlejší než dolování pravidel bez masky a filtrování výsledků až ve fázi po dolování. Vzor pravidla se skládá právě z jednoho vzoru atomu na pravé straně a z více vzorů atomu na levé straně. Rozlišujeme dva typy vzorů:

- Exact: pravidlo musí přesně odpovídat vzoru.
- Partial: pravidlo částečně odpovídá vzoru.

Vzorů lze v rámci jednoho procesu dolování definovat více a jsou vyhodnocovány jako logické NEBO. Pokud tedy máme *vzor1* a *vzor2*, potom chceme na výstupu taková pravidla, která odpovídají *vzoru1* NEBO *vzoru2*. Vzor atomu definuje rysy atomu na dané pozici pravidla. Atom se skládá ze třech částí subjekt, predikát a objekt. Každá z těchto částí může odpovídat některému vzoru z následující sady:

- Any: jakákoliv hodnota
- AnyVariable: jakákoliv proměnná
- AnyConstant: jakákoliv konstanta
- Variable(*x*): pouze proměnná *x*
- Constant(*x*): pouze konstanta *x*
- OneOf(*patterns*): alespoň jeden z definovaných vzorů
- NoneOf(*patterns*): žádný z definovaných vzorů

Další rozšíření bylo přejato z algoritmu SWARM, kde vstupuje do procesu dolování **ontologické schéma**. Ve schématu lze definovat rozsahy, typy, nadtypy, podtypy příp. alternativní identifikátory objektů dle následujících predikátů:

Tabulka 6: Podporované vlastnosti RDF schématu v nástroji rdfrules

rdfs:subClassOf	Třída je podtřídou jiné třídy.
rdfs:subPropertyOf	Vlastnost je podvlastností jiné vlastnosti.
owl:sameAs	Dva různé zdroje referují na stejný objekt.
rdfs:range	Predikát obsahuje pouze objekty daného typu.
rdfs:domain	Predikát obsahuje pouze subjekty daného typu.

Vlastnosti rdfs:range a rdfs:domain lze použít během předzpracování k rychlé detekci typů jednotlivých entit a při generování chybějících trojic typu A rdf:type B. Vlastnost owl:sameAs nám slouží k detekci duplicitních atomů a vlastnosti rdfs:subClassOf a rdfs:subPropertyOf nám můžou pomoci při odvozování nových pravidel. Uvažujme následující schéma a pravidla:

Schéma	**Pravidla**
dbo:Athlete rdfs:subClassOf dbo:Person dbo:Journalist rdfs:subClassOf dbo:Person	x rdf:type dbo:Athlete ⇒ x dbo:education dbr:Academic_degree x rdf:type dbo:Journalist ⇒ x dbo:education dbr:Academic_degree

Pokud by obě pravidla měla pokrytí sukcedentu 2 %, ale minimální práh by byl nastaven na 2,5 %, potom by se pravidla neobjevila ve výsledné množině. V případě, že máme připojeno schéma, můžeme odvodit validní pravidlo x rdf:type dbo:Person ⇒ x dbo:education dbr:Academic_degree, které může mít míru pokrytí sukcedentu vyšší než 2,5 %. Tedy díky informacím ze schématu lze objevovat další potenciálně zajímavá pravidla na základě dědičnosti tříd a vlastností.

Použití

Dolování pravidel z načteného datasetu lze provádět až po vytvoření příslušného indexu, který se nahraje do operační paměti spuštěného prostředí. Jak již bylo zmíněno, algoritmus AMIE+ dokáže rychle vyčíst očekávaná pravidla a efektivně počítat míry zajímavostí z příslušného indexu a vyžaduje tedy nahrána data v paměti 6x zreplikovaná[7]. S touto paměťovou náročností je potřeba dopředu počítat. Vytvořený index je již neměnný a tvoří vstup jak pro operaci dolování pravidel, tak pro případné dopočítávání dalších měr zajímavostí pro výstupní pravidla (viz tabulka 7).

Mimo indexu vstupují do algoritmu uživatelem definovaná omezení, vzory pravidel či minimální prahy měr zajímavostí pro redukci vyhledávacího prostoru. Nástroj pro dolování pravidel disponuje rovněž debugovacími módy, které nás informují o aktuálním stavu průběhu výpočtu. Jednotlivé fáze algoritmu jsou prováděny paralelně a snaží se využít všechna dostupná jádra procesoru; rychlost výpočtu je tedy možné vertikálně škálovat[8].

Tabulka 7: Popis základních metod třídy Index

Třída: Index		
	Metody	**Popis**
Vytváření objektu	fromDataset(*dataset*)	Vytvoří index z datasetu.
	fromCache(*is*)	Nahraje index z keše.
Akce	mineRules(*algorithm*)	Spustí se dolování pravidel dle příslušného algoritmu (aktuálně AMIE+). Vrací objekt RuleSet obsahující kolekci pravidel.
	cache()	Uložení indexu do operační paměti nebo na disk.

Zpracování výstupních pravidel

Pravidel na výstupu může být velké množství, proto by měl mít uživatel možnost pravidla řadit a filtrovat dle vzorů či velké škály nabízených měr zajímavostí. Základní implementovaná měřítka obsažená v algoritmu AMIE+ (podpora sukcedentu, pokrytí sukcedentu, spolehlivost apod.) lze explicitně dopočítávat zvlášť pro jednotlivá pravidla i po dolovací fázi. Mimoto byly do nástroje přidány dvě nové míry zajímavostí – *lift* a *pca lift*. Tyto míry určují relativní nárůst spolehlivosti pravidla oproti průměrným četnostem sukcedentu.

[7] Replikují se pouze reference na dané atomy, nikoliv kompletní data. Hashovací tabulky obsahují pouze číselné reference.

[8] Přidáváním jader lze urychlit výpočet. V současnosti ale nelze nástroj použít pro paralelní dolování na více strojích.

Nalezená pravidla lze také porovnávat a shlukovat na základě výpočtu podobnosti dvou pravidel. Každé pravidlo má právě *n* rysů $F = \{f_1, f_2, \ldots, f_n\}$. Uvažujme pravidla r_1 a r_2, potom podobnost mezi těmito pravidly se rovná funkci $sim(r_1, r_2) = \sum_i^n w_i * s(f_i(r_1), f_i(r_2))$, kde funkce $s(f_i(r_1), f_i(r_2))$ vrací podobnost dvou rysů v rozsahu $\langle 0,1 \rangle$ a hodnota w_i značí váhu *i*-tého rysu, přičemž $\sum_i^n w_i = 1$. Následně byly definovány rysy pravidel, dle kterých se počítají příslušné podobnosti:

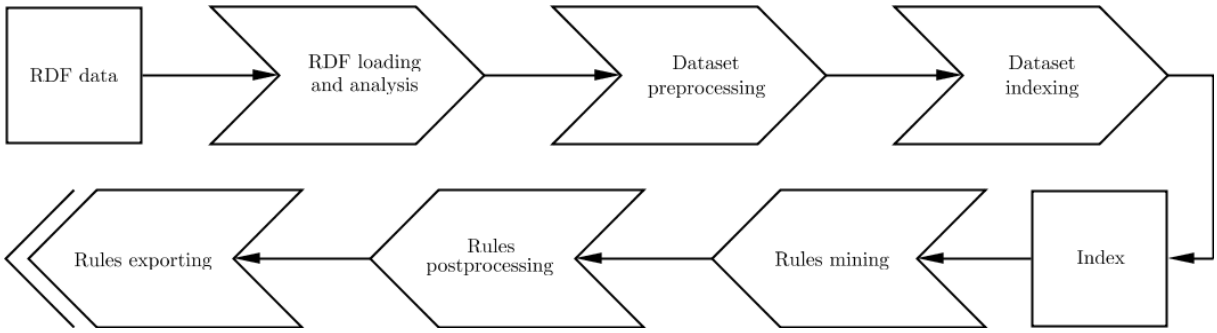
- f_1: Podobnost dle shody atomů v pravidlu.
- f_2: Podobnost dle míry pokrytí sukcedentu.
- f_3: Podobnost dle míry spolehlivosti.
- f_4: Podobnost dle délky pravidla.

Shluková analýza je prováděna algoritmem DBScan (Ester, a další, 1996), přičemž výstupem nejsou pouze shluky více pravidel, ale i anomálie – neboli odlehlá pravidla, která se liší od zbytku nalezené množiny.

Všechny výše zmíněné operace s pravidly jsou implementovány ve třídě RuleSet (viz tabulka 8), která se konstruuje během průběhu algoritmu AMIE+. Kompletní přehled všech fází procesu dobývání znalostí z RDF grafů v rámci nástroje *rdfrules* je zobrazen na obrázku 6.

Tabulka 8: Popis základních metod třídy RuleSet

Třída: RuleSet		
	Metody	**Popis**
Vytváření objektu	fromCache(*is, index*)	Vytvoří objekt z keše a indexu.
Transformace	map(*func*), filter(*func*)	Mapuje či filtruje všechny pravidla dle funkce *func.*
	take(*n*), drop(*n*), slice(*from*, *until*)	Vybírá určitou podmnožinu pravidel.
	clusters(*algorithm*)	Hledá shluky v pravidlech dle zvoleného algoritmu.
	sortBy(*func*)	Řadí pravidla dle funkce *func* pro srovnávání pravidel.
	countMeasure(*threshold*)	Dopočítává další míry zajímavostí na základě zvoleného prahu.
Akce	foreach(*func*)	Aplikuje funkci *func* na každé pravidlo v kolekci.
	export(*os*)	Exportuje pravidla do výstupního datového proudu.
	cache()	Uložení pravidel na disk pro pozdější využití.



Obrázek 6: Jednotlivé fáze dobývání znalostí z RDF grafů v nástroji rdfrules

Navazující práce

Nástroj *rdfrules* je aktuálně dostupný pouze v prostředí JVM. V následujícím období bude knihovna zabalena do R balíčku a bude ji možné používat rovněž i v prostředí R, které je určeno pro statistickou analýzu dat. Tím se stane nástroj snadno dostupný pro běžné použití datovými analytiky.

Další navazující rozšíření se týká škálování dolovacího procesu. Aktuálně lze hlavní výpočet při hledání logických pravidel škálovat pouze vertikálně. Veškerá data jsou nahrána v operační paměti a dostupná jádra procesoru dokáží paralelně hledat pravidla ve stavovém prostoru. Tento způsob výpočtu nemusí být ovšem snadno proveditelný pro velké datové množiny v rámci jednoho výpočetního stroje. V budoucím výzkumu se plánuje přizpůsobení algoritmu AMIE+ pro distribuované dolování pravidel ve výpočetním clusteru pomocí nástroje Apache Spark. Tím by bylo možné velká data a komplikovaný výpočet škálovat nejen vertikálně, ale i horizontálně.

Závěr

V tomto článku byl představen nástroj *rdfrules* využívaný pro komplexní dolování znalostí z RDF grafů v podobě logických pravidel. Nástroj je v prvé řadě možné použít pro předzpracování vstupních RDF dat a schémat dle velké škály transformačních a diskretizačních algoritmů. Jádrem nástroje *rdfrules* je algoritmus AMIE+ schopný dolovat pravidla přímo z RDF reprezentace. K tomuto algoritmu bylo implementováno mnoho rozšíření pro zvýšení efektivity výpočtu a pro nabídnutí více možností k identifikaci zajímavých pravidel. Po dolovací fázi je možné třídit a filtrovat nalezená pravidla dle nabízených měr zajímavostí; dále shlukovat pravidla a odhalovat unikátní odlehlá pravidla. Tím je uživateli usnadněno objevovat nové relevantní a zajímavé znalosti z propojených dat pro řešení jiných problémů, ať už pro vědecká nebo pro komerční využití.

Literatura

AGRAWAL, Rakesh a Ramakrishnan SRIKANT. Fast Algorithms for Mining Association Rules. *VLDB*. 1994.

BARATI, Molood, Quan BAI a Qing LIU. Mining semantic association rules from RDF data. *Knowledge-Based Systems*. 2017.

BORGELT, Christian. Efficient implementations of apriori and eclat. *FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations.* 2003.

ESTER, Martin, a další. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*. 1996.

FRIEDMAN, Jerome, Trevor HASTIE a Robert TIBSHIRANI. The Elements of Statistical Learning. *Springer series in statistics Springer*. 2001.

GALARRAGA, Luis, a další. Fast Rule Mining in Ontological Knowledge Bases with AMIE+. *The VLDB Journal*. 2015.

HAN, Jiawei a Jian PEI. Mining Frequent Patterns by Pattern-Growth: Methodology and Implications. *ACM SIGKDD explorations newsletter*. 2000.

LAVRAC, Nada a Saso DZEROSKI. Inductive Logic Programming. *WLP*. 1994.

WANG, Zhichun a Juanzi LI. RDF2Rules: learning rules from RDF knowledge bases by mining frequent predicate cycles. *arXiv preprint arXiv:1512.07734*. 2015.

WIRTH, Rudiger a Jochen HIPP. CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. 2000.

YU, Liyang. Linked Open Data. *A Developer's Guide to the Semantic Web*. 2011.

JEL Classification: O30

Summary

This article deals with a complex analysis of such datasets for automatic acquisition of new knowledge applicable to other knowledge systems. New discovered knowledge may have different forms, such as patterns or rules, which cover data with an atypical frequency. Search for these interesting and relevant rules from linked RDF data requires a multilevel analysis of facts, an efficient algorithm for the fast enumeration of important connections, and a selection of such knowledge sets which are substantial for solving business problems. This article introduces a tool called *rdfrules* used for logical rules mining from RDF datasets including pre-processing and post-processing phases. The tool uses verified algorithms for linked data processing, efficient rules mining and knowledge selection by cluster analysis.

Keywords: linked data, data mining, association and logical rules, knowledge discovery from databases, RDF.

Why Use High-Frequency Data in Quantitative Finance?

Vladimír Holý
vladimir.holy@vse.cz
Ph.D. Student of Econometrics and Operational Research
Thesis Advisor: doc. RNDr. Ing. Michal Černý, Ph.D. (cernym@vse.cz)

Abstract: Many financial time series such as foreign exchange rates, stock prices and commodity prices are in the form of high-frequency data. We review the literature assessing the impact of econometric and statistical analysis of high-frequency data in finance. We cover the main areas of quantitative finance - derivative valuation, risk management, portfolio optimization and trading strategies. The vast majority of studies report significant gains when utilizing high-frequency information even for low-frequency decisions and operations.

Keywords: High-Frequency Data, Derivative Valuation, Risk Management, Portfolio Optimization, Trading Strategies.

Introduction

Engle (2000) coined a term ultra-high-frequency data referring to irregularly spaced time series recorded at highest possible frequency corresponding to each transaction or change in bid/ask prices. Financial high-frequency time series include foreign exchange rates, stock prices and commodity prices. Over the last 20 years, there is a growing interest in high-frequency data as illustrated in Figure 1. Contrary to the rising attention from the scientific community, there are still two myths about high-frequency data in the financial industry identified by Stephanie Toper during 7th Annual Stevens Conference on High Frequency Finance and Analytics.

1. **High-frequency data are just a lot of low-frequency data.** This is not true as high-frequency data have distinct properties due to market microstructure specifics such as irregularly spaced observations, discreteness of price changes and bid-ask spread. It follows that the econometric and statistical analysis of high-frequency data requires special methods to deal with these market microstructure specifics.
2. **High-frequency data are just for high-frequency trading.** High-frequency trading naturally demands high-frequency data. Financial decisions and operations on lower frequencies, however, do benefit from high-frequency information. For example portfolio optimization carried out on a daily basis can utilize more precise daily volatility estimation based on intraday price movements.

In this paper, we focus on the refutation of the second myth. We review the literature assessing the impact of high-frequency data analysis in derivative valuation, risk management, portfolio optimization and trading strategies.

Volatility Estimation and Forecasting

High-frequency analysis primarily deals with volatility estimation and forecasting. Popular approach is to ex-post estimate quadratic variation by realized variance (Barndorff-Nielsen and Shephard, 2002a, 2002b). However, this is not as straightforward task as high-frequency data are contaminated by the market microstructure noise which makes realized variance significantly biased and inconsistent (Hansen and Lunde, 2006). In the high-frequency literature, many noise-robust estimators of quadratic variation have been proposed such as the two-scale estimator (Zhang et al., 2005), realized kernels (Barndorff-Nielsen et al., 2008), the pre-averaging estimator (Jacod et al., 2009) and the least squares estimator (Nolte and Voev, 2012). Another problem is ex-ante forecasts of quadratic variation. A popular models used for forecasting are the HAR model (Corsi, 2009) and the realized GARCH model (Hansen et al., 2012).

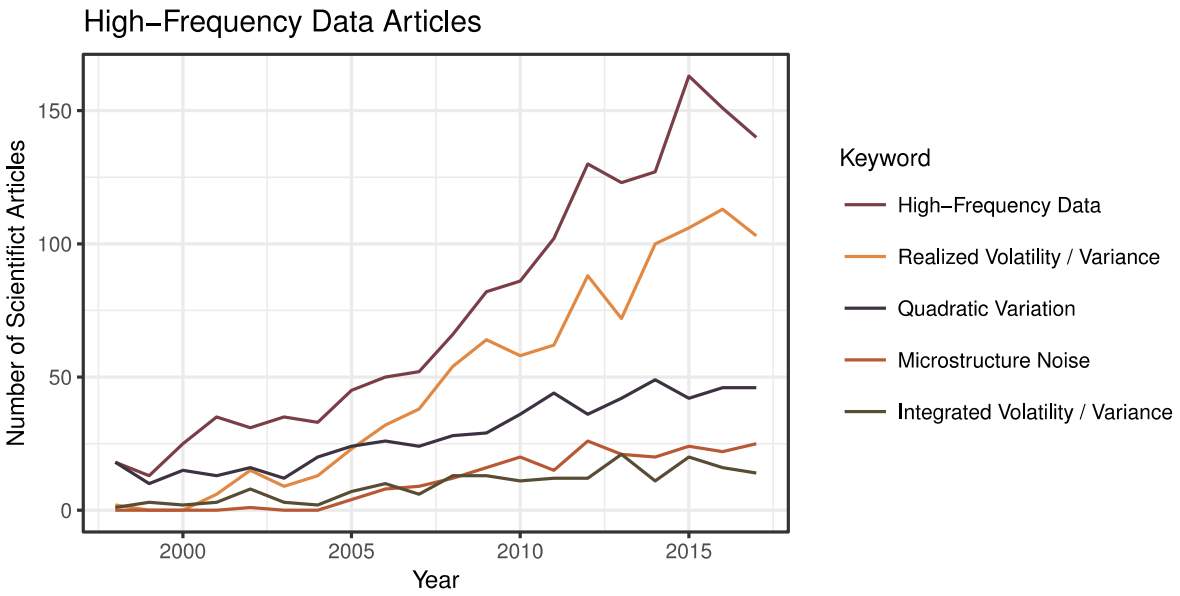


Figure 1: Number of scientific articles containing specific term in title, abstract or keywords.

Alternatively, a parametric approach can be adopted. It can be assumed that prices or some function of prices follow a specific process. Aït-Sahalia et al. (2005) estimated volatility parameter of the Wiener process in the presence of market microstructure noise while Holý and Tomanová (2017) estimated parameters of the Ornstein-Uhlenbeck process. This parametric approach can be useful in forecasting.

So far, we were interested in the values of prices. Engle and Russell (1998) analyzed durations between trading events using autoregressive conditional duration model. There are mainly three trading events that can be used in duration analysis. Duration until the price reaches some given value can be used as a proxy for volatility. Duration until the trading volume reaches some given value can be used as a proxy for liquidity. Finally, duration between transactions can be used as a proxy for trading intensity.

We illustrate the estimation of volatility using high-frequency data in the following experiment. We consider three approaches based on used frequency and methods.

1. **Use low frequency and ignore market microstructure noise.** This is the simplest approach without any major consequences as the market microstructure noise is quite negligible at lower frequencies. However, these estimates are not very precise as most of the information contained in the price process is discarded.
2. **Use high frequency and ignore market microstructure noise.** When we use the same methods for high-frequency data as for lower frequencies, we begin to face some issues caused by market microstructure noise. These estimates are significantly biased because we estimate volatility of the noise rather than the price.
3. **Use high frequency and take market microstructure noise into account.** The best option is to use the highest frequency possible and methods capable of separating the price and the noise. These estimates are unbiased and more accurate than the ones using lower frequencies.

To confirm these propositions, we simulate the Ornstein-Uhlenbeck process for 10 000 days and estimate daily volatility. In Figure 2 we estimate the quadratic variation by nonparametric methods. We compare the realized variance using 5-minute data and 20-second data with the least squares estimator of Nolte and Voev (2012) using 20-second data. In Figure 3 we estimate the volatility parameter of Ornstein-Uhlenbeck process by parametric methods. We compare the maximum likelihood estimator using 5-minute data and 20-second data with the noise-robust specification of maximum likelihood estimator of Holý and Tomanová (2017) using 20-second data. Both Figure 2 and Figure 3 show that the use of higher frequencies and noise-robust methods gives the most accurate estimates.

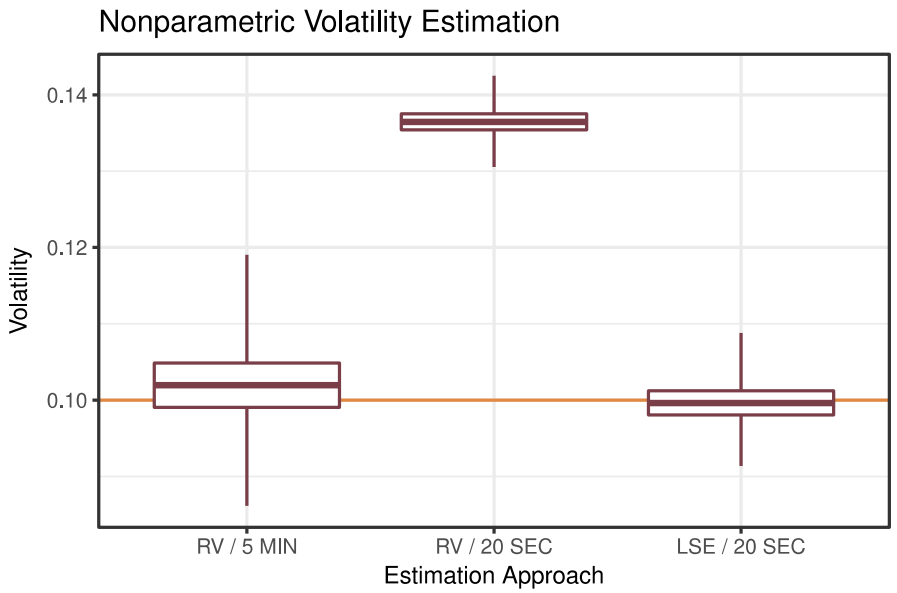


Figure 2: Illustrative simulation of volatility estimates by nonparametric methods.

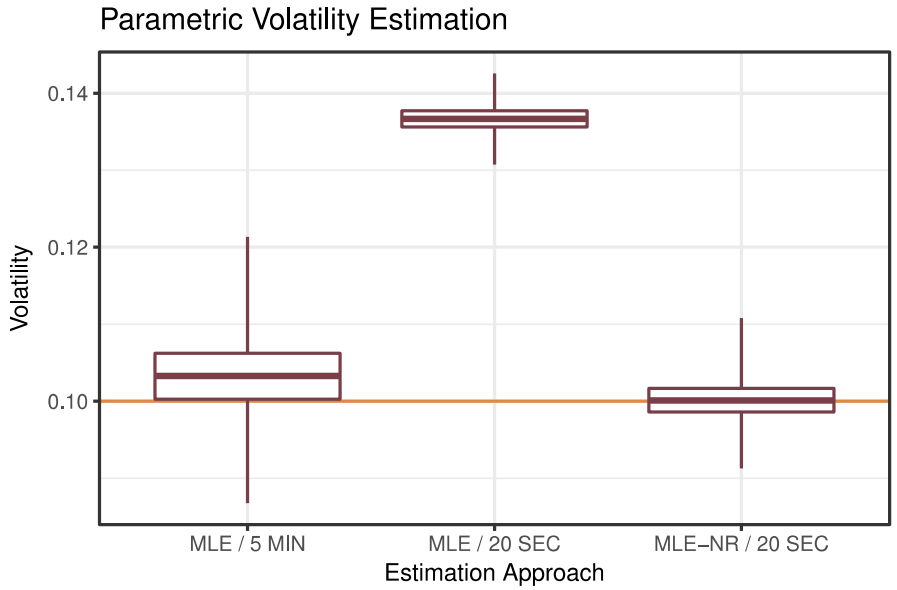


Figure 3: Illustrative simulation of volatility estimates by parametric methods.

Derivative Valuation

The first application of high-frequency data analysis we review is derivative valuation. Bollerslev and Zhang (2003) improved the multi-factor asset pricing model by incorporating high-frequency information. Their empirical study shows that high-frequency-based factor loadings yields better returns than the conventional monthly rolling regression-based estimates.

Bandi and Russell (2008) and Bandi et al. (2008a) focused on finite sample performance of several quadratic variation estimators with application to option pricing. Another comparison of quadratic variance estimators in the context of option pricing was performed by Sanfelici and Uboldi (2014).

Corsi et al. (2013) proposed an option pricing model with HAR forecasts of quadratic variation. In an empirical analysis of S&P 500 index options, they show that their model outperforms competing time-varying and stochastic volatility option pricing models.

Stentoft (2008) incorporated realized variance into option pricing model and concluded that the proposed model explains some of the mispricings found when using traditional option pricing models based on daily data.

Christoffersen et al. (2014) developed a class of option pricing models utilizing daily returns and realized variance. Their analysis of S&P 500 index showed that realized variance reduces the pricing errors of the benchmark model significantly across moneyness, maturity, and volatility levels.

Kenmoe and Sanfelici (2014) utilized high-frequency spot volatility in derivative pricing model and compared several spot volatility estimators. Empirical results showed that using intraday data rather than daily data provides smaller pricing errors.

Audrino and Fengler (2015) compared observed realized variance with realized variances implied by the Black-Scholes model, the Heston model and the Bates model. They found that there are significant deviations between the two approaches.

Singh and Vipul (2015) tested the performance of Black-Scholes model with the two-scaled realized volatility. Even with the use of high-frequency information, they found that this model is inadequate due to a negative pricing bias.

Jeon et al. (2016) evaluated the option market using GARCH-M model and its high-frequency extension in the Bayesian framework. In an empirical study, they found that their model explains a behavior of option prices close to the expiry.

Li et al. (2017) focused on an analysis of the VIX index. They captured VIX dynamics as a pure jump semimartingale with infinite jump activity and infinite variation.

Risk Management

High-frequency data can be utilized in evaluating systematic risk. Popular risk measures are the value-at-risk and expected shortfall. One of the earliest uses of high-frequency data in value-at-risk forecasting was in Beltratti and Morana (1999).

Giot and Laurent (2004) compared the realized variance ARFIMAX model with daily ARCH model for daily value-at-risk forecasts. They conclude that there is no significant improvement when using realized variance.

Kruse (2006) incorporated the realized variance in value-at-risk estimation by extreme value theory and filtered historical simulation. They found that the best performing forecasting models are hybrid specifications based on realized variance and the filtered historical simulation. Value-at-risk estimates by extreme value theory and filtered historical simulation with realized variance were also analyzed by Louzis et al. (2011).

Clements et al. (2008) compared several models for volatility and quantile forecasts using high-frequency data and found that the HAR model with empirical distribution provides the most accurate forecasts.

McMillan et al. (2008) analyzed intraday periodicity, presence of short horizon as well as long horizon dependencies and daily realized measures in the context of value-at-risk forecasting.

Brownlees and Gallo (2010) compared different volatility measures in the context of value-at-risk forecasting. They found that realized kernels are superior to realized variance, bi-power variation, two-scales realized variance and realized range in terms of value-at-risk predictive ability. Realized range was also used by Shao et al. (2009) in value-at-risk forecasting.

Herrera and Schipp (2013) estimated value-at-risk by extreme value theory in combination with autoregressive conditional duration model.

Huang and Lee (2013) incorporated high-frequency information in value-at-risk forecasting models by combining forecasts based on different intraday intervals and by combining high-frequency information into a single model.

Žikeš and Baruník (2015) modeled value-at-risk by quantile regression with HAR specification of several quadratic variation estimators.

Although the value-at-risk is dominant in high-frequency literature, Guo and Zhang (2008) estimated the expected shortfall using weighted realized variances. Watanabe (2012) utilized realized GARCH model in forecasting value-at-risk and expected shortfall. Bee et al. (2016) used realized extreme value theory for value-at-risk and expected shortfall forecasts.

Value-at-risk can also be modeled at intraday level. Several intraday high-frequency risk measures based on quantiles were proposed by Giot (2005), Giot and Grammig (2006), Dionne et al. (2009), So and Xu (2013), and Banulescu et al. (2016).

Portfolio Optimization

An important area of quantitative finance is portfolio optimization. Fleming et al. (2003) measured the economic value of high-frequency data in the context of investment decisions. Their results indicate that a risk-averse investor would be willing to pay substantial fees to capture the observed gains in portfolio performance.

Bandi et al. (2008b) evaluated the economic benefits of realized variance in dynamic portfolio choice when the prices are contaminated by the market microstructure noise.

De Pooter et al. (2008) found that when forming the mean-variance efficient stock portfolios with daily rebalancing, the optimal sampling frequency for realized variance in the presence of the market microstructure noise ranges between 30 and 65 minutes.

Liu (2009) examined the frequency of portfolio rebalancing at which the use of intraday high-frequency data is beneficial. They found that for monthly rebalancing, the use of daily data is sufficient. However, for daily rebalancing, the use of high-frequency data brings substantial improvements.

Hautsch et al. (2015) analyzed high-dimensional portfolio allocations. They found that the predictions based on high-frequency data yield a significantly lower portfolio volatility than methods employing daily returns.

Trading Strategies

There are many trading strategies whether they belong to the high-frequency trading or operate in longer time horizons. We focus on the pairs trading strategy. It is based on taking advantage of two prices exhibiting strong similarity in the long run that are out of equilibrium. Liu et al. (2017) introduced the doubly mean-reverting processes for capturing the high-frequency price differences and proposed related intraday trading strategies.

Holý and Tomanová (2017) modeled the price differences as Ornstein-Uhlenbeck process and estimated its parameters using high-frequency data contaminated by the market microstructure noise. They found that ignoring the noise would lead to twice as high estimates of volatility and speed of reversion parameters.

Conclusion

Firstly, we showed in a simple simulation that the correct use of high-frequency data improves the accuracy of daily volatility estimation from both the nonparametric and parametric perspective.

Secondly, we reviewed the literature assessing the impact of high-frequency data analysis in derivative valuation, risk management, portfolio optimization and trading strategies. We found that the vast majority of scientific articles reports significant economic gains when incorporating high-frequency data into financial models.

References

AÏT-SAHALIA, Y., MYKLAND, P. A., ZHANG, L. How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *The Review of Financial Studies*. 2005, 18, 2, s. 351-416. ISSN 0893-9454. doi: 10.1093/rfs/hhi016.

AUDRINO, F., FENGLER, M. R. Are Classical Option Pricing Models Consistent with Observed Option Second-Order Moments? Evidence from High-Frequency Data. *Journal of Banking and Finance*. 2015, 61, s. 46-63. ISSN 0378-4266. doi: 10.1016/j.jbankfin.2015.08.018.

BANDI, F. M., RUSSELL, J. R. Microstructure Noise, Realized Variance, and Optimal Sampling. *Review of Economic Studies*. 2008, 75, 2, s. 339-369. ISSN 0034-6527. doi: 10.1111/j.1467-937X.2008.00474.x.

BANDI, F. M., RUSSEL, J. R., YANG, C. Realized Volatility Forecasting and Option Pricing. *Journal of Econometrics*. 2008a, 147, 1, s. 34-46. ISSN 0304-4076. doi: 10.1016/j.jeconom.2008.09.002.

BANDI, F. M., RUSSELL, J. R., ZHU , Y. Using High-Frequency Data in Dynamic Portfolio Choice. *Econometric Reviews*. 2008b, 27, 1-3, s. 163-198. ISSN 0747-4938. doi: 10.1080/07474930701870461.

BANULESCU, D. G. et al. Forecasting High-Frequency Risk Measures. *Journal of Forecasting*. 2016, 35, 3, s. 224-249. ISSN 0277-6693. doi: 10.1002/for.2374.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models. *Journal of the Royal Statistical Society: Series B (Methodological)*. 2002a, 64, 2, s. 253-280. ISSN 1369-7412. doi: 10.1111/1467-9868.00336.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. Estimating Quadratic Variation Using Realized Variance. *Journal of Applied Econometrics*. 2002b, 17, 5, s. 457-477. ISSN 0883-7252. doi: 10.1002/jae.691.

BARNDORFF-NIELSEN, O. E. et al. Designing Realized Kernels to Measure the ex post Variation of Equity Prices in the Presence of Noise. *Econometrica*. 2008, 76, 6, s. 1481-1536. ISSN 0012-9682. doi: 10.3982/ecta6495.

BEE, M., DUPUIS, D. J., TRAPIN, L. Realizing the Extremes: Estimation of Tail-Risk Measures from a High-Frequency Perspective. *Journal of Empirical Finance*. 2016, 36, s. 86-99. ISSN 0927-5398. doi: 10.1016/j.jempfin.2016.01.006.

BELTRATTI, A., MORANA, C. Computing Value at Risk with High Frequency Data. *Journal of Empirical Finance*. 1999, 6, 5, s. 431-455. ISSN 0927-5398. doi: 10.1016/s0927-5398(99)00008-0.

BOLLERSLEV, T., ZHANG, B. Y. B. Measuring and Modeling Systematic Risk in Factor Pricing Models Using High-Frequency Data. *Journal of Empirical Finance*. 2003, 10, 5, s. 533-558. ISSN 0927-5398. doi: 10.1016/s0927-5398(03)00004-5.

BROWNLEES, C. T., GALLO, G. M. Comparison of Volatility Measures: A Risk Management Perspective. *Journal of Financial Econometrics*. 2010, 8, 1, s. 29-56. ISSN 1479-8417. doi: 10.1093/jjfinec/nbp009.

CHRISTOFFERSEN, P. et al. The Economic Value of Realized Volatility: Using High-Frequency Returns for Option Valuation. *Journal of Financial and Quantitative Analysis*. 2014, 49, 03, s. 663-697. ISSN 0022-1090. doi: 10.1017/S0022109014000428.

CLEMENTS, M. P., GALVÃO, A. B., KIM, J. H. Quantile Forecasts of Daily Exchange Rate Returns from Forecasts of Realized Volatility. *Journal of Empirical Finance*. 2008, 15, 4, s. 729-750. ISSN 0927-5398. doi: 10.1016/j.jempfin.2007.12.001.

CORSI, F. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*. 2009, 7, 2, s. 174-196. ISSN 1479-8417. doi: 10.1093/jjfinec/nbp001.

CORSI, F., FUSARI, N., LA VECCHIA, D. Realizing Smiles: Options Pricing with Realized Volatility. *Journal of Financial Economics*. 2013, 107, 2, s. 284-304. ISSN 0304-405X. Doi: 10.1016/j.jfineco.2012.08.015.

DE POOTER, M., MARTENS, M., VAN DIJK, D. Predicting the Daily Covariance Matrix for S&P 100 Stocks Using Intraday Data: But Which Frequency to Use? *Econometric Reviews*. 2008, 27, 1-3, s. 199-229. ISSN 0747-4938. doi: 10.1080/07474930701873333.

DIONNE, G., DUCHESNE, P., PACURAR, M. Intraday Value at Risk (IVaR) Using Tick-by-Tick Data with Application to the Toronto Stock Exchange. *Journal of Empirical Finance*. 2009, 16, 5, s. 777-792. ISSN 0927-5398. doi: 10.1016/j.jempfin.2009.05.005.

ENGLE, R. F. The Econometrics of Ultra-High-Frequency Data. *Econometrica*. 2000, 68, 1, s. 1-22. ISSN 0012-9682. doi: 10.1111/1468-0262.00091.

ENGLE, R. F., RUSSELL, J. R. Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*. 1998, 66, 5, s. 1127-1162. ISSN 0012-9682. doi: 10.2307/2999632.

FLEMING, E., KIRBY, C., OSTDIEK, B. The Economic Value of Volatility Timing Using "Realized" Volatility. *Journal of Financial Economics*. 2003, 67, 3, s. 473-509. ISSN 0304-405X. doi: 10.1016/s0304-405x(02)00259-3.

GIOT, P. Market Risk Models for Intraday Data. *European Journal of Finance*. 2005, 11, 4, s. 309-324. ISSN 1351-847X. doi: 10.1080/1351847032000143396.

GIOT, P., GRAMMIG, J. How Large is Liquidity Risk in an Automated Auction Market? *Empirical Economics*. 2006, 30, 4, s. 867-887. ISSN 03777332. doi: 10.1007/s00181-005-0003-z.

GIOT, P., LAURENT, S. Modelling Daily Value-at-Risk Using Realized Volatility and ARCH Type Models. *Journal of Empirical Finance*. 2004, 11, 3, s. 379-398. ISSN 0927-5398. doi: 10.1016/j.jempfin.2003.04.003.

GUO, M.-Y., ZHANG, S.-Y. Study on CVaR Forecasts Based on Weighted Realized Volatility. In *2008 International Conference on Management Science & Engineering*, s. 91-96, ISBN 978-1-4244-2387-3. doi: 10.1109/icmse.2008.4668899.

HANSEN, P. R., LUNDE, A. Realized Variance and Market Microstructure Noise. *Journal of Business & Economic Statistics*. 2006, 24, 2, s. 127-161. ISSN 0735-0015. doi: 10.1198/073500106000000071.

HANSEN, P. R., HUANG, Z., SHEK, H. H. Realized GARCH: A Joint Model for Returns and Realized Measures of Volatility. *Journal of Applied Econometrics*. 2012, 27, 6, s. 877-906. ISSN 0883-7252. doi: 10.1002/jae.1234.

HAUTSCH, N., KYJ, L. M., MALEC, P. Do High-Frequency Data Improve High-Dimensional Portfolio Allocations? *Journal of Applied Econometrics*. 2015, 30, 2, s. 263-290. ISSN 1099-1255. doi: 10.1002/jae.2361.

HERRERA, R., SCHIPP, B. Value at Risk Forecasts by Extreme Value Models in a Conditional Duration Framework. *Journal of Empirical Finance*. 2013, 23, s. 33-47. ISSN 0927-5398. doi: 10.1016/j.jempfin.2013.05.002.

HOLÝ, V., TOMANOVÁ, P. Ornstein-Uhlenbeck Process Contaminated by the White Noise: Effects, Estimation and Application. 2017. Working Paper.

HUANG, H., LEE, T.-H. Forecasting Value-at-Risk Using High-Frequency Information. *Econometrics*. 2013, 1, 1, s. 127-140. ISSN 2225-1146. doi: 10.3390/econometrics1010127.

JACOD, J. et al. Microstructure Noise in the Continuous Case: The Pre-Averaging Approach. *Stochastic Processes and their Applications*. 2009, 119, 7, s. 2249-2276. ISSN 0304-4149. doi: 10.1016/j.spa.2008.11.004.

JEON, S., CHANG, W., PARK , Y. An Option Pricing Model Using High Frequency Data. *Procedia Computer Science*. 2016, 91, s. 175-179. ISSN 1877-0509. doi: 10.1016/j.procs.2016.07.035.

KENMOE, R. N., SANFELICI, S. An Application of Nonparametric Volatility Estimators to Option Pricing. *Decisions in Economics and Finance*. 2014, 37, 2, s. 393-412. ISSN 1593-8883. doi: 10.1007/s10203-013-0150-1.

KRUSE, R. Can Realized Volatility Improve the Accuracy of Value-at-Risk Forecasts? 2006. Working Paper.

LI, J., LI, L., ZHANG, G. Pure Jump Models for Pricing and Hedging VIX Derivatives. *Journal of Economic Dynamics and Control*. 2017, 74, s. 28-55. ISSN 0165-1889. doi: 10.1016/j.jedc.2016.11.001.

LIU, B., CHANG, L.-B., GEMAN , H. Intraday Pairs Trading Strategies on High Frequency Data: The Case of Oil Companies. *Quantitative Finance*. 2017, 17, 1, s. 87-100. ISSN 1469-7688. doi: 10.1080/14697688.2016.1184304.

LIU, Q. On Portfolio Optimization: How and When Do We Benefit from High-Frequency Data? *Journal of Applied Econometrics*. 2009, 24, 4, s. 560-582. ISSN 0883-7252. doi: 10.2307/40206292.

LOUZIS, D. P., XANTHOPOULOS-SISINIS , S., REFENES, A. P. Are Realized Volatility Models Good Candidates for Alternative Value at Risk Prediction Strategies? 2011. Working Paper.

MCMILLAN, D. G., SPEIGHT, A. E., EVANS, K. P. How Useful Is Intraday Data for Evaluating Daily Value-at-Risk? Evidence from Three Euro Rates. *Journal of Multinational Financial Management*. 2008, 18, 5, s. 488-503. ISSN 1042-444X. doi: 10.1016/j.mulfin.2007.12.003.

NOLTE, I., VOEV, V. Least Squares Inference on Integrated Volatility and the Relationship Between Effcient Prices and Noise. *Journal of Business & Economic Statistics*. 2012, 30, 1, s. 94-108. ISSN 0735-0015. doi: 10.1080/10473289.2011.637876.

SANFELICI, S., UBOLDI, A. Assessing the Quality of Volatility Estimators via Option Pricing. *Studies in Nonlinear Dynamics and Econometrics*. 2014, 18, 2, s. 103-124. ISSN 1081-1826. doi: 10.1515/snde-2012-0075.

SHAO, X. D., LIAN, Y. J., YIN , L. Q. Forecasting Value-at-Risk Using High Frequency Data: The Realized Range Model. *Global Finance Journal*. 2009, 20, 2, s. 128-136. ISSN 1044-0283. doi: 10.1016/j.gfj.2008.11.003.

SINGH, S., VIPUL. Performance of Black-Scholes Model with TSRV Estimates. *Managerial Finance*. 2015, 41, 8, s. 857-870. ISSN 0307-4358. doi: 10.1108/MF-06-2014-0177.

SO, M. K. P., XU , R. Forecasting Intraday Volatility and Value-at-Risk with High-Frequency Data. *Asia-Pacific Financial Markets*. 2013, 20, 1, s. 83-111. ISSN 1387-2834. doi: 10.1007/s10690-012-9160-1.

STENTOFT, L. Option Pricing Using Realized Volatility. 2008. Working Paper.

WATANABE, T. Quantile Forecasts of Financial Returns Using Realized GARCH Models. *Japanese Economic Review*. 2012, 63, 1, s. 68-80. ISSN 1352-4739. doi: 10.1111/j.1468-5876.2011.00548.x.

ZHANG, L., MYKLAND, P. A., AÏT-SAHALIA, Y. A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data. *Journal of the American Statistical Association*. 2005, 100, 472, s. 1394-1411. ISSN 0162-1459. doi: 10.2307/27590680.

ŽIKEŠ, F., BARUNÍK, J. Semi-Parametric Conditional Quantile Models for Financial Returns and Realized Volatility. *Journal of Financial Econometrics*. 2015, 14, 1, s. 185-226. ISSN 1479-8417. doi: 10.1093/jjfinec/nbu029.

JEL Classification: C22, G10.

Acknowledgements: This paper has been prepared under the support of the project of the University of Economics, Prague - Internal Grant Agency, project No. F4/93/2017.Shrnutí

Proč používat vysokofrekvenční data v kvantitativních financích?

Mnoho finančních časových řad jako jsou směnné kurzy cizích měn, ceny akcií a ceny komodit je ve formě vysokofrekvenčních dat. Zaměříme se na literaturu posuzující vliv ekonometrické a statistické analýzy vysokofrekvenčních dat ve financích. Pokryjeme hlavní oblasti kvantitativních financí - oceňování derivátů, řízení rizik, optimalizaci portfolia a obchodní strategie. Převážná většina studií uvádí významné zisky při využívání vysokofrekvenční informace i při rozhodování a operování na nízkofrekvenční úrovni.

Klíčová slova: Vysokofrekvenční data, Oceňování derivátů, Řízení rizik, Optimalizace portfolia, Obchodní strategie.

Measuring dependency between stock behavior and financial news by machine learning methods

Kirill Odintsov

kodintsov@hotmail.com

Doktorand oboru ekonometrie a operační výzkum

Školitel: Pelikan Jan, prof. RNDr., CSc., (jan.pelikan@vse.cz)

Abstract: In this text we showed an end to end application of how to use articles from Reuters to predict future behavior of a stock prices. For that we downloaded around 9 000 000 articles from Reuters. Using "Bag of words" document representation with Latent sematic analysis and K-mean clustering we classified our relevant documents.
Then we created features over the classified documents and using logistic regression on WOE transformed variables we predicted the probabilities that the stock will fall by one or more percent in next 1 and 5 working days.
The methodology described above was demonstrated on Microsoft stock price prediction.

Key words: Stock price behavior prediction, Text mining, Latent sematic analysis, clustering

Introduction

Beginning with the article from H.Markowitz (Markowitz, 1952) the topic of portfolio optimization became very attractive for the researchers. Many advanced and interesting algorithms were developed starting from utilization of higher moments and more advanced risk measures like in (Krokhmal et al, 2002) or (Pflug, Gaivoronski, 2005) to stochastic dominance introduced in (Dantcheva, Ruszczynski, 2003) and continued in (Dantcheva, Ruszczynski, 2006).

All of above mentioned algorithm are computationally very difficult and thus can be applied from only very limited number of stock, but in reality, the number of stocks is very high. In most of the literature it is assumed that the short list of pre-selected stocks is done by some expert. The experts usually use their extensive knowledge of the subject that is sustained by reading the related financial news.

With recent progresses in artificial intelligence and text mining we must ask ourselves whether the machine would not be able to surpass the expert in the matter as the number of news one can potentially read is nearly endless and only machine can truly go through all the news. Twitter, Financial articles from various websites and other texts are freely available to analyze. The idea to utilize text mining to predict stock behavior is not very new either. There are also quite a lot of articles about this topic. For very extensive summary of application for stock price prediction please look at (Beckmann, 2017) and for utilization of text mining for forex prediction please look at (Nassirtoussi et al, 2015).

Another rational behind utilization of text mining for predicting stock behavior is the well-known Efficient market hypothesis presented at (Fama, 1965) which is saying that all the historical prices and their changes of stocks are already reflected in the current stock prices and thus statistical prediction is not as valuable. This raises a question whether even information derived is already in the current stock prices.

In this text we would like to provide and apply simplified end to end approach how to predict negative future behavior of certain stocks. Since this topic is very interdisciplinary we do not aim to provide complete and robust methodology. Our aim is to show the feasibility and demonstrate on simple example. The robust version will be presented as part of my dissertation thesis

In this text we will predict the probability that Microsoft stocks will fall significantly (by more than 0.01 percent in next period) by using only the articles from Reuters (www.reuters.com).

As described in (Beckmann, 2017) and (Nassirtoussi et al, 2015) in most texts the authors do the similar task, but they either already have their document or at least some part of the documents pre-

classified. In cases where the documents are not pre-classified simpler TF-IDF methodology is used. (Fung et al, 2003) uses Singular Value Decomposition (SVD) for classification but they used already pre-classified documents and the SVD is only for classification of bad vs good topic rather than classification of documents.

Methodology presented in this text and shown on Microsoft stock uses Latent sematic analysis (LSA) that is based on SVD. Unlike in other publications we do not have any pre-classified documents as we would like to have no manual labor in predicting of each separate stock from different textual data. We are also using variable clustering on the outputs of LSA and then creating automatically the features based on frequency of articles from each topic. On these features we build 2 WOE logistic regression models predicting the probability that the stocks will decrease in value by at least 1% in next one/five business day/s.

As we already mentioned the task at hand is very interdisciplinary, so we had to use 3 statistical soft wares to leverage the advantage of each. For web scraping we used Python, for text mining R and for the final modelling we used SAS Miner.

Classification of the documents

In this part we will describe everything from acquiring the data to final classification to topics of every document.

We started by getting the data from www.reuters.com. For this we used python libraries “request” – performing HTTP requests and “BeautifulSoup4” for HTML processing. We downloaded all articles extended headlines from last 11 years. For more details about downloaded please refer to Table 1.

Table 1: Loaded data from Reuters Source: Author

Number of articles:	9 164 292
Start date:	01.01. 2007
End date:	27.01. 2018
Days where at least one article:	4 045
Average length of article:	64.7

As you can see from the table one the number of articles was immense, but we needed to take long time period to have enough data for modelling for observing the stocks changes.

Initially we wanted to classify all documents together as in future for multiple stocks it would allow us to use both modelling approaches (the both modelling approaches will be discussed later in the text) where we build automatic models for each stock separately or build one common model for each stock, but as in this case we wanted to demonstrate only on one stock and the computational difficulty was not fully solved yet we limited our data set to only articles about Microsoft. In this case we used simplified pseudo-manual approach where we manually identified that every document that consists at least one of the terms "Microsoft”,” Gates” or “MSFT” is about Microsoft. In the future automatic algorithm should be developed (probably by using text mining on different documents – Owner information, company stock name...) to do this. One note from the author is that this was very surprising as SAS text miner tool has capabilities to handle such problems with ease from author’s experience, but the author had no access to proper license from the SAS text miner tool for this project

To see how data are looking after filtering only Microsoft related documents please refer to Table 2.

Table 2: Only Microsoft data Source: Author

Number of articles:	20315
Start date:	05.01. 2007
End date:	25.01. 2018
Days where at least one article:	2 751
Average length of article	74.7

As we can see on Figure 1 all the Microsoft related documents are more or less equally distributed for all days in our sample and the distribution is stable. If the distribution would significantly change or our documents would be highly concentrated on few days that would be a big problem

Figure 1: Number of news articles about Microsoft daily Source: Author

At this point the problem of falling data quantity for not so well-known stocks about which not many articles are written is obvious. Also, some global events are not even related to Microsoft directly but still can cause their stocks to fall. Both these issues would be solved by working with on all the documents.

After the initial data download we must convert the unstructured data to some structured form. At this point we have chosen the simplified "Bag of words" approach. This approach is simplified as it ignores the ordering of the words and thus do not capture all the information. Please for more information about the "Bag of words" approach and alternatives refer to (Aggarwal, Zhai, 2012). For the "Bag of words" approach the typical data representation is as the literature calls it "Term document matrix" (please refer to (Kumar Paul, 2016)). It is a matrix of *(m*n)* where *m* is number terms and *n* is number of documents. Where the fields of the matrix how many times is *i* term presented in the *j* document.

The "Term document matrix" in this way would be too big and sparse thus we need to normalize our data as described in (Kumar, Paul, 2016). This normalization is done by R package "tm" and it consists of

- Removal of empty spaces, punctuation and other extra symbols,
- Removal of the stop words, which are the words that on they own make no sense. Like for example articles,
- Removal of numbers and case sensitivity,
- Stemming – some words have different forms like "dog", "doggy", "dogs" with quite similar meaning, the stemming joins these words to one
- Removal of the sparse terms

Please for exact algorithms to do above mentioned please refer to (Feinerer, Hornik, 2008). To better understand what kind of terms we have in our documents after the normalization please look Figure 2 that represent by R world cloud the most frequent words in all the documents.

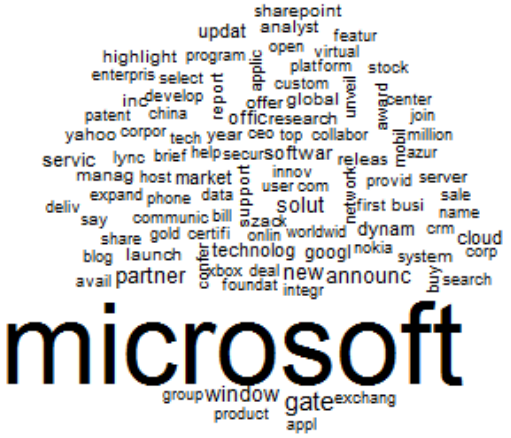


Figure 2: Graphical representation of most common words – after text normalization Source: Author

Even after the normalization the final matrix is fairly big. To reduce dimensionality, we use Latent sematic analysis (LSA). Before the LSA could be applied we needed to apply a transform to our "Term document matrix" to account for Zipf's law. More details about transformation and Zipf's Law can be found in (Kumar Paul, 2016)). The LSA uses basic algebraic principle called Singular value decomposition (that can be found in many algebraic or statistical textbooks) that helps us project our documents (vectors) to lower dimensional spaces while keeping the interrelationship between them. For more information about LSA please refer to (Kumar Paul, 2016)), and (Wild, 2015). In our case we reduced the dimensions to 30. This number was not optimized it was just expert choice that if we have slightly more dimensions it is not a big problem but less dimensions would be problematic.

Theoretically after getting the reduced dimension space we could have already use the coordinates of the documents in the new lower dimensional space as predictors, but as we wanted to have the topics separated we used K-means to cluster the documents by their distance in the new lower dimensional space. For this we chose 15 clusters. This choice should in future be more investigated. Thus, we assigned one topic to each of the documents. It is noteworthy that we do not know what the topics are. But we do not need to know as we will use them in our statistical model that based on the target will decide how they will influence the model. In theory it could be helpful to understand the topics, but we would like more automatized approach and we believe this will not harm the predictability as much.

Prediction of stock behavior

Now that we have all documents classified into 15 topics we need to first create features on them and build a model to know the importance of each feature and being able to predict the behavior of the stock.

If we would have more stocks, we would have two options

- To model on each stock separately
- To build one general model for all of our stocks

The general model approach has a huge advantage that it helps application on less known stocks, stocks with shorter history, the data is not only time series and we will have much more data for modelling allowing us not to over-fit and to use advance machine learning approaches like XGBoost (Extreme Gradient Boosting) and/or Deep Learning which are both heavily data quantity dependent. On the other hand, it brings additional challenges like for example we will need to distinguish between general document and document related to each of the stocks as if we wouldn't do it the prediction will be independent of the stock, which is definitely not any good for stock selection problem.

For separate model for each stock there is a problem of low number of data but on the other hand we can use all created features on all documents as the model can distinguish the important topics. Even though even here some flag whether the document is or is not directly relevant to the stock will probably increase the predictability significantly when used in interaction with other predictors

For the modelling we decided to use Logistic regression on Woe (Weight of evidence) transformed variables this is because we had only around 2764 observation to model on and this includes both train and validation.

Creation of features from the classified document

To get the predictors from the topics we for each day in our interval calculated how many articles from each topic were released on that day, last 5 days and last 30 days. Then we also computed 3 extra predictors of how many news were release about Microsoft that day, last 5 days, last 30 days. As such we have created 48 predictors.

Modeling

For proper modelling we also need the target. For this we downloaded all Microsoft adjusted close prices from 01.01.2007 to 27.01.2018. We had around 2700 days in our sample as stock exchange is closed on some days and, we did not take the 01.01.2007 into consideration as we would not have all predictors for this data. For the days where we had data the data we calculated 2 targets. One is more short term and it was 1 if the stock fell by more than 1% in at the next business day if not it was 0. The second target had almost same definition, but instead of one we used 5 business days.

We defined two targets mainly because we wanted to see long term and short term and midterm influence of article on stock price. We should mention that in most of the literature they look at stock price change within hours from publishing of the article, but sadly we did not have such data

The Stock data were taken directly from Yahoo Finance. For downloading of more stocks web scraping would have to be used.

Please note that our input data ignore the time series nature of the stock and the correlation to other stock indexes. The ideal model would combine even this to the modelling but for now we will ignore this.

Please look at the event rate (occurrence of 1) by different targets in Table 3.

Table 3: Event rate by target Source: Author

Target	**Event rare**
Stock fell next business day	19%
Stock fell in 5 business days	30%

Before the modelling we split randomly our sample into 70% train and 30%. In this case it would be ideal to use only cross validation, and bootstrapping on results, but to simplify we did not do it.

As we want to model probability we chose logistic regression. Before applying the logistic regression, we did the WOE transformation on our variables to get rid of outliers, being able to have nonlinear trend of certain predictors. We could have used full logistic regression (instead of predictors use flags for each of their category) to achieve both benefits, but we do not enough observations for that as this method is more observation demanding. There are other benefits of WOE but they are not so relevant for us in this specific case.

As we have 58 predictors we were faced of how to select proper subset of predictors that would be predictive and does not lead to overfitting. From this we could have used regularization criterions, but we have decided to use “Stepwise” with cross validation error as criterion. The big advantage of this as unlike in standard stepwise that is based on tests that turn wrong in case of multicollinearity the Cross-validation error should not have such issue. So, we didn’t need to deal with multicollinearity problems.

Result

Please see Table 4 for final selected predictors for each target. This result was received after some tuning and improving the model parameters and grouping that we do not feel is necessary to mention for this article.

Table 4: Final features for both models Source: Author

Predicting behavior in next day	Predicting behavior in 5 days
Topic 10 Last 30	Topic 10 Last 30
	Topic 13 Last 1
	Topic 13 Last 30
	Topic 1 Last 30
Topic 7 Last 30	Topic 7 Last 1
Topic 5 Last 1	

We can see that the number of documents with topics 10 and 7 are important for both models. It is very interesting that the aggregation per longer time period tends to be quite strong even for shorter target. This means that the quantity of information is more important that its's freshness for our both targets. This could be quite different if we would have chosen even shorter targets.

Please look at the Table 5 to see what the influence of each topic on our prediction is.

Table 5: The influence of topics on stock return Source: Author

Topic of the article	Influence on stock return
Topic 1	positive
Topic 5	negative
Topic 7	positive in 5 days, negative in 1 day
Topic 10	negative
Topic 13	negative

Please look at the Table 6 to see the final performance of the model. The performance is measured by so called Gini index, which is standard performance measure for logistic regression models. We can see that the both models are over-fitted this is cause by the fact that we have low number of observations. Even after some effort we did not manage to get rid of the overfitting.

Table 6: Performance of the final models Source: Author

	Gini train	Gini validate
Predicting behavior in next day	24	17
Predicting behavior in 5 days	29	11

We can see that the model on shorter targets is much stronger and is less over-fitted.

Conclusion

We showed the whole process of utilization of Reuter's articles to predict the negative stock behavior. We can see that the process truly has some predictive power. The prediction on to next day proven to be stronger but we would not say that this is conclusive result as it was tried on only small data and on one stock. It was quite surprising that even older documents had still value over only using the recent documents.

Since this text is more or less feasibility study and simple demonstration there are quite a lot of problems that should be improved. Namely in text mining part we should figure out how to run on bigger data. What is the optimal number of dimension and of topics? Should other data representation than" Bag of Words" be used? Comparison of other approaches then just LSA should be done. We should find more options how to merge synonyms. We should utilize more data sources like articles from Yahoo finance, tweets and so on. Also, we should be able to automatically assign articles to stock indexes.

In modeling part, we should look closer on feature creation. What would be the result if we would build the model on the aggregated coordinates? How to properly aggregate coordinates. We should try improving feature creation from the classified document. Why did we look only on 1, 5 and 30 days? We should solve how to model with multiple stocks. Should every stock have separate model, or should we have one model for all stocks and the difference of prediction for every stock will be from taking only stock specific document for this stock. If we manage to get more data apply more sophisticated methods than logistic regression.

And most importantly we need to be able to figure out the proper target and utilization in standard portfolio model pre-selection or even in final portfolio model selection.

Literature

AGGARWAL, C.C., ZHAI C. A Survey of Text Classification Algorithms. **In**: AGGARWAL, C.C., ZHAI C. *Mining Text Data*. 2012. Boston: Springer, 2012, 6, pp.163-222. ISBN: 978-1-4614-3223-4.

BECKMANN, M. *Stock price change prediction using news text mining.* Rio Jainero. 2017. Dissertation thesis. Coppe UFRJ

DANTCHEVA, D., RUSZCZYNSKI, A. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization.* 2003, 14, pp. 548-566.

DANTCHEVA, D., RUSZCZYNSKI, A. Portfolio optimization with stochastic dominance constraints. *Journal of Banking and Finance.* 2006, 30, pp. 433-451.

FAMMA, E. The Behavior of Stock-Market Prices. Journal of Business. 1965, 38, 1, pp. 34-105.

FEINERER, I., HORNIK, K., MAYER, D. Text Mining Infrastructure in R. *Journal of Statistical Software.* 2008, 25, 5, s. 1-54. URL: http://www.jstatsoft.org/v25/i05/.

FUNG, G. P. C., YU, J. X., LAM, W. Stock prediction: Integrating text mining approach using real-time news. *Computational Intelligence for Financial Engineering.* 2003, pp. 395-403, DOI: 10.1109/CIFER.2003.1196287.

KROKHMAL**,** P., URYASEV, PALMQUIST, J. Portfolio optimization with conditional Value-at-Risk objective and constraints. *Journal of risk. 2002.*

KUMAR, A., Paul, A. *Mastering Text Mining with R.* Mumbai – Birmingham: Packt Publishing Ltd., 2016. ISBN 978-1-78355-181-1.

MARKOWITZ, H. Portfolio selection. *The Journal of Finance.*1952, 7, pp. 77-91.

NASSIRTOUSSI, A. K., AGHABOZORGI, S., WAH, T.Y., NGO, D. C. L. Text mining pf news-headlines for FOREX market prediction: Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*. 2015, 42, pp. 306-324.

PFLUG, G., GAIVORONSKI, A. Value-at-Risk in portfolio optimization: properties and computational approach. *Journal of risk*. 2005.

JEL Classification: C58.

Shrnutí

Zkoumání závislosti chování akcií na finnačních zprávach s využítím machine learningu

V tomto textu jsme chtěli demonstrovat možnost využití článků z databáze Reuters k predikci cen akcií. K tomuto účelu jsme pomocí tak zvaného web scrapingu stáhli 9 000 000 článků z www.reuters.com. Textové dokumenty jsem pomocí přístupu „Bag of words“ a normalizace textu převedli na řídkou mnoho dimenzionální matici. Dále jsme pomoci Latent sematic analysis, které je založeno na algebraickém Singular value decomposition, redukovali počet dimenzí a tak i řídkost. Naše dokumenty jsme tedy promítli do prostoru o meně dimenzích a dále jsme použili v tomto prostoru shlukování pomocí K-means metodologie.

Z klasifikovaných dokumentů jsme dále vytvořili prediktory, které po WOE transformaci vstupovali do logistické regrese. V této regresi jsme na základě prediktorů predikovali pravděpodobnost, že akcie v horizontu 1 (resp. 5) pracovních dnů klesne na ceně o více než 1 procento.

Celá výše popsaná metodologie byla aplikována na akcie Microsoftu.

Klíčová slova: Predikce chování cen akcií, Text mining, LSA, shluková analýza dokumentů.

Two Heuristics for Vehicle Routing with Three Kinds of Carriers: A Computational Study

Petr Štourač

stop02@vse.cz

PhD student of econometrics

Supervisor: doc. RNDr. Ing. Michal Černý, Ph.D., (cernym@vse.cz)

Abstract: We consider the following version of Vehicle Routing. We are given a family of indivisible batches with known weights located in City 1 (depot), where three carriers are located. The batches are to be delivered into known destinations on a given graph with weighted edges, where the weights correspond to distances. There are no restrictions on delivery times or assumptions on distances. Carrier I has a fleet of vehicles with known capacities which can be used in the TSP-style (i.e., a vehicle can serve a sequence of nodes and returns to City 1) and the price of the delivery is proportional to the total distance traveled by all vehicles. Carrier II works in a similar way, with the exception that he has a single vehicle with an unlimited capacity and that the price is proportional to the distance traveled and weight transported. Carrier III is a kind of `post-office' - the costs are proportional to the weight transported only. The question is how to assign the batches to the carriers in order to minimize overall shipping costs. We design two heuristic algorithms for the problem. We compare their performance to the exact approach: we formulate a natural ILP model and use CPLEX to solve it.

Keywords: Vehicle routing, binary integer programming, heuristic.

Introduction

The „Vehicle Routing Problem“ (VRP) is one of the most widely studied topics in Operations Research. The VRP studies how to serve, from a central depot, a set of customers using a fleet of vehicles with varying capacities. This problem was first formulated in article (Clarke and Wright, 1964, p. 568-581). A methodology for classifying the literature of the Vehicle Routing Problem is presented in (Eksioglu, Vural and Reisman, 2009, p. 1472-1483; Braekers, Ramaekers and Van Nieuwenhuyse, 2016, p. 300-313). An effective heuristic algorithm for the Traveling Salesman Problem is described in (Lin and Kernighan, 1973, p. 498-516), an efficient insertion heuristics for Vehicle Routing Problem are described in (Campbell and Savelsbergh, 2004, p. 369-378).

We consider Vehicle Routing with more kinds of carriers distinguished by different cost functions, the issue of more carriers is also addressed in the article (Chu, 2005, p. 657-667). A carrier is characterized by a fleet of vehicles (with either bounded or infinite capacities) and a cost function. The cost function of a carrier can depend on the distance traveled and on the weights of items to be delivered. Even further factors can play a role in practice, such as the ordering of edges visited or arrival times to nodes, but these are not taken into account in this study (e.g. the problem of the VRP with time windows is discussed in (Baldacci, Mingozzi and Roberti, 2012, p. 1-6; Jiang, Ng, Poh and Teo, 2014, p. 3748-3760)).

This paper is focused on the distinction of carriers according to their cost functions. We consider three types: costs depending only on the distance traveled, costs depending only on the weights loaded in the depot and on a combination of both. This distinction does not play a crucial role in the ILP formulation of the problem (simply, the objective is the sum of the corresponding cost functions), but plays a role in the design of heuristics.

To formalize the problem more precisely, assume that we are given a family of indivisible batches with known weights. The batches are to be delivered into known nodes on a given graph with weighted edges, where weights correspond to distances. There are no assumptions on the distances. In the same node three different carriers are also located. There are no restrictions on delivery times. All batches are located in a City 1 (depot node).

Carrier I has a fleet of vehicles with known capacities. This fleet can be used in the TSP-style (i.e., a vehicle can pass through a sequence of nodes and return to City 1) and the price of the delivery is

proportional to the sum of distances traveled by the vehicles. Carrier II works in a similar way, except for that he has a single vehicle with an unlimited capacity and the price is proportional to the distance traveled plus total weight loaded in the depot. Carrier III is a kind of `post-office' - the costs are proportional to the weight transported only. The goal is to minimize the overall costs.

ILP model

Let $G = \{V;E\}$ be a complete undirected graph with $V = \{1; 2; \dots ; n\}$, where node 1 is the depot.

Input data:

- m = number of batches,
- l = number of vehicles of Carrier I,
- d_{ij} =distance between nodes i, j (generally, the matrix (d_{ij}) need not be symmetric),
- w_k = weight of batch k,
- c_q = capacity of the q-th vehicle of Carrier I,
- p_q = costs per km of the q-th vehicle of Carrier I,
- p_{Carr_II} = costs per km of the vehicle of Carrier II,
- r_{Carr_II} = costs per ton loaded by Carrier II,
- r_{Carr_III} = costs per ton transported by Carrier III,
- $z_{ki} \in \{0; 1\}$, $z_{ki} = 1$ iff node i is the destination for batch k.

Model variables:

- $y_k^q \in \{0; 1\}$, $y_k^q = 1$ iff batch k is transported by vehicle q of Carrier I,
- $x_{i,j}^q \in \{0; 1\}$, $x_{i,j}^q = 1$ iff vehicle q of Carrier I travels from node i to node j,
- $y_k^s \in \{0; 1\}$, $y_k^s = 1$ iff batch k is transported by Carrier II,
- $x_{i,j}^s \in \{0; 1\}$, $x_{i,j}^s = 1$ iff vehicle of Carrier II travels from node i to node j,
- $y_k^t \in \{0; 1\}$, $y_k^t = 1$ iff batch k is transported by Carrier III,
- u_i^q are auxiliary variables in anti-cyclic constraints for Carrier I,
- u_i^s are auxiliary variables in anti-cyclic constraints for Carrier II.

Objective function:

$$\min \sum_{q,i,j} p_q d_{ij}^q x_{ij}^q + p_{Carr\_II} \sum_{i,j} d_{ij}^s x_{ij}^s + r_{Carr\_II} \sum_k w_k y_k^s + r_{Carr\_III} \sum_k w_k y_k^t \qquad (1)$$

The objective is to minimize the sum of cost functions of the three carriers.

Constraints:

$$\sum_i x_{ij}^q = \sum_i x_{ji}^q \qquad \forall q, j, \qquad (2)$$

$$\sum_k x_{ij}^s = \sum_i x_{ji}^s \qquad \forall j, \qquad (3)$$

$$\sum_k w_k y_k^q \leq c_q \qquad \forall q, \qquad (4)$$

$$\sum_q y_k^q + y_k^s + y_k^t = 1 \qquad \forall k, \qquad (5)$$

$$\sum_{j,j\neq i} x_{ji}^{q} \leq \sum_{k} y_{k}^{q} z_{ki} \qquad \forall q, \forall i \geq 2, \qquad (6)$$

$$\sum_{j,j\neq i} x_{ji}^{s} \leq \sum_{k} y_{k}^{s} z_{ki} \qquad \forall i \geq 2, \qquad (7)$$

$$\sum_{j,k,j\neq i} z_{ki} x_{ji}^{q} \geq \sum_{k} y_{k}^{q} z_{ki} \qquad \forall q, \forall i \geq 2, \qquad (8)$$

$$\sum_{j,k,j\neq i} z_{ki} x_{ji}^{s} \geq \sum_{k} y_{k}^{s} z_{ki} \qquad \forall i \geq 2, \qquad (9)$$

$$\sum_{j} x_{1j}^{q} \leq 1 \qquad \forall q, \qquad (10)$$

$$u_{i}^{q} - u_{j}^{q} + n x_{ij}^{q} \leq n-1 \qquad \forall q, i, \forall j \geq 2, i \neq j, \qquad (11)$$

$$u_{i}^{s} - u_{j}^{s} + n x_{ij}^{s} \leq n-1 \qquad \forall i, \forall j \geq 2, i \neq j, \qquad (12)$$

Constraint (2) states that a vehicle of Carrier I must enter and leave a node. Constraint (3) states same for the vehicle the Carrier II. Inequality (4) assures that capacity of *q*-th vehicle of Carrier I is not exceeded. Equation (5) tells that all batches must be delivered. Constraint (6) prevents the *q*-th vehicle of Carrier I from entering a node if the node is not served by the vehicle. Similarly, constraint (7) states the same for Carrier II. Conversely, constraints (8) and (9) require entry to node by vehicle if this vehicle serves the node for both carriers I and II. At least one leave the City 1 is possible for each vehicle of the carriers I state (10). Anti-cyclic conditions are in (11) and (12).

Heuristics

In this section, we design two heuristics for the problem. Both the heuristics take batches one by one and consecutively assign them to carriers and vehicles, while constructing the routes for vehicles of the distance-aware carriers. In each iteration, both the heuristics assign at least one batch unassigned so far and update the so-far-constructed routes. The difference is that the first heuristic select the first undelivered batch from the list to be assigned, while the second takes all unassigned batches into account in every iteration.

For Carrier I and Carrier II, a kind of the Insert Heuristic is used to update the route. Carrier III need not care about the vehicle position: we can assume that his batches are delivered directly from the depot.

For the description of the heuristics, the following notation will be useful.

Route: The route is a finite sequence of nodes from V where the first and last node is the depot.

- $(a_1^q, a_2^q, \cdots, a_{N_q}^q \equiv a_1^q)$ = route of the *q*-th vehicle of Carrier I, and,
- $(b_1, b_2, \cdots, b_M \equiv b_1)$ = route of the vehicle of Carrier II.

Delivered batches: The delivered batches is a set of batches which are delivered by a vehicle.

- Y^q = delivered batches by the *q*-th vehicle of Carrier I,
- Y^s = delivered batches by the vehicle of Carrier II, and
- Y^t = delivered batches by the vehicle of Carrier III.

Configuration: The configuration is a set of current routes and current assigned batches of all vehicles.

Cost: The cost is the lowest cost of delivering selected batch k with respect the current configuration.

- μ_k^q = the lowest cost of inserting batch *k* to the route of *q*-th vehicle of Carrier I,
- v_k = the lowest cost of inserting batch *k* to the route of the vehicle of Carrier II,
- ω_k = the cost of using the vehicle of Carrier III for delivering batch *k*, and
- σ = the lowest cost of delivering at a given configurations.

The actual $\mu_k^q, v_k, \omega_k \; and \; \sigma$ are to be specified for each of the heuristics individually.

Heuristic I

The batches are processed one-by-one in an initial ordering. For each batch, the cheapest way of shipping is found based on the routes of individual carrier vehicles planned so far. In case that the batch is transported by vehicle of the Carrier I, if possible, additional batches that have not yet been delivered to the destination are added to the car. The heuristic ends after processing the last undelivered batch from the list.

Step 1: Initialize start parameters. All routes are initialized to the value (1,1), i.e start and stop node must be City 1 (depot).

Step 2: Calculate delivery costs of batch *k* for every vehicle. A batch k is selected from the list of the unassigned batches. The costs for each vehicle are calculated as follows:

- $\mu_k^q = \infty \; if \; w_k > c_q - \sum_{l,l \in Y^q} w_l, else \; \mu_k^q = \min_{a_j} \left\{ p_q \left(d_{a_j i_k} + d_{i_k a_{j+1}} - d_{a_j a_{j+1}} \right) \right\}$ where $a_j \in \left(a_j^q\right) \wedge j < N_q, i_k$ = the destination node for batch *k*,
- $v_k = \min_{b_j} \left\{ p_{Carr\_II} \left(d_{b_j i_k} + d_{i_k b_{j+1}} - d_{b_j b_{j+1}} \right) + w_k r_{Carr\_II} \right\}$ where $b_j \in \left(b_j\right) \wedge j < \mathrm{M}, i_k$ = the destination node for batch *k*,
- $\omega_k = w_k r_{Carr\ III}$.

Step 3: Choose the carrier's vehicle with minimal costs. The costs are calculated as follows:

- $\sigma = \min_q \left\{ \mu_k^q, v_k, \omega_k \right\}$.

For the minimizing carrier and vehicle, do the following:

Carrier I: add additional undelivered batches to the vehicle for the destination (if it can carry), add another node to the vehicle route $\left(a_j^q\right)$, mark laden batches as delivered. Update the set of delivered batches Y^q for the vehicle *q* with laden batches.

Carrier II: add another node to the vehicle route $\left(b_j\right)$, mark laden batch as delivered. Update the set of delivered batches Y^s with laden batches.

Carrier III: mark laden batch as delivered. Update the set of delivered batches Y^t with laden batches.

If there is still an undelivered batch, take the first one and proceed to Step 2, otherwise the processing will stop.

Heuristic II

This heuristic looks for the batch with the cheapest possible transport among the batches unallocated so far. The calculation is based on the actual configuration. In case the batch is transported by a vehicle of Carrier I, additional batches that have not yet been delivered (if any) to the destination are added to the vehicle. In case the batch is transported by a vehicle of Carrier II, additional batches with the same destination are added to the vehicle under the condition that their transport is cheaper than if they were transported by Carrier III. The heuristics ends after processing the last batch.

Step 1: Initialize start parameters. All routes are initialized to the value (1,1), i.e start and stop node must be City 1 (depot).

Step 2: Calculate delivery costs for every vehicle. The costs for each vehicle and each undelivered batch are calculated as follows:

- $\mu_k^q = \infty \; if \; w_k > c_q - \sum_{l,le Y^q} w_l, else \; \mu_k^q = \min_{a_j} \left\{ p_q \left(d_{a_j i_k} + d_{i_k a_{j+1}} - d_{a_j a_{j+1}} \right) \right\}$ where $a_j \in \left(a_j^q\right) \wedge j < N_q, i_k$ = the destination node for batch *k*,
- $v_k = \min_{b_j} \left\{ p_{Carr\_II} \left(d_{b_j i_k} + d_{i_k b_{j+1}} - d_{b_j b_{j+1}} \right) + w_k r_{Carr\_II} \right\}$ where $b_j \in \left(b_j\right) \wedge j < \mathrm{M}, i_k$ = the destination node for batch *k*,
- $\omega_k = w_k r_{Carr\ III}$.

For each vehicle minimal costs are selected over all undelivered batches:

- $\mu_{k_I}^q = \min_k \{\mu_k^q\}$ where the decision on the minimum is made on the basis of expression μ_k^q / w_k,
- $v_{k_{II}} = \min_k \{v_k\}$,
- $\omega_{k_{III}} = \min_k \{\omega_k\}$.

Step 3: Choose the carrier's vehicle with minimal costs. The costs are calculated as follows:

- $\sigma = \min_q \left\{ \mu_{k_I}^q, v_{k_{II}}, \omega_{k_{III}} \right\}$.

For the minimizing carrier and vehicle, do the following:

Carrier I: add additional undelivered batches to the vehicle for the destination (if it can carry), add another node to the vehicle route $\left(a_j^q\right)$, mark laden batches as delivered. Update the set of delivered batches Y^q for the vehicle *q* with laden batches.

Carrier II: add additional undelivered batches to the destination on the condition that their delivery is cheaper than if they were delivered by Carrier III, add another node to the vehicle route $\left(b_j\right)$, mark laden batch as delivered. Update the set of delivered batches Y^s with laden batches.

Carrier III: mark laden batch as delivered. Update the set of delivered batches Y^t with laden batches.

If there is still an undelivered batch, take the first one and proceed to Step 2, otherwise the processing will stop.

Performance of the heuristics

The model and both heuristics were tested on six datasets. Their parameters are described in Table 1. The samples are intentionally small so that we can get either an optimum solution or a tight lower bound from a solver (we used CPLEX here) and compare it to the output of the heuristic methods.

We used a computer with the following parameters: Intel Core i5-4200U CPU @ 1.60 GHz 2.30 GHz, 4 GB RAM, Windows 10 (x64). The ILP model is solved by CPLEX 12.6 and both heuristics are implemented in Python 3.4.

Results summarized in Table 2. Note that Heuristic I is sensitive to the initial ordering of batches, which also affects the search for the routes of vehicles.

Table 1. Summary data for test instances. The columns stand for the number of nodes, number of batches, number of vehicles of Carrier I, number of variables and number of constraints of the ILP model. In Ex. 1, batches are sorted at random, in Ex. 1-v2, batches are sorted in the descending order according to the weights, and in Ex. 1-v3, batches are sorted by distances to the target node. Ex. 2 v2 is a modification of Ex. 2 with lower per-km costs of Carrier III.

	vert's	batches	No. I	var's	constr's
Example 1	5	10	2	115	96
Example 1 v2	5	10	2	115	96

	vert's	batches	No. I	var's	constr's
Example 1 v3	5	10	2	115	96
Example 2	10	20	5	740	693
Example 2 v2	10	20	5	740	693
Example 3	30	40	8	8 500	8 446

Table 2. Results of experiments: the optimal values of objective function found by CPLEX (in Ex. 3, only a lower bound was available due to excessive memory requirements), the values found by Heuristics I and II and the corresponding ratios (HI/C, HII/C, HI/HII).

	Cplex	Heur. I	HI/C	Heur. II	HII/C	HI/HII
Example 1	68 100	71 343	1.05	70 283	1.03	0.99
Example 1 v2	68 100	74 833	1.10	70 283	1.03	0.94
Example 1 v3	68 100	73 221	1.07	70 283	1.03	0.96
Example 2	128 187	162 366	1.27	133 790	1.04	0.82
Example 2 v2	67 988	78 581	1.16	70 928	1.04	0.90
Example 3	≈334 237	360 311	1.08	346 349	1.04	0.96

Example 1 in detail

In Example 1 we have $n = 5$ (number of nodes), $m = 10$ (number of batches), $l = 2$ (number of vehicles of Carrier I), the distance matrix $D = (d_{ij})$ has the form

$$D = \begin{pmatrix} 0 & 13 & 9 & 23 & 17 \\ 13 & 0 & 16 & 8 & 25 \\ 9 & 16 & 0 & 30 & 28 \\ 23 & 8 & 30 & 0 & 15 \\ 17 & 25 & 28 & 15 & 0 \end{pmatrix},$$

and parameters of batches are as follows:

Batch No.	1	2	3	4	5	6	7	8	9	10
Weight	20	35	19	57	74	10	8	11	39	52
Target node	2	4	3	4	3	4	5	5	3	2

Capacities c_1; c_2 of the vehicles of Carrier I are 50; 100, and their per-km costs p_1; p_2 are 6; 12, respectively. Further we have $p_{Carr\_II} = 11$ (per-km costs of Carrier II), $r_{Carr\_II} = 380$ (per-ton costs of Carrier II) and $r_{Carr\_III} = 450$ (per-ton costs of Carrier III).

The results of Example 1 are presented in Table 3 and in Figure .

Table 3. Results (Example 1): The ratio of the weight of batches transported by vehicles of Carrier I to its capacity found by CPLEX and Heuristics I and II.

Filling	Cplex	Heur. I	Heur. II
Carrier I - Vehicle 1	50/50	49/50	50/50
Carrier I - Vehicle 2	100/100	100/100	93/100

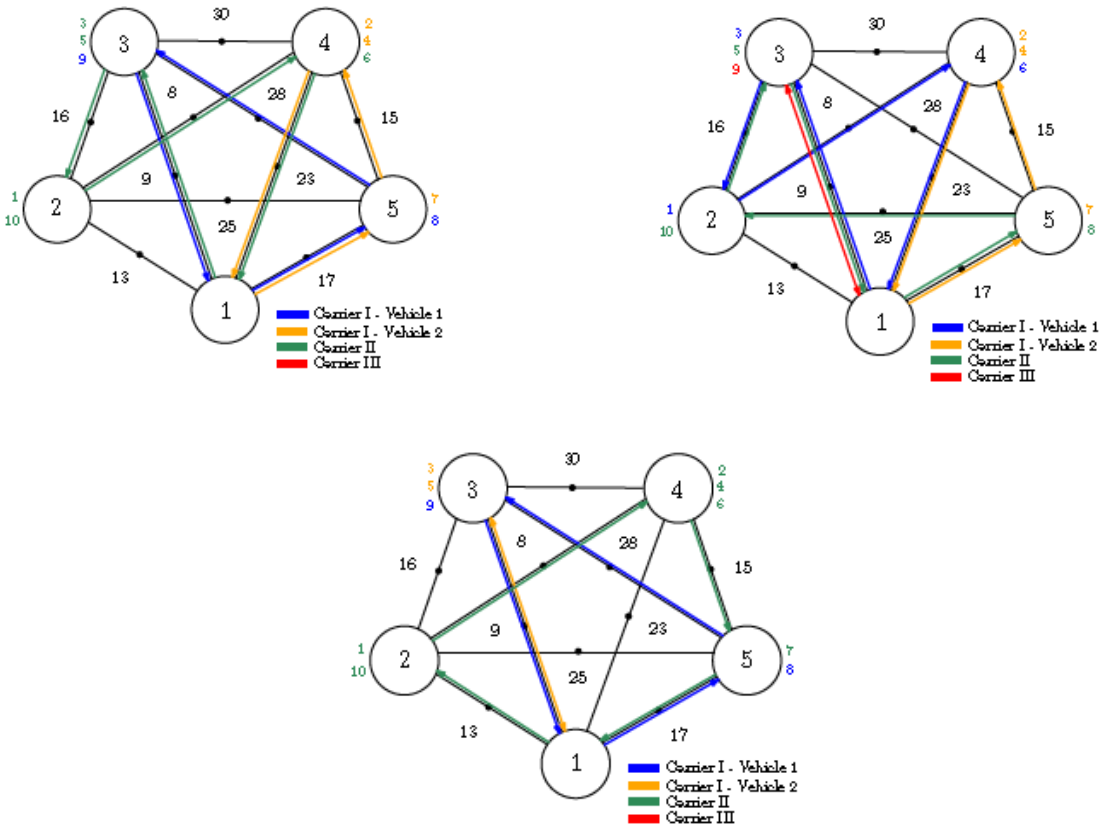


Figure 1. Results (Example 1): The optimum solution found by CPLEX, Heuristic I and Heuristic II.

Conclusions

We considered the case of vehicle routing where carriers differ by their cost functions. We considered three types of carriers, whose costs depend only on the distance traveled, only on the weight transported and on both. We formulated the problem as an ILP and showed that even small instances are not solvable exactly (which is not surprising); therefore we designed heuristic approaches. The heuristics utilize the particular forms of cost functions of the carriers (but note that they can be useful for a broader class of cost functions). It turns out that Heuristic II outperforms Heuristic I and, on average, Heuristic II constructs feasible solution just about 3 % - 4 % worse compared to the true optimal value (at least with the tested instances where the optimum solution could be found).

References

Baldacci, R., Mingozzi, A., and Roberti, R., 2012. *Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints*. European Journal of Operations Research. Vol. 218. ISSN 0377-2217.

Braekers, K., Ramaekers, K., and Van Nieuwenhuyse, I., 2016. *The vehicle routing problem: State of the art classification and review*. Computers & Industrial Engineering. Vol. 99. ISSN 0360-8352.

Campbell, A.M., and Savelsbergh, M., 2004. *Efficient insertion heuristics for vehicle routing and scheduling problems*. Transportation Science. Vol. 38 (3). ISSN 0041-1655.

Chu, Ch-W., 2005. *A heuristic algorithm for the truckload and less-than-truckload problem*. European Journal of Operations Research. Vol. 165. ISSN 0377-2217.

Clarke, G., and Wright, J.W., 1964. *Scheduling of Vehicles from a Central Depot to a Number of Delivery Points*. Operational research. ISSN 1109-2858.

Eksioglu, B., Vural, A. V., and Reisman, A., 2009. *The vehicle routing problem: A taxonomic review*. Computers & Industrial Engineering. Vol. 57. ISSN 0360-8352.

Jiang, J., Ng, K. M., Poh, K. L., and Teo, K. M., 2014. *Vehicle routing problem with a heterogeneous fleet and time windows*. Expert Systems with Applications. Vol. 41. ISSN 0957-4174.

Lin, S., and Kernighan, B. W., 1973. *An Effective Heuristic Algorithm for the Traveling Salesman Problem*. Operations Research. Vol. 21. ISSN 1109-2858.

JEL Classification: C61

Shrnutí

Dvě heuristiky pro rozvozní úlohu se třemi druhy dopravců: výpočetní studie

Byl uvažován případ rozvozní úlohy, kde se dopravci liší podle nákladových funkcí. Byli zvažováni tři druhy dopravců, jejichž náklady závisí pouze na ujeté vzdálenosti, pouze na přepravované hmotnosti a na obou typech nákladů. Byl formulován problém jako ILP a bylo ukázáno, že ani malé případy nejsou přesně řešitelné (což není překvapující); proto byly navrženy heuristické přístupy. Heuristiky využívají konkrétní formy nákladových funkcí dopravců (ale je třeba poznamenat, že mohou být využitelné pro širší třídu nákladových funkcí). Ukazuje se, že Heuristika II je lepší než Heuristika I a v průměru Heuristika II konstruuje reálné řešení jen asi o 3 % až 4 % horší ve srovnání se skutečnou optimální hodnotou (alespoň u testovaných případů, kdy bylo možné najít optimální řešení).

Klíčová slova: rozvozní úloha, binární celočíselné programování, heuristika.

Time-varying quantiles and their applications

Petra Tomanová

petra.tomanova@vse.cz

Ph.D. student of Econometrics and Operations Research

Advisor: prof. RNDr. Ing. Michal Černý, Ph.D., (cernym@vse.cz)

Abstract: This paper reviews methods for modeling and estimating time-varying quantiles and summarizes their applications. We focus on methods based on quantile regression, especially on Conditional autoregressive value-at-risk (CAViaR) models and their modifications and briefly discuss methods based on copula functions and others alternatives for time-varying quantiles modelling. Further, models utilizing high-frequency information in the form of realized measures are reviewed. Well-documented applications such as the value-at-risk estimation, measuring co-movements, testing for the financial contagion and the asset allocation problem are discussed.

Key words: financial contagion, high-frequency data, time-varying quantiles, value-at-risk.

Introduction

In finance, the essential question is: *Are past returns good proxy for future returns?* Mutual fund disclaimer that past performance does not guarantee future result. On the other hand, if we are unwilling to make any assumptions then uncertainty of returns is unmeasurable and we cannot assess important market features such as market risk. So, what are reasonable assumptions regarding asset returns? Cont (2001) discussed stylized facts of asset returns, i.e. empirical properties common across many assets, markets and time periods and observed by independent studies. For example, author pointed out that the density of the returns tends to be sharp peaked and heavy tailed and student distribution with four degrees of freedom displays a tail behavior similar to many asset returns. Once the Gaussian assumption is dropped, the question arises as to why the focus should be on mean and variance. Admittedly, the mean is rather basic, but the attraction of the variance is limited because attention is typically on certain quantiles or indeed the whole distribution (Harvey, 2013).

As authors De Rossi et al. (2006) pointed out, in a cross-section, the changes in sample quantiles over time are not difficult to estimate and they are easily interpretable. On the other hand, in time series analysis, the estimation of time-varying quantile becomes much more difficult exercise but still of a considerable practical importance. One of the reasons for the interest in quantiles is that they define, or help to define, certain measures of risk. For example, widely used value at risk (VaR) is simply a particular quantile of future portfolio values, conditional on current information. The conditional time-varying quantile q_t for a given confidence level $\theta \in (0,1)$ and for a random variable y_t at time t conditional on $\mathcal{F}_{t-1}$ is defined as

$$Pr[y_t \leq q_t | \mathcal{F}_{t-1}] = \theta.$$

For tabulated distributions, such as the normal and t, the quantiles can be calculated directly and the one-step-ahead conditional distribution is at hand. The quantile function for a given distribution function, $F(y_t)$, is $F^{-1}(\theta)$. Multistep predictive distributions can be simulated, and hence quantiles can be calculated to a required degree of precision (Harvey, 2013).

What happens if we are not willing to assume a specific distribution? An approach to relaxing the dependence on distribution assumptions is to develop nonparametric methods for time series data. Harvey (2013) discussed the model based on kernels, where the whole distribution is tracked as it changes over time, and at the same time, features of the distribution, such as quantiles, can be extracted. However, proceeding in this way raises various issues.

In order to obtain a more precise description of time series than non-parametric can provide we will sometimes resort to semi-parametric methods which, without completely specifying the form of the

price process, imply the existence of a parameter which describes a property of the process, e.g. the most of methods based on quantile regression.

In this paper, we strive to review and summarize the methods which can be used for time-varying quantile estimation, either directly by specifying the form how the quantiles evolve in time or indirectly by assuming specific distribution, estimating its parameters and computing quantiles by inverting the distribution function. The difficulty of latter category of approaches lies in the assumption of specific distribution because financial returns usually exhibit volatility clustering, peaked and fat tailed distribution and marginal skewness (Huang et al., 2010). Further, we focus on those methods which utilize the high-frequency data to get more precise estimators of time-varying quantiles.

Generally, time-varying quantiles describe movements, dispersion and asymmetry in a time series and thus they are useful for a wide range of application, especially in finance. One of the most well-known and straightforward application is the value-at-risk estimation, computing risk measures for risk management in general. Secondly, time-varying quantiles can be used for an estimation of co-movements between time-series and consequently in an asset allocation problem. We discussed those application in the penultimate section. Last section concludes.

Methods

In this section, we review and summarize methods for modeling and estimating time-varying quantiles. First, we focus on methods based on quantile regression, especially on Conditional autoregressive value-at-risk (CAViaR) models and their modifications and then briefly discuss methods based on copula functions and others alternatives for time-varying quantiles modelling.

The quantile regression models

The quantile regression (Koenker et al., 1978, 2006) solution involved the criterion function that, when minimized, returned the 'optimal' quantile regression estimator. This criterion required no distributional assumption, as such quantile regression is often regarded as a non-parametric method, despite mostly parametric forms being assumed for the regression relationship between the response and the regressors; most examples of its use thus make it semi-parametric in nature (Gerlach et al., 2011).

The general dynamic quantile regression problem may be specified as

$$y_t = f_t(\boldsymbol{\beta}, \boldsymbol{x_{t-1}}) + u_t,$$

where y_t is a dynamic observation at time t, $\boldsymbol{x_{t-1}}$ is a set of explanatory variables, which could include lagged values of the response y_{t-k}, $k > 0$, $\boldsymbol{\beta}$ are the unknown parameters and u_t is an unknown error term, with a generally unspecified distribution. The conditional quantile at probability level θ, is then

$$q_\theta(y_t|\boldsymbol{\beta}, \boldsymbol{x_{t-1}}) = f_t(\boldsymbol{\beta_\theta}, \boldsymbol{x_{t-1}}),$$

where $\boldsymbol{\beta_\theta}$ is the solution to

$$\min_{\boldsymbol{\beta}} \sum_t \rho_\theta\big(y_t - f_t(\boldsymbol{\beta}, \boldsymbol{x_{t-1}})\big).$$

The function $\rho(\cdot)$ is a loss function, usually specified as $\rho_\theta(e) = e\big(\theta - I(e < 0)\big)$. The function $f_t(\cdot)$ defines the dynamic link between the response y_t and the explanatory information $\boldsymbol{x_{t-1}}$ (Gerlach et al., 2011).

Quantile regression models have several advantages. Quantile regression estimators may be more efficient than least-square estimators when error terms are not gaussian. The entire conditional distribution can be characterized and different relationships between the regressor and the dependent variable may arise at different quantiles (Buchinsky, 1998).

In a frequently cited paper Engle et al. (2004), authors proposed conditional autoregressive value-at-risk via regression quantiles (CAViaR) models. The tail quantile is estimated by a quantile regression and then the one-step ahead prediction is performed. Unlike the linear quantile autoregression models, CAViaR models are able to capture the behavior of quantiles in returns as nonlinear functions of past observations.

A generic CAViaR model can be specified as

$$q_\theta(y_t|\boldsymbol{\beta}) = \beta_{\theta,0} + \sum_{k=1}^{v} \beta_{\theta,k} q_\theta(y_{t-k}|\boldsymbol{\beta}) + \sum_{l=1}^{w} \beta_{\theta,v+l} f(\boldsymbol{x_{t-l}})$$

where $f(\boldsymbol{x_{t-l}})$ is a function of a finite number of lagged observable variables of order l and $q_\theta(y_t|\boldsymbol{\beta})$ is an empirical specification for the q_θ where $\boldsymbol{\beta}$ is the vector of parameters of the required dimension. Autoregressive terms $\beta_{\theta,k} q_\theta(y_{t-k}|\boldsymbol{\beta})$, $k = 1, \dots, v$, ensure that the quantile changes smoothly over time and function $f(\boldsymbol{x_{t-l}})$ links $q_\theta(y_t|\boldsymbol{\beta})$ to observable variables that belong to the information set.

A several practical issues arise when the CAViaR approach is used. The most severe problem with CAViaR approach is the so-called quantile crossing problem, i.e. violation of monotonicity assumption. The fundamental one is choosing suitable specification of the model. Authors Engle et al. (2004) proposed an in-sample and out-of-sample dynamic quantile test which helps to select the most suitable functional form. Figure 1 shows the estimated time-varying quantiles using equity returns on Morgan Stanley Capital International (MSCI) index for the Netherlands and $\theta = \{0.10, 0.25, 0.75, 0.90\}$.

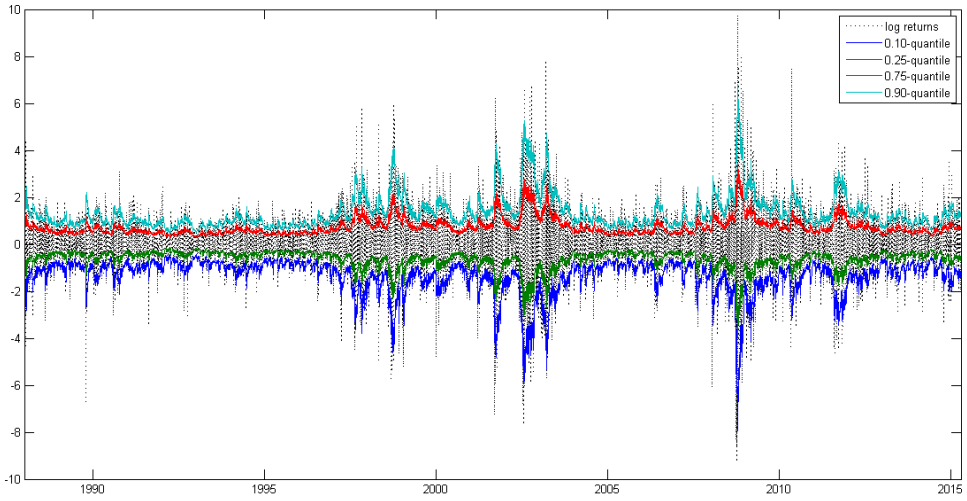


Figure 1: Equity returns on MSCI index for the Netherlands (dotted line) and estimated time-varying quantiles for $\boldsymbol{\theta = \{0.10, 0.25, 0.75, 0.90\}}$.

Kuester et al. (2006) compared the out-of-sample performance of several methods for predicting VaR with CAViaR method using more than 30 years of the daily return data on the NASDAQ Composite Index. Authors reported that CAViaR models perform inadequately, though an extension to a particular CAViaR model outperformed the others. Pilar et al. (2014) reported that asymmetric extensions of the CaViaR method provide promising results for VaR prediction.

Huang et al. (2010) formulated a time-varying CAViaR model whose parameters vary according to the evolution of the returns on an index. White et al. (2015) proposed a vector autoregressive extension to CAViaR models, so framework simultaneously accommodates models with multiple random variables, multiple confidence levels, and multiple lags of the associated quantiles.

Besides CAViaR models, the quantile regression was used for a time-varying quantile single-index model estimation (Li et al., 2018). The model is suitable for complex data, in which the responses and covariates are longitudinal/functional, with measurements taken at discrete time points. Bouyé et al. (2009) proposed a general approach to nonlinear quantile regression modelling that is based on the specification of the copula function that defines the dependency structure between the variables of interest.

The copula function

An extensive survey on time-varying copulas can be found in the paper of Manner et al. (2012), as well as in the survey of Patton (2012). The important issue is to choose the right copula family and use a flexible specification for the time-variation of the dependence parameter otherwise the measure such that VaR might be inaccurate. We mentioned several dynamic copula models that might be suitable for time-varying quantiles estimation.

Harvey (2008) proposed a method for estimating time-varying quantiles based on copula theory. The methodology of changing copulas is based on a pre-filtering and construction of indicator variables. Estimates have different properties than previous methods but the advantage of the methodology is that it is less computationally demanding. However, it is restricted to stationary or close to stationary time series. Despite this drawback it is suitable for wide range of application, for example in returns analyses.

Ye et al. (2017) extend a copula-based quantile-specific conditional association regression model proposed in the paper of Li et al. (2014). The model centers around the quantile-specific odds ratio which computes the ratio that two random variables are simultaneously below their respective quantile.

Chen et al. (2009) proposed parametric copulas for quantile autoregressive models for nonlinear time-series. Bernardi et al. (2015) introduce an approach based on a Bayesian quantile regression framework, Markov chain Monte Carlo algorithm, exploiting the representation of the asymmetric Laplace distribution as a location-scale mixture of normals.

Other methods

De Rossi et al. (2006, 2009) proposed a modified state space signal extraction algorithm for fitting a time-varying quantile to a sequence of observations when a time series model for the corresponding population quantile is formulated. Estimated quantiles are useful not only for description of the time series but also for predictions. One of the suitable properties is that time-varying quantiles have the appropriate number of observations above and below. Moreover, the quantile crossing problem is not issue here.

Nonparametric techniques based on kernel and weighted observations using schemes derived from time series modeling was proposed in the paper of Harvey et al. (2009). A probability density function is estimated by maximum likelihood. The tricky part is bandwidth selection which is thoroughly describe in paper. Authors stated that the technique should be combined with filter for scale and/or location since tracking the distribution is only viable when it changes relatively slowly. The advantage is that the estimating the cumulative distribution function converges relatively fast.

GARCH-type models with parametric specified error distributions are also very useful in this area (Kuester et al., 2006). Further, we focus rather on realized GARCH models and other models utilizing high-frequency data since the realized measures are proven to contribute to accuracy of the estimators.

High-frequency data

The methods which utilize realized volatility measure for either parametric or nonparametric estimation of conditional distributions have been proposed in the literature, e.g. Andersen et al. (2003), Giot et al. (2004), and Clements et al. (2008). Brownlees et al. (2009), Shephard and et al. (2009), and Maheu et al. (2011) introduced methodologies for predictions of density using parametric return-based volatility models.

Žikeš et al. (2014) investigated how the conditional quantiles of future returns and volatility of financial assets vary with various measures of ex post variation in asset prices and option-implied volatility. By relying on nonparametric measures of volatility the restrictive assumptions on the dynamics of the underlying conditional distributions are not needed. Authors showed that simple linear quantile regressions for returns and heterogenous quantile autoregressions for realized volatility perform very well in capturing the dynamics of the respective conditional distributions, both in absolute terms as well as relative to the CAViaR and the Autoregressive fractionally integrated moving average (ARFIMA)-based lognormal-normal mixture of Andersen et al. (2003).

The CAViaR models can be easily extended to utilize high-frequency information by adding realized measures into quantile specification (Tomanová, 2017).

The following alternatives to quantile estimation utilizing high-frequency information can be found in the literature. Avdulaj et al. (2015) combined generalized autoregressive score copula functions with high frequency data that allows them to capture and forecast the conditional time-varying joint distribution of the oil-stocks pair. Banulescu et al. (2013) proposed intraday high frequency risk measures for VaR in the case of irregularly spaced high-frequency data and later, they proposed methodology for their forecasting (Banulescu et al., 2016). Bee et al. (2016) introduced a two-step approach where returns are first pre-whitened with a high-frequency based volatility model, and then an extreme-value-theory-based model is fitted to the tails of the standardized residuals.

Beltratti et al. (1999) compared GARCH and FIGARCH models and computed the multi-horizon volatility forecasts for periods up to 10 days. Watanabe (2012) applies the realized GARCH model, which incorporates the GARCH model with realized volatility, to quantile forecasts of financial returns. Louzis et al. (2011) assessed the VaR prediction accuracy and efficiency of six ARCH-type models, six realized volatility models and two GARCH models augmented with realized volatility regressors. Dionne et al. (2009) proposed a log-ACD–ARMA–EGARCH model which specify the joint density of the marked point process of durations and high-frequency returns.

Conditional VaR might be calculated by utilizing weighted realized volatility (Guo et al., 2008) or quantile regression (Taylor, 2017). Huang et al. (2013) examine methods that incorporate the high frequency information either indirectly, through combining forecasts (using forecasts generated from returns sampled at different intraday interval), or directly, through combining high frequency information into one model. Authors consider subsample averaging, bootstrap averaging, forecast averaging methods for the indirect case, and factor models with principal component approach, for both direct and indirect cases.

Kruse (2006) combines extreme value theory and filtered historical simulation with Autoregressive fractionally integrated moving average time series models for realized volatility. So et al. (2013) developed modeling tools to forecast VaR and volatility with investment horizons of less than one day.

Herrera et al. (2013) proposed extreme value models in a conditional duration framework. This Birnbaum–Saunders autoregressive conditional duration model is the first autoregressive conditional duration (ACD) model to integrate the concept of conditional quantile estimation into an ACD model by specifying the time-varying model dynamics in terms of the conditional median duration, instead of the conditional mean duration (Bhatti, 2010).

Applications

The joint distributional characteristics of asset returns are pivotal for many issues in financial economics. They are useful for the pricing of financial instruments, and they speak directly to the risk-

return tradeoff central to portfolio allocation (when returns are non-Gaussian), performance evaluation, and managerial decision-making. Moreover, they are essential for risk measurement and management (VaR), figuring prominently in both regulatory and private-sector initiatives, and market-timing strategies where the sign of future prices changes is to be predicted (Christoffersen et al., 2006; Žikeš et al., 2014; Andersen, 2003).

Co-movements

Identifying and quantifying the underlying co-movements in financial data have important implications for asset allocation, portfolio valuation, regulatory enforcement, and policy analysis. It is evident that the movements in a time series may be described by time-varying quantiles.

Liu et al. (2015) proposed stock trading strategies based on fundamental and technical analysis utilizing time-varying quantiles and applied it to the stock market in the People's Republic of China. Comparing results for both the buy-and-hold strategy and a popular NARX-based neural network trading strategy showed that this strategy performed well.

Financial contagion, which is defined as the transmission of shocks to other markets and a significant increase in the cross-market dependence beyond fundamental level, has attracted great interest among academics, policy makers, and practitioners due to its profound implications for the real economy, risk management, and asset valuation. In the literature, a large and growing number of models and different frameworks have been proposed and applied in the investigation of the dependence or the change in dependence between economic variables during crises (Ye et al., 2017). It is essential to estimate dependencies during crisis properly since the assumption that statistical properties of financial data during stable periods remain (almost) the same as during crisis is not correct. Statistical analysis made in times of stability does not provide much guidance in times of crisis (Danielsson, 2002). It is caused by the fact that in normal times people act individually, some are selling while others buy. In contrast, during crisis actions of people become more similar, there is a general flight from risky assets to safer properties. Herding instincts causes people to behave in a similar way. Moreover, in an extensive survey across data classes and risk models, the empirical properties of current risk forecasting models are found to be lacking in robustness while being excessively volatile. For regulatory use, the VaR measure is lacking in the ability to fulfill its intended task, it gives misleading information about risk, and in some cases may actually increase both idiosyncratic and systemic risk (Danielsson, 2002). Hence it is essential to investigate crisis and presence of financial contagion.

Authors of the paper Cappiello et al. (2014) proposed methodology of measuring common movements between equity returns based on time-varying quantiles which is useful for investigating whether (possibly asymmetric) financial contagion is present during financial crises and showed that it is not conditional on market volatility (Forbes et al., 2002). The approach is based on the conditional probability that a random variable is lower than a given quantile, when the other random variable is also lower than its corresponding quantile. Their results document significant increases in equity return co-movements during crises of the 1990s on the major Latin American equity markets returns consistent with the presence of financial contagion. Tomanova (2016) used the methodology for measuring financial contagion in Europe.

Value-at-risk

A well-known measure of market risks is value-at-risk (VaR). This risk measure is translation invariant, law-invariant, monotonic and satisfy positive homogeneity. VaR is the most common risk measure used in banking and finance (e.g. Basel committee recommendation) and it is easy to interpret.

Application of the CAViaR models to VaR estimation can be found in the papers of Kouretas et al. (2005), Bao et al. (2006), Kuester et al. (2006). Chen et al. (2012) discussed Bayesian Value-at-Risk in their paper and proposed methodology for forecasting based on the asymmetric Laplace distribution.

The work of Ghorbel et al. (2014) is concerned with the statistical modeling of the dependence structure between three energy commodity markets (WTI crude oil, natural gas and heating oil) using the concept of copulas and proposes a method for estimating the VaR of energy portfolio based on the combination of time series models with models of the extreme value theory before fitting a copula.

Comparison study of predictive performances of the models can be found in the papers of Kuester (2006).

Conclusion

In this paper, the methods for modeling and estimating time-varying quantiles were reviewed and their applications were summarized. We focused on methods based on quantile regression, especially on Conditional autoregressive value-at-risk (CAViaR) models and their modifications and made some notes on methods based on copula functions and others alternatives for time-varying quantiles modelling. We also discussed models utilizing high-frequency information in the form of realized measures.

In the future work, we will focus on the quantile estimation based on high-frequency data and on data that arrive sequentially. Since the data set is large, we will develop a model which takes into account that the data cannot be stored. Hence the quantiles have to be computed in a real-time setting. In such settings, classical estimators that require storing the whole history of the data (or stream) cannot be deployed (Yazidi, 2016).

Further, we would like to focus on the class of observation-driven time series models referred to as generalized autoregressive score (GAS) models or DCS models, where the mechanism to update the parameters over time is the scaled score of the likelihood function. The GAS approach provides a unified and consistent framework for introducing time-varying parameters in a wide class of nonlinear models (Creal et al., 2012).

References

ABAD, Pilar; BENITO, Sonia; LÓPEZ, Carmen. A comprehensive review of Value at Risk methodologies. *The Spanish Review of Financial Economics*, 2014, 12.1: 15-32.

ANDERSEN, Torben G., et al. Modeling and forecasting realized volatility. *Econometrica*, 2003, 71.2: 579-625.

AVDULAJ, Krenar; BARUNIK, Jozef. Are benefits from oil–stocks diversification gone? New evidence from a dynamic copula and high frequency data. *Energy Economics*, 2015, 51: 31-44.

BANULESCU, Denisa; COLLETAZ, Gilbert; HURLIN, Christophe, et al. High-Frequency Risk Measures. In: *Lyon Meeting*. 2013.

BANULESCU, Denisa, et al. Forecasting High- Frequency Risk Measures. *Journal of Forecasting*, 2016, 35.3: 224-249.

BAO, Yong; LEE, Tae- Hwy; SALTOGLU, Burak. Evaluating predictive performance of value- at- risk models in emerging markets: a reality check. *Journal of forecasting*, 2006, 25.2: 101-128.

BEE, Marco; DUPUIS, Debbie J.; TRAPIN, Luca. Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective. *Journal of Empirical Finance*, 2016, 36: 86-99.

BELTRATTI, Andrea; MORANA, Claudio. Computing value at risk with high frequency data. *Journal of empirical finance*, 1999, 6.5: 431-455.

BERNARDI, Mauro, et al. Bayesian tail risk interdependence using quantile regression. *Bayesian Analysis*, 2015, 10.3: 553-603.

BHATTI, Chad R. The Birnbaum–Saunders autoregressive conditional duration model. *Mathematics and Computers in Simulation*, 2010, 80.10: 2062-2078.

BOUYÉ, Eric; SALMON, Mark. Dynamic copula quantile regressions and tail area dynamic dependence in Forex markets. *The European Journal of Finance*, 2009, 15.7-8: 721-750.

BROWNLEES, Christian T.; GALLO, Giampiero M. Comparison of volatility measures: a risk management perspective. *Journal of Financial Econometrics*, 2009, 8.1: 29-56.

BUCHINSKY, Moshe. Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, 1998, 88-126.

CAPPIELLO, Lorenzo, et al. Measuring comovements by regression quantiles. *Journal of Financial Econometrics*, 2014, 12.4: 645-678.

CHEN, Qian; GERLACH, Richard; LU, Zudi. Bayesian Value-at-Risk and expected shortfall forecasting via the asymmetric Laplace distribution. *Computational Statistics & Data Analysis*, 2012, 56.11: 3498-3516.

CHEN, Xiaohong; KOENKER, Roger; XIAO, Zhijie. Copula- based nonlinear quantile autoregression. *The Econometrics Journal*, 2009, 12.s1.

CHRISTOFFERSEN, Peter F.; DIEBOLD, Francis X. Financial asset returns, direction-of-change forecasting, and volatility dynamics. *Management Science*, 2006, 52.8: 1273-1287.

CLEMENTS, Michael P.; GALVÃO, Ana Beatriz; KIM, Jae H. Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, 2008, 15.4: 729-750.

CONT, Rama. Empirical properties of asset returns: stylized facts and statistical issues. 2001.

CREAL, Drew; KOOPMAN, Siem Jan; LUCAS, André. Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 2013, 28.5: 777-795.

DANIELSSON, Jón. The emperor has no clothes: Limits to risk modelling. *Journal of Banking & Finance*, 2002, 26.7: 1273-1296.

DE ROSSI, Giuliano; HARVEY, Andrew. Quantiles, expectiles and splines. *Journal of Econometrics*, 2009, 152.2: 179-185.

DE ROSSI, Giuliano, et al. *Time-varying quantiles*. University of Cambridge, Faculty of Economics, 2006.

DIONNE, Georges; DUCHESNE, Pierre; PACURAR, Maria. Intraday Value at Risk (IVaR) using tick-by-tick data with application to the Toronto Stock Exchange. *Journal of Empirical Finance*, 2009, 16.5: 777-792.

ENGLE, Robert F.; MANGANELLI, Simone. CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 2004, 22.4: 367-381.

FORBES, Kristin J.; RIGOBON, Roberto. No contagion, only interdependence: measuring stock market comovements. *The journal of Finance*, 2002, 57.5: 2223-2261.

GERLACH, Richard H.; CHEN, Cathy WS; CHAN, Nancy YC. Bayesian time-varying quantile forecasting for value-at-risk in financial markets. *Journal of Business & Economic Statistics*, 2011, 29.4: 481-492.

GHORBEL, Ahmed; TRABELSI, Abdelwahed. Energy portfolio risk management using time-varying extreme value copula methods. *Economic Modelling*, 2014, 38: 470-485.

GIOT, Pierre; LAURENT, Sébastien. Modelling daily value-at-risk using realized volatility and ARCH type models. *Journal of Empirical Finance*, 2004, 11.3: 379-398.

GUO, Ming-yuan; ZHANG, Shi-ying. Study on CVaR forecasts based on weighted realized volatility. In: *Management Science and Engineering, 2008. ICMSE 2008. 15th Annual Conference Proceedings., International Conference on*. IEEE, 2008. p. 91-96.

HARVEY, Andrew C. Dynamic distributions and changing copulas. 2008.

HARVEY, Andrew C. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*. Cambridge University Press, 2013.

HARVEY, Andrew C.; ORYSHCHENKO, Vitaliy. Local kernel density estimation from time series data (Preliminary and incomplete). 2009.

HERRERA, Rodrigo; SCHIPP, Bernhard. Value at risk forecasts by extreme value models in a conditional duration framework. *Journal of Empirical Finance*, 2013, 23: 33-47.

HUANG, Dashan, et al. Index-Exciting CAViaR: a new empirical time-varying risk model. *Studies in Nonlinear Dynamics & Econometrics*, 2010, 14.2.

HUANG, Huiyu; LEE, Tae-Hwy. Forecasting value-at-risk using high-frequency information. *Econometrics*, 2013, 1.1: 127-140.

KOENKER, Roger; BASSETT JR, Gilbert. Regression quantiles. *Econometrica: journal of the Econometric Society*, 1978, 33-50.

KOENKER, Roger; XIAO, Zhijie. Quantile autoregression. *Journal of the American Statistical Association*, 2006, 101.475: 980-990.

KOURETAS, Georgios P., et al. Conditional Autoregressive Value at Risk by Regression Quantiles: Estimating market risk for major stock markets. 2005.

KRUSE, Robinson. Can realized volatility improve the accuracy of Value-at-Risk forecasts. In: *Leibniz University of Hannover Working Paper*. 2006.

KUESTER, Keith; MITTNIK, Stefan; PAOLELLA, Marc S. Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 2006, 4.1: 53-89.

LI, Jianbo, et al. Estimation and testing for time-varying quantile single-index models with longitudinal data. *Computational Statistics & Data Analysis*, 2018, 118: 66-83.

LI, Ruosha; CHENG, Yu; FINE, Jason P. Quantile association regression models. *Journal of the American Statistical Association*, 2014, 109.505: 230-242.

LIU, Tiantian; QIU, Ning; GU, Wentao. A stock trading strategy based on time-varying quantile theory. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 2015, 19.3: 417-422.

LOUZIS, Dimitrios P.; XANTHOPOULOS-SISINIS, Spyros; REFENES, Apostolos N. Are realized volatility models good candidates for alternative Value at Risk prediction strategies?. 2011.

MAHEU, John M.; MCCURDY, Thomas H. Do high-frequency measures of volatility improve forecasts of return distributions?. *Journal of Econometrics*, 2011, 160.1: 69-76.

MANNER, Hans; REZNIKOVA, Olga. A survey on time-varying copulas: Specification, simulations, and application. *Econometric Reviews*, 2012, 31.6: 654-687.

PATTON, Andrew J. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 2012, 110: 4-18.

SHEPHARD, Neil; SHEPPARD, Kevin. Realising the future: forecasting with high- frequency- based volatility (HEAVY) models. *Journal of Applied Econometrics*, 2010, 25.2: 197-231.

SO, Mike KP; XU, Rui. Forecasting intraday volatility and value-at-risk with high-frequency data. *Asia-Pacific Financial Markets*, 2013, 20.1: 83-111.

TAYLOR, James W. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, 2017, 1-13.

TOMANOVÁ, Petra. Measuring Comovements by Univariate and Multivariate Regression Quantiles: Evidence from Europe. In: *Mathematical Methods in Economics*. 2016, 863--868.

TOMANOVÁ, Petra. Measuring Co-movements Based on Quantile Regression using High-Frequency Information: Asymmetric Tail Dependence. In: *Mathematical Methods in Economics (MME 2017)*, 2017, 801-806.

WATANABE, Toshiaki. Quantile forecasts of financial returns using realized GARCH models. *The Japanese Economic Review*, 2012, 63.1: 68-80.

WHITE, Halbert; KIM, Tae-Hwan; MANGANELLI, Simone. VAR for VaR: Measuring tail dependence using multivariate regression quantiles. *Journal of Econometrics*, 2015, 187.1: 169-188.

YAZIDI, Anis; HAMMER, Hugo. Quantile estimation in dynamic and stationary environments using the theory of stochastic learning. *ACM SIGAPP Applied Computing Review*, 2016, 16.1: 15-24.

YE, Wuyi; LUO, Kebing; LIU, Xiaoquan. Time-varying quantile association regression model with applications to financial contagion and VaR. *European Journal of Operational Research*, 2017, 256.3: 1015-1028.

ŽIKEŠ, Filip; BARUNÍK, Jozef. Semi-parametric conditional quantile models for financial returns and realized volatility. *Journal of Financial Econometrics*, 2014, 14.1: 185-226.

JEL Classification: C10, C22.

Acknowledgments: This work was supported by the research grant VŠE IGA F4/93/2017, Faculty of Informatics and Statistics, University of Economics, Prague. This support is gratefully acknowledged.

Shrnutí

Time-varying kvantily a jejich aplikace

V této práci se zabýváme utříděním metod pro modelování a odhad time-varying kvantilů a jejich aplikací. Zaměřujeme se na metody využívající kvantilovou regresi, zejména na Conditional autoregressive value-at-risk (CAViaR) modely a jejich modifikacemi a stručně se zabýváme copula modely a další alternativy pro modelování time-varying kvantilů. Dále se zaměřujeme na modely využívající informaci z vysokofrekvenčních dat, jako jsou realizované míry volatility. Nakonec shrneme jejich použití pro odhad hodnoty value-at-risk, měření co-movementů, testování na přítomnost finanční nákazy a problém alokace aktiv.

Klíčová slova: finanční nákaza, vysokofrekvenční data, time-varying kvantily, value-at-risk.

Hospitalizace pacientů s Alzheimerovou chorobou

Popisné charakteristiky první hospitalizace a následných rehospitalizací

Hana M. Broulíková

xsmrh00@vse.cz

Doktorand oboru statistika

Školitel: doc. Ing. Markéta Arltová, Ph.D., (marketa.arltova@vse.cz)

Abstrakt: Cílem tohoto příspěvku je popsat hospitalizace alzheimerovské populace v České republice. Příspěvek vychází z dat získaných propojením národních registrů hospitalizací a úmrtí spravovaných Ústavem zdravotnických informací a statistiky. Registr hospitalizací obsahuje data na úrovni jednotlivců za všechny hospitalizace proběhlé v období 1994–2014. Registr úmrtí obsahuje záznam o každém zesnulém ve stejném období. Tyto dva registry je možné propojit za pomoci šifrovaného rodného čísla. Je definována kohorta pacientů, kteří v roce 2000 podstoupili první hospitalizaci pro Alzheimerovu chorobu. Délka sledovaného období umožňuje zpětnou kontrolu, zda se v roce 2000 skutečně jednalo o první hospitalizaci pro základní diagnózu Alzheimerovy choroby, a zároveň lze sledovat průběh rehospitalizací až do chvíle úmrtí pacienta.

Kohorta čítá 1224 pacientů, z nichž jednu třetinu tvoří muži. Většina 939 pacientů absolvuje pouze první hospitalizaci. Nejvyšším počtem rehospitalizací ve vzorku je 13. Průměrná doba trvání hospitalizace dosahuje 122 dní (medián 32 dní). Většina (62,5 %) pacientů byla hospitalizována v psychiatrických nemocnicích. Průměrný věk v době první hospitalizace činil 75,5 (76) let u mužů a 78 (78,5) let u žen. Pacienti umírali krátce po přijetí k první hospitalizaci ve věku 77 (77) let u mužů, respektive 80,5 (80) let u žen. Tento výsledek je v souladu s tvrzením České alzheimerovské společnosti o tom, že čeští pacienti jsou často diagnostikováni a léčeni až těsně před svou smrtí.

Klíčová slova: zdravotnická statistika, registr hospitalizací, registr úmrtí, Alzheimerova choroba, demence, kohortní analýza

Úvod

Cílem tohoto příspěvku je popsat charakteristiky pacientů poprvé hospitalizovaných v České republice pro Alzheimerovu chorobu. Předmětem zkoumání jsou především demografické údaje, údaje o formě hospitalizace a případných rehospitalizací. Alzheimerova choroba je nemoc neurodegenerativního charakteru způsobující pokles kognitivních schopností pacienta. Jedná se o nejčastější příčinu demence (Mátl and Mátlová 2015). V následku Alzheimerovy choroby ztrácejí pacienti schopnost nezávislého života a jsou v pokročilejších fázích částečně či zcela odkázáni na pomoc okolí ať již ve formě domácí péče, péče poskytované institucí sociální péče či na relativně dlouhodobou hospitalizaci v zařízení zdravotní péče.

Důsledkem celosvětově i národně se zvyšující incidence a prevalence Alzheimerovy choroby ve stárnoucí společnosti je vysoká ekonomická zátěž pro pacienty a jejich rodiny, ale také pro veřejné rozpočty zdravotní a sociální péče (Mátl et al. 2016; Prince et al. 2016). Nedávné výzkumy ukazují, že jedním z možných způsobů snížení této zátěže je například včasná diagnóza Alzheimerovy choroby a její léčba (Getsios et al. 2012; Weimer and Sager 2009). Přestože současná medicína nedokáže Alzheimerovu chorobu vyléčit (Fisher 2008), dokáže zpomalit její průběh a tím umožnit pacientům strávit relativně delší období pouze s mírnou či střední formou demence (Lopez et al. 2005). Léčba tedy nejen zvyšuje kvalitu života pacientů, ale také snižuje celospolečenské náklady v důsledku vyhnutí se části vysokých nákladů na sociální péči. V České republice poslední odhady získané simulačními metodami naznačují, že by úspory včasné detekce a léčby Alzheimerovy choroby mohly dosahovat až půl milionu korun ušetřených v průběhu života jednoho pacienta (Broulíková et al. 2017a).

Odhad pro Českou republiku však neuvažuje náklady hospitalizací z důvodu nedostatku statisticky relevantních informací epidemiologického a demografického charakteru. Na tuto překážku ve výzkumu a plánování péče svorně upozorňují Česká alzheimerovská společnosti (Mátl, Mátlová and

Holmerová 2016) i Národní akční plán pro Alzheimerovu nemoc a další obdobná onemocnění (Ministerstvo zdravotnictví ČR 2016b). Náklady hospitalizací jsou přitom pravděpodobně nezanedbatelné vzhledem k vysokým nákladům na jeden ošetřovací den přesahujícím tisíc Kč (Broulíková et al. 2017b). Celkový význam nákladů na hospitalizaci a například také úspory z jejich předcházení v případě včasné diagnózy a ambulantní léčby závisí na prevalenci, délce hospitalizací a následných rehospitalizací alzheimerovské populace. Ústav zdravotnických informací a statistik (ÚZIS) spočítal, že průměrná délka hospitalizace v letech 2008–2012 dosahovala 90 dnů (ÚZIS 2013). Česká alzheimerovská společnost odhaduje nutnost hospitalizace u 6% pacientů trpících Alzheimerovou chorobou. Z dostupných (převážně případových) studií se zdá, že pacienti bývají většinou hospitalizováni až v pokročilé fázi nemoci (Lužný et al. 2014) nedlouho před smrtí (Mátl, Mátlová and Holmerová 2016). Mnoho z hospitalizovaných je až při této příležitosti diagnostikováno s Alzheimerovou chorobou, což vypovídá o prostoru pro zlepšení diagnostiky a léčby tohoto onemocnění.

Širším účelem tohoto příspěvku zmapovat možnosti popisu alzheimerovské populace za pomoci národních registrů. Tyto údaje přispějí k rozšíření vědomostí o české alzheimerovské populaci a zároveň naleznou uplatnění v konstrukci statistických modelů. Zejména znalost období první hospitalizace a počtu rehospitalizací je pro složitější analýzu klíčová.

Metodologie

Tento příspěvek využívá kohortní analýzu pro zjištění základních popisných statistik o hospitalizacích pacientů s Alzheimerovou chorobou v České republice. Tyto údaje byly vypočteny z individuálních dat získaných z Ústavu zdravotnických informací a statistiky ČR ze dvou národních registrů hospitalizací a úmrtí. Registr hospitalizací obsahuje záznamy o všech hospitalizovaných v letech 1994–2014 včetně jejich demografických charakteristik, druhu zdravotnického zařízení, základní a vedlejší diagnózy a délky pobytu. Kompletní registr hospitalizací za dané období obsahuje desítky milionů pozorování. Registr úmrtí eviduje všechna úmrtí v ČR za období 1994–2014 včetně demografických charakteristik zesnulých a obsahuje přes dva miliony pozorování. Oba registry je možné propojit na základě zakódovaných rodných čísel pacientů a zemřelých.

Konstrukce zkoumaného souboru

Prvním krokem k popisu hospitalizačního chování pacientů s Alzheimerovou chorobou bylo vytvoření kohorty, jež absolvovala v daném roce první hospitalizaci, a sledování vývoje stavu těchto pacientů z hlediska rehospitalizací až do jejich úmrtí. Byla zvolena kohorta s první hospitalizací v roce 2000, neboť u této kohorty je možné předpokládat, že bude vzhledem k délce časové řady a předpokládané době dožití pacientů s Alzheimerovou chorobou možné retrospektivně odstranit pacienty, pro které se v roce 2000 nejednalo o první hospitalizaci, a zároveň bude možné tuto kohortu sledovat v čase z hlediska rehospitalizací až do doby úmrtí pacientů.

Alzheimerovští pacienti jsou pro potřeby tohoto výzkumu definováni jako pacienti, jejichž základní diagnóza je kódována jako G30.x podle Mezinárodní klasifikace nemocí (MKN-10) (WHO 2011), kde x zastupuje jednočíselné označení čtvrtého místa subkapitoly. Diagnóza G30 obecně označuje Alzheimerovu chorobu, která se dále dělí na její čtyři specifické typy, kterými jsou Alzheimerova nemoc s časným začátkem do 65 let (G30.0), Alzheimerova choroba s pozdním začátkem po 65. roce (G30.1), jiná Alzheimerova nemoc (G30.8) a nespecifikovaná Alzheimerova nemoc (G30.9). Toto vymezení Alzheimerovy choroby je poměrně úzké a vzhledem k existenci smíšených demencí a problému s diagnostikou jsou v některých publikacích zabývajících se demencemi (ÚZIS 2013) uvažovány také kódy psychiatrických diagnóz F00 Demence u Alzheimerovy choroby, F01 Vaskulární demence a F03 Neurčená demence. Širší vymezení však není předmětem tohoto výzkumu a v prvním kroku úpravy dat byly z registru hospitalizací filtrovány pouze základní diagnózy G30.x. Na tomto místě je dále vhodné uvést význam proměnné základní diagnóza. Jedná se o diagnózu, kvůli níž hospitalizace proběhla. Tento příspěvek tedy nebere v úvahu hospitalizace pacientů, u kterých byla Alzheimerova choroba uvedena jako některá z dalších diagnóz. Proměnná vedlejší diagnózy je vhodná především pro studium hospitalizací plynoucích z komorbidit Alzheimerovy nemoci.

Vzhledem k rozsahu registru hospitalizací rozděluje ÚZIS případy podle let propuštění pacientů. Propuštěním se rozumí také propuštění z důvodu úmrtí pacienta. Pro konstrukci kohorty poprvé hospitalizovaných je však určující datum přijetí, nikoli datum propuštění. Jelikož hospitalizace nezřídka přesáhnou dobu jednoho roku, nelze předpokládat u většiny propuštěných v roce 2000 datum přijetí ve stejném roce, a zároveň mnoho propuštění z pozdějších let mohlo být důsledkem hospitalizace započaté již v roce 2000. V druhém kroku bylo zapotřebí sloučit všechna pozorování propuštěných v období let 2000-2014 a vyfiltrovat pouze ta pozorování s datem začátku v roce 2000.

Za účelem sledování charakteristik hospitalizací a potřeb rehospitalizací alzheimerovských pacientů definuji kohortu, která v daném roce zažívá první hospitalizaci, a dále ji sleduji v čase. Soubor pacientů s počátkem hospitalizace v roce 2000 vytvořený ve druhém kroku však obsahuje i pacienty, u kterých se při přijetí v roce 2000 nejednalo o první hospitalizace, nýbrž o rehospitalizaci. Tito pacienti jsou ve třetím kroku ze souboru vyřazeni na základě zpětného pohledu do let 1994–1999 v případě, že již byli pro základní diagnózu G30.x v minulosti hospitalizováni. Jako výsledek získáváme kohortu prvních hospitalizací v roce 2000. Je třeba uvědomit si, že někteří pacienti mohli být rehospitalizováni již v rámci roku 2000. Při konstrukci popisných statistik o kohortě je tedy nutné tyto rehospitalizace odfiltrovat a zachovat pouze unikátní záznam ke každému rodnému číslu. Unikáních rodných čísel bylo s první hospitalizací v roce 2000 spojeno 1224. Ve čtvrtém kroku jsou k souboru připojeny rehospitalizace s propuštěním v následujících letech 2001-2014.

V pátém kroku je soubor propojen s registrem úmrtí. Ačkoli se jedná o výraznou menšinu 111 případů, je nutné upozornit na existenci pacientů, kteří do roku 2014 dle záznamů registru úmrtí nezemřeli, či zemřeli vícekrát. Jinými slovy registr úmrtí obsahuje malé procento chybných záznamů. Pro účely této studie byly záznamy pacientů, kteří byli i po roce 2014 vedeni jako nezesnulí, považovány za chybné a byly ze souboru vyřazeny. Důvody k využití tohoto předpokladu jsou dva. U většiny 73 pacientů je vzhledem k jejich věku při první hospitalizaci při znalosti úmrtnostních tabulek ČR (ČSÚ 2016a) a zvýšeného rizika úmrtí při diagnóze AD (Fitzpatrick et al. 2005) nepravděpodobné jejich dožití roku 2014. U menšiny 37 případů se jedná o pacienty s brzkým či velmi brzkým propuknutím Alzheimerovy choroby (diagnóza 30.0). Ačkoli je možné, že tito pacienti dožili roku 2014, Alzheimerova choroba s brzkým nástupem bývá spojena s odlišným průběhem od častější formy Alzheimerovy choroby s pozdním nástupem (30.1), která je hlavním zájmem výzkumu v mé disertační práci. Proto byli vzhledem k nejasnostem tito pacienti vyřazeni. Druhým problémem spojeným s registrem úmrtí byla existence duplikovaných úmrtí. V tomto případě se však jednalo pouze o jednotky případů, které byly ve všech případech vyřešeny vyřazením pozorování, u nichž bylo datum úmrtí starší než datum přijetí k hospitalizaci.

Duplikáty pozorování byly přítomny také v případě záznamu o hospitalizacích. Jednalo se konkrétně o šestnáct pozorování shodných v proměnné rodné číslo, datum a čas přijetí. Pokud byl celý záznam shodný i v ohledu zbylých proměnných, bylo ponecháno vždy pouze jedno takové pozorování. V případech, kdy se jednalo o záznamy se shodným datem a časem přijetí, avšak rozdílným datem propuštění, byla ponechána ta pozorování s delší dobou hospitalizace. Důvodem k tomuto kroku je možnost chyby vzniklé při předávání pacienta mezi jednotlivými odděleními či zařízeními, kdy mohl být chybně opsán původní údaj o přijetí k hospitalizaci. Konečný soubor čítá 1702 hospitalizací kohorty prvně hospitalizovaných v roce 2000.

Charakteristiky souboru

Výpočetní část tohoto příspěvku spočívala především v konstrukci popisných charakteristik, jakými jsou průměr či medián souboru, s cílem vytvořit představu o způsobu hospitalizací alzheimerovských pacientů v České republice. Kromě popisných charakteristik týkajících se celé kohorty byli pacienti z důvodu rozdílné očekávané doby dožití a průběhu nemoci dále rozděleni podle pohlaví. Další dělení se týká podrobnější specifikace diagnóz či poskytovatele péče.

Výsledky

První hospitalizace

Kohorta prvně hospitalizovaných čítá 1224 pacientů v průměrném věku 77,5 roku a mediánem 77 let, přičemž se tito pacienti dožívají průměrně pouze o dva roky déle po první hospitalizaci, a sice 79,5 roku. Medián úmrtí je 79 let. První hospitalizace trvá v průměru 117,5 dne, její délka je však vysoce heterogenní od 1 do 3218 dnů (směrodatná odchylka 294), čemuž odpovídá také výrazně kratší mediánová doba hospitalizace 30 dní. Převážná většina 62,5 % byla poprvé hospitalizována v psychiatrické nemocnici (PN), 22 % ve všeobecné či fakultní nemocnici (VN, FN) a 10,5 % v léčebně dlouhodobě nemocných. (LDN) V závislosti na zařízení, v němž hospitalizace proběhla, se výrazně liší délka pobytu. Nejdelší hospitalizace vykazují psychiatrické nemocnice s průměrnou ošetřovací dobou 162,5 dne (medián 47 dnů). První hospitalizace ve všeobecných a fakultních nemocnicích trvají 22 (12) dnů a v léčebnách dlouhodobě nemocných 69,5 (37) dne.

Kohorta prvních hospitalizací čítá 389 mužů (32 %) a 835 žen (68 %). První hospitalizace proběhla u mužů průměrně v 75,5 (76) letech, přičemž muži umírali průměrně v 77 (77) letech. Ženy byly poprvé hospitalizovány ve věku 78,5 (78) let a umíraly o dva roky později v 80,5 (80) letech. Délka hospitalizace u mužů 102,5 (28) dnů byla v průměru kratší než 124,5 (33) dnů u žen. Vzorec v rozmístění mužů a žen do různých typů lůžkové péče je obdobný. Zajímavá je informace, že průměrná délka hospitalizace u mužů je nižší než u žen, přestože muži jsou mírně nadproporčně zastoupeni v psychiatrických nemocnicích (65 %) než ženy (61,5 %). Ženy však naopak vykazují vyšší zastoupení v léčebnách dlouhodobě nemocných (12,22 %) oproti mužům (7 %).

Z hlediska rozdělení pacientů podle diagnóz jsou v kohortě téměř z poloviny (45 %) zastoupeni pacienti s diagnózou Alzheimerovy choroby s pozdním nástupem (G30.1). Čtvrtina (25,5 %) pacientů byla diagnostikována s jinou Alzheimerovou nemocí (G30.8) a téměř stejné množství (22 %) s nespecifikovanou Alzheimerovou nemocí (G30.9). Zbylých 7,5 % pacientů trpělo Alzheimerovou nemocí s časným začátkem (G30.0). Pacienti s časným začátkem nemoci byli poprvé hospitalizování v průměru v 66,5 (65) letech a umírali o dva roky později ve věku 69,5 (69,5) let. Průměrný věk hospitalizace byl v případě ostatních diagnóz obdobný a to 79 (79) let u pacientů s pozdím začátkem i jinou Alzheimerovou chorobou (78). Pacienti s nespecifikovanou formou Alzheimerovy nemoci byli poprvé hospitalizování v průměru v 76,5 (76) letech. Zajímavým zjištěním je fakt, že průměrná doba dožití od první hospitalizace je ve všech případech blízko dvěma letům. Konkrétně se u diagnóz G30.1 a G30.8 dožívali pacienti průměrně 81 (80) let a 78 (78) let v případě diagnózy G30.9.

Výsledky pro první hospitalizace jsou shrnuty v tabulce 1.

Tabulka 1: První hospitalizace Zdroj: Vlastní výpočty.

Kohorta	Počet pozorování	Podíl (%)	Věk hospitalizace (roky)		Věk úmrtí (roky)		Délka hospitalizace (dny)	
			průměr	medián	průměr	medián	průměr	medián
Všichni	1224		77,5	77	79,5	79	117,5	30
Dle pohlaví								
Muži	389	32	75,5	76	77	77	102,5	28
Ženy	835	68	78,5	78	80,5	80	124,5	33
Dle diagnózy								
G30.0	90	7,5	66,5	65	69,5	69,5	157	25
G30.1	552	45	79	79	81	80	122,5	33
G30.8	312	25,5	79	78	81	80	154	43
G30.9	270	22	76,5	76	78	78	51	17,5
Dle zařízení								
FN	50	4	75	77	76	77,5	27,7	16
VN	220	18	75	75	77,5	77	21	11,5
PN	766	62,5	78	78	80	80	162,5	47

Kohorta	Počet pozorování	Podíl (%)	Věk hospitalizace (roky)		Věk úmrtí (roky)		Délka hospitalizace (dny)	
LDN	129	10,5	79,5	79	81	81	69,5	37
Ostatní	59	5	–	–	–	–	–	–

Rehospitalizace

Většina 939 pacientů (77 %) prodělala jednu hospitalizaci, zatímco ostatní byli rehospitalizováni. Nejvyšší počet rehospitalizací dosáhl 13, průměrem je jedna rehospitalizace, mediánem nula. Rozložení rehospitalizovaných pacientů podle pohlaví odpovídá poměru pozorovaném pro první hospitalizaci. I zde tvoří třetinu (32,5 %) rehospitalizovaných muži, dvě třetiny rehospitalizací (67,5 %) připadají na ženy. Průměrná rehospitalizace trvala 130,5 dne, medián je však i v tomto případě výrazně nižší 35 dní. Souhrnný údaj pro všechny hospitalizace včetně rehospitalizací je jen o něco nižší, když průměrná délka hospitalizace dosahuje 122 dní a medián 32 dní. Rehospitalizace přichází v průměru jeden rok po první hospitalizaci, mediánová hodnota vypovídá dokonce o období kratším jednoho roku.

Nadpoloviční většina (52,5 %) rehospitalizací je poskytována psychiatrickými nemocnicemi. Třetina (34 %) rehospitalizovaných pacientů je léčena ve všeobecných či fakultních nemocnicích a 7,5 % v léčebnách dlouhodobě nemocných. Zbylí pacienti jsou rozloženi mezi ostatní zdravotnická zařízení. Oproti prvním hospitalizacím tedy dochází k odklonu pacientů z psychiatrických nemocnic a léčeben dlouhodobě nemocných k fakultním či všeobecným nemocnicím.

Rozdělení pacientů podle diagnóz se v případě rehospitalizací mírně liší od prvních hospitalizací. Nejčetněji jsou opět zastoupeni alzheimerovští pacienti s diagnózou pozdního nástupu choroby (G30.1), jejich podíl však klesl o pět procentních bodů na 40 %. Následují pacienti s nespecifikovanou Alzheimerovou chorobou (G30.9), jejichž podíl vzrostl o pět procentních bodů na 27 % a předčil četnost diagnózy jiné Alzheimerovy choroby (G30.8), jež dosahuje 18 %. Zároveň také vzrostl podíl pacientů s časným začátkem téměř na dvojnásobných 14 %. Zajímavým zjištěním je, že rehospitalizovaní pacienti s G30.0 se dožívali pouze 65 (64) let, což je asi o pět let méně než v případě těch, kteří rehospitalizaci nepotřebovali. Obdobný pohled se však naskýtá v případě rehospitalizací u dalších diagnóz (G30.1: 80/78 let; G30.8: 81/79; G30.9 76/76) pro něž se taktéž zkracuje doba dožití.

Výsledky za rehospitalizace jsou shrnuty v tabulce 2.

Tabulka 2: Rehospitalizace Zdroj: Vlastní výpočty.

Kohorta	Počet pozorování	Podíl (%)	Věk rehospitalizace (roky)		Věk úmrtí (roky)		Délka hospitalizace (dny)	
			průměr	medián	průměr	medián	průměr	medián
Hospitalizace	1702		76,5	76	78,5	79	122	32
Pouze jedna	939	55	78,5	78	80	80	127,5	34
Rehospitalizace	763	45	74	75	77	77	115,5	31
Rehospitalizace Podrobněji								
1. hospital.	285	17	74,5	75	77	78	91	25
2. hospital.	285	17	75,5	76	77	78	145	35
3. hospital.	106	6	75	75	77	77	125	30
>3 hospital.	87	5	73	75	76	76	89,5	42
Rehospitalizace Dle pohlaví								
Muži	248	32,5	73,5	75	75	76	100	28
Ženy	515	67,5	75,5	76	78	78	123	32
Dle diagnózy								
G30.0	108	14	63	63	66	65	111	19,5
G30.1	297	39	78	77	80	79	110	34

Kohorta	Počet pozorování	Podíl (%)	Věk rehospitalizace (roky)		Věk úmrtí (roky)		Délka hospitalizace (dny)	
G30.8	145	19	77,5	77	80,5	79	199	47
G30.9	213	28	74	75	76	76	69	21
Dle zařízení								
FN	32	7	71,5	70	72,5	71	32	17
VN	129	27	75	76	77	77	85	34
PN	251	52,5	75	76	77	77	180,5	37
LDN	35	7,5	76,5	75	77,5	76	84	36
Ostatní	31	6	–	–	–	–	–	–

Závěr

Tento příspěvek přinesl první pohled na údaje o hospitalizacích a následných rehospitalizacích pacientů s Alzheimerovou chorobou v České republice. Výsledky jsou v souladu s popisem současného stavu Českou alzheimerovskou společností a některými případovými studiemi, které hovoří o hospitalizaci v pokročilém stádiu nemoci a krátce před úmrtím pacienta. Zároveň odhalil prostor pro postup v budoucí práci s registrovými daty za účelem popisu potřeby hospitalizací alzheimerovské populace. Vzhledem k předpokládané době dožití s Alzheimerovou chorobou zhruba jedné desítky let a délce dostupné časové řady, je možné obdobným způsobem konstruovat kohorty pro první hospitalizace v dalších letech okolo přelomu tisíciletí. Tento přístup by měl poskytnout dostatečně velký vzorek pro hlubší analýzu a testy statistické významnosti rozdílů mezi skupinami pacientů. Nabízejícími se otázkami je rozdělení pacientů a pacientek v době první hospitalizace z hlediska věku, vlivu věku na pravděpodobnost rehospitalizace a obecně na délku pobytů v lůžkových zařízeních.

Dále je zapotřebí pro dobré porozumění problematice výsledky mezioborově konzultovat se specialisty v oblasti gerontologie a pokusit se pozorované anomálie vysvětlit, či alespoň dostatečně zohlednit v popisu alzheimerovské populace. Pro ilustraci uveďme příklad diagnózy Alzheimerovy choroby s časným začátkem (G30.0). Je ke zvážení, zda se tito pacienti neliší od běžnější alzheimerovské populace s diagnózami G30.1–G30.9 natolik, že jejich zařazení ubírá validitě výsledků z hlediska jejich využití pro stavbu statistických modelů nejen v rámci připravované disertační práce. Při pohledu z opačné strany může být zapotřebí zohlednit také pacienty s diagnózami F00–F03.

Vedle popisu hospitalizací se jejich registr při propojení s registrem úmrtí nabízí jako vhodný podklad pro tvorbu úmrtnostních tabulek pacientů s Alzheimerovou chorobou. Zjištění o kratší době dožití rehospitalizovaných pacientů může být do jisté míry vysvětlitelné nejen závažnějším průběhem nemoci, ale také vyšším výskytem komorbidit. Studium komorbidit by mělo být umožněno jak na základě propojení pacienta hospitalizovaného pro Alzheimerovu chorobu s jeho dalšími hospitalizacemi pro jiné diagnózy, tak i za pomoci proměnné sekundárních diagnóz.

Literatura

BROULÍKOVÁ, H. M., V. SLÁDEK, M. ARLTOVÁ AND J. ČERNÝ. Economic Impact of Alzheimer´s Disease Early Detection in Czechia. In T. LOSTER AND T. PAVELKA. The 11th International Days of Statistics and Economics. Prague: Melandrium, 2017a.

BROULÍKOVÁ, H. M., P. WINKLER, L. KONDRÁTOVÁ AND M. PÁV. Costs of mental health services in Czechia. In.: National Institute of Mental Health, 2017b (unpublished).

ČSÚ. Life tables. In. Prague: Czech Statistical Office, 2016a.

FISHER, A. Cholinergic treatments with emphasis on m1 muscarinic agonists as potential disease-modifying agents for Alzheimer's disease. Neurotherapeutics, 2008, 5(3), 433-442.

FITZPATRICK, A. L., L. H. KULLER, O. L. LOPEZ, C. H. KAWAS, et al. Survival following dementia onset: Alzheimer's disease and vascular dementia. Journal of the neurological sciences, 2005, 229, 43-49.

GETSIOS, D., S. BLUME, K. J. ISHAK, G. MACLAINE, et al. An economic evaluation of early assessment for Alzheimer's disease in the United Kingdom. Alzheimer's & Dementia, 2012, 8(1), 22-30.

LOPEZ, O. L., J. T. BECKER, J. SAXTON, R. A. SWEET, et al. Alteration of a clinically meaningful outcome in the natural history of Alzheimer's disease by cholinesterase inhibition. Journal of the American Geriatrics Society, 2005, 53(1), 83-87.

LUŽNÝ, J., I. HOLMEROVÁ, P. WIJA AND I. ONDREJKA Dementia Still Diagnosed Too Late—Data from the Czech Republic. Iranian journal of public health, 2014, 43(10), 1436.

MÁTL, O. AND M. MÁTLOVÁ Zpráva o stavu demence 2015. Edtion ed. Praha: Czech Alzheimer Society, 2015. 18 p. ISBN 978-80-86541-45-7.

MÁTL, O., M. MÁTLOVÁ AND I. HOLMEROVÁ Zpráva o stavu demence 2016. Edtion ed. Praha: Czech Alzhiemer Society, 2016. ISBN 978-80-86541-50-1.

MINISTERSTVO ZDRAVOTNICTVÍ ČR. Národní akční plán pro Alzheimerovu nemoc a další obdobná onemocnění na léta 2016 - 2019. In MZ ČR. 2016b.

WORLD HEALTH ORGANIZATION. International statistical classification of diseases and related health problems. - 10th revision, edition 2010. In. Malta, 2011.

PRINCE, M., A. COMAS-HERRERA, M. KNAPP, M. L. GUERCHET, et al. The World Alzheimer Report 2016. London: 2016.

ÚZIS. Health care of patients treated for dementia in out-patient and in-patient facilities in the Czech Republic in 2008–2012 Prague: 2013.

WEIMER, D. L. AND M. A. SAGER Early identification and treatment of Alzheimer's disease: social and fiscal outcomes. Alzheimer's & Dementia, 2009, 5(3), 215-226.

JEL Classification: I18, J11, J14

Dedikace: Podpořeno grantem č. F4/38/2017 Interní grantové agentury Vysoké školy ekonomické v Praze.

Summary

Hospitalisation of Alzheimer's Patients

This study investigates demographical characteristics of inpatients suffering from Alzheimer's disease in the Czech Republic as well as parameters of their hospitalization such as its length or number of rehospitalisation cases. This empirical analysis is based on two national registers maintained by the Institute of Health Information and Statistics. The National Register of Hospitalised Patients contains individual data about all hospitalizations in the period 1994-2014. The National Register of Death Causes comprises data about all individuals deceased in the same period. I integrated these two sets of data on the basis of a unique identifier and constructed a cohort of patients who were hospitalized for the first time in the year 2000. The length of the period allows to make sure that a given patient was not hospitalised before 2000 and moreover to follow the rehospitalisation pattern until the individual dies.

The cohort consists of 1224 individuals of which one third are men. Most of the patients (939) were hospitalised only once without any rehospitalisation. The maximum number of rehospitalisation cases reached 13. The average length of inpatient stay was 122 days (median 32). Majority (62.5 %) of patients were hospitalised in psychiatric hospitals. The average age of first hospitalisation was 75.5 (median 76) for men and 78 (78.5) for women. Patients tended to die shortly after the hospitalisation; men aged 77 (77) and women 80.5 (80) years. This result supports the hypothesis that Czech Alzheimer's patients are often diagnosed and treated only shortly before death.

Key words: health statistics, register of death causes, register of hospitalised patients, Alzheimer's disease, dementia, cohort analysis.

Rozptyl odhadu PACF založenom na useknutí časového radu AR(*p*)

Samuel Flimmel

samuel.flimmel@vse.cz

Doktorand oboru statistika

Školitel: doc. RNDr. Ivana Malá, CSc., (malai@vse.cz)

Abstrakt: Boxová – Jenkinsová metodológia vyžaduje v prvom kroku identifikáciu rádov časového radu ARMA(*p*,*q*). K odhadnutiu rádov *p* a *q* sa navrhuje použitie autokorelačnej funkcie (ACF) a parciálnej autokorelačnej funkcie (PACF). Následne je potrebné určiť hranice významnosti, pomocou ktorých sa identifikuje významnosť ACF/PACF pre jednotlivé rády. Pre ACF bola navrhnutá Bartlettová aproximácia a pre účely PACF bola navrhnutá zase Quenouilleová aproximácia. Tieto aproximácie však platia len pre výberové odhady ACF a PACF, čiže v prípade iných odhadov je potrebné tieto hranice odvodiť. To sa týka aj prípadu odhadu PACF založenom na useknutí pôvodného časového radu. Doposiaľ sa nepodarilo odvodiť rozdelenie tohto odhadu, a preto ako jedna z možností sa ponúka aspoň jeho aproximácia založená na simulačnej štúdii. V tomto článku sa budeme danou simuláciou zaoberať a navrhneme „pravidlo palca“, ktoré nám môže výrazne pomôcť pri rôznych aplikáciach. Taktiež ukážeme praktické využitie tohto pravidla a jeho dopad pri identifikácii rádu časového radu AR(*p*).

Kľúčové slová: PACF, rozptyl odhadu, AR(*p*)

Úvod

Odhad parciálnej autokorelačnej funkcie je dôležitá súčasť procesu identifikácie časového radu. Ide o jeden z krokov v známej Boxovej – Jenkinsovej metodológii, viď (Box a spol., 1975). Na základe autokorelačnej funkcie (ACF) a parciálnej autokorelačnej funkcie (PACF) sa následne odhadnú rády *p* a *q* v sledovanom ARMA(*p*,*q*) časovom rade. Pre odhad rádu sa hľadá taká hodnota rádu ACF/PACF, že všetky vyššie rády ACF/PACF už nie sú štatisticky významné. Inými slovami sa nezamieta nulová hypotéza o ich nulovosti (podobne ako pri určovaní významnosti jednotlivých parametrov v modele lineárnej regresie).

Pri odhadoch významnosti jednotlivých rádov sa používajú tzv. aproximácie, Bartlettová pre ACF (Bartlett, 1946) a Quenouilleová pre PACF (Quenouille, 1949). Tieto aproximácie slúžia k určeniu hranice významnosti, ktorá vyjadruje kritickú hodnotu pre spomínanú nulovú hypotézu o nulovosti daného rádu ACF/PACF. Pokiaľ je hodnota ACF/PACF v tomto ráde vyššia ako hranica významnosti, tak zamietame nulovú hypotéze, tj. na hladine významnosti 5% prehlásime ACF/PACF v danom ráde za nenulovú. V opačnom prípade, tj. keď je hodnota ACF/PACF v tomto ráde nižšia ako hranica výzmanosti, tak nezamietame nulovú hypotézu na hladine významnosti 5% a na daný rád ACF/PACF sa pozeráme ako na nevýznamný (s nulovou hodnotou). Bohužiaľ, Bartlettová a Quenouilleová aproximácia platí len pre výberové odhady ACF a PACF, pretože vychádzajú z pravdepodobnostných rozdelení výberových odhadov. V prípade iných odhadov ACF a PACF je potrebné odvodiť požadované pravdepodobnostné rozdelenia týchto odhadov, aby sme na ich základe mohli určiť hranice významnosti.

Pre odhad ACF založený na useknutí časového radu sa autorovi tohto článku podarilo odvodiť obdobu Barttletovej aproximácie, teda asymptotické rozdelenie tohto odhadu. V tomto článku sa preto zameriame na obdobu Quenouilleovej aproximácie, špeciálne na odhad PACF založený na useknutí časového radu pre prípad časového radu AR(*p*).

Ako v článku uvidíme, odhad PACF založený na useknutí časového radu je svojou konštrukciou komplikovaný a nie je vôbec jednoduché odvodiť jeho skutočné či aspoň asymptotické rozdelenie. Aby sme disponovali aspoň nejakou pomôckou pri určovaní hranice významnosti, tak sme v tomto článku pomocou simulačnej štúdie vytvorili „pravidlo palca“. Toto pravidlo následne využijeme v praxi pri identifikácii rádu časového radu AR(*p*).

Definície a značenia

Najskôr definujeme biely šum ako časový rad $\{\varepsilon_n, n \in N_0\}$, pre ktorý platí, že ε_n sú centrované, navzájom nekorelované náhodné veličiny s neznámym konštantným rozptylom $\sigma_\varepsilon^2 > 0$.

Následne definujeme autoregresívny časový rad AR(p) pomocou rovnice:

$$X_n = \varphi_1 X_{n-1} + \varphi_2 X_{n-2} + \cdots + \varphi_p X_{n-p} + \varepsilon_n,$$

kde $\varphi_1, \varphi_2, \ldots, \varphi_p \in R$ sú parametre, $\{\varepsilon_n, n \in N_0\}$ je biely šum a $\varphi_p \neq 0$.

Ďalej definujeme autokovariančnú funkciu R(k) centrovaného stacionárneho procesu $\{X_n, n \in N_0\}$ rádu k ako:

$$R(k) = E(X_k X_0).$$

Pomocou autokovariančnej funkcie definujeme autokorelačnú funkciu (ACF) $\rho(k)$ centrovaného stacionárneho radu $\{X_n, n \in N_0\}$ rádu k ako:

$$\rho(k) = \frac{R(k)}{\sigma_X^2},$$

kde σ_X^2 je rozptyl časového radu.

Využitím autokorelačnej funkcie definujeme parciálnu autokorelačnú funkciu (PACF) $\alpha(k)$ centrovaného stacionárneho radu $\{X_n, n \in N_0\}$ rádu k ako:

$$\alpha(1) = \rho(1)$$
$$\alpha(k) = corr\left(X_k - \widetilde{X}_k, X_0 - \widetilde{X}_0\right), k > 1,$$

kde *corr* značí koreláciu a $\widetilde{X}_0$ ($\widetilde{X}_k$) je projekcia X_0 (X_k) na Hilbertov priestor vymedzený náhodnými veličinami $X_1, X_2, X_3, \ldots, X_{k-1}$.

Na záver si definujeme useknutý časový rad $\{Z_n, n \in N_0\}$, ktorý vzniká useknutím pôvodného časového radu $\{X_n, n \in N_0\}$ v hranici useknutia $h=0$:

$$Z_n = 0, \qquad X_n < 0,$$
$$Z_n = 1, \qquad X_n \geq 0.$$

Vlastnosti ACF a PACF

Odhad PACF založený na useknutí časového radu získame pomocou ACF useknutého časového radu. Využijeme analytický vzťah medzi ACF pôvodného a ACF useknutého časového radu, (Kedem, 1980):

$$\rho_X(k) = \frac{2}{\pi} \arcsin\left(\rho_Z(k)\right),$$

kde $\rho_X(k)$ je ACF pôvodného časového radu a $\rho_Z(k)$ je ACF useknutého časového radu.

Pre získanie odhadu PACF následne použijeme vzťah medzi ACF a PACF, (Yaffe a McGee, 2009):

$$\alpha(k)=\frac{\begin{vmatrix} 1 & \rho(1) & \Lambda & \rho(k-2) & \rho(1) \\ \rho(1) & 1 & \Lambda & \rho(k-3) & \rho(2) \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} & \mathrm{M} \\ \rho(k-1) & \rho(k-2) & \Lambda & \rho(1) & \rho(k) \end{vmatrix}}{\begin{vmatrix} 1 & \rho(1) & \Lambda & \rho(k-1) \\ \rho(1) & 1 & \Lambda & \rho(k-2) \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \rho(k-1) & \rho(k-2) & \Lambda & 1 \end{vmatrix}}, k>1,$$

kde $|\cdot|$ symbolizuje determinant.

Posledným vzťahom, ktorý si v tejto kapitole formulujeme je už spomínaná Quenouilleová aproximácia, (Quenouille, 1949). Pokiaľ platí, že $\alpha(k)=0$ pre $k>k_0$, potom:

$$\hat{\alpha}(k)\sim N\left(0,\frac{1}{m}\right), k>k_0,$$

kde $\hat{\alpha}(k)$ je výberová PACF a *m* je počet pozorovaní, z ktorých výberovú PACF počítame.

Odhad založený na useknutí časového radu

Odhad PACF, založený na useknutí časového radu, získame následovným postupom:

z pôvodného časového radu $\{X_n, n\in N_0\}$ získame useknutý časový rad $\{Z_n, n\in N_0\}$,

spočítame výberovú ACF z useknutého časového radu $\{Z_n, n\in N_0\}$,

využitím vzťahu medzi ACF pôvodného a ACF useknutého časového radu spočítame ACF pôvodného časového radu $\{X_n, n\in N_0\}$,

využitím vzťahu medzi ACF a PACF spočítame PACF pôvodného časového radu $\{X_n, n\in N_0\}$.

Vychádzajúc z Quenouilleovej aproximácie sa pre výberovú PACF hľadá také k_0, pre ktoré platí:

$$|\hat{\alpha}(k)|>2\sqrt{\frac{1}{m}}, k>k_0.$$

Odhad výberovej PACF má rozptyl $\frac{1}{m}$ a samotná hranica významnosti $2\sqrt{\frac{1}{m}}$ predstavuje 97,5% kvantil normálneho rozdelenia s daným rozptylom. Ide o 5% hladinu významnosti, ale zároveň obojstrannú alternatívnu hypotézu. Hodnota 2 je aproximácia hodnoty 97,5% kvantilu normálneho rozdelenia, ktorá je pri zaokrúhlení na 2 desatinné miesta 1,96.

Samozrejme, že nemôžeme očakávať rovnakú hodnotu rozptylu odhadu pri odhade založenom na useknutí časového radu ako pri výberovej PACF. Useknutím časového radu sa stráca množstvo informácie, preto je prirodzené, že rozptyl odhadu založenom na useknutí bude väčší ako rozptyl výberového odhadu PACF. Vytvorili sme preto simulačnú štúdiu, aby sme skúmali rozptyl nášho odhadu PACF založenom na useknutí v pomere ku rozptylu výberového odhadu PACF.

Simulačná štúdia

Simulačná štúdia bola vykonaná v štatistickom programe R. Bola rozdelená do 3 základných prípadov autoregresného časového radu, tzn. AR(1), AR(2) a AR(3). Tieto 3 prípady sa pokúsime pokryť čo

najlepšie, preto vytvoríme kombinácie parametrov tak, aby reprezentovali čo najviac prípadov. Sledovali multiplikatívnu zmenu rozptylu odhadu PACF pomocou metódy založenej na useknutí časového radu voči rozptylu pre výberovú PACF.

Pre prípad AR(1) sme sledovali ešte aj dopad počtu pozorovaní na skúmaný pomer rozptylu odhadu založenom na useknutí časového radu voči rozptyl výberového odhadu.

Každý jeden prípad hodnoty parametra φ sme simulovali 2000 krát a pracovali sme s jednotkovým rozptylom pre biely šum. Rozptyl sme merali len pre odhady PACF rádu vyššieho ako je rád autoregresného časového radu p (pre tieto hodnoty platí Quenouilleová aproximácia) a maximálne do 6. rádu.

Prípad AR(1)

Pre AR(1) model sme simulovali hodnoty parametru $\varphi = -0{,}9, -0{,}8, \mathrm{K}\ ,0{,}8, 0{,}9$. Ide o relatívne hustú sieť parametrov, preto v tomto prípade by malo byť pokrytie dostatočné. Pre každý prípad sme uvažovali 1000 pozorovaní a následne 10 000 pozorovaní, aby sme dokázali vyhodnotiť závislosť multiplikatívneho dopadu na rozptyl od počtu pozorovaní.

Pre každý prípad je odhadnutá výberová autokorelačná funkcia a autokorelačná funkcia založená na useknutí časového radu až do 6. rádu. Následne je z 2000 simulácií spočítaný výberový rozptyl týchto odhadov pre každú hodnotu parametra a pre každý rád autokorelačnej funkcie. Na záver sme spočítali pomery rozptylov odhadov získaných metódou založenou na useknutí voči rozptylom výberových odhadov. Tieto pomery prezentujeme v tabuľkách uvedených nižšie.

Tabuľka 1: AR(1), *n* = 1000

φ	***k*=2**	***k*=3**	***k*=4**	***k*=5**	***k*=6**
-0,9	10,07	14,63	19,90	25,29	30,58
-0,8	7,41	8,64	9,78	12,01	12,92
-0,7	5,30	6,00	6,84	7,50	7,42
-0,6	4,48	4,62	5,03	5,53	5,80
-0,5	3,53	3,95	3,86	4,31	4,55
-0,4	3,32	3,31	3,40	3,16	3,51
-0,3	2,91	2,85	2,82	2,97	3,07
-0,2	2,60	2,54	2,64	2,66	2,62
-0,1	2,41	2,69	2,67	2,41	2,56
0,0	2,29	2,53	2,47	2,27	2,72
0,1	2,56	2,56	2,49	2,62	2,58
0,2	2,58	2,75	2,75	2,67	2,71
0,3	2,76	2,97	2,73	3,04	3,01
0,4	3,26	3,42	3,34	3,26	3,50
0,5	3,86	3,91	4,25	4,21	4,04
0,6	4,21	4,89	4,86	5,18	5,16
0,7	5,50	6,75	7,13	7,04	8,10
0,8	6,91	8,92	9,66	10,99	12,84
0,9	11,10	14,53	17,95	21,45	29,68

Zdroj: vlastný

Tabuľka 2: AR(1), *n* = 10 000

φ	***k*=2**	***k*=3**	***k*=4**	***k*=5**	***k*=6**
-0,9	10,45	13,79	15,05	16,89	18,63
-0,8	7,02	8,27	9,04	9,59	10,46
-0,7	5,84	6,25	6,55	6,25	7,03
-0,6	4,59	4,69	5,46	4,93	4,67
-0,5	3,58	3,69	3,83	4,21	3,72
-0,4	3,17	3,22	3,31	3,37	3,41
-0,3	2,90	2,93	2,89	2,86	2,95
-0,2	2,78	2,49	2,82	2,66	2,64
-0,1	2,50	2,57	2,73	2,44	2,40
0,0	2,57	2,33	2,34	2,45	2,56
0,1	2,40	2,58	2,62	2,65	2,67
0,2	2,89	2,50	2,76	2,47	2,48
0,3	2,66	3,08	2,92	3,01	3,00
0,4	3,30	3,50	3,40	3,43	3,37
0,5	3,78	4,16	3,84	3,87	3,98
0,6	4,14	4,88	4,25	5,08	5,28
0,7	5,00	6,18	6,41	6,67	7,00
0,8	6,93	8,57	8,56	9,62	10,33
0,9	10,53	13,02	16,13	18,31	17,92

Zdroj: vlastný

Ako vidíme z priložených tabuliek, tak počet pozorovaní nemá výrazny vplyv na multiplikatívny dopad rozptylu. Výnimku tvoria okrajové prípady, tj. prípady s parametrom blízko nestacionarite alebo veľkého rádu PACF.

Aj keď vezmeme v úvahu tieto okrajové prípady, môžeme prehlásiť, že z našej simulačnej štúdie nám vyšlo, že rozptyl je vždy väčší viac ako 2-násobne. Keby sme naopak neuvažovali parametre blízko 0, ktoré sú väčšinou ťažko pozorovateľné, tak by sme mohli prehlásiť odvážnejšie tvrdenie, a to že rozptyl vychádza takmer vždy viac ako 3-násobne.

Prípad AR(2)

Pre AR(2) model sme simulovali hodnoty parametrov $\varphi_1 = -0{,}9, -0{,}7, \mathrm{K}\,, 0{,}7, 0{,}9$ a $\varphi_2 = -0{,}9, -0{,}7, \mathrm{K}\,, 0{,}7, 0{,}9$, pričom pri každej dvojici parametrov je najskôr zisťované či spĺňa podmienky stacionarity. Pre každý prípad sme uvažovali 1000 pozorovaní, keďže sme v predchádzajúcej kapitole vyhodnotili, že počet pozorovaní nemá významný vplyv na výsledok.

Tabuľka 3: AR(2), *n* = 1000

φ_1	φ_2	***k*=3**	***k*=4**	***k*=5**	***k*=6**
-0,9	-0,9	112,08	209,33	338,47	335,51
-0,7	-0,9	74,99	111,71	181,70	252,44
-0,5	-0,9	50,26	63,82	117,63	175,93
-0,3	-0,9	38,47	32,31	66,15	97,94
-0,1	-0,9	32,80	16,75	42,63	42,90
0,1	-0,9	33,44	17,38	38,63	44,31
0,3	-0,9	40,02	29,96	60,81	77,93
0,5	-0,9	50,64	54,60	89,60	146,95
0,7	-0,9	73,38	98,02	162,92	223,25
0,9	-0,9	108,35	174,88	280,67	312,59
-0,9	-0,7	24,05	39,70	66,26	117,12
-0,7	-0,7	16,49	24,04	36,31	59,01
-0,5	-0,7	10,72	9,99	15,60	23,40
-0,3	-0,7	9,56	7,27	10,34	10,87
-0,1	-0,7	8,69	6,19	7,80	7,34
0,1	-0,7	8,59	5,97	7,71	7,45
0,3	-0,7	9,38	7,03	11,48	10,64
0,5	-0,7	11,26	11,44	15,73	23,48
0,7	-0,7	18,54	21,91	35,13	55,08
0,9	-0,7	22,07	36,37	57,36	93,19
-0,9	-0,5	12,95	18,93	25,08	38,80
-0,7	-0,5	8,98	10,75	13,70	17,98
-0,5	-0,5	6,29	6,66	7,02	8,52
-0,3	-0,5	4,85	4,85	5,12	5,20
-0,1	-0,5	4,74	3,85	3,80	4,34
0,1	-0,5	4,17	3,93	4,02	4,24
0,3	-0,5	4,87	4,64	4,91	5,47
0,5	-0,5	6,21	6,62	6,69	8,42
0,7	-0,5	8,62	10,43	12,82	16,26
0,9	-0,5	13,16	16,95	25,20	34,95
-0,9	-0,3	9,92	12,84	15,53	21,70
-0,7	-0,3	6,71	7,12	7,45	8,86
-0,5	-0,3	4,34	4,38	4,13	5,22
-0,3	-0,3	3,60	3,37	3,54	3,68
-0,1	-0,3	3,09	2,95	2,82	3,22
0,1	-0,3	2,96	2,85	3,30	3,03
0,3	-0,3	3,71	3,22	3,55	3,49
0,5	-0,3	4,39	4,79	4,54	4,28
0,7	-0,3	6,73	7,48	8,17	8,55
0,9	-0,3	9,95	13,04	14,42	18,81
-0,9	-0,1	11,77	12,52	15,48	21,20
-0,7	-0,1	6,06	6,13	7,00	7,41
-0,5	-0,1	3,78	4,00	3,89	4,12
-0,3	-0,1	3,05	3,04	3,13	2,85
-0,1	-0,1	2,59	2,86	2,69	2,75
0,1	-0,1	2,68	2,57	2,54	2,69
0,3	-0,1	2,84	2,90	2,99	3,02
0,5	-0,1	3,74	3,81	4,13	4,52
0,7	-0,1	5,97	6,03	6,85	6,97
0,9	-0,1	10,49	12,95	15,49	19,19
-0,7	0,1	6,97	7,92	9,39	9,58
-0,5	0,1	3,90	4,52	4,23	4,26
-0,3	0,1	3,10	3,37	3,05	3,13
-0,1	0,1	2,43	2,83	2,71	2,59
0,1	0,1	2,59	2,50	2,54	2,70
0,3	0,1	3,29	2,88	3,00	2,85
0,5	0,1	4,30	4,12	4,38	4,20
0,7	0,1	7,18	7,74	7,94	9,54
-0,5	0,3	5,38	6,10	6,41	6,73
-0,3	0,3	3,58	3,41	4,05	3,93
-0,1	0,3	3,03	2,95	2,95	3,07
0,1	0,3	3,10	2,96	3,01	2,91
0,3	0,3	3,34	3,33	3,48	3,85
0,5	0,3	5,57	6,40	6,13	6,78
-0,3	0,5	4,99	5,17	5,35	6,97
-0,1	0,5	3,77	3,93	4,11	4,18
0,1	0,5	3,60	3,55	4,32	4,35
0,3	0,5	5,08	5,36	5,52	5,82
-0,1	0,7	5,71	6,31	7,65	7,12
0,1	0,7	5,74	5,44	7,86	6,44

Zdroj: vlastný

Pre AR(2) vychádzajú trochu volatilnejšie hodnoty, ale stále platí, že rozptyl je vždy väčší ako 2-násobne. Taktiež platí, že keby sme nebrali v úvahu parametre blízko 0, tak by sme mohli rozptyl prehlásiť až za takmer vždy viac ako 3-násobný. Extrémne vysoké hodnoty dostávame v prípade, že φ_2 vychádza blízko 1 v absolútnej hodnote.

AR(3)

Pre AR(3) model sme simulovali hodnoty parametrov $\varphi_1 = -0{,}9, -0{,}6, \mathrm{K}\ , 0{,}6, 0{,}9$, $\varphi_2 = -0{,}9, -0{,}6, \mathrm{K}\ , 0{,}7, 0{,}9$ a $\varphi_3 = -0{,}9, -0{,}6, \mathrm{K}\ , 0{,}7, 0{,}9$, pričom pri každej trojici parametrov je najskôr zisťované či spĺňa podmienky stacionarity. Pre každý prípad sme uvažovali 1000 pozorovaní.

Prípadov, ktoré sú stacionárne, je 163. Kvôli vyššiemu počtu by už nebolo veľmi praktické výsledky zobrazovať do tabuľky. Výsledky vyšli veľmi podobne ako pre prípad AR(2). Najnižšia hodnota vyšla 2,25 a najvyššia zase 585,05. Celkovo vyšlo len 14 hodnôt z 489, ktoré boli menšie ako 3. Z uvedených hodnôt teda vyplýva rovnaký záver ako pre prípad AR(1) a prípad AR(2).

Vyšli nám teda 2 varianty pravidla. Opatrnejší variant pravidla by bol počítať s dvojnásobne väčším rozptylom, pretože žiadna z hodnôt nebola menšia ako dvojnásobok. Pokiaľ by sme kládli väčší dôraz na vyššie hodnoty parametrov, tak by sme mohli pracovať s pravidlom trojnásobného rozptylu.

Aplikácia

Nové pravidlo palca si demonštrujeme pri identifikácii časového radu AR(3), ktorý je infikovaný aditívnymi odľahlými pozorovaniami. Pravdepodobnosť, že dané pozorovanie bude odľahlé je 5%. Najskôr pre jeho odhad budeme pracovať so štandardnou Quenouilleovou aproximáciou a následne do nej pridáme multiplikatívnu konštantu *c*, ktorú sme sa snažili získať v simulačnej štúdii.

Pre našu aplikáciu sme vytvorili 5000 simulácii, v ktorých sme generovali absolútnu hodnotu parametrov časového radu AR(3) náhodne podľa rovnomerného rozdelenia, tj. $\varphi_i \sim U(0{,}2, 1{,}0)$. Hodnoty blízko 0 nie sú brané v úvahu, pretože sú zvyčajne veľmi ťažko identifikovateľné. Znamienko parametrov bolo určené taktiež náhodne, a to s alternatívnym rozdelením s pravdepodobnosťou úspechu π = 0,5. Po vygenerovaní potrebných parametrov bola následne skontrolovaná stacionarita časového radu a v prípade nestacionarity boli vygenerované nové parametre.

Každá simulácia bola identifikovaná ako AR(*p*) s rádom *p* medzi 1 až 6, MA(*q*) s rádom *q* medzi 1 až 6 alebo ARMA. Pokiaľ by mali vychádzať vyššie rády *p* a *q* ako 6, tak bolo uvažované, že taký časový rad nemá useknutú PACF, resp. ACF. Na záver boli dané všetky výsledky pre MA(*q*) dané dohromady, aby mali výsledky čitateľnejšiu hodnotu. Výsledky prinášame v nasledujúcej tabuľke.

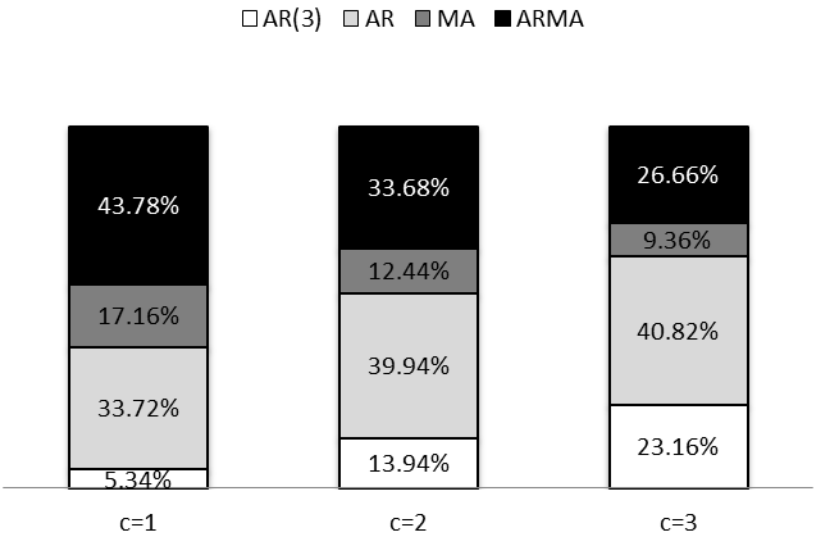
Tabuľka 4: Aplikácia

v %	***c*=1**	***c*=2**	***c*=3**
AR(1)	0,00	0,02	0,10
AR(2)	0,08	0,42	1,04
AR(3)	5,34	13,94	23,16
AR(4)	6,22	10,00	11,54
AR(5)	10,84	13,46	14,22
AR(6)	16,58	16,04	13,92
MA	17,16	12,44	9,36
ARMA	43,78	33,68	26,66

Zdroj: vlastný

Z priloženej tabuľky vidíme, že multiplikátor výrazne zvyšuje správnosť identifikácie časového radu AR(3). V prvom stĺpci vidíme hodnoty pre prípad nepoužívania multiplikátora, resp. keď sa rovná 1. Je vidieť, že s rastúcou hodnotou multiplikátora sa presúva podiel simulácii vyhodnotených ako ARMA, tj. bez useknutia PACF smerom k AR(*p*), a to najvýraznejšie práve k skutočnej hodnote rádu. Mohlo by sa dokonca zdať, že v prípade vyššieho multiplikátoru, by mohlo dojsť ešte k zlepšeniu identifikácie, avšak v tom prípade by sme už riskovali podhodnocovanie veľkosti rádu pre viaceré simulácie.

Na záver si zobrazíme výsledky našej aplikácie ešte graficky.



Obrázok 1: Aplikácia Zdroj: vlastný

V grafe vidíme jednoznačný presun z ARMA do AR(3) s rastúcou hodnotou multiplikátoru.

Záver

Pomocou simulačnej štúdie sme ukázali, že metóda odhadu PACF založenom na useknutí má niekoľko násobne vyšší rozptyl odhadu v porovnaní s Quenouilleovou aproximáciou pre výberovú PACF. Ukázali sme, že počet pozorovaní nehrá rolu, avšak hodnoty parametrov očividne rolu hrajú. Napriek tomu, že sa nám nepodarilo ukázať explicitný vzťah pre rozptyl odhadu, vyslovili sme pravidlo palca o dvojnásobnej a trojnásobnej hodnote rozptylu.

Využitie tohto pravidla sme demonštrovali na odhade rádu časového radu AR(3), kde využitie spomínaných pravidiel palca malo za následok výrazne zlepšenie identifikácie časového radu.

Okrem tejto aplikácie môže nájsť toto pravidlo uplatnenia aj v ďalších prípadoch. Pomocou neho sme schopný lepšie určiť hranicu významnosti PACF pre jednotlivé rády, čo má za následok lepšiu diagnostiku časového radu a samotných dat. Lepšia diagnostika časového radu vedie ku konštrukcii lepšieho modelu, ktorý následne dáva dôveryhodnejšie predpovede.

Literatúra

[1] BARTLETT, Maurice S. On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society*. 1946, č. 1, s. 27 – 41.

[2] BOX, George E. P., Gwilym M. JENKINS a Gregory C. REINSEL. Time series analysis: forecasting and control. 4th ed. Hoboken: John Wiley, c2008. Wiley series in probability and statistics. ISBN 978-0-470-27284-8.

[3] KEDEM, Benjamin. *Binary Time Series*. New York: Marcel Dekker, 1980.

[4] QUENOUILLE, Maurice H.. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society: Series B*. 1949, č. 3, s. 68-84

[5] YAFFE, Robert. A. a Monnie McGEE. *Introduction to time series analysis and forecasting: with applications of SAS and SPSS*. San Diego: Academic Press, 2009.

JEL Classification: C22, C24.

Dedikacia: Príspevok je spracovaný ako súčasť projektu IG410027.

Summary

Název článku v angličtině Variance of the PACF estimator based on clipping AR(p)

Firstly, for Box – Jenkins methodology, it is demanding to estimate orders for process ARMA(p,q). To obtain orders p and q, it recommends to estimate autocorrelation function (ACF) and partial autocorrelation function (PACF). Subsequently, we need to determine a critical value, which can help us to determine significance of ACF/PACF for specific orders. For this matter, Bartlett's approximation was proposed for ACF and Quenouille's approximation for PACF. Unfortunately, these approximations are working only for the sample estimator of ACF and PACF, so for other estimators we need to derive their own ciritical values. Especially, it holds for the case of estimator of PACF based on the clipped original time series. So far, nobody has derived distribution of the estimator, and so we can at least take some rule of thumb from simulation study. In this paper, we design such a simulation study and we propose such a rule of thumb that could help us in practice. Demostrantion of a practical use is provided with its impact on the order estimation of AR(p).

Key words: PACF, estimate variance, AR(p)

Analýza přístupů pro posouzení výnosnosti produktů životního pojištění z pohledu klienta

Jan Fojtík

xfojj00@vse.cz

Doktorand oboru statistika

Školitel: prof. Ing. Richard Hindls, CSc., dr. h. c., (richard.hindls@vse.cz)

Abstrakt: Finanční společnosti vytváří nové produkty cílící zejména na retailové investory. Tyto produkty jsou vytvářeny s ohledem na požadavky klientů po straně výnosů či rizika. Díky velké flexibilitě se výsledné produkty často stávají komplikované a běžný klient může mít problém se v nich orientovat. Za účelem zvýšení transparentnosti investičních produktů bylo vydáno nové legislativní nařízení PRIIP's. Součástí tohoto nařízení je strukturovaný dokument, který by měl jednotně popsat investiční produkty zejména po stránce výnosu a rizika. Mezi investiční produkty, na které se tato legislativa vztahuje, patří investiční životní pojištění. Cílem tohoto příspěvku je analýza legislativy PRIIP's, jako vhodného ukazatele pro porovnání výkonosti životního pojištění a prozkoumání možných problémů při finální intepretaci směrem k běžnému klientovi. Výsledky této analýzy jsou konfrontovány s výsledky autorovy diplomové práce, která byla zaměřena na vytvoření ukazatele pro porovnání produktů životního pojištění. V úvodu práce je definován pojem optimálního ukazatele a vysvětleny základní vlastnosti, které by optimální ukazatel měl mít. V praktické části práce je analyzováno, do jaké míry ukazatel sestrojený dle legislativy PRIIP's splňuje vlastnosti optimálního ukazatele. Výsledná analýza studuje citlivost výsledků dle PRIIP's na změnu základních pojistných parametrů, jako je výše pojistného plnění, pojistné schéma, úroková míra nebo frekvence placení pojistného. Z výsledků práce vyplývá, že použití PRIIP's je dobrá cesta k zvýšení transparentnosti finančních produktů a ochraně klientů, ale pro investiční životní pojištění může být v určitých případech matoucí.

Klíčová slova: životní pojištění, PRIIPS, investiční produkty, porovnání výnosnosti

Úvod

V dnešní době finanční instituce nabízejí širokou škálu investičních produktů cílících zejména na retailové klienty. Mezi nejčastěji nabízené produkty pro běžného klienta patří zejména investiční fondy, strukturované investiční produkty, strukturované vklady nebo investiční životní pojištění. Specifikum těchto produktů spočívá ve značné flexibilitě a schopnosti přizpůsobit se klientům po stránce výnosu i rizika. Velká flexibilita na jednu stranu pomáhá klientovi nastavit si produkt podle svých potřeb, ale na straně druhé se díky své komplexnosti finální produkty stávají složitými a obtížně pochopitelnými. Z tohoto důvodu klienti často investují do produktů, kterým nerozumí, nebo nejsou zcela obeznámeni s podstupovaným rizikem či nákladovostí. Výsledné nenaplnění očekávaných investičních cílů je dále odrazuje od dalšího využívání těchto produktů a nevrhá dobré světlo na kapitálový trh, který je nedílnou součástí tržního hospodářství. Výrazný posun ve zvýšení transparentnosti a srozumitelnosti investičních produktů nabízených drobným investorům je důležitým krokem pro ochranu drobných investorů a obnovení důvěry ve finanční trhy. V návaznosti na zpřehlednění finančních produktů lze v několika minulých letech pozorovat nárůst internetových srovnávačů umožňujících porovnat výnosnost či nákladovost finančních produktů. Tyto srovnávače cílí především na nejčastěji používané finanční produkty, jako je spotřební úvěr, hypotéka nebo pojistné produkty jako je povinné ručení a životní pojištění. Nástup srovnávačů přinesl klientům rychlý a jednoduchý přehled o nabízených produktech. Otázkou však zůstává transparentnost a důvěryhodnost srovnávacího mechanismu. Za provozovateli těchto srovnávačů se mohou skrývat finanční zprostředkovatelé nebo samotné finanční instituce, které mohou upřednostňovat svoje produkty nebo produkty s vyšším výnosem nebo provizí. Výsledky srovnávačů proto nemusí odpovídat skutečným potřebám klienta, který si může opět pořídit nevhodný produkt. Následky špatně zvoleného finančního produktu mohou běžného klienta finančně zatížit na mnoho let. Následné vyvázání se z dlouhodobých produktů typu životního pojištění, hypotéky nebo většího spotřebního úvěru nemusí být jednoduché a může být i finančně náročné díky vysokým storno poplatkům. Speciální pozornost si díky své důležitosti zaslouží především pojistné produkty, které jsou svým

charakterem nejsložitější. U ostatních finančních produktů je riziko špatně zvoleného produktu spojeno pouze s finanční ztrátou. Jestliže klient plně nerozumí životnímu produktu, tak v případě vzniku pojistné události nemusí obdržet pojistné plnění, na které se původně zamýšlel pojistit. Tento fakt může mít velký společenský dopad zejména pro mladé rodiny nebo osoby, které by důsledkem úrazu nebo úmrtí přišly o příjem. Výše zmíněné důvody ukazují, že nalezení vhodného ukazatele pro porovnání pojistných produktů je velmi důležitý krok ochrany klientů před nákupem nevhodných životních produktů. Vytvoření metody, kterou lze vhodně porovnat různé pojistné produkty by mohlo přispět k lepšímu pochopení pojistných produktů širokou veřejností a zejména ulehčit rozhodování při výběru životního pojištění.

Aby bylo možné korektně porovnávat produkty životního pojištění, musí daný ukazatel splňovat určité charakteristiky. Pokud bude tento ukazatel vhodně vystihovat tyto charakteristiky, lze jej považovat za „*optimální*“. Optimální ukazatel by měl **popsat produkty životního pojištění jako celek**, tedy včetně často opomíjené rizikové složky. Nedostatkem většiny dosavadních ukazatelů je orientace pouze na výnos nebo náklady. Nákladové ukazatele nemusí dostatečně popisovat produkt a klienti můžou nabývat mylně dojmu, že levnější produkt znamená lepší. Obdobně to platí pro výnos, produkt, který slibuje největší investiční výnos, nemusí být nutně nejvhodnější jejich potřebám. Zahrnutí všech tří složek -výnos, náklady a riziko - do ukazatele je důležitou podmínkou pro porovnání mezi pojistnými produkty. Životní pojistky se zpravidla neřadí mezi krátkodobé produkty a mnohdy jsou uzavírané i na desetiletí. Zejména pokud jsou tyto produkty sjednávány jako krytí hypoték, tak pojistná doba často dosahuje i třicet let. V jiných případech mohou být tyto produkty sjednávány doživotně nebo do důchodového věku. Schopnost **popsat produkty životního pojištění i z dlouhodobého hlediska** by měla být nedílnou vlastností optimálního ukazatele. V praxi lze pořídit širokou škálu pojistných produktů, které se mohou lišit zejména typem výplaty pojistného plnění nebo frekvencí pojistného. Někteří klienti potřebují produkt, který pokryje pouze riziko, jiní zase chtějí spíše zhodnocovat svoje prostředky. Například pokud je pojistná smlouva svázána s hypotečním úvěrem, tak pojistná částka klesá úměrně dluhu. V jiných případech může být hodnota výplaty pojistného plnění odvozena od úspěšnosti investiční strategie. Dalším způsobem, jak lze rozdělit životní produkty, je frekvence placení pojistného. Pojistné lze platit měsíčně, čtvrtletně, ročně nebo formou jednorázového pojistného při sjednání produktu. Optimální ukazatel by měl být schopen **porovnat odlišné produkty životního pojištění**. Poslední a zásadní vlastností optimálního ukazatele je **jednoznačná, nezavádějící a srozumitelná intepretace**. Smyslem optimálního ukazatele je vytvořit rychle pochopitelnou pomůcku porovnávající vícero pojistných produktů. Snadno si lze představit, že široká veřejnost se neorientuje v pojistných nebo finančních produktech na odborné úrovni. Ukazatel, jehož hodnota je matematicky nebo finančně složitě interpretovatelná, proto nemusí být klienty kladně přijímán.

V České republice vznikla v posledních několika letech iniciativa (Dlouhá, 2014, TANK. 2011 nebo Produktové listy.cz, 2012) s úmyslem rozkrýt složitou a netransparentní poplatkovou strukturu pojistných produktů. Výsledkem je 7 ukazatelů v čele s ukazatelem SUN, který Česká asociace pojišťoven doporučuje jako nejvhodnější (Česká asociace pojišťoven, 2014). Tyto ukazatele jsou převážně nákladového charakteru a nemusí být úplně vhodné pro popsání životních produktů. Podrobnější analýza a porovnání kvalit jednotlivých ukazatelů je popsána v (Fojtík, 2016). Za účelem odstranění nedostatků v transparentnosti a srozumitelnosti investičních produktů vydala Evropská unie nově nařízení Evropského Parlamentu o sděleních klíčových informací týkajících se strukturovaných retailových investičních produktů a pojistných produktů s investiční složkou (Eur-lex.europa.eu., 2014). S tímto nařízení se lze setkat pod zkratkou PRIIP's z anglického Packaged Retail and Insurance-based Investment Products. Smyslem PRIIP's je snaha o sjednocení publikovaných informací o investičních produktech napříč všemi finančními institucemi na území Evropské unie, mezi které se řadí i investiční životní pojištění. Finanční instituce jsou nově povinny publikovat informace o svých produktech v jednotné šabloně. Klient by nyní měl mít lepší přehled o základních vlastnostech a měl by si udělat lepší představu o vhodnosti a výnosnosti zvažovaného produktu. PRIIP's metodika vytvořila všeobecný přístup, který jednotně popisuje všechny investiční produkty včetně investičního životního pojištění. Cílem práce je prozkoumat, jestli je tato metoda vhodná pro posouzení investičního životního pojištění a zda budou její výsledky korektně popisovat charakter pojistných produktů. Výsledky z PRIIP's metodologie budou porovnány s ukazatelem založeným na analýze peněžních toků představené v (Fojtík, 2016).

Základní myšlenka a metodika PRIIP‘s

Hlavním smyslem nové legislativy PRIIP‘s je sjednocení publikovaných informací o investičních produktech. Finanční instituce jsou nově pro produkty s investiční složkou povinny vytvořit jednotnou šablonu zvanou Key Informational Dokument (KID). Jednotné struktura KID by měla přispět k lepší orientaci mezi typy, rizikem nebo výnosem všech investičních produktů. Pro lepší orientaci by KID neměl přesáhnout 3 strany, které jsou rozděleny na 6 oddílů. Jednotlivé oddíly mají za úkol informovat klienta o základních vlastnostech produktu. Jmenovitě se jedná o:

1. „**Produkt**“ obsahující informace o názvu a tvůrci produktu spolu s kontaktními informacemi.
2. „**O jaký produkt se jedná?**“ obsahující základní rysy produktu a pro koho je určen.
3. „**Jaká podstupuji rizika a jakého výnosu bych mohl dosáhnout?**“ obsahující stručný popis očekávaných rizik a výnosů včetně různých vývojových scénářů.
4. „**Co se stane, když tvůrce produktu není schopen uskutečnit výplatu?**“ obsahující informace o nároku na odškodnění a záruky.
5. „**S jakými náklady je investice spojena?**“ obsahující informace o nákladové struktuře produktu.
6. „**Jak dlouho bych měl investici držet? Mohu si peníze vybrat předčasně?**“ obsahující informace o doporučené době držení a o dopadech předčasného ukončení produktu.

Výše uvedený seznam je koncipován tak, aby potencionální klienty informoval o očekávaném výnosu a nejdůležitějších rizicích spojených s držením či odstoupením od smlouvy. Zásadním faktorem úspěchu KID bude především forma a správnost prezentovaných výsledků. Je racionální, že klienti budou preferovat produkty s vyšším výnosem, proto by měl být kladen velký důraz na jeho realistický a přesný odhad. Výpočet míry výnosu pojistného produktu dle legislativy PRIIP’s je založen na metodologii vnitřního výnosového procenta. Vnitřní výnosové procento se aplikuje na vektor peněžních toků spojených s projekcí pojistné smlouvy, kde na straně nákladů stojí zaplacené pojistné a na straně příjmů hodnota fondu v případě dožití nebo odkupu. Vzorec pro výpočet míru výnosu dle PRIIP’s pomocí vnitřního výnosového procenta lze zapsat jako

$$VVP\left(\left(-pojistné_1\right)+\left(-pojistné_2\right)+...+Fond_N\right).$$

Hlavní složkou z vektoru peněžních toků, kterou má klient možnost ovlivňovat, je hodnota fondu dle investičního úspěchu. Čím větší bude hodnota fondu, tím větší lze očekávat i výnosnost pojistného produktu. Z tohoto důvodu bude mít přepokládaný výnos podkladového aktiva nebo fondu přímý dopad do výnosnosti produktu dle PRIIP’s. Metodika pro výpočet výnosu podkladového aktiva je definována také legislativou PRIIP’s a je založena na analýze historických výnosů z posledních pěti let. Výsledkem jsou čtyři výnosové scénáře: stresový, nepříznivý, průměrný a příznivý. Tyto scénáře jsou postaveny na základě 1%, 10%, 50% a 90% kvantilu rozdělení historických výnosů. Otázkou zůstává, jak kvalitně dokáže analýza pětileté historie predikovat výnos na 30 let dopředu.

Alternativní ukazatel výnosnosti životních produktů

V práci (Fojtík, 2016) je navržen ukazatel, který lze považovat dle výše uvedených kritérií za optimální. Tento ukazatele je založen na analýze očekávaných peněžních toků pojistného produktu. Hodnota očekávaných peněžních toku lze získat součtem všech očekávaných výplat a příjmů spojených s produktem. Pro přehlednost bude výnos dle (Fojtík, 2016) značen IRR a výnos dle PRIIP’s bude značen PRIIP’s. Rozdíl mezi těmito dvěma ukazateli je, že PRIIP’s zohledňuje pouze pojistné a hodnotu fondu a IRR ukazatel zohledňuje jak pojistné, hodnotu fontu tak i hodnotu výplatu za pojistnou. Hodnotu míry výnosu dle IRR ukazatele lze získat pomocí vnitřního výnosového procenta jako

$$VVP\left(ESmrt+EStorno+EFond-EPojistne\right),$$

kde *ESmrt* je vektor očekávaných hodnot pojistného plnění v případě smrti, *EStorno* je vektor očekávaných hodnot odbytného. *EFond* je vektor očekávaných hodnot fondu a *EPojistné* je vektor očekávaných hodnot zaplaceného pojistného.

Testování PRIIP's na vlastnosti optimálního ukazatele

Smyslem této části je prezentovat výsledky ukazatele PRIIP's a pozorovat jeho citlivost na změnu základních parametrů pojistných produktů, které jsou běžně nabízeny. Citlivost bude pozorována na změně těchto parametrů:

1. Změna pojistné částky
2. Změna výnosové míry podkladového aktiva
3. Změna výplatního schématu pojistného plnění
4. Změna frekvence placení pojistného

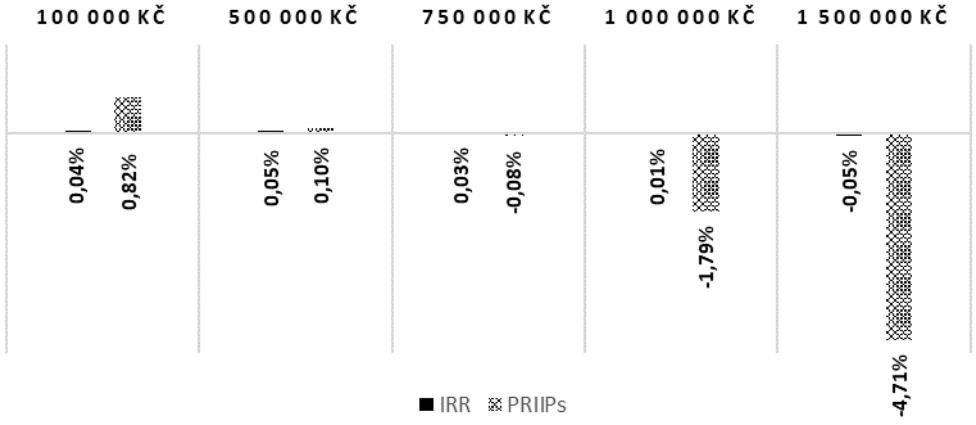
Pro analýzu byl zvolen modelový přiklad investičního životního pojištění se základními parametry:

- Muž ve věku 30 let
- Doporučená doba držení 30 let
- Roční pojistné 12 000 Kč s měsíční frekvencí
- Konstantní pojistná částka 1 000 000 Kč.
- Roční očekávané zhodnocení podkladového aktiva 2 %

Pro srovnání bude na výsledných grafech zobrazen výsledek jak za PRIIP's ukazatel, tak i za IRR.

Změna pojistné částky

V předchozí kapitole bylo naznačeno, že zásadním faktorem ovlivňujícím PRIIP's ukazatel je hodnota fondu na konci doby držení. Na tuto hodnotu nemá vliv pouze výsledek investiční strategie, ale také velikost sjednané pojistného krytí. Při sjednání vyšší pojistného krytí dochází k nárůstu rizikového pojistného, které je strháváno z hodnoty fondu. Zvyšováním pojistného krytí dochází na jedné straně ke snižování hodnoty fondu, a na straně druhé by klientovi v případě pojistné události byla vyplacena vyšší částka. Optimální ukazatel by proto měl tyto dva protichůdné jevy zohlednit. Z obrázku 1 je patrné, že ukazatel PRIIP's reaguje dle očekávání a klesá přímo úměrně růstu pojistného krytí. Tento závěr nelze potvrdit u alternativního ukazatele IRR, který ve výpočtu zohledňuje nárůst pojistného krytí skrze očekávanou výplatu v případě smrti. Tento fakt je viditelný zejména při nárůstu pojistné částky z 100 000 Kč na 500 000 Kč, kdy dochází ke zvyšování očekávaného výnosu dle IRR ukazatele.



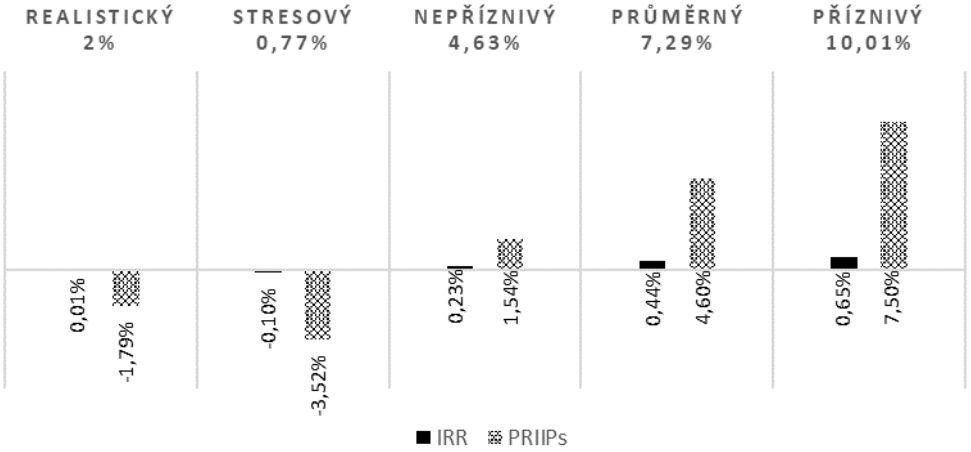
Obrázek 1: Vliv různé hodnoty pojistné částky na ukazatel PRIIP's Zdroj: Vlastní zpracování

Zajímavým poznatkem je vysoká sensitivita PRIIP'S ukazatele, který z pozitivního výnosu 0,82 % při pojistné částce 100 000 Kč klesá až na – 4,71 %. Citlivost PRIIP's na velikost pojistného krytí může být z interpretačního hlediska směrem ke klientovi velmi složitá, protože klient se bude snažit automaticky produktům se záporným výnosem vyhýbat, i když mu v případě pojistné události

přislibují vyšší výplatu. Ukazatele PRIIP's proto může klienty nabádat ke koupi produktů s nízkým nebo nedostatečným pojistným krytím.

Změna výnosové míry podkladového aktiva

Důležitým aspektem investičního životního pojištění je možnost ovlivnit výnos podkladového aktiva skrze výběr investiční strategie. Výnos podkladového aktiva potom přímo ovlivňuje velikost fondu spolu s PRIIP's ukazatelem. Pro lepší přehled o možném investičním dopadu prezentuje PRIIP's výnos na čtyřech možných scénářích vývoje podkladového aktiva. Ukazatel PRIIP's je pro výše zmíněné strategie zobrazen na obrázku 2. Z výsledků vyplývá, že ukazatel PRIIP's je přímo úměrný výši očekávaného zhodnocení podkladového aktiva.



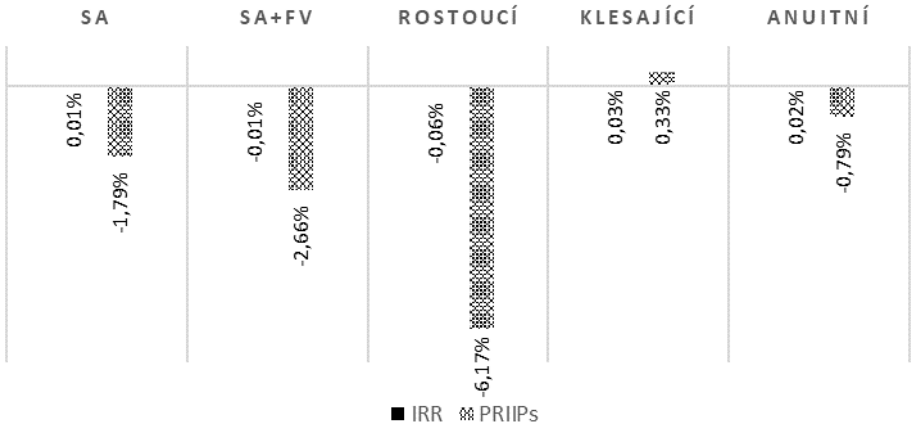
Obrázek 2: Vliv různých výnosových měr na ukazatel PRIIP's Zdroj: Vlastní zpracování

Dále si lze povšimnout, že průměrný investiční výnos podkladového aktiva je odhadován na poměrně vysoké číslo, 7,29 % p.a. Zkušený analytik na první pohled pozná, že tato míra je nadhodnocená a z dlouhodobého hlediska nereálná. Při průměrném scénáři je očekávaný výnos pojistného produktu dle PRIIP's 4,6 %. Při konzervativnějším zhodnocení podkladového aktiva 2 % p.a. je očekávaný výnos pojistného produktu dle PRIIP's -1,79 %. Je pochopitelné, že v reálném světě nebude velké množství klientů, kteří si budou chtít koupit produkt slibující záporný výnos. Produkty se záporným výnosem by nepůsobily příliš důvěryhodně a většina potencionálních klientů by se jich po představení PRIIP's ukazatele stranila. Vyšší očekávaný výnos je zpravidla spojen i s vyšším rizikem. Legislativa PRIIP's zpracovala i metodiku, která by měla produkt charakterizovat i po straně rizika.

Důvodem nadhodnocení očekávaného výnosu podkladového aktiva je jeho metoda výpočtu, která je založena na analýze krátkého historického horizontu 5 let. Jestliže si klient sjedná produkt na konci ekonomického růstu, tak očekávaný výnos podkladového aktiva bude pravděpodobně nadhodnocený. Jakmile dojde k poklesu výkonosti ekonomiky a skutečný výnos bude mnohem menší nebo dokonce záporný, klient se může cítit poškozen, protože nedosahuje výnosu, který mu byl „*přislíben*". Metoda výpočtu očekávaného výnosu podkladového aktiva dává větší smysl při investování do krátkodobějších investičních produktů. Pokud budou pojistné produkty nadále zahrnuty do legislativy PRIIP's, tak by bylo vhodné revidovat výpočet očekávaného výnosu podkladového aktiva.

Změna výplatního schématu pojistného plnění

V praxi se lze často setkat s produkty, které mají různé schéma pojistného plnění tak, aby se maximálně přizpůsobily požadavkům klientů. Pojistné plnění se může v průběhu času lineárně navyšovat nebo případně snižovat. Jestliže je produkt použit na krytí hypotečního úvěru, tak pojistná částka může klesat úměrně dluhu. Pro další analýzu bylo vybráno pět výplatních schémat: konstantní výplata (SA), konstantní výplata plus hodnota fondu (SA + FV), rostoucí, klesající a anuitní (hypoteční) schéma pojistné částky.

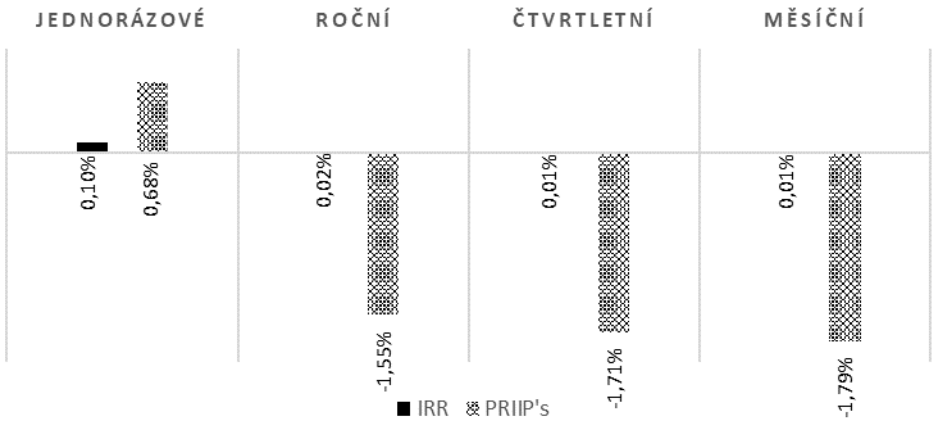


Obrázek 3: Vliv odlišného výplatního schématu na ukazatel PRIIP's Zdroj: Vlastní zpracování

Výsledky ukazují, že ukazatel PRIIP's je velmi citlivý na typ výplatního schématu. Ukazatel na základě PRIIP's opět upřednostňuje produkt, u kterého dochází k větší akumulaci hodnoty fondu. Z obrázku 3 vyplývá, že nejvýnosnější je produkt s klesající pojistnou částkou. Tento produkt je ale velmi specifický. Mladý klient, který má nízkou pravděpodobnost úmrtí, by v případě úmrtí dostal plnou výši pojistného plnění. Čím bude klient starší, tím méně peněz by v případě úmrtí dostal. Postupem času se z tohoto produktu stává spíše produkt spořící než pojistný. Tento produkt nebude dávat pro běžného klienta, který hledá spíše pojistnou ochranu než zhodnocování prostředků, příliš velký smysl. Jestliže klienti nebudou dobře rozumět principu pojištění nebo samotnému produktu, tak by na základě PRIIP's ukazatele mohli volit produkty, které se zdají být ekonomicky výnosné, ale z hlediska poskytování pojistné ochrany nesmyslné.

Změna frekvence placení pojistného

Jedním ze základních parametrů pojistné smlouvy je frekvence placení pojistného. Následující analýza bude provedena na čtyřech častých frekvencích, a to jednorázová, roční, čtvrtletní a měsíční frekvence placení pojistného. Obrázek 4 zobrazuje porovnání ukazatele PRIIP's pro stejný produkt lišící se ve frekvenci placení pojistného. Z tohoto výstupu je patrné, že jednorázový typ frekvence placení výrazně vystupuje z řady zbylých ukazatelů. Tento rozdíl je opět způsoben akumulací a zhodnocením pojistného do fondu. V případě jednorázového vkladu pojistného při sjednání smlouvy se všechny jednotky pojistného zhodnocují po celou dobu projekce. Naopak v případě měsíčního pojistného se každé další zaplacené pojistné zhodnocuje pouze po zbylou dobu do konce smlouvy. Tedy, první zaplacené pojistné se zhodnocuje po celou dobu smlouvy a poslední zaplacené pojistné se zhodnotí už jen jeden měsíc. Jak se ukazuje, tak ukazatel PRIIP's je citlivý i na frekvenci pojistného.



Obrázek 4: Vliv různé frekvence pojistného na ukazatel PRIIP's Zdroj: Vlastní zpracování

Závěr

Cílem nařízení PRIIP's je zlepšit transparentnost a vzájemnou srovnatelnost investičních produktů pro běžného klienta pomocí jednotného a strukturovaného dokumentu KID. Mezi investiční produkty, na které toto nařízení vztahuje, patří i investiční životní pojištění. Životní pojištění je všeobecně velmi komplikovaný produkt, jehož výnosnost nelze popsat jednoduchými finančními metodami používaných například na investiční fondy. Typickou charakteristikou pojistných produktů je široký výběr hlavní pojistného krytí s možností připojistit se na další rizika. Tyto skutečnosti komplikují tvorbu univerzálního ukazatele výnosnosti pojistných produktů. Běžný klient se v pojistných produktech příliš neorientuje a často může být do koupě nevhodného produktu dotlačen na základě neúplných informací. Nový ukazatel na základě PRIIP's legislativy se jeví jako dobrý krok ke zvýšení transparentnosti a důvěryhodnosti investičních produktů. V případě pojistných produktů je stále nedokonalý a určitá novelizace tohoto nařízení by byla vítána. Zejména proto, že všechny pojišťovny na Evropském trhu jsou povinny KID svých produktů připravit a publikovat.

Nedokonalost PRIIP's ukazatele pojistných produktů spočívá zejména ve velké citlivosti na změnu parametrů pojistného produktu, jako je investiční strategie, velikost nebo typ pojistného plnění. Legislativa PRIIP's je vytvořena pouze na porovnání produktů investičního životního pojištění. To znamená, že KID nebudou poskytovány ke všem životním produktům, ale pouze k těm s investiční složkou. Další problém vzniká při intepretaci výsledných výnosů jednotlivých produktů. Do výnosu pojistných produktů se na rozdíl od čistě investičních produktů promítne i rizikové pojistné. Z tohoto důvodu může být prezentovaný výnos velmi malý nebo záporný. Tato skutečnost nutně neznamená, že produkt není výnosný, protože jeho hlavním účelem není vydělávat peníze, ale krýt riziko. Výše zmíněné důvody mohou vést k zavádějícím výsledkům a pokud nebude klientům poskytnuta pří výběru vhodného pojistného produktu odborná asistence, můžou opět kupovat nevhodné produkty.

Literatura

Česká asociace pojišťoven. 2014. *Standardizovaný ukazatel nákladovosti*. Česká asociace pojišťoven. Dostupné z: http://www.cap.cz/odborna-verejnost/samoregulacni-standardy-cap/standardizovany-ukazatel-nakladovosti.

Dlouhá, Petra. 2014. *Kolik stojí investiční životní pojištění*. Ukazatele nákladovosti. peníze.cz. Dostupné z: http://www.penize.cz/investicni-zivotni-pojisteni/287667-kolik-stoji-investicni-zivotni-pojisteni-ukazatele-nakladovosti.

Eur-lex.europa.eu., 2014. *EUR-Lex - 32014R1286 - EN*. Dostupné z: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014R1286.

Fojtík, Jan. 2016. *Metody hodnocení produktů životního pojištění*. Vysoká škola ekonomická, Praha.

Produktové listy.cz., 2012. *Produktové listy*. Dostupné z: http://www.produktovelisty.cz/pojisteni-osob/clanky.html.

TANK., 2011. *Ukazatel TANK. atank.cz*. Tým TANK, Dostupné z: http://atank.cz/ukazatel-TANK-nakladovost-IZP/.

JEL Classification: C38, C63, G22.

Summary

Analysis of evaluation methods for life insurance product from client's perspective

The retail investors are nowadays offered mostly packaged retail and insurance-based investment products. Due to the complexity of these products it can be difficult to understand the risk profile of this products. The proper methods to evaluate and compare these product is a good way to improve transparency and protect client. The new regulation PRIIP's offers methods to evaluate investments product with unified template KID. This paper analyses the usability of PRIIP's as an evaluation method of life insurance products. The PRIIP's evaluation method is compared with author's paper dealing with the topic of life insurance evaluation. The analysis provided in this paper consists of testing the PRIIP's sensitivity on the change of basic insurance parameters such as sum assured, pay-off scheme, interest rate or premium frequency.

Key words: Life insurance, Investments for retail investor, Evaluation methods, PRIIP's

Frequency Modelling within Claim Reserving Using Micro Data

Michal Gerthofer

michal.gerthofer@vse.cz

Doktorand oboru statistika

Školitel: doc. Ing. Iva Pecáková, CSc., (iva.pecakova@vse.cz)

Abstract: Estimation of claims reserves, which should be held by the insurer so as to be able to meet expected future claims arising from policies currently in force and policies written in the past, presents an important task for insurance companies to predict their liabilities. A common approach to the reserving problem is based on annual aggregated models. This article proposes a frequency part of the micro model, which combines maximum likelihood estimator of incurred but not reported (IBNR) number of claims with estimator of distribution for time from reporting to closing of the claim. Since all the available information from the data is utilized by this method, more accurate behavior of the data can be captured. A real data example together with a diagnostic analysis is provided to demonstrate an objective illustration of the potential benefits of the presented approach.

Key words: claims reserving, non-life insurance, claim frequency, Cox survival regression, survival analysis, micro modelling.

Introduction

Claims reserving is a classical problem in non-life insurance, sometimes also called general insurance (in UK) or property and casual insurance (in USA). A non-life insurance policy is a contract between the insurer and the insured. The insurer receives a deterministic amount of money, known as premium, from the insured in order to obtain a financial coverage against well-specified randomly occurred events. If such an event (claim, loss) happens, the insurer is obliged to pay in respect of the claim a claim amount, also known as loss amount.

Claims reserving now means, that the insurance company puts sufficient provisions from the premium payments aside, so that it is able to settle all the claims that are caused by these insurance contracts. The main issue is how to determine or estimate these claims reserves, which should be held by the insurer. A number of various methods based on aggregated data has been invented in this field. Since insurance companies store all the information about the claim settlement, methods based on micro data gained increasing attention.

Nevertheless, claims reserving is a worldwide phenomenon that is of high interest among people from the insurance business. Over the recent years, there has been increasing interest in the utilization of insurance data at a more granular level, and to model claims using stochastic processes. The Solvency II directive brings the stochastic claims reserving problem not only inside the insurance companies (and this is done by law in many countries). The claim by claim stochastic reserving becomes an issue for the auditors as well as for the supervisors. Such an important economic problem is more than relevant not only for the academic society, but also for many practitioners.

Current situation

The first reserving method the so-called Mack chain ladder was proposed by Mack (1993). Later, as already mentioned, a number of various stochastic methods also based on aggregated data were developed, see England and Verrall (2002), Wüthrich and Merz (2008) or Hudecová and Pešta (2013) for an overview.

To make use of all the information about the claim settlement stored by insurance companies, methods based on micro data gained more attention and several works dedicated to this topic appeared. Micro models are usually split into severity and frequency part. Antonio and Beirlant (2007) suggest the generalized linear mixed models (GLMM) for severity modelling. Further, Harej (2017) proposed

recent work dedicated to severity modeling using neural network. However, this model was based exclusively on simulated data. On the other hand, regarding the frequency modelling, Avanzi (2016) introduced model based on a Cox process. Despite being progressive, all of these works focus solely on frequency or severity part of the micro model instead of total prediction of reserve.

Structure of the work

As mentioned before frequency model is split into two parts. The prediction of incurred but not reported (IBNR) number of claims, which is described in Section 1. Section 2 is further dedicated to the estimation of distribution of closing time, i.e., time from reporting to closing of the claim.

Database

Dataset comes from the Czech Insurers´ Bureau, an organization of insurance companies that is authorized to provide liability insurance for damage caused by vehicles (car insurance). The dataset contains 17658 records about all claims that occurred between years 2005 and 2016. This data only deals with material claims. Dataset provides granular information about individual claim such as accident and reporting date as well as information about all the reserve and payment transactions.

The fact that the focus is on material claims, which do not have a long reporting delay and data contains 12 years of history, see incremental run off triangle of number of claims in Table 1, allows us to assume that all claims with accident year between 2005 and 2010 have already been reported. In other words, we do not expect any claim to be reported after 2016. Using this assumption, data highlighted by grey can be split into calibration (brighter grey) and validation part (darker grey), i.e., claims which have reporting date before 1.1. 2011 and another where reporting date is 1.1. 2011 or later, respectively. Thus, comparison to reality can be performed in order to assess precision of the prediction.

Table 1: Run-off triangle for incremental number of IBNR claim amounts N_{it}.

Accident year i	Development year t											
	1	2	3	4	5	6	7	8	9	10	11	12
2005	2612	617	24	12	0	0	0	0	0	0	0	0
2006	2405	622	30	8	1	0	0	0	0	0	0	
2007	2416	587	41	5	6	0	0	0	0	0		
2008	2619	563	42	10	4	1	0	0	0			
2009	2081	429	39	8	3	1	0	0				
2010	1980	415	61	16	0	0	0					
2011	1745	354	45	13	1	0						
2012	1635	322	27	4	1							
2013	1752	284	17	9								
2014	1705	253	13									
2015	1608	219										
2016	1667											

Estimator of IBNR Number of Claims based on Distribution of Reporting Delay

This part describes an innovative claim frequency approach, which combines survival analysis for right truncated data with maximum likelihood estimator of incurred but not reported (IBNR) number of claims. Since all the available information is utilized, exact behavior of the data can be captured by non-parametric methods and consequently the maximum likelihood is applied to obtain final prediction of IBNR number of claims. Diagnostic analysis and model selection are provided to demonstrate an objective illustration of the potential benefits of the presented approach

First of all, initial notation must be introduced. For the sake of the clarity, basic variables are displayed in Figure 1. Time 0 denotes inception time of the model and time τ represents the current time. The first part of the figure is dedicated to the claim where occurrence time of the claim T_i as well as reporting time $T_i + D_i$ are lower than the current time. In other words it means that claim has already been reported. Lower part of the figure displays the situation where the claim occurred in the past, however, it has not yet been reported. Since the reporting time $T_i + D_i$ is later than the current time τ, we deal with truncated data where truncation time is τ. It should be noted that T_i, D_i are observable since $T_i + D_i < \tau$. Therefore, the total number of claims N can be split into number of claims which have been already reported, $N^{(1)}$, and number of claims incurred but not reported, $N^{(2)}$, the so-called IBNR number of claims.

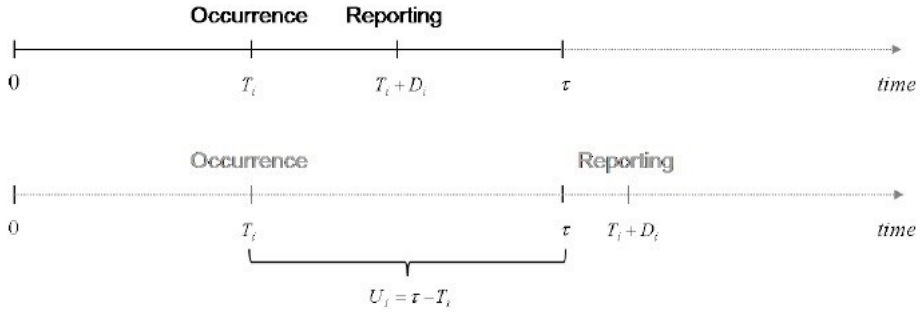


Figure 1: Illustration of the claim settlement.

It is assumed that the number of claims in a particular time $t \leq \tau$ follows a process

$$N(t) \equiv \sum_{i=1}^{N} \chi\{T_i \leq t\}$$

with the variable intensity $\lambda(t)$. We will also consider the partition $\delta_k, k = 1, \ldots, K$ of the time interval $(0, \tau)$. It is assumed that within a short time interval δ_k, the intensity of the claims process is constant, i.e. $\lambda(t) = \lambda_k \text{ for } t \in \delta_k$. Let us denote the number of claims that occurred within the time interval δ_k as

$$N_k \equiv \sum_{i=1}^{N} \chi\{T_i \in \delta_k\}.$$

It is also assumed that each of these claims within the interval δ_k has a certain (identical) probability to be reported until the current time τ. Let us denote this probability as p_k. Further, the number of

claims that occurred within δ_k and were reported until τ is denoted as $N_k^{(1)}$ and has the binomial distribution

$$N_k^{(1)} \sim Bi(N_k, p_k),$$

with a probability density function

$$P\left(N_k^{(1)} = x\right) = \binom{N_k}{x} p_k^x (1-p_k)^{N_k - x}.$$

Firstly, let us assume that that p_k is known. At the time τ, the number of reported claims $N_k^{(1)}$ is observable. The unknown parameter is the total number of claims N_k occurred within δ_k. This is a rather unusual setup. In common binomial distribution tasks, the sample size is usually known and the success probability has to be estimated. Nevertheless, an estimator of N_k is derived in the following text. In order to use notation common for the binomial distribution we simplify the notation for a given k, as follows

$$n \equiv N_k, p \equiv p_k, q \equiv (1-p_k).$$

and we can get the usual probability density function

$$P\left(N_k^{(1)} = x\right) = \binom{n}{x} p^x q^{n-x}.$$

Maximum Likelihood Estimator

Now, we proceed to the maximum likelihood estimation of the parameter n based on the knowledge of x and p_k. We start with a description of the estimator in case of an integer . Thus, we want to maximize the likelihood function

$$l(n) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{(n-x)!\,x!} p^x (1-p)^{n-x},$$

where n is an integer, $n \geq x$, and $p \in [0, 1]$. Since we are interested in the integer value of n, argument of the maximum for given likelihood function can be found easily.

The probability p_k is usually unknown in practical tasks. Therefore, an estimate of p_k has to be plugged in. Some possibilities of estimating p_k based on the reporting delay data are investigated in following section.

Estimator of Distribution of the Reporting Delay

This section is dedicated to the distribution of the reporting delay D_i, which is defined as the time from the occurrence of the claim until its reporting time as displayed in Figure 1. Estimation of distribution requires special attention and techniques due to the truncated data that we deal with. To remind the notation, time 0 represents the inception time of the model and time τ denotes the current time. As previously pointed out, the times D_i and T_i are observed only if the i-th claim occurred and has been reported

$$T_i + D_i \leq \tau,$$

which means D_i is observable only if

$$D_i \leq U_i,$$

where $U_i \equiv \tau - T_i$ is called the truncation time. This phenomenon is known as the right truncated data.

Nonparametric Estimator

Nonparametric estimators for the truncated or censored data are usually called the Kaplan Meier (K-M) estimators or the product-limit estimates. Details can be found in Kaplan (1958) or in the monograph Lawless (2003).

Let us denote the distribution function of the reporting delay as $F_D(\cdot)$ and the observations $(u_{i'}, d_{i'}), i' = 1, \dots, n'$, where n' is the number of observations. Then, $d_{(1)} < d_{(2)} < \cdots < d_{(J)}$ are ordered distinct values of $d_1, d_2, \dots, d_{n'}$. Observations may be recorded in discrete time intervals e.g. in days. Therefore, it is possible that some observations will have identical value and $J \leq n'$. A practical limitation of the Kaplan Maier method for truncated data is that it provides an estimation only for the conditional distribution

$$F_D^*(d) = \frac{F_D(d)}{F_D(d_{(J)})},$$

where $d \leq d_{(J)}$ .We denote

$$r_{(j)} \equiv \sum_{i'=1}^{n'} \chi(d_{i'} = d_{(j)}),$$

$$H_{(j)} \equiv \sum_{i'=1}^{n'} \chi(d_{i'} \leq d_{(j)} \leq u_{i'}),$$

where $j = 1, \dots, J$ and $\chi(Z)$ is the indicator of the event Z. Thus, the conditional distribution function $F_D^*(\cdot)$ can be estimated as

$$\hat{F}_D^*(d) = \prod_{d_{(j)} > d} \left(1 - \frac{r_{(j)}}{H_{(j)}}\right).$$

Practical Application of the Models

Proposed framework is applied to a very unique, real dataset which contains sufficient history of incurred claims, i.e., there exist accident years where no further claims are expected to be reported. This allows us to split data on calibration and validation part, so we can compare reality with predictions.

First of all, nonparametric Kaplan Maier estimation of the reporting delay distribution are performed. Consequently, these estimates are used for prediction of IBNR number of claims using moment and maximum likelihood methods. The results are then compared to reality as well as to the Mack chain ladder estimates which are widely used in the insurance industry.

Estimation of Distribution of Reporting Delay Based on Real Data

As mentioned before, distribution of reporting delay must be estimated before the prediction of the IBNR number of claims is computed. Since calibration data is right truncated, modification of the classical Kaplan Maier stated in the previous chapter is used. Figure 2 displays the difference between estimation of distribution using Kaplan Maier assuming truncated data and classical Kaplan Maier. We

can see that the difference between distributions is almost negligible but an impact on the total estimator of the IBNR number of claims is significant as it is shown later. Therefore, the choice of a proper estimation of reporting delay distribution is very important.

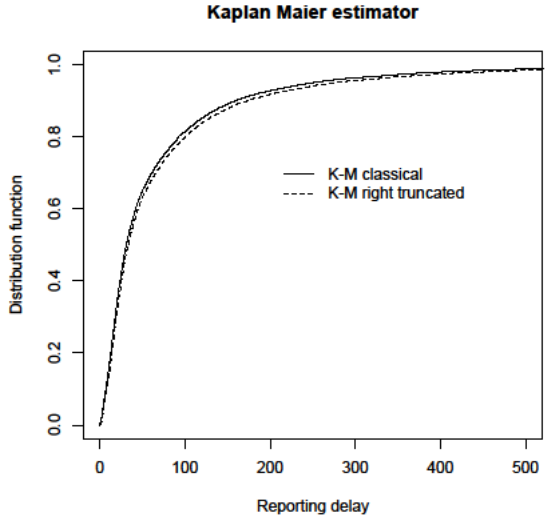


Figure 2: Estimation of reporting delay distribution using Kaplan Maier.

Predictions of IBNR Number of Claims versus Reality

In this part, maximum likelihood method is combined with aforementioned estimations of reporting delay distribution, namely classical Kaplan Maier and modification of Kaplan Maier for truncated data. In addition Mack chain ladder predictions are performed based on aggregate annual as well as weekly data.

As it was pointed out, estimated distributions of reporting delay are based on daily data. However, number of claims already reported, $N_k^{(1)}$, entering maximum likelihood estimator are weekly counts $N_{week_k}^{(1)}$. This partial aggregation is performed due to the instability of the estimator in the last two observed accident days. These two days contain sparse number of observations, $N_{acc.day_{last-1}}^{(1)}$, $N_{acc.day_{last}}^{(1)}$, and even slight change causes significant impact on the final prediction. Various smoothing methods were used in order to handle this issue. However, all these approaches depend heavily on subjective setting of the smoothing parameters. Using weekly aggregation predictions became stable and reasonable. It should be noted that all these claims are assumed to be reported in the middle of each week.

Prediction of the IBNR number of claims of all proposed models together with real IBNR number of claims are displayed in Table 2. Mack chain ladder predictions based on aggregated annual as well as weakly data were computed. It should be emphasized that if the assumptions of Mack chain ladder hold, the estimate is unbiased. However, according to the residual diagnostic in Figure 3 (especially the two plots in the middle), assumptions of the model do not clearly hold, which can be caused by an insufficient number of observation.

Table 2: Real and predicted IBNR number of claims.

Real	Mack	Predictions	MLE	Predictions
564	Annual	538	K-M classic	447
	Weekly	561	K-M truncated	542

Nevertheless, prediction of the annual Mack chain ladder is very close to the real IBNR number of claims even though verification of assumptions was not clear. Moreover, the weekly Mack chain ladder prediction is even more precise.

Maximum likelihood predictions built on modification of Kaplan Maier provide very reasonable estimates, even better than annual Chain ladder. On the other hand maximum likehood estimator based on classical Kaplan Maier provide the worst prediction. This is manly caused by the impact of truncated data.

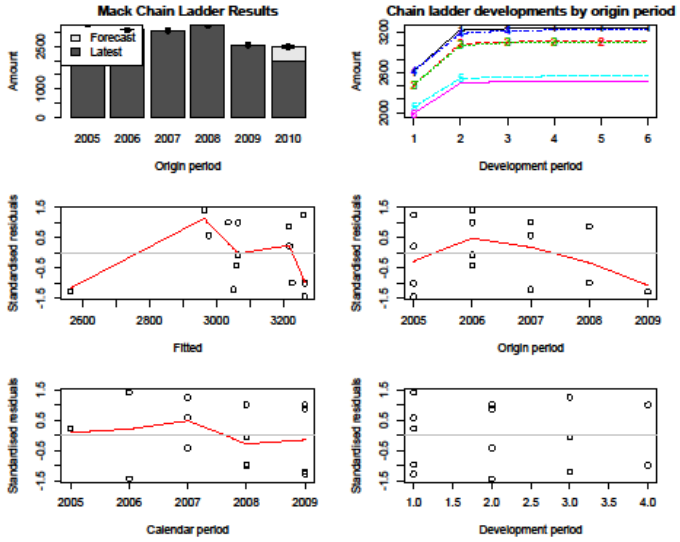


Figure 3: Residual diagnostics for annual Mack chain ladder.

In order to assess accuracy of the predictions more closely, predictions by accident years are computed in Table 3. Estimate for classical Kaplan Maier is excluded due to inappropriate method for given data. Weekly Mack chain ladder prediction slightly overestimated the last accident year and underestimated the rest of accident years. The same holds also for annual Mack chain ladder but with lower gap. Maximum likelihood predictions also highly underestimate the IBNR number of claims in the first five accident years. Overall weekly Mack chain ladder predicted the data the best.

Table 3: Predicted IBNR number of claims by accident years (numbers are rounded to the integer).

Accident year	Real	Mack annual	Mack weekly	MLE K-M
2005	**0**	0	0	0
2006	**0**	0	0	0
2007	**6**	0	1	0
2008	**15**	9	8	0
2009	**51**	35	36	12
2010	**492**	494	516	530
Total	**564**	538	561	447
Error		-26	-3	-17

Nevertheless, this analysis indicates that proposed micro model provide reasonable and comparable predictions with the well-known Mack chain ladder method. Moreover, the advantage of these model is ability to predict IBNR number of claims from given accident week and particular development day or even hour. Additionally, we have to be aware of the specific assumptions that vary significantly model by model and their violation might cause misleading results. In order to assess the robustness, and precision of the estimators in general simulation analysis must be performed. This will be the goal of the following research.

Estimator of Distribution of Closing Time

In this section our attention is dedicated to the distribution of closing time. Let us denote variable C_i as time from reporting to closing of the claim and define $V_i \equiv \tau - (T_i + D_i)$. It is important that reported claims can be divided into the claims which have been already closed and claims which are still opened. Lower part of Figure 4 illustrates situation where closing time is lower than current time. This phenomenon is known as censored data where V_i is censoring time for investigated variable C_i. In this case we do not know the exact value of the variable C_i but we know that it is higher than V_i.

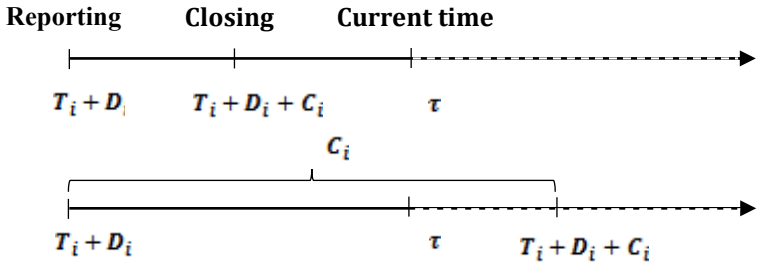


Figure 4: Illustration of the claim settlement.

In previous section modification of Kaplan Maier for truncated data was used. However, here modification of Kaplan Maier for censored data must be used.

Nonparametric Estimator for Censored Data

Let us denote the distribution function of the reporting delay as $F_C(\cdot)$ and the observations $(k_i, \delta_i), i = 1, \dots, n$, where k_i is either time from reporting to closing or in case claim hasn't closed yet, censoring time. Further, δ_i is indicator of closed claim and n is the number of observation, in our case number of reported claims. Suppose that there are K $(K \leq n)$ distinct ordered observations $c_1 < c_2 < \cdots < c_K$ of variable C_i. Define also $d_j \equiv \sum_{i=1}^{n} \chi(k_i = c_j, \delta_i = 1)$ representing number of closed claims at c_j. The product-limit estimate of survival function $S(t) \equiv P(C_i \geq t)$ is defined as

$$\hat{S}(t) = \sum_{j:c_j<t} \frac{n_j - d_j}{n_j},$$

where $n_j \equiv \sum_{i=1}^{n} \chi(k_i \geq c_j)$ is number of claims at risk at c_j, that is, the number of claims not closed and uncensored just prior to c_j. If a censoring time and n lifetime are recorded as equal, we adopt the convention that the censoring time is infinitesimally larger. Thus, any individuals with censoring times recorded as equal to c_j are included in the set of n_j claims at risk, as are claims which closed at c_j.

Semi-parametric Estimator for Censored Data

Next step from the aforementioned non-parametric Kaplan Maier estimator is semi-parametric Cox proportional hazard regression. This approach proposed in paper Regression Models and Life Tables by Cox (1972) is one of the most frequently cited journal articles in statistics and medicine. Cox regression belongs to the semi-parametric models in contrary to non-parametric Kaplan Meier (product-limit) or parametric models.

First of all, the hazard function $\lambda(t)$ is defined as the instantaneous rate of closure of the claim at time t, given that a claim has opened until at least time t

$$\lambda(t) = \lim_{h \to 0+} \frac{P(t \leq C_i < t + h | C_i \geq t)}{h} = \frac{f(t)}{S(t)},$$

where $f(t)$ is probability density function of variable C_i for all i. Consequently, regression model has the following form

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)\exp\{X_{i,1}\beta_1 + \cdots + X_{i,p-1}\beta_{p-1} + X_{i,p}\beta_p\}$$

where $\lambda_0(t)$ is the baseline hazard function, describing how the risk of an event per time unit changes over time at baseline levels of covariates. The effect of parameters, $\exp\{\boldsymbol{X}_i\boldsymbol{\beta}\}$, describes how the hazard varies in response to explanatory covariates.

Let us take the notation defined in the previous part. Ignoring ties for the moment, conditioned upon the existence of a unique closure at some particular time c_j the probability that the closure occurs in the claim with time c_j is

$$L_j(\boldsymbol{\beta}) = \frac{\theta_j}{\sum_{i:k_i \geq c_j} \theta_i},$$

where $\theta_j = \exp\{\boldsymbol{X}_i\boldsymbol{\beta}\}$. Observe that the factor $\lambda_0(\cdot)$ that would be present in both the numerator and denominator have canceled out. Treating the events as if they were statistically independent, the joint probability of all realized events conditioned upon the existence of events at those times is the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{K} \frac{\theta_j}{\sum_{i:k_i \geq c_j} \theta_i},$$

The corresponding log partial likelihood can be maximized over $\boldsymbol{\beta}$ to produce partial likelihood estimates of the model parameters. According to the score function and Hessian matrix, the partial likelihood can be maximized using the Newton-Raphson algorithm. The inverse of the Hessian matrix, evaluated at the estimate of $\boldsymbol{\beta}$ can be used as an approximate variance-covariance matrix for the estimate, and used to produce approximate standard errors for the regression coefficients.

Crucial assumption of the model is proportional hazards which means that the hazard for any individual is a fixed proportion of the hazard for any other individual. Further assumption we have to be aware of is a multiplicative risk structure. For further details about properties of the estimator see Cox (1972), Lawless (2003) or Klein (2003).

Practical Application of the Models

In this part, we use all the available data (29332), i.e., data corresponding to accident year 2005 to 2016 and no split to validation and calibration part. In order to capture possible dependency between variable C_i and D_i in the model, reporting delay is assumed to be the only covariate. However, to capture nonlinearity, we split reporting delay into several factors, $D_{i,0-1year}$, $D_{i,1-2years}$, $D_{i,2-3years}$ and $D_{i,3-more\,years}$, where , $D_{i,0-1year} = 1$ if , $0 \leq D_i < 366$ otherwise zero and so on.

The model with assuming reporting delay as continuous covariate has the following form

$$\lambda(t|D_i) = \lambda_0(t)\exp\{D_i\alpha\}.$$

In Figure 5 we can see estimates of survival function $S(t)$ using information about the censoring (black line) and estimate assuming censoring time as closing one (red line). During model selection, one has to be aware of the fact that using inappropriate models can cause significant impact on the final prediction.

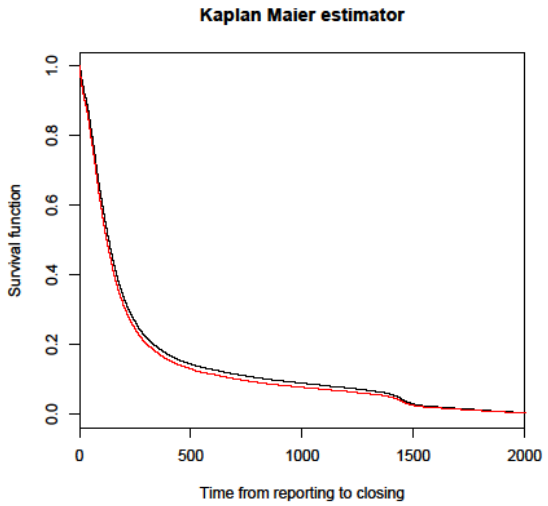


Figure 5: Kaplan Maier estimator.

Consequently, the model with factor covariate has the following form

$$\lambda(t|\boldsymbol{D}_{i,factors}) = \lambda_0(t)\exp\{D_{i,0-1year}\beta_1 + D_{i,1-2years}\beta_2 + D_{i,2-3years}\beta_3 + D_{i,3-more\ years}\beta_4\}.$$

However, results from the Kaplan-Maier for censored data in Figure 6 indicate that the assumption of proportionality does not hold due to the crossing survival functions.

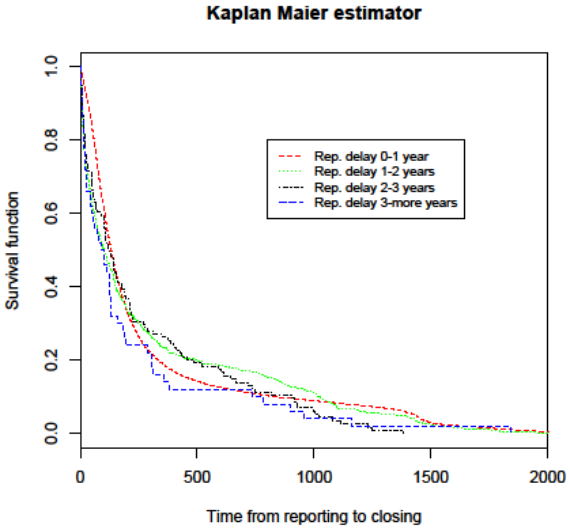


Figure 6: Kaplan Maier estimator for different factors.

Despite the violation of the assumption, we applied Cox proportional regression but as we can see in Figure 7 semi-parametric Cox regression dos not capture data well. Therefore, parametric estimator in Figure 5 or 6 could be the right choice. Further research can be focused on models which can cope with the violation of proportionality assumption.

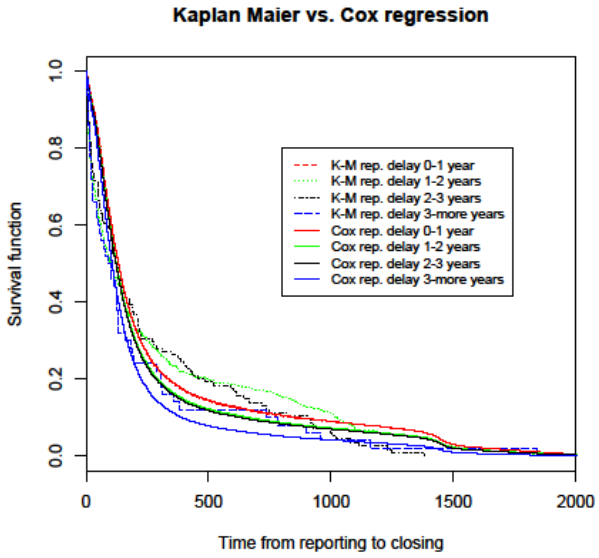


Figure 7: Kaplan Maier versus Cox regression estimator for different factors.

Conclusions

In this work, we proposed an innovative approach for prediction of the IBNR number of claims combining various approaches for estimation of reporting delay distribution based on right truncated data using maximum likelihood. Classical methods such as Mack chain ladder provide predictions with powerful properties. However, they are based on aggregated data, which use only a piece of the available information. On the other hand, proposed models utilize all the available information from the data and allow us to estimate not only number of claims in a given development year but also in a given development day or even an hour. This property might be further useful for the severity part of a micro model framework.

The first contribution of this work is implied by the data used, which allow us to apply proposed methods and compare them to the reality. Most of the published papers on this topic are based on simulated data which usually do not reflect the reality. This makes our conclusions unbiased and objective.

The goal of this analysis was also to develop a frequency micro model, which would provide comparably precise prediction as classical Mack chain ladder. Asset of the last part was estimation of distribution of time from reporting to closing where Cox regression was not suitable choice and Kaplan Maier for censored data was chosen.

References

Antonio, K., Beirlant, J. (2007). "Actuarial statistics with generalized linear mixed models". *Insurance: Mathematics and Economics*, Vol. 40, No. 1, pp. 58–76.

Avanzi, B., Wong, B., Yang, X. (2016). "A micro-level claim count model with overdispersion and reporting delays". *Insurance: Mathematics and Economics*, Vol. 71, pp. 1-14.

England, P. D., Verrall, R. J. (2002). "Stochastic Claims Reserving in General Insurance." *British Actuarial Journal*, Vol. 8, No. 3, pp. 443–518.

Cox, D. R. (1972). "Regression models and life-tables". Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2, pp. 187–220.

Harej, B., Gächter, R., Jamal, S. (2017). "Individual Claim Development with Machine Learning Report 2017". Available at http://www.actuaries.org/ASTIN/Documents/ASTIN_ICDML_WP_Report_final.pdf (2.1.2018).

Hudecová, Š., Pešta, M. (2013). "Modeling Dependencies in Claims Reserving with GEE." Insurance: Mathematics and Economics, Vol. 53, No. 3, pp. 786–794.

Klein, J. P. (2003). *SURVIVAL ANALYSIS Techniques for Censored and Truncated Data.* 2nd ed. Springer, New York.

Kaplan, E.L. and Meier, P. (1958). "Nonparametric estimation from incomplete observations". J. Amer. Stat. Assoc. Vol. 53, pp. 457-481.

Lawless, J.F. (2003). Statistical Models and Methods for Lifetime Data, 2nd ed. Wiley, USA, New Jersey.

Mack, T. (1993). "Distribution-free Calculation of the Standard Error of Chain Ladder Reserve Estimates". *ASTIN Bulletin,* Vol. 23, pp. 213–225.

Wüthrich, M. V, Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Hoboken: Wiley Finance.

JEL Classification: C130, C180, C230, C330, C460, C510, G220.

Ackowledgement: The research was supported by the project IGA No. 70/2016.

Shrnutí

Model frekvence počtu škod v rezervování neživotního pojištění s využitím mikro dat

Odhad výše rezerv, kterou by mněla pojišťovna držet s cílem dospět všem svým budoucím závazkům vyplývajícím z aktuálně platných pojistek a pojistek upsaných v minulosti, hraje významnou úlohu pro pojišťovny ve smyslu predikce jejich budoucích závazků. Běžné přístupy k tomuto problému jsou založeny na základě ročních agregovaných modelů. Tato práce navrhuje důležitou část mikro modelu a to modelování frekvence škod, která kombinuje odhad počtu nastalých ale zatím nenahlášených škod (IBNR) s odhadem pravděpodobnostní distribuce času od nahlášení po uzavření škody. Výsledky mohou být dále použity jako vstup modelu výše škod. Protože tyto metody využívají veškerou informaci dostupnou z dat, modely by měly lépe zachytit skutečnou povahu dat. S cílem objektivní ukázky možných výhod navržených modelů jsou postupy demonstrovány na reálných datech spolu s diagnostickou analýzou.

Klíčová slova: rezervování v neživotním pojištění, model frekvence počtu škod, model přežívání, mikro modelování.

Analýza dopadů národních ekvivalenčních škál založených na různých přístupech na české ukazatele

Michaela Jirková

Michaela.jirkova@vse.cz

Doktorand oboru statistika

Školitel: doc. Ing. Jakub Fischer, Ph.D., (jakub.fischer@vse.cz)

Abstrakt:

Ekvivalenční škála spotřebních jednotek zohledňuje úspory z rozsahu různých typů domácností podle jejich velikosti a struktury. Cílem tohoto příspěvku je porovnat úspory z rozsahu uvažované v mezinárodních škálách s odhadem úspor z rozsahu v jednotlivých zemích na základě jejich specifických podmínek a reálné ekonomické situace jejich domácností. Tento příspěvek je zaměřen především na situaci v České republice, ale je zde uvedeno i srovnání s jinými evropskými zeměmi. Odhad úspor z rozsahu může být založen na různých přístupech. Jeden z alternativních přístupů ke škálám spotřebních jednotek v některých zemích je založen na výdajích domácností. Tato metoda by mohla využívat údaje ze Statistiky rodinných účtů (HBS). Další metoda využívá užitkovou funkci domácností, která je nepřímo měřena použitím subjektivních údajů o finanční spokojenosti. Analýza vychází z údajů z evropského Šetření o příjmech a životních podmínkách (EU-SILC). Spotřební jednotky připravené podle obou přístupů jsou odhadovány pomocí regrese. Dopad využití národní ekvivalenční škály na některé české ukazatele je významný, mezi indikátory příjmů obzvlášť u míry ohrožení příjmovou chudobou. Změna struktury lidí pod hranicí příjmové chudoby je pozorovatelná, zejména podle jejich společenského postavení. Analýza dopadů v České republice i ve srovnání s evropskými zeměmi je uvedena v tomto příspěvku.

Klíčová slova: úspory z rozsahu, spotřební jednotky, ekvivalenční škála, výdaje domácností, užitková funkce, míra ohrožení příjmovou chudobou

Úvod

Hodnocení příjmové situace mezi domácnostmi je ovlivněno stanovením úspor z rozsahu realizovaných českými domácnostmi. Ukazatele příjmů jsou ovlivněny použitou stupnicí spotřebních jednotek (SJ). Některé stupnice byly harmonizovány pro mezinárodní srovnání. Pro hodnocení v evropských zemích je Eurostatem běžně používána modifikovaná škála OECD. Ta udává stejné úspory z rozsahu pro všechny evropské země na úrovni 0,5 pro dalšího dospělého v domácnosti a 0,3 pro dítě. Takový systém vah se nazývá ekvivalenční škála a pro Eurostat tedy představuje (1; 0,5; 0,3). Navzdory mnoha výhodám této metodologicky harmonizované koncepce je doporučován odhad národní stupnice spotřebních jednotek. Národní ekvivalenční škála by mohla přesněji odrážet národní ekonomické a sociální podmínky podle reálné velikosti úspor z rozsahu realizované českými domácnostmi. V literatuře je uvedena řada obecných i alternativních přístupů k odhadu národních úspor z rozsahu. Cílem této práce je odhadnout velikost národních úspor z rozsahu podle různých přístupů a porovnat mezi sebou odhadované ekvivalenční škály odpovídající podmínkám českých domácností. Tato analýza se primárně zaměřuje na výdajový a užitkový přístup. Oba odhady ekvivalenční škály jsou také posuzovány ve vztahu k běžně používané odborné stupnici. Vliv úspor z rozsahu je patrný především srovnáním životních podmínek různých typů domácností. Všechny ukazatele příjmů, včetně míry ohrožení příjmovou chudobou, jsou ovlivněny mírou úspor z rozsahu. Zvláště struktura lidí pod hranicí chudoby je velmi závislá na ekvivalenční škále. Proto je zde uveden také dopad národních stupnic na míru ohrožení příjmovou chudobou u různých typů domácností a na závěr jsou uvedeny některé poznatky a doporučení.

Literatura poskytuje řadu přístupů k měření úspor z rozsahu. Přehled důležitých metod zavádí Buhmann a kol. (1988). V tomto článku jsou prezentovány dva směry odhadu. První z nich je odborný přístup k vytváření odborných stupnic a druhý obecný přístup je založen na výsledcích šetření, které vedly ke škálám založeným na výběrových šetřeních. Mezi odborné stupnice se počítají zejména váhy pro účely mezinárodního srovnání. Některé z nich publikují Chanfreau a Burchardt (2008). V této

skupině je zahrnuta varianta druhé odmocniny navržená Lucemburským příjmovým studiem (LIS) a dále známá ekvivalenční škála od OECD včetně její modifikované verze. První z nich je Oxfordova stupnice původně doporučená OECD a druhá z nich je odvozena. Tato stupnice byla zpracována autory Hagenaars, De Vos a Zaid (1994) a v této době ji běžně používá Eurostat. Výhodou této škály je zohlednění rozdílných potřeb členů domácnosti ve vztahu k demografickým charakteristikám lidí. Tyto odborné stupnice byly navrženy odborníky evropských nebo jiných mezinárodních institucí s cílem uplatňovat společný přístup ve všech zemích a zajišťuje určitou srovnatelnost údajů o životní úrovni mezi zeměmi.

Vedle tohoto přístupu by mohly být na vnitrostátní úrovni použity i jiné metody zohledňující potřeby specifické pro jednotlivé země. Buhmann a kol. (1988) uvádí obecný přístup založený na údajích ze šetření u domácností o výdajích na spotřebu. Zde doporučuje pro přípravu stupnic regresní analýzou údajů ze šetření specifikovat vztah mezi velikostí domácnosti a celkovými výdaji. Čím větší je ekvivalenční elasticita *e*, která se pohybuje mezi 0 a 1, tím menší jsou úspory z rozsahu předpokládané ekvivalenční škálou. Vztah mezi výdajovými potřebami a velikostí domácnosti by měl být vyjádřen nějakou rovnicí. Jiní autoři doporučují v rámci rovnice výdajů domácností zohlednit i jiné proměnné. Podle autorů Van der Gaag a Smolenský (1982) je třeba rozlišovat mezi domácností s dětmi i bez nich. Dopady na rodiny s dětmi by měly být vyšší než u domácnosti dospělých. Rozsah ekvivalence by měl odrážet jak úspory z rozsahu, tak rozdíly v charakteristikách domácností. Vzhledem k velikosti domácnosti se elasticita sníží s počtem dětí (Schwarze, 2003). Podle (Dudel, 2015) jsou odhady neparametrických mezí na ekvivalenčních škálách pro páry s jedním dítětem a bezdětnými páry jako referenční kategorií mezi (1,16; 1,46), takže váha spotřební jednotky pro dítě by měla být v rozmezí od 0,16 do 0,46. Všechny tyto metody využívající údaje ze šetření mohou být podle Buhmann a kol. (1988) rozděleny do dvou alternativních přístupů. Je možné najít úspory z rozsahu buď z úrovně spotřeby domácností pomocí údajů o jejich výdajích, nebo ze subjektivního vnímání úrovně spotřeby členů domácnosti. Tento první výdajový přístup je založen na metodě lineární regrese pomocí některé z rovnic výdajů. Nejdůležitější rovnice představují autoři Van der Gaag a Smolensky (1982). V této práci je použita lineární regrese včetně významných vstupních proměnných, s nimiž je pracováno například i v analýze Bishop (2015). Regresní koeficienty zde znamenají zvýšení výdajů přidáním dalšího člena domácnosti. V literatuře byla také zavedena řada dalších přístupů založených na druhém přístupu k subjektivnímu vnímání úrovně spotřeby. V článku autorů Bütikofer a Gerfin (2009) je prezentován užitkový přístup. Tato metoda založená na užitkových funkcích využívá subjektivní finanční spokojenost jako prostředek pro nepřímé funkce. Zjednodušeně tato metodika předpokládá, že úspory z rozsahu společného života se uskutečňují, když součet jednotlivých užitků z osobní spotřeby všech členů domácnosti přesahuje celkový užitek domácnosti. Údaje nám umožňují porovnávat užitkové funkce (vyjádřené finanční spokojeností) jednotlivců a osob žijících v páru a následně stanovit úspory z rozsahu.

Míra úspor z rozsahu je důležitým faktorem ovlivňujícím ukazatele porovnávající životní podmínky domácností. Hodnocení jednotek spotřeby ovlivňuje především ukazatele příjmů. Vzhledem k tomu, že jakékoliv spotřební jednotky používané namísto počtu členů domácnosti zvyšují průměrný osobní příjem, příjmy na spotřební jednotku (ekvivalizovaný příjem) budou vyšší než příjem na osobu. Dopad aplikování spotřebních jednotek, které disponují celkovým příjmem domácnosti, na rozdělení příjmů je diskutován Malou (2015). Rozsah ekvivalence mění rozdělení příjmů, a tudíž i nerovnost příjmů a všechny ukazatele závislé na příjmech, zejména hranici chudoby a míru ohrožení příjmovou chudobou (Förster, 1994). Podle autorů De Vos a Zaidi (1997) je hranice chudoby velmi citlivá na ekvivalenční škálu. V důsledku toho je ovlivněna nejen míra, ale také struktura lidí pod hranicí chudoby. Předpokládá se, že především struktura lidí podle jejich věku a ekonomické aktivity by se významně změnila. Největší dopad bude zejména na starší osoby v důchodu kvůli tomu, že se svými příjmy soustřeďují v blízkosti hranice chudoby (Bishop & al., 2017a). Použitím nižších úspor z rozsahu pro dalšího dospělého se míra chudoby starších lidí snižuje, zatímco míra chudoby dětí se zvyšuje, což podle Bishopa a kol. (2017b) vede ke zvýšení rozdílu mezi nimi. V tomto článku jsou také zvažovány různé váhy pro další členy podle jejich pořadí v domácnosti.

Data a metodologie

Výdajová metoda

Podle autorů van der Gaag a Smolensky (1982) by měly být výdaje domácností modelovány určitou nějakou rovnicí. Obecná rovnice výdajů má vzorec:

$$q = \alpha_0 + \alpha(z) + \beta \ln x, \quad (1)$$

Následující rovnice je adaptací obecného vzorce (1) pro potřeby výdajové metody:

$$q_i = b_0 + b_1 \times adult_i + b_2 \times adult_i^{\wedge 2} + b_3 \times children_i, \quad (2)$$

kde q_i značí výdaje, $adult_i$ počet dospělých členů a $children_i$ počet dětí v domácnosti i. Rovnici lze odhadnout pomocí regresní analýzy. Regresní koeficienty pak označují zvýšení výdajů způsobené přidáním dalších pozorování (v našem případě dalších členů domácnosti). Parametr d_i, který nám umožňuje nalézt systém vah pro další členy konkrétního typu domácnosti i s počtem dospělých ($adult_i$) a počtem dětí ($children_i$), lze odvodit z následujícího vzorce:

$$d_i = ((b_0 + b_1 \times adult_i + b_2 \times adult_i^{\wedge 2} + b_3 \times children_i) - (b_0 + b_1 + b_2))/(b_0 + b_1 + b_2). \quad (3)$$

Pro odhad úspor z rozsahu pomocí tohoto výdajového přístupu je možné použít údaje z šetření Statistika rodinných účtů (HBS), které shromažďuje informace o výdajích domácností. Data jsou každoročně poskytována Českým statistickým úřadem a pro tuto analýzu byla využívána databáze za rok 2013, vůči kterému byly vztaženy všechny další analýzy (CZSO_HBS, 2013). Výdajový přístup byl již aplikován v předchozím autorském výzkumu Brázdilová a Musil (2016). Potvrdilo se, že výdaje domácností závisí na počtu dospělých a počtu dětí v domácnosti. Byla zjištěna výsledná rovnice a odhadované úspory z rozsahu, jakož i ekvivalenční škála podle výdajového přístupu. V následující publikaci Brázdilová a Musil (2017) byl prezentován vliv odhadované ekvivalenční škály podle tohoto výdajového přístupu na národní indikátory příjmů a jejich důvěryhodnost v rámci České republiky.

Užitková metoda

Tradičně je ekvivalenční škála definována jako poměr výdajů (nebo příjmů) dvou různých typů domácnosti se stejnou životní úrovní. Formálně to odpovídá poměru nákladových funkcí dvou typů domácností, které dosahují stejné úrovně užitku (Bütikofer & Gerfin, 2009). To vyžaduje analýzu užitkových funkcí v různých domácnostech. Takový užitkový přístup založený na funkci nepřímého užitku byl aplikován autory Bütikofer a Gerfin (2009). Autoři specifikovali kolektivní model domácnosti, který se pokouší zachytit jak úspory z rozsahu ve spotřebě domácností, tak nerovné rozdělení příjmů v rámci domácnosti. Tato analýza se zaměřuje na první fázi, tedy odhad úspor z rozsahu za předpokladu rovného rozdělení zdrojů. Individuální preference jsou zahrnuty do užitkové funkce domácnosti za předpokladu specifikace určitých pravidel. Funkce nepřímého užitku osoby i má podobu:

$$V_i = \alpha(z_i) + \beta \ln x_i + \varepsilon_i, \quad (4)$$

kde V_i označuje úroveň užitku, z_i pozorovatelné charakteristiky a x_i celkové výdaje na spotřebu osoby i. Vysvětlovaná proměnná V_i představuje finanční spokojenost osoby za vysvětlující proměnné z_i jsou považovány veličiny: věk, dummy proměnná pro úroveň vzdělání a také dummy proměnné pro označení páru a také přítomnosti dítěte v domácnosti. Dále mezi vysvětlující proměnné patří logaritmus příjmů, které jsou zde považovány za spotřební výdaje x_i. U jednotlivců se předpokládá, že svůj příjem spotřebují v každém období, tj. $x = y^h$, kde y^h znamená příjem domácnosti. Souvislost mezi příjmy a výdaji popisuje publikace autorů Berisha a Meszaros (2016). Funkce nepřímého užitku pro domácnosti jednotlivců je tedy následující:

$$V = \alpha(z) + \beta \ln y^h + \varepsilon. \quad (5)$$

Pro páry je tato rovnice složitější, neboť závisí na pravidle sdílení spotřeby v rámci daného páru. Úspory z rozsahu ze společného bydlení existují, pokud celková soukromá spotřeba obou členů domácnosti *f* a *m* přesahuje příjem domácností y^h. Tento efekt je zachycen v následujícím vzorci:

$$x_m + x_f = \tau\, y^h. \tag{6}$$

Skalár τ představuje princip transformace příjmů domácností y^h na celkovou spotřebu domácností. Nazývá se faktor celkových úspor z rozsahu. Pokud τ = 1, potom nejsou žádné úspory z rozsahu (veškerá spotřeba je soukromá). Logická horní hranice pro τ je 2 (veškerá spotřeba je veřejná). Pravidlo sdílení je klíčovým faktorem, který určuje, jaký podíl je přidělen jednomu členu páru. Pokud existuje omezení týkající se rovnoměrného sdílení spotřeby u párů, potom spotřeba jednoho člena je:

$$x = \frac{\tau\, y^h}{2}. \tag{7}$$

Užitková funkce každého člena páru je následující:

$$V = \alpha(z) + \beta \ln \frac{\tau\, y^h}{2} + \varepsilon. \tag{8}$$

Polovina faktoru celkových úspor z (τ) / 2 představuje ekvivalenční škálu za předpokladu rovného sdílení spotřeby mezi členy domácnosti. Určuje podíl příjmů páru, který jednotlivec potřebuje, aby dosáhl stejného užitku. Na základě tohoto postupu lze odhadnout regresní funkci užitku (měřenou spokojeností s příjmem) pomocí logaritmu příjmů a dummy proměnné pro označení páru. Tradiční přístup k ekvivalenční škále je založen na odhadu využívajícího obecného modelu užitkových funkcí pro každý typ domácnosti *h*:

$$V^h = \alpha(z) + \beta \ln x^h + \delta C, \tag{9}$$

kde *C* je dummy proměnná pro označení páru a δ je efekt užitku pro takovou dvoučlennou domácnost. Pro domácnosti jednotlivců Bütikofer a Gerfin (2009) redukují tento model také na individuální užitkovou funkci danou již výše uvedenou rovnicí (4). Následně má výsledný logaritmus ekvivalenční škály, za předpokladu domácnosti páru jako referenční, podobu:

$$\ln x^s - ln\ x^c = \frac{\delta}{\beta}, \tag{10}$$

kde x^c označuje výdaje domácnosti páru a x^s jsou výdaje domácnosti jednotlivce. Podíl δ / β se nazývá parametr posunu. Tímto lze parametr ekvivalenční škály (τ) / 2 odhadnout jako exp (δ / β). To představuje podíl výdajů (nebo příjmů) domácností páru, který domácnost jednotlivce potřebuje pro dosažení stejné úrovně užitku, což je vyjádřeno podílem x^s / x^c:

$$\frac{x^s}{x^c} = \frac{\tau}{2} = e^{\frac{\delta}{\beta}}. \tag{11}$$

Za předpokladu domácnosti jednotlivce jako referenční domácnosti by potřebný podíl x^c / x^s mohl být odhadnut pomocí převrácené hodnoty jako 1 / exp (δ / β). To představuje násobek, který domácnost páru musí spotřebovat, aby dosáhla stejné úrovně užitku jako jednotlivec. Hodnoty jsou v intervalu <1; 2>, kde 1 znamená, že úspory z rozsahu jsou maximalizovány a 2 představují nulové úspory ze společného žití. Vyjadřuje počet spotřebních jednotek pro dvoučlennou domácnost. Díky tomu lze odhadnout ekvivalenční škálu, protože váha pro první dospělou osobu je 1 a koeficient pro druhou osobu vznikne jednoduše rozdílem. Tato metodika založená na užitkové funkci byla již použita pro výpočet úspor z rozsahu ve výzkumu Jirková a Musil (2017), kde byla odhadnuta ekvivalenční škála podle užitkového přístupu.

Pro odhad úspor z rozsahu pomocí užitkového přístupu je nutné použít údaje ze šetření Příjmy a životní podmínky (EU-SILC), které je zaměřeno na příjmy a ekonomickou a finanční situaci domácností. Údaje jsou každoročně poskytovány Českým statistickým úřadem a pro účely této analýzy je využita databáze za rok 2013 především z důvodu modulových otázek v SILC 2013 zaměřených na finanční spokojenost domácností (CZSO_SILC, 2013). Navíc díky použití databáze ze stejného roku u obou uvedených přístupů je možné srovnání odhadovaných úspor z rozsahu podle

obou přístupů. Pro následné hodnocení dopadů národních ekvivalenčních škál na různé ekonomické ukazatele a indikátory příjmů je třeba využít také dat ze šetření EU-SILC. Na nich se sledují ukazatele jako nerovnost příjmů a míra ohrožení příjmovou chudobou (Šustová a Zelený, 2013). Kvůli porovnání s oficiálními hodnotami byla také zvolena databáze z roku 2013. Díky harmonizaci tohoto šetření napříč státy Evropské unie je navíc možné mezinárodní srovnání dopadů ekvivalenčních škál na příjmové ukazatele.

Výsledky

Výsledné odhady národní ekvivalenční škály podle alternativních přístupů

Parametry výsledné výdajové rovnice nám umožňují vypočítat celkové výdaje pro každý konkrétní typ domácnosti s přihlédnutím ke struktuře domácností. V tabulce 1 jsou uvedeny celkové počty spotřebních jednotek pro každý konkrétní typ domácnosti a odhady spotřebních jednotek, které patří dalšímu členu domácnosti na základě zvýšení výdajů, které daná osoba potřebuje navíc vůči domácnosti jednotlivce. Druhý dospělý člověk v domácnosti způsobuje nárůst celkových výdajů zhruba o 76 %, u třetího a dalšího dospělého je to mnohem méně na základě dat z roku 2013. První dítě zvýší celkové výdaje zhruba o 34 %. Spotřeba dětí není tak velká, jako je tomu u dospělých. Celkové výdaje vícečlenných domácností s dětmi vykazují vyšší úspory z rozsahu než vícečlenné domácnosti dospělých.

Tabulka 1: Odhad spotřebních jednotek (SJ) podle výdajového přístupu Zdroj: výpočty autora

	Jednotlivci	**Páry**	**Děti**
Celkové výdaje domácností (Kč/měsíc)	14 600	25 676	19 630
Počet SJ	1	1,76	1,34
Váha	0	0,75	0,30

Podle detailnější analýzy jsou spotřební jednotky v průběhu času stabilní a průměrné výsledky mohou být považovány za váhy dalších členů domácnosti. Stejná analýza byla totiž provedena pro roky 2010−2014 s velmi podobnými výsledky. Dalším dospělým osobám v domácnosti je přiřazena váha 0,75 a dalšímu dítěti do 14 let váha 0,3. Podle tohoto výdajového přístupu by národní ekvivalenční škála pro Českou republiku byla (1; 0,75; 0,3). Znamená to, že další dospělý v domácnosti v České republice dosahuje nižších úspor z rozsahu, než je průměr za evropské domácnosti. Ovšem úroveň spotřeby dítěte (do 14 let) znamená pouze 0,3 z celkové spotřeby jednotlivce, což odpovídá evropskému průměru. Úspory z rozsahu jsou tedy vyšší pro domácnosti s dětmi než pro domácnosti (se stejnou velikostí domácnosti) dospělých.

Podobně byly také parametry regresní analýzy užitkových funkcí transformovány na spotřební jednotky. Pro tuto analýzu byla využita data ze SILC 2013. Výsledky regresní analýzy založené na nepřímých užitkových funkcích jsou uvedeny v následující tabulce 2. Výsledky regresní koeficienty (parametry β a δ) se využívají k odhadu spotřebních jednotek. Parametr posunu se vypočítá jako podíl hodnot δ a β. Parametr ekvivalenční škály je definován jako exponenciální parametr posunu. Následně jsou tabulce 2 uvedeny odhady počtu spotřebních jednotek vah pro další členy domácnosti, zaokrouhlené na jedno desetinné místo.

Tabulka 2: Odhad spotřebních jednotek (SJ) podle užitkového přístupu Zdroj: výpočty autora

	Jednotlivci	**Páry**	**Děti**
Parametr posunu = δ/β	-	-0,317	-0,236
Parametr ekv. škály = exp(δ/β)	-	0,728	0,790
Počet SJ = 1/ exp(δ/β)	1	1,373	1,266
Váha = 1/ exp(δ/β)- 1	0	0,4	0,3

Tato metoda vede k závěru, že váha pro dalšího dospělého je 0,4, zatímco váha pro dítě je 0,3 na základě užitkové funkce. Další dospělý člověk tedy představuje 40 % spotřeby a dítě 30 % spotřeby prvního dospělého v domácnosti. Odhadované národní škály je možné srovnávat s mezinárodně uznávanými škálami včetně analýzy rozdílů. V tabulce 3 jsou srovnávány ekvivalenční škály na základě různých přístupů. Tyto stupnice se významně liší. Přístup „na hlavu“ nenachází využití,

neboť nepředpokládá žádné úspory ze společného hospodaření, přestože existují. Podobné výsledky jsou pozorovány u dětí, protože s výjimkou přístupu OECD všechny ostatní stupnice očekávají spotřebu dítěte na úrovni přibližně 30 % spotřeby prvního dospělého v domácnosti. Úroveň spotřeby druhého a dalšího dospělého se však liší. Nejnižší úspory z rozsahu (25 %) jsou předpokládány ve výdajové metodě. Naopak, nejvyšší (60 %) jsou očekávány v užitkovém přístupu.

Tabulka 3: Srovnání ekvivalenčních škál Zdroj: výpočty autora

Škála SJ	První dospělý	Další dospělý	Další dítě
Na hlavu	1	1,00	1,00
OCED	1	0,70	0,50
Modifikovaná OECD	1	0,50	0,30
Odhad dle výdajové metody	1	0,75	0,30
Odhad dle užitkové metody	1	0,40	0,30

Dopad národní ekvivalenční škály na příjmové ukazatele

Volba ekvivalenční škály ovlivňuje rozložení ekvivalizovaných příjmů a následně i hranice chudoby pro odhad míry ohrožení příjmovou chudobou. Je zřejmé, že výběr ekvivalenční škály má vliv na ukazatele chudoby. Hranice příjmové chudoby je definována jako 60 % mediánu národního ekvivalizovaného disponibilního příjmu domácností. Údaje o příjmech domácností poskytuje Český statistický úřad, který provádí národní verzi mezinárodně harmonizovaného šetření EU-SILC o životních podmínkách domácností. Dále je míra ohrožení příjmovou chudobou ovlivněna ekvivalenční škálou, protože vyjadřuje podíl lidí pod touto hranicí chudoby. Je vhodné se zaměřit nejen na celkovou míru ohrožení příjmovou chudobou, ale také na její hodnoty podle sociální skupiny domácností. Vzhledem k tomu, že se složení domácností může lišit v každé sociální skupině, dopad na každou sociální skupinu může být odlišný. V této kapitole je prezentováno srovnání vlivu různých ekvivalenčních škál na tento ukazatel. Srovnány jsou zde běžně používané odborné mezinárodní škály a odhadované národní ekvivalenční škály podle obou alternativních přístupů.

Míra ohrožení příjmovou chudobou pro Českou republiku v roce 2013 činí 8,6 %. Tento běžně používaný ukazatel je založený na modifikované škále OECD. Zatímco pokud by pro jeho výpočet byla využita původně navrhovaná škála OECD, byla by úroveň tohoto indikátoru pro Českou republiku za daný rok o něco nižší. Naopak při zohlednění pouze příjmu na obyvatele by tato míra dosahovala pro rok 2013 až úrovně 12,6 %. Tento ukazatel je tedy citlivý na volbu ekvivalenční škály a při jakékoliv volbě jeho konstrukce založené na ekvivalizovaných příjmech bude jeho hodnota vždy nižší než bez uvažování jakékoliv ekvivalizace. Výsledky tohoto ukazatele příjmů pomocí různých odhadů národních ekvivalenčních stupnic jsou uvedeny v následující tabulce. Spotřební jednotky zohledňující vyšší úspory z rozsahu indikují vyšší osobní ekvivalizovaný příjem, a tím i vyšší hranici chudoby. V důsledku toho zůstává více lidí pod touto hranicí, což zvyšuje míru chudoby. Oba odhady národních stupnic by změnily míru ohrožení příjmovou chudobou v opačném směru. Škála podle výdajového přístupu by znamenala pokles na úroveň 8,15 %, zatímco škála podle užitkového přístupu by zvýšila tuto míru na úroveň 9,15 %. První odhadovaná ekvivalenční škála bere v úvahu nižší úspory z rozsahu pro domácnost s více dospělými osobami ve srovnání s upravenou měrou OECD, což snižuje jejich ekvivalizovaný příjem a rozložení příjmů by bylo pravděpodobně více rovnoměrné. Na druhou stranu odhad založený na užitkovém přístupu předpokládá vyšší úspory z rozsahu, vyšší ekvivalizovaný příjem a pravděpodobně vyšší příjmovou nerovnost.

Tabulka 4: Dopad škály SJ na ukazatel chudoby v roce 2013 Zdroj: výpočty autora

Škála SJ	Hranice chudoby	Počet osob pod hranicí chudoby	Míra ohrožení příjmovou chudobou
Na hlavu	79 254	1 295 969	12,57 %
OECD	98 015	877 385	8,51 %
Modifikovaná OECD	**116 093**	**885 947**	**8,60 %**
Odhad dle výdajové metody	100 375	840 354	8,15 %
Odhad dle užitkové metody	125 349	942 891	9,15 %

Kromě hodnot ukazatelů je ovlivněna i jejich variabilita. Proto je ovlivněn nejen počet, ale také struktura osob ohrožených chudobou. Různé skupiny lidí jsou ekvivalenční škálou ovlivňovány různým způsobem. S ohledem na ekonomickou aktivitu lidí dochází na jedné straně k nejvýznamnějším změnám ve skupině dětí a na straně druhé u skupiny důchodců, protože častěji žijí v domácnostech s více členy, resp. samostatně. Velikost domácnosti je nejdůležitějším faktorem, který způsobuje význam dopadu ekvivalenční škály. Příjmová situace vícečlenných domácností je nejvíce ovlivněna nejvýše stanovením ekvivalenční škály, ale změna je zaznamenána i pro domácnost jednotlivce. Důvodem je pohyb celkové distribuce příjmů, která způsobuje relativní změnu pořadí všech domácností podle jejich příjmů.

Srovnání běžně používané modifikované stupnice OECD a obou odhadů národních stupnic je uvedeno v tabulce 5. Podle Eurostatem zveřejněného ukazatele míry ohrožení příjmovou chudobou pouze 8,7 % osob pod hranicí chudoby jsou děti, zatímco důchodci tvoří 19,9 %. Při použití obou odhadů stupnic lze pozorovat změnu počtu osob pod hranicí chudoby. Jejich struktura podle ekonomické aktivity je také mírně odlišná. Mezi lidmi pod hranicí chudoby vypočítané z ekvivalizovaného příjmu podle výdajového přístupu je dokonce až 12,1 % dětí a pouze 13,9 % důchodců. Děti obvykle žijí ve vícečlenných domácnostech, které podle odhadované míry založené na výdajovém přístupu realizují menší úspory z rozsahu. Proto je jejich příjmová situace horší ve srovnání s jinými typy domácností. Více dětí se tedy dostává do chudoby i navzdory nižší hraniční hodnotě. Příjmová situace důchodců žijících obvykle samostatně zůstává nezměněna. Nicméně, právě oni se díky jiné distribuci příjmů dostanou nad hranici chudoby. Změna procentuálního zastoupení mezi chudými je patrná i u ekonomicky aktivních osob (samostatně výdělečně činných a zaměstnanců s nižším nebo vyšším vzděláním) a také nezaměstnaných, kteří v této struktuře osob pod hranicí podle výdajové metody zastupují vyšší podíl ve srovnání s běžně používanou hranicí chudoby. Vezmeme-li v úvahu hranici podle užitkového přístupu, bylo by pod touto novou hranicí chudoby méně dětí (7,3 %) a více důchodců (25,6 %). Rovněž zastoupení ekonomicky aktivních osob a nezaměstnaných by bylo nižší než při použitím klasické hranice. Struktura lidí pod touto hranicí se tedy také změnila oproti původní metodě výpočtu, avšak tentokrát opačným směrem než v předchozím případě.

Tabulka 5: Osoby pod hranicí chudoby podle ekonomické aktivity v roce 2013 Zdroj: výpočty autora

	Populace		**Struktura osob pod hranicí chudoby (%)**		
	Absolutní struktura	**Relativní struktura (%)**	**Hranice dle modif.OECD škály**	**Hranice dle výdajové metody**	**Hranice dle užitkové metody**
Děti	702 413	6,8	8,7	12,1	7,3
Zaměstnanci s niž. vzd.	1 560 548	15,2	9,2	10,3	8,8
Samostatně činní	834 999	8,1	8,2	8,5	7,2
Zaměstnanci s vyš. vzd.	2 219 852	21,5	4,0	5,3	3,8
Důchodci	2 475708	24,0	19,9	13,9	25,6
Nezaměstnaní	539 960	5,2	25,3	27,0	23,4
Ostaní	1 973 324	19,2	24,7	22,9	23,9
Celkem	**10 306 805**	**100,0**	**100,0**	**100,0**	**100,0**

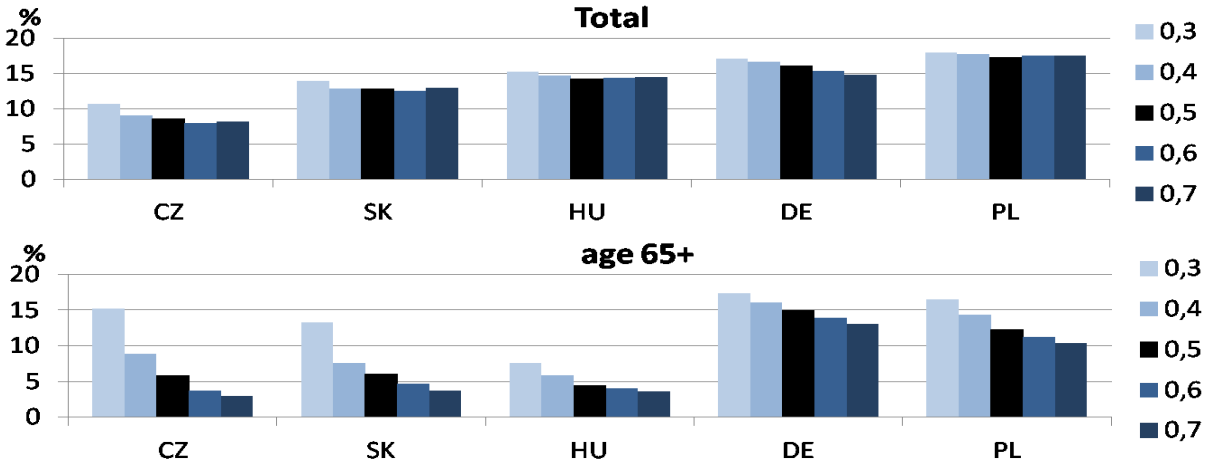
Vedle změny struktury lidí pod hranicí chudoby jsou pozorovatelné i určité změny v míře ohrožení příjmovou chudobou u různých skupin lidí podle jejich ekonomické aktivity. V tabulce 6 jsou tyto míry pro sociální skupiny osob porovnány podle použití různých ekvivalenčních škál. Celková míra vycházející z přístupu Eurostatu pro celou populaci je mezi úrovněmi vyplývajícími z národních odhadů. Je to způsobeno především výraznými změnami těchto měr u dětí a důchodců, zatímco míry pro ostatní skupiny zůstaly stabilní. Míra ohrožení příjmovou chudobou u dětí je vyšší (o 3,5 p.b.) podle stupnice z výdajového přístupu a nižší (o 1,3 p.b.) podle stupnice založené na užitkovém přístupu. Oproti tomu míry pro důchodce klesá o 2,4 procentních bodů u míry dle výdajové metody a zvyšuje se o 2,6 procentních bodů u užitkové metody. Změna struktury osob ohrožených příjmovou chudobou může mít významný vliv na politická rozhodnutí v sociální a rodinné politice.

Tabulka 6: Ukazatel chudoby podle ekonomické aktivity osob (v %) v roce 2013 Zdroj: výpočty autora

	Míra chudoby dle modif. OECD škály	Míra chudoby dle výdajové metody	Míra chudoby dle užitkové metody
Děti	11,0	14,5	9,7
Zaměstnanci s niž. vzd.	5,2	5,5	5,3
Samostatně činní	8,7	8,5	8,1
Zaměstnanci s vyš. vzd.	1,6	2,0	1,6
Důchodci	7,1	4,7	9,7
Nezaměstnaní	41,6	42,1	40,9
Ostaní	11,1	9,8	11,4
Celkem	**8,6**	**8,2**	**9,2**

Výsledky této analýzy prokazují, že národní úspory z rozsahu se liší od těch, které jsou zohledněny mezinárodními harmonizovanými odbornými škálami. Další člen české domácnosti, zejména dospělý, má jinou váhu, než se předpokládá v modifikované škále OECD. Oba alternativní přístupy k odhadu národních úspor z rozsahu jsou založeny na datech z výběrových šetření. Podle výdajového přístupu vypočítaného pomocí výdajů domácností by měla být váha dalších dospělých na úrovni 75 % prvního člena domácnosti, zatímco přístup založený na subjektivním vnímání finanční spokojenosti ukazuje, že váha je naopak nižší, je to jen asi 40 %. Úroveň spotřeby dítěte v domácnosti zůstává na úrovni mezinárodní modifikované stupnice OECD 30 %.

Ukazatele chudoby se v jednotlivých zemích liší, míra ohrožení příjmovou chudobou v České republice je na nejnižší úrovni mezi evropskými zeměmi, zatímco u ostatních postkomunistických zemí je mírně vyšší. Na druhou stranu evropský průměr i německý průměr jsou dokonce mnohem vyšší. Významnější rozdíly jsou pozorovány ve srovnání podle věku osob. Zatímco v Německu je úroveň tohoto ukazatele stejná pro všechny věkové skupiny, ve všech čtyřech postkomunistických zemích jsou pozorovatelné rozdíly, kdy je míra pro osoby ve věku 65 a více let (důchodci) mnohem nižší než u osob mladších 18 let (děti). Dopad volby ekvivalenční škály je pak významný především při rozlišení věkových skupinu než pro celkovou míru, pro kterou nebyla zjištěna žádná významná změna. Navýšení váhy dalšího dospělého v domácnosti znamená nižší úspory z rozsahu, čímž celková míra klesá a naopak. U věkové skupiny 65 a více je zjištěn stejný dopad, ale ve větším měřítku. V Německu nevznikají tak velké rozdíly, ale ve všech čtyřech postkomunistických zemích se tyto rozdíly prohlubují. Zejména v České republice je dopad obrovský, protože nárůst váhy na 0,7 by vedl ke dvojnásobnému poklesu míry, a naopak pokles váhy na 0,3 by představoval trojnásobné zvýšení míry. Dopad ekvivalenční škály na míru ohrožení příjmovou chudobou se v jednotlivých zemích liší, a to zejména podle věkových kategorií. Hlavním zjištěním je, že největší dopad změny ekvivalenční škály by byl sledován u důchodců v České republice.



Obrázek 1: Dopad ekvivalenční škály (váhy dalších dospělých) na celkovou míru ohrožení příjmovou chudobou celkem a u důchodců v 5 evropských zemích Zdroj: výpočty autora

Závěr

Stanovení ekvivalenční škály má významný dopad na ukazatele sociální statistiky, konkrétně na ukazatele příjmů a následných charakteristik příjmové chudoby. Národní statistické organizace používají modifikovanou škálu OECD, aby vytvořily konzistentní údaje o příjmové chudobě v Evropě. Předpokládá se, že jednotná škála zajistí mezinárodní srovnatelnost, kterou uživatelé spravedlivě vyžadují. Nebylo však dosud prokázáno, že tato stupnice je vhodná ve všech zemích, kde se uplatňuje. Nadto existují důkazy o tom, že v některých zemích se úspory z rozsahu domácnosti liší. Je otázkou, zda použití jednotné škály vede k srovnatelným výsledkům, nebo zda by harmonizace měla být prováděna na úrovni metodologie odhadu národní stupnice namísto doporučované společné škály. Je zřejmé, že druhá varianta by byla náročnější, protože je třeba překonat mnoho překážek. Přinejmenším šetření poskytující zdrojová data by musela být harmonizována, což je případ SILC, nikoliv HBS. Dalším bodem je metodika nebo přístup, který je třeba aplikovat, jelikož různé postupy mohou vést k rozdílným výsledkům. V této práci jsou uvedeny dva přístupy s různými výsledky. Výdajový přístup je založen na spotřebě, která úzce souvisí s účelem rozsahu stupnice spotřeby, tj. popisuje úspory vyplývající ze společného života. HBS však není harmonizováno z mnoha hledisek: velikost vzorku, technika výběru, definice výdajů (např. včetně započtení imputovaného nájemného). Užitkový přístup využívá sice údaje založené na harmonizovaném průzkumu (SILC), ale může být sporné využívat subjektivní ukazatele, neboť postoj k životu se může v jednotlivých zemích lišit.

Jedním z výsledků tohoto výzkumu je zjištění, že mezinárodní modifikovaná škála OECD není pro Českou republiku vhodná. V této analýze byl využit výdajový a užitkový přístup, přičemž každý z nich vede k různým odhadům ekvivalenční škály. Podle výdajového přístupu další dospělý realizuje menší úspory z rozsahu, než je předpokládáno modifikovanou stupnicí OECD, jelikož koeficient činí 0,75. Nicméně přístup založený na užitkové funkci vede k jinému závěru, protože odhadovaný koeficient činí 0,4. Na druhou stranu všechny metody vedou ke stejnému závěru pro dítě, protože koeficient je ve všech případech stejný - 0,3. Dále byl prokázán dopad ekvivalenční škály na ukazatele příjmů a zejména na míru ohrožení příjmovou chudobou. První přístup by snížil míru z běžně vypočtené úrovně 8,6 % v roce 2013 na 8,2 %, zatímco druhý by ji zvýšil na hodnotu 9,2 %. Tato míra se liší zejména mezi skupinami lidí podle jejich ekonomického postavení, kde jsou nejvyšší rozdíly mezi důchodci a dětmi. Domácnosti s více osobami (s dětmi) by byly pravděpodobněji více ohroženy příjmovou chudobou v případě, že by se uvažovaly nižší úspory z rozsahu podle prvního přístupu než jednotlivci, např. důchodci žijící samostatně. Podle druhého přístupu je situace naprosto opačná. Tímto způsobem bude ovlivněna i struktura lidí pod hranicí chudoby.

Literatura

BRÁZDILOVÁ, Michaela a Petr MUSIL. Estimate of Scale of Consumption Units in the Czech Republic Based on Expenditures of Czech Households. In: *Applications of Mathematics and Statistics in Economics – AMSE 2016 [online]*. Banská Stiavnica, 2016, s. 58-65. ISBN 978-80-89438-04-4.

BRÁZDILOVÁ, Michaela a Petr MUSIL. Impact of Consumption Unit's Scale on Credibility of the Income Indicators in the Czech Republic. In: *Statistika, iss. 97, č. 2*. 2017, s. 15–24. ISSN0322-788X. *https://www.czso.cz/documents/10180/45606525/32019717q2015.pdf/7c57049b-b0d1-43c9-ad87-e40914f8f660?version=1.0.*

BERISHA, Edmond a John MESZAROS. Household debt, economic conditions, and income inequality: A state level analysis. In: *The Social Science Journal, 2017, iss. 54, č. 1*. 2016, s. 93-101. *DOI: 10.1016/j.soscij.2016.11.002.*

BISHOP, John A., I. A. ZARCO, A. GRODNER a H. LIU. 2015). Subjective poverty equivalence scales for Euro Zone countries. In: *Journal Of Economic Inequality, 2014, vol. 12, č. 2.* 2015, s. 265-278. *DOI: 10.1007/s10888-013-9254-7.*

BISHOP, John. A., J. LEE a L. A ZEAGER. Improving the Supplemental Poverty Measure: Two Simple Proposals. In: *ECINEQ, Society for the Study of Economic Inequality in its series Working Papers, 2017, č. 429.* 2017. *http://www.ecineq.org/milano/WP/ECINEQ2017-429.pdf.*

BISHOP, John. A., J. LEE a L. A. ZEAGER. Adjusting for family size in the supplemental poverty measure. In: *Journal: Applied Economics Letters, 2017. DOI: 10.1080/13504851.2017.1343444.*

BUHMANN, B., L. RAINWATER, G. SCHMAUS a T. M. SMEEDING. Equivalence Scales, Well-Being, Inequality, And Poverty - Sensitivity Estimates Across 10 Countries Using The Luxembourg Income Study (Lis) Database. In: *Review Of Income And Wealth, č. 2.* 1988, s. 115-142. *http://www.roiw.org/1988/115.pdf.*

BÜTIKOFER, Aline a Michael GERFIN. The Economies of Scale of Living Together and How They Are Shared: Estimates Based on a Collective Household Model. In: *Discussion Paper č. 4327.* 2009. *http://ftp.iza.org/dp4327.pdf.*

CHANFREAU, Jenny a Tania BURCHARDT. Equivalence scales: rationales, uses and assumptions. 2008. *Available at: http://www.gov.scot/resource/doc/933/0079961.pdf.*

CZSO_HBS. 2013. *https://www.czso.cz/csu/czso/statistika-rodinnych-uctu-metodika [online]*.

CZSO_SILC. 2013. *https://www.czso.cz/csu/czso/prijmy-a-zivotni-podminky-domacnosti [online]*.

DE VOS, Klaas a M. Asgar ZAIDI. Equivalence scale sensitivity of poverty statistics for the member states of the european community. In: *Review Of Income And Wealth, č. 3.* 1997, s. 319-333. *http://www.roiw.org/1997/319.pdf.*

DUDEL, Christian. Nonparametric bounds on equivalence scales. In: *Economics Bulletin, iss. 35, č. 4.* 2015, s. 2161-2165.

FÖRSTER, Michael F. Measurement Of Low Incomes And Poverty In A Perspective Of International Comparisons. In: *Labour Market And Social Policy, OECD, Directorate For Education, Employment, Labour And Social Affairs, č.14.* 1994.

HAGENAARS, A. J. M., K. DE VOS a M. A. ZAIDI. *Poverty Statistics in the Late 1980s: Research Based on Micro-data*. Office for Official Publications of the European Communities, Luxembourg, OECD, 1994.

JIRKOVÁ, Michaela a Petr MUSIL. Analysis of economies of scale in czech households based on different approaches. In: *Applications of Mathematics and Statistics in Economics 2017– AMSE [online]*. Szklarska Poręba, Poland, 2017, s. 227-234. ISBN 978-83-7695-693-0.

MALÁ Ivana. Vliv definice spotřebních jednotek na ekvivalizované příjmy domácností v České republice (The Impact of Definition of Consumption Units on Equivalised Household Income in the Czech Republic). In: *Acta Oeconomica Pragensia, 12, č. 3.* 2015, s. 53-67. *DOI 10.18267/j.aop.536.*

SCHWARZE, Johannes. Using Panel Data On Income Satisfaction To Estimate Equivalence Scale Elasticity. *In: Review Of Income And Wealth, č. 3.* 2003, s. 359-372.

VAN DER GAAG, Jacques a Eugene SMOLENSKY. True household equivalence scales and characteristics of the poor in the United States. 1982. *http://www.roiw.org/1982/17.pdf.*

ŠUSTOVÁ, Šárka a Martin ZELENÝ. At Work…, and Poor? A Look at the Czech Working Population in the Living Condition Survey (EU-SILC). In: *Statistika, 93, č.1.* 2013, s. 56 - 70.

JEL Classification: D12, D31

Summary

Analysis of the Impact of National Equivalence Scales Based on Different Approaches on Czech Indicators

Equivalence scale of consuption units takes into account the economies of scale of different households types according thein size and structure. The aim of this paper is to compare the economies of scale considered in international equivalence scale with estimate of the economies of scale in individual countries based on their specific conditions and real economic situation of their households. This paper is focus primarily on situation in the Czech Republic, but some comparison to other European countries is provided. The estimate of economies of scale could be based on different approaches. One of the alternative approaches to consumption unit scales applied in some countries is based on household expenditures. This method could use data from Household

Budget Surveys (HBS). Another method uses utility function of households, which is indirectly measured by using subjective data of financial satisfaction. The analysis is based on data from European Income and Living Conditions Surveys (EU-SILC). Consumption units prepared according to both approaches are estimated applying regression. The impact of application of national equivalence scale on some Czech indicators is huge, among income indicators especially on at-risk-of-poverty rate. The change of structure of people below poverty threshold is observable, particularly by their social status. The analysis of impacts in the Czech Republic in comparison to European countries is presented in this paper.

Key words: economies of scale, consumption units, equivalence scale, household expenditures, utility function, at-risk-of-poverty rate.

Využití vybraných metod shlukové analýzy v bankovním propenzitním modelu

Sergej Sirota

xsirs00@vse.cz

Doktorand oboru statistika

Školitel: prof. Ing. Hana Řezanková, CSc., (hana.rezankova@vse.cz)

Abstrakt: Pro odhadování bankovních propenzitních modelů se v praxi nejčastěji využívá logistická regrese, jedná se tedy o klasifikační úlohu (klasifikaci klientů) se známým počtem skupin. Článek si klade za cíl představit a porovnat možnosti využití vybraných metod shlukové analýzy, jež slouží ke klasifikaci bez známého počtu skupin, k vylepšení bankovního propenzitního modelu odhadnutého logistickou regresí. Jsou porovnávány možnosti metody *k*-průměrů, která se řadí do metod rozkladu, a dvoukrokové metody shlukování, která byla navržena pro velké datové soubory. Porovnání se týká jednak kvality vytvořených shluků (při shlukování objektů) obrysovým koeficientem, jednak úspěšnosti samotných nově vytvořených propenzitních modelů na základě míry úspěšnosti. Představí se celkem tři přístupy, které shodně v prvním kroku analýzy aplikují metody shlukové analýzy pro rozdělení objektů v datové matici na 5, 10 a 15 uvažovaných shluků. Tímto rozdělením vznikne i nová proměnná představující příslušnost objektu do shluku (shluková proměnná). První přístup v druhém kroku využívá vytvořenou shlukovou proměnnou jako vysvětlující v modelu logistické regrese. Druhý a třetí přístup odhaduje model logistické regrese pro každý vytvořený shluk, kde již ale nevstupuje shluková proměnná jako vysvětlující proměnná. Rozdíl spočívá v přístupu odhadu propenzitního modelu.

Klíčová slova: propenzitní model, metody shlukové analýzy, logistická regrese, míra úspěšnosti

Úvod

Propenzitní modely jsou často využívány v bankovnictví pro klasifikaci klientů s cílem pomoci naplnit obchodní požadavky, které se týkají především zvýšení měr prodejů bankovních produktů a snížení míry odchozích klientů. Propenzita znamená sklon k určité události, přičemž tento sklon se vyjadřuje pomocí skóre, které je přímo úměrné pravděpodobnosti nastání události odhadované pomocí modelu. Pro odhad modelu se nejčastěji využívá logistická regrese, popřípadě rozhodovací stromy. Existují články (např. Arpino; Cannas, 2016, Rudolph et al., 2016, Yang et al., 2016), které popisují možnosti využití shlukové analýzy pro zlepšení statistického modelu odhadovaného pomocí logistické regrese. Ideou je zkoumané objekty v datové matici rozdělit na základě shlukové analýzy do *k* shluků a následně pro každý shluk pomocí regresní funkce odhadnout vysvětlovanou proměnnou. Cílem tohoto článku je představit a porovnat tři možné způsoby využití vybraných metod shlukové analýzy při odhadu bankovního propenzitního modelu pomocí logistické regrese. Aplikace prvního přístupu již byla popsána v (Sirota a Řezanková, 2017) a dále bude uvedena pro srovnání. V článku je zkoumán dopad využití metody *k*-průměrů a dvoukrokové metody shlukování na úspěšnost nově vytvořených propenzitních modelů (měřených pomocí míry úspěšnosti) pro spotřebitelské půjčky odhadnutými logistickou regresí. Tato úspěšnost je porovnávána vůči míře úspěšnosti současného nasazeného propenzitního modelu na stejné modelovací bázi. V rámci analýz jsou rovněž hodnoceny vytvořené shluky v datové matici obrysovým koeficientem (Löster, 2016). Veškeré výpočty byly provedeny v programu *IBM SPSS Modeler*.

Použitá metodologie

Sklon k nastání události je vyjádřen pomocí skóre, které se odhaduje nejčastěji modelem logistické regrese. Skóre leží v intervalu $\langle 0,1 \rangle$ a v praxi se pak následně objekty v datové matici rozřadí do 10 skupin se stejně širokými intervaly skóre od nejlepších po nejhorší. Na základě takovéhoto rozřazení objektů lze hodnotit odhadnutý model pomocí *míry úspěšnosti*. Tato míra vyjadřuje podíl objektů, u kterých sledovaná událost skutečně nastala, na celkovém počtu vybraných objektů. Je ovlivněna počtem vybraných objektů. V našem případě jsou objekty představovány klienty a nastání

události představuje koupě produktu spotřebitelského financování. Pro porovnání úspěšností nově odhadnutých propenzitních modelů vůči současnému se budou uvažovat klienti s hodnotou skóre alespoň 0,7, což odpovídá zavedené praxi.

Logistická regrese

V současnosti je logistická regrese nejběžnější statistická metoda v bankovnictví pro odhad sklonu nastání události, protože se předpokládá alternativní vysvětlovaná proměnná Y, kde $Y_i = 1$ znamená nastoupení sledovaného jevu u i-tého objektu, a $Y_i = 0$ nenastoupení jevu. Jinými slovy to v našem případě znamená, zda klient produkt koupil či nekoupil. Na rozdíl od klasické lineární regrese, kde vysvětlovaná proměnná je z celého oboru reálných čísel, se uvažuje omezený interval $\langle 0,1 \rangle$. V modelu logistické regrese se odhaduje pravděpodobnost, že vysvětlovaná proměnná Y_i nabude hodnoty 1 nebo 0, kde vztah s vektorem hodnot vysvětlujících proměnných $\mathbf{x}$ je uvažován nelineární. Tento vztah lze popsat známým modelem logistické regrese:

$$P(Y = 1|\mathbf{x}) = \pi = \frac{e^{\boldsymbol{\beta}\mathbf{x}}}{1 + e^{\boldsymbol{\beta}\mathbf{x}}},$$

kde

$\boldsymbol{\beta}$ je vektor parametrů regresní funkce.

Odhad parametrů lze získat například metodou maximální věrohodnosti (Agresti, 2002).

Vybrané metody shlukové analýzy a hodnocení kvality shlukování

Aplikací shlukové analýzy lze zkoumat vztahy mezi objekty v datové matici. Obecný postup shlukové analýzy je takový, že se klasifikují objekty do určitého počtu skupin tím způsobem, aby uvnitř jedné skupiny byly objekty co nejvíce podobné, a zároveň objekty mezi skupinami byly co nejvíce rozdílné. Metody shlukové analýzy se rozlišují podle způsobu klasifikace na metody hierarchické a na metody rozkladu. V tomto článku je aplikována metoda k-průměrů, patřící do shlukových metod rozkladu, a dále dvoukroková metoda shlukování, která byla navržena pro velké datové soubory a patří do hierarchického přístupu. Kvalita vytvořených shluků je hodnocena obrysovým koeficientem.

Metoda *k*-průměrů

Mezi nejznámější metody rozkladu patří metoda k-průměrů, která je vhodná v případě, kdy objekty v datové matici jsou charakterizovány kvantitativními proměnnými. Objekty jsou ve výsledku pevně přiřazeny do určitých shluků, jejichž počet je třeba zadat před analýzou. Algoritmus k-průměrů je iterativní, kde se nejprve pro každý počáteční shluk spočte centroid, který představuje průměrné hodnoty jednotlivých proměnných v daném shluku. Následně se minimalizuje funkce:

$$f_{\mathrm{KP}} = \sum_{h=1}^{k}\sum_{i=1}^{n} u_{ih} \left\| \mathbf{x}_i - \bar{\mathbf{x}}_h \right\|^2,$$

kde

u_{ih}= 1, pokud i-tý objekt (i = 1, 2, …, n, kde n značí počet objektů) patří do h-tého shluku (h = 1, 2, …, k, kde k značí počet shluků) – v opačném případě u_{ih} = 0,

$\mathbf{x}_i$ je vektor hodnot vysvětlujících proměnných pro i-tý objekt,

$\bar{\mathbf{x}}_h$ je vektor průměrných hodnot vysvětlujících proměnných pro každý h-tý shluk,

$\| . \|$ je euklidovská vzdálenost.

Uvedená funkce se minimalizuje za uvedených podmínek:

$$\sum_{h=1}^{k} u_{ih} = 1 \text{ pro } i = 1, 2, \ldots, n,$$

$\sum_{i=1}^{n} u_{ih} > 0$ pro $h = 1, 2, \ldots, k$ (Hebák et al., 2013).

Dvoukroková metoda shlukování

Tato metoda byla navržena pro shlukování objektů ve velkých datových souborech a z tohoto důvodu byla vybrána pro shlukování klientů v modelovacím souboru, který je velmi rozsáhlý. Způsob shlukování je postaven na algoritmu BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), kde v prvním kroku se vytvářejí pomocné shluky, které se následně ve druhém kroku shlukují již klasickým hierarchickým způsobem. I v tomto případě se musí stanovit předem počáteční počet shluků. Podrobněji o této metodě lze nalézt v (Zhang, 1996).

Obrysový koeficient

Pro vyhodnocení kvality vytvořených shluků lze použít obrysový koeficient (Löster, 2016). Nabývá hodnot v intervalu $\langle -1,1 \rangle$, kde větší hodnoty tohoto koeficientu znamenají lepší rozřazení objektů do vytvořených shluků. Základem je zhodnocení kvality přiřazení jednotlivých objektů do přiřazeného shluku, které pro každý *i*-tý objekt v datové matici lze vyjádřit následovně:

$$\psi_i = \frac{\mu_i - \eta_i}{\max\{\eta_i, \mu_i\}},$$

kde

$$\eta_i = \frac{\sum_{j \in C_h} d_{ij}}{n_h - 1},$$

$$\mu_i = \min_{g \neq h} \left(\frac{\sum_{j \in C_g} d_{ij}}{n_g} \right),$$

d_{ij} je euklidovská vzdálenost mezi *i*-tými a *j*-tými objekty,

n_h (n_g) představuje počet objektů v *h*-tém (*g*-tém) shluku.

Výsledný obrysový koeficient pro dané uspořádání shluků představuje průměrnou hodnotu přes všechny objekty v datové matici:

$$\psi = \frac{\sum_{i=1}^{n} \psi_i}{n}.$$

Vybrané přístupy aplikace metod shlukové analýzy při odhadu propenzitního modelu logistickou regresí

Datový soubor pro odhad modelu pro spotřebitelské půjčky je velmi rozsáhlý, protože obsahuje okolo 2000 proměnných, které představují produktové a transakční charakteristiky pro 154 113 klientů. Vysvětlovaná proměnná Y_i představuje informaci o tom, zda *i*-tý klient si sjednal nebo nesjednal spotřebitelský úvěr.

Z důvodu rozsáhlosti datového souboru bylo pro účely shlukové analýzy, tak i pro logistickou regresi vybráno ze souboru celkem 100 nejlepších kvantitativních proměnných z hlediska vysvětlované proměnné. Tuto možnost výběru nabízí přímo program *IBM SPSS Modeler* a je založen na porovnání korelace mezi vysvětlovanou a vysvětlující proměnnou. Získaný výběr obsahuje i devět kvantitativních vysvětlujících proměnných obsažených v nasazeném modelu odhadnutým logistickou regresí.

Shluková analýza byla aplikována na celý datový soubor a pro účely logistické regrese byl datový soubor rozdělen na trénovací a testovací část v poměru 70:30. Nasazený model slouží jako referenční pro porovnání úspěšností nově vytvořených modelů (popřípadě přístupů) s využitím metod shlukové analýzy a je označen jako *aktuální propenzitní model*. Úspěšnost propenzitních modelů je porovnávána pomocí míry úspěšnosti na testovací bázi, kde do výběru vstupují klienti s odhadnutou pravděpodobností sjednání produktu alespoň 0,7. Aktuální propenzitní model na testovací bázi dostahuje míry úspěšnosti 77,91 %.

Před odhadem nových propenzitních modelů je nejdříve aplikována shluková analýza, kde pro algoritmus *k*-průměrů a dvoukrokové metody shlukování je uvažováno 5, 10 a 15 existujících shluků v datovém souboru. Touto aplikací je vytvořena nová proměnná značící příslušnost klienta do daného shluku a dále bude nazvána jako *shluková proměnná*.

Přístup 1

Jak již bylo uvedeno, tento přístup byl představen v (Sirota a Řezanková, 2017). Nové modely se odhadly logistickou regresí, kde vstupující proměnné byly proměnné z aktuálního propenzitního modelu (celkem devět), a navíc se přidala nově vytvořená shluková proměnná – celkem tedy 10 vysvětlujících proměnných. Jelikož se uvažovaly dvě metody shlukové analýzy s 5, 10 a 15 existujícími shluky v datovém souboru, odhadlo se celkem šest nových modelů.

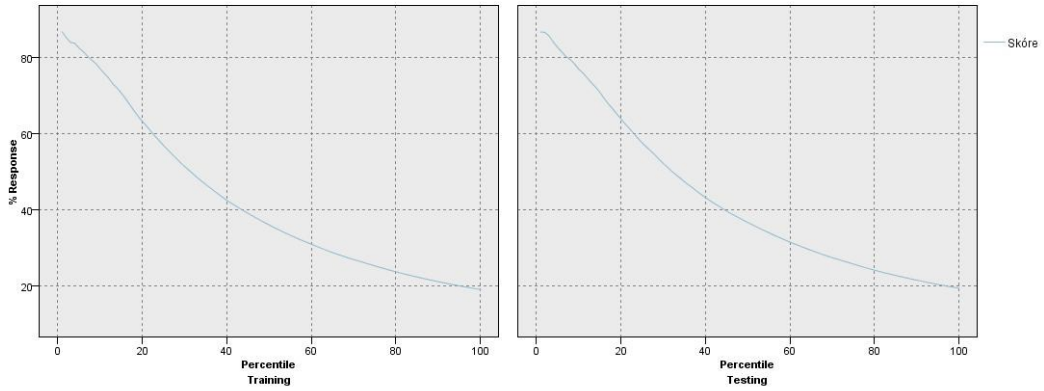
Přístup 2

Druhou možností bylo na základě shlukové analýzy rozdělit modelovací bázi na vytvořené shluky (5, 10 a 15) a pro každý shluk odhadnout model logistické regrese se vstupujícími proměnnými zaktuálního propenzitního modelu. Celkem vzniklo 60 modelů, protože ač jsou vysvětlující proměnné vždy stejné, tak koeficienty jsou v modelech rozdílné. Pro porovnání míry úspěšnosti se ale výsledky modelů pro danou shlukovou metodu a daný počet shluků agregovaly. Pro příklad u metody *k*-průměrů s uvažováním pěti shluků v datovém souboru: z každého shluku se vybrali klienti na základě modely odhadnuté pravděpodobnosti sjednání alespoň 0,7, a na těchto klientech se vypočítala míra úspěšnosti.

Přístup 3

Třetí přístup je podobný jako předchozí s tím rozdílem, že pro každý shluk se odhadly nové modely logistické regrese, kde vysvětlující proměnné byly zvoleny nově ze 100 kvantitativních proměnných, a to na základě uvedených podmínek:

- statisticky významné Waldovy testy o nulovosti hodnot jednotlivých regresních koeficientů na 5% hladině významnosti,
- párové Pearsonovy korelační koeficienty vysvětlujících proměnných menší nebo rovny 0,8 (v absolutní hodnotě),
- křivka míry úspěšnosti na trénovací a testovací množině by se měla co nejvíce blížit průběhu znázorněnému na obrázku 1.



Obrázek 1: Ideální průběh křivky míry úspěšnosti na trénovací a testovací množině
Zdroj: vlastní zpracování v programu *IBM SPSS Modeler*

Uvedený přístup byl velmi pracný, protože se jednalo o vytvoření celkem 60 zcela nových modelů z hlediska výběru vysvětlujících proměnných. Pro porovnání míry úspěšnosti se výsledky modelů opět agregovaly pro danou shlukovou metodu a daný počet shluků jako v přístupu 2.

Výsledky

U všech popsaných přístupů byla v prvním kroku nejprve aplikována shluková analýza, která je společná pro uvedené přístupy. V tabulce 1 a 2 (členěné dle uvažované metody shlukové analýzy) jsou shrnuty výsledky jak měření kvality shluků, tak i úspěšnosti jednotlivých přístupů pro odhad sklonu nastání události měřených na základě míry úspěšnosti.

Hodnoty obrysového koeficientu se pohybovaly v rozmezí 0,2–0,5 (pro 5, 10 a 15 shluků),což lze považovat za kvalitní rozložení objektů do shluků. Při porovnání metod shlukové analýzy vyšly hodnoty lépe u metody *k*-průměrů, kde největší hodnota koeficientu byla dosažena při uvažování 5 a 15 existujících shluků v datové matici. U dvoukrokové metody shlukování vyšlo nejlépe shlukování do pěti shluků.

U hodnocení již samotných nově vytvořených modelů nejlépe dopadl přístup 3 s využitím metody *k*-průměrů a klasifikací objektů do pěti shluků, kde míra úspěšnosti činila 81,19 %. V případě využití dvoukrokové metody shlukování byla dosažena nejvyšší míra úspěšnosti opět v přístupu 3, ovšem při uvažování 15 shluků. Míra úspěšnosti při využití aktuálního propenzitního modelu byla 77,91 %.

Tabulka 1: Úspěšnost jednotlivých přístupů s využitím algoritmu *k*-průměrů

Počet shluků	Obrysový koeficient	Míra úspěšnosti (v %)		
		Přístup 1	Přístup 2	Přístup 3
5	0,5	79,20	79,57	81,19
10	0,4	79,25	79,74	80,23
15	0,5	79,16	79,76	79,48

Zdroj: vlastní zpracování

Tabulka 2: Úspěšnost jednotlivých přístupů s využitím algoritmu dvoukrokové metody shlukování

Počet shluků	Obrysový koeficient	Míra úspěšnosti (v %)		
		Přístup 1	Přístup 2	Přístup 3
5	0,3	78,49	78,52	78,48
10	0,2	78,94	79,10	77,99
15	0,2	78,85	79,06	79,54

Zdroj: vlastní zpracování

Závěr

V příspěvku byly představeny tři možné způsoby využití metody *k*-průměrů a dvoukrokové metody shlukování pro zlepšení odhadu propenzitního modelu logistickou regresí. V prvním přístupu se uvažovaly při odhadu modelu vysvětlující proměnné z aktuálního propenzitního modelu, kde se navíc přidala do odhadu nová shluková proměnná. Dosáhlo se nepatrně lepších hodnot měr úspěšnosti, kde nejlepší výsledky poskytla metoda s využitím algoritmu *k*-průměrů při 10 vytvořených shlucích v datovém souboru. Druhý a třetí přístup je podobný, protože se datový soubor rozdělil na vytvořené shluky a pro každý shluk se odhadoval model logistické regrese. Rozdíl byl v tom, že u druhého přístupu byly vysvětlující proměnné převzaty z aktuálního propenzitního modelu, kdežto u třetího přístupu se proměnné vybíraly nově. Třetí způsob aplikace přinesl nejlepší výsledky, kde nejvyšší míru úspěšnosti poskytla metoda s využitím algoritmu *k*-průměrů při rozdělení datového souboru na pět shluků. Z výsledků vyplývá, že využití *k*-průměrů dává lepší výsledky – ostatně i samotná průměrná kvalita vytvořených shluků byla lepší než u dvoukrokové metody shlukování.

Výsledky uvedených přístupů naznačují, že má smysl pokračovat ve zkoumání dalších možností využití metod shlukové analýzy pro zlepšení úspěšnosti odhadovaného propenzitního modelu logistickou regresí. Novým přístupem může být aplikace algoritmu regresního shlukování pro odhad propenzitního modelu (Gawrysiak et al., 2001).

Literatura

SIROTA, Sergej; ŘEZANKOVÁ, Hana. Application of clustering methods in bank's propensity model. In LÖSTER, T. – PAVELKA, T. (Eds.). *11th International Days of Statistics and Economics (MSED)*, Praha: Melandrium, 2017, s. 1431-1439. ISBN 978-80-87990-12-4.

AGRESTI, Alan. Categorical data analysis (2. vyd.). NYC: John Wiley & Sons, 2002.

ARPINO, Bruno; CANNAS, Massimo (2016). Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. *Statistics in Medicine*, 2016, 35(12), 2074-2091.

ZHANG, Tian, et al. BIRCH: an efficient data clustering method for very large databases. In: ACM Sigmod Record. ACM, 1996. p. 103-114.

GAWRYSIAK, Piotr; at al. Regression—Yet Another Clustering Method. In: Intelligent Information Systems 2001. Physica, Heidelberg, 2001. p. 87-95.

HEBÁK, P., et al. Statistické myšlení a nástroje analýzy dat (1. vyd.). Praha: Informatorium. ISBN 978-80-7333-105-4, 2013.

YANG, Shu, et al. Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics, 2016, 72.4: 1055-1065.

RUDOLPH, Kara E., et al. Optimally combining propensity score subclasses. Statistics in medicine, 2016, 35.27: 4937-4947.

LÖSTER, Tomáš. Determining the optimal number of clusters in cluster analysis. In LÖSTER, T. – PAVELKA, T. (Eds.). *10th International Days of Statistics and Economics* (MSED), Praha: Melandrium, 2016, s. 1078-1090.

JEL Classification: C25, C38, D12

Summary

Use of Selected Clustering Methods in Bank's Propensity Model

Bank's propensity model is mostly estimated by using logistic regression. This paper introduced three possible uses of *k*-means and TwoStep clustering method for increasing success rate of propensity model, which is estimated by logistic regression. There were considered 5, 10 and 15 clusters in the modeling base for each clustering method. *K*-means algorithm had better results than TwoStep algorithm.

Keywords: propensity model, clustering methods, logistic regression, success rate