SBORNÍK

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

Vědecký seminář se uskutečnil dne 9. února 2017
pod záštitou děkana FIS doc. RNDr. Luboše Marka, CSc.

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

Nakladatelství Oeconomica – Praha 2017

ISBN 978-80-245-2199-2

OBSAH

Předmluva

Na počátku února roku 2017 se konal tradiční seminář „Den doktorandů“ Fakulty informatiky a statistiky VŠE v Praze. Seminář slouží jako platforma pro presentaci výsledků vědecké a odborné práce studentů všech doktorských oborů fakulty. Pro mnohé z nich je to první vystoupení před odbornou veřejností, na němž získávají cenné zkušenosti a mohou si tak vytříbit schopnosti formulace svých názorů a hypotéz nebo argumentů na jejich podporu. Nedílnou součástí jejich vystoupení je presentace závěrů výzkumné práce a diskuse nad jejími výsledky. Vlastní příprava vystoupení pak umožní doktorandům zvykat si na fakt, že výsledky své často dlouhodobé práce musí stěsnat do relativně krátkého časového úseku presentace, v níž by měli být schopni jasně, srozumitelně, přehledně a poutavým způsobem přednést to, na čem pracovali.

Do dvacátého druhého ročníku „Dne doktorandů“ se přihlásilo celkem dvacet dva účastníků. Jeden z nich se omluvil. Na oboru „Aplikovaná informatika“ to bylo devět (z toho jeden omluvený) příspěvků, na oboru „Ekonometrie a operační výzkum“ sedm příspěvků a na oboru „Statistika“ šest příspěvků. Semináře se zúčastnili studenti obou forem studia, jak presenční, tak i kombinované.

Za práci v komisích bych chtěl poděkovat všem jejím členům, kteří pracovali pod vedením předsedů. 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 doc. Ing. Ota Novotný, CSc. (KIT) a doc. Ing. Vlasta Střížová, CSc. (KSA). Pro studijní program „Kvantitativní metody v ekonomice“, studijní obor „Ekonometrie a operační výzkum“ pracovala komise ve složení předseda pan prof. Ing. Roman Hušek, CSc. (KEKO), členy jeho komise pak byli doc. Ing. Tomáš Cahlík, CSc. (KEKO) a doc. RNDr. Ing. Michal Černý, Ph.D. (KEKO). Pro obor statistika byla předsedkyní komise paní prof. Ing. Hana Řezanková, CSc. (KSTP), a členové byli doc. Ing. Dagmar Blatná, 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: Neuděleno
2. místo: Ing. Jiří Hradil: Ergonomie software – identifikace a původ vzorů uživatelských rozhraní
Ing. Richard Novák: Big data from the perspective of data sources and legal regulation
3. místo: Ing. Jana Fortinová: Aktuální otázky řízení eGovernmentu
Ing. Ivan Jelínek: Text mining na sociálních sítích: analýza politické události

Studijní program – Kvantitativní metody v ekonomice
obor Ekonometrie a operační výzkum
1. místo: Ing. Lukáš Frýd: Analýza šikmosti finančních aktiv
Mgr. Vladimír Holý: Market Microstructure Noise
2. místo: Neuděleno
3. místo: Ing. Michal Švarc: Vliv profesní mobility na hodnocení pracovní spokojenosti
Ing. Ondřej Sokol: Statistiky nad intervalovými daty

obor Statistika
1. místo: RNDr. Petr Sotona: Porovnání kalibrace modelů úmrtnosti na českých datech
2. místo: Ing. Jan Zeman: Využití metod smíšených frekvencí při konstrukci čtvrtletního ukazatele výkonnosti zpracovatelského průmyslu ČR
3. místo: Ing. Martin Matějka: Využití stavově-prostorových modelů v demografii

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

Leveraging the Enterprise Architecture used in conjunction with the Business Process Management

Štěpán Alexa
Stepan.alexa@vse.cz
Doktorand oboru aplikovaná informatika
Školitel: Prof. Ing. Václav Řepa, (repa@vse.cze)

Abstract: Growing complexity of the enterprise ecosystem along with availability of various approaches that can constitute holistic information asset in an enterprise bring number of challenges. Also the digital industry has passed over past two decades through rapid evolution triggered both by availability of new technologies and business models. This trend implies that enterprises and organizations have to remain flexible when executing their business by maintaining the business and infrastructure alignment in constantly changing ecosystem. It has been widely recognized that both the enterprise architecture as well as process driven organization approaches provide the tool used by organizations to explain how business, resources and other elements within the organization are related to each other in way organizations do their work. In this paper it is discussed the role and associated value that the enterprise architecture as well as process driven approach represent when describing what constitutes an enterprise. At the same time it elaborates on constructs to be used for building the holistic layer of enterprise architecture that emphasizes process driven approach and in second phase to build metamodel of the enterprise on that basis.

Keywords: Enterprise Architecture, Process driven organization, Business process, WEA (Whole of Enterprise Architecture), Service

Introduction

An integrated approach that helps to express the way the elements of the enterprise are arranged and orchestrated to achieve business goals have been topic for scholars as well as enterprise domain stakeholders all the time. The gradual convergence of information technology with daily business expresses the increasing challenge in orchestrating business processes under the growing complexity of organization's infrastructure such as IT, human and capital resources and so forth (Silvius, 2013). However from practical standpoint the challenge for stakeholders of the organization has seemed to remain over past decades the same. That is, with concisely articulated and grounded information asset, to enable rational decision making about elements of the enterprise such as processes, infrastructure and people (Kluge, 2006). Number of concepts and approaches were introduced over time to address this topic (Giachetti, 2010). Ross, Weill and Robertson define "organizing logic" of the business processes and Information technology infrastructure with aim to execute requirements for standardization and integration based on operating model used by the organization (Ross et al, 2006). In their view enterprise architecture in fact fulfills the role of organizing logic by providing complete view on systems, resources, business processes and underlying technologies that is aligned with strategic objectives rather than only focusing on current organization's needs. The value of organizing logic from that perspective therefore consists in its contribution to achievement of organization's business goals as concluded by others, too (Osterwalder, 2004), (Repa, 2012), (Nightingale, 2012).

This paper therefore focuses on understanding the driving principles of process driven organization and enterprise architecture as well as way they are constituted and how they can be used in conjunction to express how enterprises do their work holistically. It aims to identify the baseline of the enterprise architecture metamodel in the enterprise in sense of the WEA model (Hrabě, 2014) and determine the way it should be used. The essence of problem statement consists in deriving the foundation of how to orchestrate the infrastructure and business processes in the organization towards achieving the business objective by using right essences from both approaches rather than focus on comparison of both approaches. The paper is organized as follows. Further section provides information about research design. Next section is dedicated to literature review while subsequent sections provide grounded baseline for identification of constructs and the principles for the proposed model that expresses the synergy of both approaches. That is followed by a section where proposed model is presented. Final section of the paper provides conclusion of the paper.

Research Design

The research presented in this paper focuses on scholar cognition at the intersection of Business Process Management and Enterprise Architecture approaches. Those, by definition, are large domains with number of overlaps. The objective of the research is to identify the federated model explaining arrangement of business process as well as other enterprise elements in the enterprise. Since the author of this paper has 15 years of practical experience with management both IT and business of the enterprise. Because of nature of researched topic and availability of qualitative heuristics the author used the research methodology on a basis of Kolbe's experimental cycle that enables linkage of the qualitative heuristics with modelling of theoretical concepts (Molnar et al, 2012), (Gala, 2014). Furthermore, this approach was used in successfully defended dissertation of similar topic by Gála (Gála, 2014). This experimental cycle uses qualitative research and it is used in way that firstly tests the relevance of principles and constructs, formulated on a basis of literature research and empirical observation with key stakeholders representing target enterprises and then they are used to build the model. This phase is realized using structured interview with the stakeholders. Second, the principles and constructs are adjusted according to the findings of the research. Third, the formulated concept is being validated through the multiple case study, within the area or multiple areas of the enterprise in scope of proposed model and method for its use. All the steps are executed iteratively. The research methodology is explained on Figure 1. This approach also enables to combine deduction and induction scholar methods.

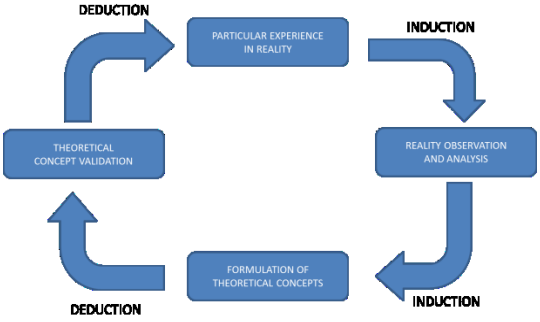


Figure 1. Kolbe's experimental cycle Source: (Molnár et al, 2012)

Literature Review

During the past two decades the enterprise architecture has become a well-recognized and matured discipline evolved to offer an in-depth view on the elements of the enterprise. Traditional taxonomic approach set by Zachman in 1997 describes enterprise architecture as a logical concept of the organization's elements mapping as well as relations between them and emphasizes "order and control mechanisms for the development of information systems" (Zachman, 1997). Later this definition was extended so that it has been enterprise architecture objective to provide holistic concept enabling to describe how companies perform their work (Ross et al., 2006), (Janssen, 2009). Enterprise architecture has been a model-based methodology in meaning that schematic description makes up the core of its approach. On the other hand Lankhorst mentions that enterprise architecture has been a concept and managerial discipline to provide multi-layered view on the organization through integrated view of business, applications, actors and underlying infrastructure (Lankhorst, 2009). According to Chiprianov the enterprise architecture approach consists of a set of models describing the structure and functions of an enterprise (Chiprianov, 2011). Lankhorst and Osterwalder express the positioning of enterprise architecture within an organization in way that enterprise architecture facilitates the one-way link between Business and IT strategy represented by particular goals on one hand and organization's daily operations on the other hand (Lankhorst, 2009), (Osterwalder, 2004). Edhah and Zafar are more specific and mention that enterprise architecture "translate the broader principles, capabilities, and the defined business objectives in the strategies into processes that allow the enterprise to realize the objectives" (Edhah and Zafar, 2016).

From the perspective of process based organization approach explained by Repa, Hammer and Davenport (Repa, 2011), (Hammer and Champy, 1993), (Davenport, 2005) , business can be expressed as a way processes of money earning logic, in case of an enterprise (Osterwalder, 2004) or value delivery in case of an organization in general (Murman et al., 2002), (Nightingale, 2009), are organized and executed to perform daily routine tasks. As the organizing logic tends to be executed to reflect business objectives therefore it can be regarded as a major contributor to the overall business strategy (Edhah and Zafar, 2016). Compared to Enterprise Architecture approach DeToro and McCabe foresee the Business Process Management as the new way of managing the enterprise (DeToro and McCabe, 1997), which is different when compared to traditional functional, hierarchical management. This standpoint is acknowledged by Pritchard and Armistead whose research expresses BPM "as a

'holistic' approach to the way in which organizations are managed" (Pritchard and Armistead, 1999). Scheer emphasizes the role of BPM as an information asset similarly to Van der Aalst who points out the descriptive capabilities of BPM to support lifecycle of the business processes (Scheer, 2012), (Aalst Van Der et al, 2003) . It is visible that both approaches address the same problem from different perspectives. Enterprise architecture by definition addresses the gap between "as-is" towards "to-be" state of the art architecture that is compliant with business objectives. BPM, in turn, aims to orchestrate the business processes and underlying infrastructure in way they support changes at any point of time.

At the high level the business architecture defines the value increments achieved through business processes. For instance, during the sales process adds value by converting an "opportunity" into a "deal". The Business Architecture in fact describes the business model of the enterprise. Osterwalder (Osterwalder, 2004) sees business model as a key tool to demonstrate the linkage between business goals, business re-quirements and its return on investment for key stakeholders. Federal Enterprise Ar-chitecture Framework (FEAF) state that its Business Reference Model (BRM) and associated elements "form a key part in delivering expected outcomes and business value to an organization" (USFCIOC, 2013). According to the MMABP (Repa, 2012) the Business Architecture should conform to business strategy and reflect related challenges. In that respect enterprise architecture as well as BPM provide methods and techniques used to produce architecture models representing different views on the enterprise.

Enterprise Architecture in the organization

The information about an organization covering „all objects of an organization without exception" is expressed by holistic layer of the enterprise architecture (Hrabě, 2014). Gála states that the ontological model of an enterprise includes all relevant actualities of an enterprise whilst „ontological model instance represents qualitative and quantitative characteristics of individual enterprise entities" (Gála et al., 2012). Model WEA (Whole of Enterprise Architecture) expresses the enterprise architecture layers determined by level of detail and depth of the enterprise actualities (Hrabě, 2014). WEA consists of following layers and sublayers:

- Enterprise Architecture
  - Architectonic vision
  - Holistic architecture models of and enterprise
  - Segment and domain architecture models of the enterprise
- Solution architecture
  - Domain and project solution architectures
- Solution design
  - Design and implementation of particular solution elements

Holistic layer of Enterprise architecture represents full view on the organization in all aspects of its being. It is expressed by metamodel of the organization that describes elements making up the organization and relations between them. According to Hrabě an organization can be in that respect viewed as a system:

- That executes the work
- That needs direction and rules
- That needs resources

This is depicted on figure 2 by orange, blue and green boxes while author mentions breakdown in every category.

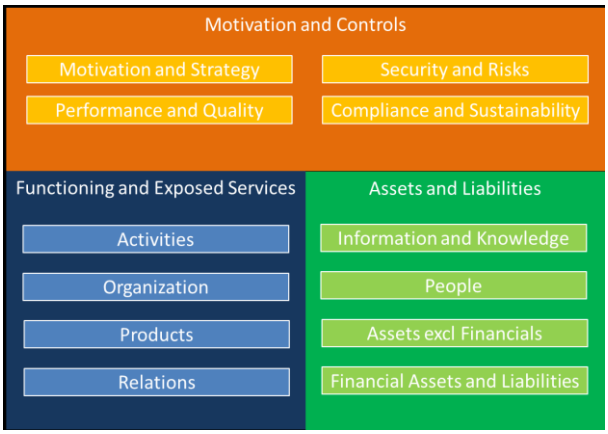


Figure 2 – Domain segregation of holistic architecture of an enterprise Source: (Hrabě, 2014)

The domain of work execution (blue) is represented by business processes heading towards creation of value added for process consumers (Řepa, 2011), (Hammer, 2007), (Davenport, 2005). As per theory of enterprise economy the resources of the organization can be expressed as a production factors and inputs to the business process. The relation between business process and resources that can be referred to as a part of organization's infrastructure will be elaborated further on in this paper. Enterprise architecture model presented by Gála defines construct „people in the enterprise architecture" (Gála, 2014) that includes interested stakeholders and architectonic team. Real enterprise architecture of existing organization possesses dynamic nature that is determined on a basis how has an organization been functioning over time. Ross, Weill and Robertson accent process approach in their definition of enterprise architecture as well (Ross et al., 2006). According to them enterprise architecture is an organizing logic of business processes and information technology infrastructure reflecting requirements for integration and standardization based on the operating model of an organization. It provides such view on business processes, systems and technologies so that individual activities create capabilities of an organization aligned with long term objectives rather than fulfill immediate goals only. According to the author the above findings therefore can be rephrased to constitute following recourses:

- Enterprise architecture as a reflection of real functioning of an organization concentrates all elements of an organization and internal as well as external relations between them. It can be expressed as a principle of completeness.
- Enterprise architecture provides transformation of strategic objectives into daily routine activities of the organization. It can be expressed as a principle of expediency.
- Enterprise architecture provides an information asset used in way that its purpose is fulfilled. It can be expressed as a principle of efficiency.
- There is a level of relation between enterprise architecture and operating model of an organization. It can be expressed as a principle of conformity
- The Enterprise architecture in the organization support changes in arrangement of enterprise elements as well as in structure of individual elements both as a consequence of internal and external influence over time. It can be expressed as a principle of flexibility.

Process driven organization

The first complete explanation of the idea of process management as a style of managing an organization was published in (Hammer and Champy, 1993). The major reason for the process-orientation in management is the vital need for the dynamics in the organization's behavior (Repa, 2011). That means that any process in the organization should be linked to the customer needs as directly as possible (OpenSoul, 2016). Hammer emphasizes that the customer need should be represented by concrete objective linked to a process (Hammer, 2007) while Davenport adds that it is desirable to express the objective using quantitative measure (Davenport, 2005). From that perspective the long term process capability to fulfill its objective under the influence of external (according to the process) factor changes can be expressed as a principle of sustainability. The general classification of processes in the organization distinguishes mainly between:

- Key processes, i.e. those processes in the organization which are linked directly to the customer, covering the whole business cycle from expression of the customer need to its satisfaction with the product / service. The rationale of their existence is achievement of the organization's goals (Voříšek, 2011).
- Supporting processes, which are linked to the customer indirectly - by means of key processes which they are supporting with particular products / services. Řepa mentions that the rationale of their existence is based on premise that „these supporting processes provide a service to other processes" in an organization (Řepa, 2012).

The abovementioned logic can be set as a principle of process existence within a process based organization.

Figure 3 shows different problem areas connected with the process based organization. All three viewpoints at the figure together address all substantial parts of the organization's life: content, technology, and people. Each particular point of view is characterized by typical questions which should be answered by the methodology in that field.

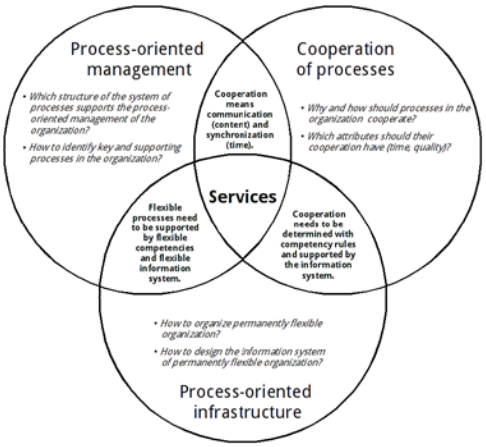


Fig. 3. Service as a common denominator of content, technical, and human aspects of the organization management Source: (Řepa, 2011)

Process arrangement of an organization is made up by structure of the processes – „Detail process model" (Řepa, 2012) and their mutual process relations as expressed by „Global processes model" (Řepa, 2012). It is also influenced by a business model that Osterwalder defines as a „a description of a value a company offers to one or several segments of customers and the architecture of the firm and its network of partners for creating, marketing and delivering this value and relationship capital, in order to generate profitable and sustainable revenue streams" (Osterwalder, 2004). A value the process brings to process consumers is key factor for design of the respective process in an organization and it represents common denominator of different approaches like Lean process framework (Nightingale 2009), Value based framework (Murman et al., 2002), Process based organization (Hammer, 2007), (Repa, 2011), (Davenport, 2005) - to name a few. Therefore it is expressed in this paper as a Principle of existence of added value to the business process.

Process expression of functioning of an organization

Key and supporting processes usually coexist in a determined mutual relation. That is visible from interpretation of figure 2. It says that processes may influence each other. The influence can be direct – when processes communicate with each other over a service – or indirect – when processes depend on each other without explicit communication between them. This relation can be set as:

- Structural alignment representing synchronization of a process instance execution with instances of other processes
- Content-wise alignment representing signalization of data between two processes
- This definition can be set as principle of relation existence between processes. In reference to expression of key processes, supporting processes and relation between them there is still question whether the structure and content of businesses processes in the organization can be conceptualized.

Davenport describes the unifying view on behavioral aspect of the organization as a „Process commoditization" (Davenport, 2005). Different view on the same topic offer process models APQC (American Productivity and Quality Center) and EFQM (European Foundation of Quality Management) (APQC, 2016), (EFQM, 2016).

Based on literature research the content of business process can be expressed into one of below categories:

- Business processes of leadership and strategic management providing strategy, planning and company administration
- Business processes of organization's operations providing value added to production by organization's stakeholders
- Business processes of organization's support providing infrastructure for above defined categories
- Structural view expresses the relation between process elements on a basis of object modeling. It determines that there exist mutually related levels of detail that can be used to describe activities of an organization. Some of the authors define the relation between views on different level of detail as a composition whilst others use hierarchical structure. Using the above explained findings the logic of the content-wise as well as the structure-wise expression of a business process can be derived. Basic artifact in sense of process driven approach that provides interface between the processes is a service that is integral part of global system of processes in the enterprise (Zimmermannová, 2011), (Řepa, 2011). For sake for further use this finding will referred as principle of cooperation between processes. When comparing recourses of cooperation between the processes as depicted on figure 2 with both structural

and content-wise viewpoints it is visible that same principle can be applied for analysis of business processes per se as well as analysis of relation between them.This idea is ilustrated in following table.

Table 1 – Structural and content viewpoint expressing unified process view on the organization
source: (author)

	Business Process	**Cooperation between BP**
Structural viewpoint	Expresses views on business process on different layers of abstraction	Expresses synchronization of process instance execution with other processes
Content viewpoint	Expresses views on business process from perspective of its content and value added it brings for its consumers	Expresses content alignment by signalization of data between the processes

Proposed model

Since topic of this paper puts together the Enterprise architecture and the process approach then it implies that the constitution of the model itself will be based on process driven organization. According to already described layered model of WEA – the proposed model of holistic layer of the enterprise refers to a second layer of WEA. It therefore reflects the requirement to provide holistic view on the enterprise in order to truly explain elements of the enterprise and way they are linked and orchestrated. The constructs are expressed in below table.

Table 1 – Constructs of metamodel of the enterprise ource: author

Element	**Purpose**	**Source**
Processes	The enterprise needs act to achieve purpose of its existence	Balanced Scorecard, Process driven organization, eTOM, Model WEA, APQC
Resources	The enterprise needs resources in order to be able to act	IFRS, Model WEA, APQC, Theory of enterprise economy
Motivators	Motivators determine the framework within which the enterprise can act	Balanced Scorecard, Enterprise Governance, WEA, Value based framework, Maturity models
Services	Providing of the value to cooperating processes or stakeholders	Process driven organization, eTOM, Model WEA
Stakeholders	Stakeholders represent entities influencing how the enterprise acts towards achieving the objectives that have been set and way they are achieved.	Value based framework, eTOM, Enterprise architecture as a service, Process driven organization
Relations between the elements	To explain how the enterprise elements influence each other	Enterprise ontology, Business ontology, Enterprise architecture as a service, Model WEA

There were set six general principles of Enterprise architecture expressing real functioning of the enterprise in section of this paper that elaborates on the Enterprise architecture. Therefore it is necessary that proposed model follows those principles.

- Principle of completeness (requiring that the enterprise architecture is expressed holistically)
- Principle of expediency (requiring that focus of the enterprise architecture is set right according to purpose of its existence)
- Principle of efficiency (requiring that use the enterprise architecture leads to achievement of the purpose of its existence)
- Principle of conformity (requiring that the enterprise architecture is compliant with set operating model of the enterprise and its long term objectives)
- Principle of flexibility (requiring that the enterprise architecture provides flexible capability of the enterprise to innovate its processes and mitigate the negative influence of events coming from the external ecosystem of the enterprise

Apart from abovementioned principles linked to the Enterprise architecture itself the proposed model also need to comply with principles explained in section of the paper dedicated to process based approach. They are summarized in below table.

Table 2 – Derived recourses for constitution of holistic EA layer source: (author)

Name	Expression of purpose
Principle of existence of relation between processes	Express how processes influence one another
Principle of existence of an individual process value added within the global system of processes in the enterprise	Express which value brings an individual process for stakeholder of other process
Principle of existence of a process in process driven organization	Express how a process contributes to overall achievement of set objectives of the enterprise
Principle of cooperation between processes	Express content-wise and technical-wise scope of a service as a result of cooperation between processes
Principle of process sustainability	Express long term capability of the process to fulfill set objective

According to the opinion of author of this paper the proposed holistic layer of the enterprise architecture it essentially conforms with WEA concept whilst proposed contents fits the logic the second layer of WEA.

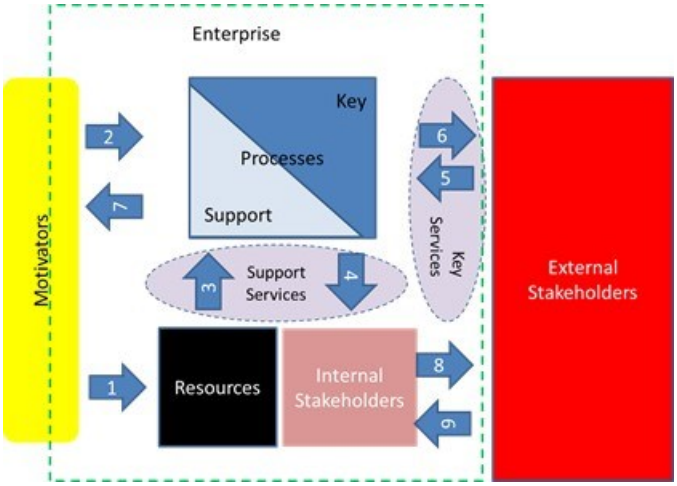


Figure 4 – Proposed holistic layer of the Enterprise architecture based on process approach
source: author

Figure 4 represents the proposed model and expresses the orchestration of constructs depicted in table 2. Activity of the enterprise is expressed by a system of its processes. In compliance with concept of process driven organization all the processes in the enterprise can be segregated into key and supporting processes. Key processes add value for customer that is monetized and therefore the enterprise can exist only if they are fulfilled. On the contrary the purpose of supporting processes is to provide a service to other processes. Processes influence one another and this is valid for interaction between processes of the same type and for interaction between processes of different type as well.

These features can be further elaborated in ontological model of the enterprise itself. The services in proposed model are derived by applying the principle of cooperation between processes. In the model the service serves as the interface between processes and also as the interface between processes and other enterprise elements. By using this logic model expresses that the supporting service enable:

- Use of enterprise resource within process
- Access of stakeholder to a process and vice versa

The example supporting service can be exposure of enterprise reports within financial controlling process. Another supporting service can represent access to IT helpdesk service for other employees of the enterprise within the process of incident management.

In turn, a business service as per model enable bi-directional mediation of value between business process and external (from viewpoint of modeled enterprise) stakeholder. The stakeholder can be represented by a physical person, legal entity or business process of external enterprise or organization. Business service thus can be the download of bank statement realized over application within process of customer care. Another business service can be convergent payment enabling single payment for multiple services of different providers realized over uniform payment channel within process of Billing. Arrangement of business and supporting services in the enterprise expresses the Service architecture of the enterprise.

Another element in the model is made up by enterprise resources. Arrangement of the enterprise resources demonstrates static character and the resources can be exhaustively classified although it is more challenging to express use of these resources in processes of the enterprise. The proposed model provides clue for that challenge because it leverages definition of service to facilitate access of a process or stakeholder to an enterprise resource.

The purpose of motivator in the model is to set prerequisites for execution of processes in the enterprise. The arrangement of motivators within the enterprise is represented by the Architecture of motivators. The motivators can be owned either by the enterprise itself or external ecosystem i.e. regulatory authority. The model expresses that the link between motivator and resource can be direct. The example of such motivator can be the policy that sets limits of provisions for „just in time“ manufacturing, capital type ratio per revenue and similar.

Description of the relations:

- Relation 1 – Motivators can influence resources and internal stakeholders directly. Apart from already mentioned examples other motivators can be ethical code, code of business conduct, incentive scheme and so forth.
- Relation 2 – Motivators directly influence processes. The example can be governance policy that sets the way processes need to be managed.
- Relation 3 – Supportive service provide exposure of resources to processes and stakeholders. The example is IT helpdesk service, ombudsperson service, meeting booking service and similar.
- Relation 4 – Condition of resources or stakeholders determine the way processes and service are designed and executed. New technology can lead to use of new service or deployment of new process. The example is deployment of new production line enabling sales on newly acquired market.
- Relation 5 – The enterprise realize value through exposed business service. Such a service can be payment gateway or order intake..
- Relation 6 – External stakeholder realizes value through business process. Example can be a service informing customer about availability of goods on stock or online validation ID of sold device.
- Relation 7 – Condition of business process (i.e. fulfillment of set objective) influences motivators.
- Relation 8 – Internal stakeholders (i.e. shareholder, employee) are in contact with partners and customers. In desired case it influences the architecture of motivators and condition of resources or processes.
- Relation 9 – same as previous point but reciprocally

Conclusion and further research

Availability and capability of new technologies, rapidly evolving business models and dynamically changing enterprise environment are factors that motivate organizations to constantly innovate their business processes while keeping them aligned with their business objectives. Consequently, the Enterprise Architecture as well as concept of process driven organization are approaches expected by the company executives to provide the highly valuable assets and management tools used to ensure continuous innovation and to master the content that fits proven practices of the industry and be reflective of its changes. In this paper, it was described the problem statement leading to identification of key constructs of the Holistic layer of the Enterprise architecture based on process approach and rationale behind the presented concept. It is author’s conviction that the proposed model on a basis of the principles of process driven organization used in conjunction with enterprise architecture’s recourses as set in this paper help to bridge the gap in focus of these two approaches by emphasizing on expression of the way organizations are doing their business. Therefore the proposed model addressed though not solved fully the problem statement mentioned at the beginning of the paper. To fully solve the problem an ongoing research should be focused on two areas, to derive ontological model of the enterprise on a basis of proposed model in this paper and to propose a method used by the enterprise that serves as a guideline for model use in way it produces benefit for the enterprise.

Literature

Aalst Van Der W.M.P., Hofstede Ter A.H.M., Weske M. (2003). Business Process Management: A Survey, Lecture Notes in Computer Science, Vol. 2678, Springer

Chiprianov, Vanea, et al. Telecommunications Service Creation: Towards Extensions for Enterprise Architecture Modeling Languages. In: ICSOFT (1). 2011. p. 23-28.

Davenport T., 2005 The Coming Commoditization of Processes, Harward Business School Press, Boston.

DeToro, I. & McCabe, T. (1997). How to Stay Flexible and Elude Fads. Quality Progress, vol. 30, no. 3, pp. 55-60

Edhah, B. S., & Zafar, A. (2016). Enterprise Architecture: A Tool for IS Strategy Formulation. continuity, 12, 13.

Gála L., Buchalcevová A., Jandoš J. (2012); Enterprise Architecture. ISBN 978-80-904661-6-6

Gala, L. Enterprise Architecture as a Service. Praha, 2014. Dissertation Thesis. University of Economics in Prague

Giachetti, Ronald E. 2010. Design of enterprise systems: theory, architecture, and methods. Boca Raton, Fla.: CRC Press, c2010, xvii, 429 p. ISBN 14-398-1823-1.

Hammer, M., Champy, J.: "Re-engineering the Corporation: A Manifesto for Business Revolution", Harper Business, New York, NY, 1993.

Hammer, M. 2007. The Process Audit, Harvard Business Review

Hrabě, P.: Koncepce podnikové architektury pro reformu veřejné správy ČR, disertační práce, VŠE-FIS, Praha, 2014

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

Kluge C., Dietzch A., Rosemann M. (2006) How to realize corporate value from enterprise architecture, Association for Information Systems : European Conference on Information Systems 2006 Proceeding

Lankhorst M., Enterprise Architecture at Work, Modelling, Communication and Analysis, 2009 ISBN 978-3-642-01309-6, Springer-Verlag Berlin Heidelberg

Molnar, Z., Mildeova S., Rezankova H., Brixi R., Kalina J. 2012. Pokročilé metody vědecké práce. Praha: Profess Consulting, 2012, 170 s. ISBN 978-80-7259-064-3.

Murman, E., Allen, T., Bozdogan, K., Cutcher-Gershenfeld, J., McManus, H., Nightingale, D. & Warmkessel, J. (2002). Lean enterprise value. Insights From Mit's Lean.

Nightingale D.J. (2009), Principles of Enterprise Systems, MIT Lean Advancement Initiative, Second International Engineering Systems Symposium

Nightingale D. J. (2012). Roadmap for Enterprise Transformation, IIE Enterprise Transformation Conference 2012, Atlanta, Georgia

OpenSoul Project: http://opensoul.panrepa.org, 2000 - 2016.

Osterwalder A., The Business Model Ontology – a proposition in a design science approach , Phd Thesis , UNIVERSITE DE LAUSANNE, 2004, http://www.hec.unil.ch/aosterwa/phd/osterwalder_phd_bm_ontology.pdf

Pritchard, J.-P., & Armistead, C. (1999). Business process management - lessons from European business. Business Process Management Journal, 5(1), 10-32.

Řepa, V.: "Information Modelling of Organizations". Bruckner, Prague, 2012. ISBN 978-80-904661-3-5.

Řepa, V.: "Role of the Concept of Service in Business Process Management". In: Information Systems Development. New York: Springer, LNCS, pp. 623–634, 2011, ISBN 978-1-4419-9790-6.

Ross J. W., Weill P, Robertson, D; Enterprise Architecture as Strategy: Creating a Foundation for Business Execution, Harvard Business Review Press (2006), ISBN-10: 1591398398

Scheer A.W., 2012, Business Process Engineering: Reference Models for Industrial Enterprises, Springer-Verlag, ISBN 13-978-3-642-79144-4

SILVIUS, A. J. G., et al. Business and IT alignment in context. 2013. Dissertation work, University of Utrecht

U.S. Federal CIO Council, 2013, Federal Enterprise Architecture Framework, https://www.whitehouse.gov/omb/e-gov/fea [accessed 3.4.2016]

Zachman, J. A. (1997). Enterprise architecture: The issue of the century. Database Programming and Design, 10, 44-53.

Zimmermannová, M. & Řepa, V., 2011. Vztah procesního řízení a Enterprise Architecture. Systémová integrace, 18(4), pp. 7 - 21.

JEL classification: M15

Shrnutí

Použití procesního přístupu při popisu podnikové architektury v organizaci

Článek pojednává o tématu zpracování celostní vrstvy podnikové architektury s využitím Hraběteho konceptu WEA a zaměřuje se na aspekty procesně řízené organizace při odvození konstruktů zmíněné vrstvy. Obsah tohoto článku se váže na předmět autorovy disertační práce a představuje model celostní vrstvy podnikové architektury na základě konceptu WEA.

Klíčová slova: Podniková architektura, Procesně řízená organizace, Podnikový proces, WEA (Whole of Enterprise Architecture), Služba.

Using the International Patent Classification System in Competitive Technical Intelligence

Jan Černý
cerj07@vse.cz
Doktorand oboru aplikovaná informatika
Školitel: prof. Ing. Zdeněk Molnár, CSc.

Abstract: Patent information is widely known as a valuable source for Competitive Technical Intelligence focused on competitor activity analyses in the technological field. Moreover, patent classification schemes provide us with a unique possibility to discover industry trend directions and key players in a particular field. This paper will describe the structure of the International Patent Classification together with search methods that can lead us to required analytical outputs. The author has used the Global Patent Index bibliographic database to present search syntaxes with classification codes and analysis methods. For the purposes of this paper, we have chosen to explore the key inventors in the field of forms of protection against or prevention of injuries to drivers, passengers or pedestrians in the event of accidents or other traffic hazards for personal car and motorcycle and truck industry purposes.

Keywords: Competitive Technical Intelligence, International Patent Classification, IPC, competitor analysis, traffic safety

Introduction

Competitive Technical Intelligence (CTI) as a subset of Competitive Intelligence can be defined as collecting, analyzing and disseminating data, information and knowledge from an external company environment where technology is a common aspect (Coburn 1999; Porter et al. 2007). The main tasks of CI focus on the following technical intelligence topics:

- Strategic planning of R&D
- Early Technical Warnings (ETW)
- Key players
- Commercialization

We can effectively use patent classifications for gathering for all four topics. Strategy planning information needs focus on industry trends, commercialization departments monitor the process of the ageing of each of the inventions, ETWs bring the information about new classification notations as the brand new technology field information signals, and last but not least we can extensively analyze the key player activities in detail. (Porter et al. 2007; Pičman 2009)

For the purposes of this study we have chosen to analyze International Patent Classification (IPC). This will then be used in the Global Patent Index bibliographic database that allows us to build search syntax by using this classification scheme. We will work with the specified field of innovations – protection and prevention during possible negative traffic events.

The paper will follow this structure:

- International Patent Classification characteristics
- Using Patent Systems for Key Player and Industry Trends Search
- Results
- Conclusions

International Patent Classification

The first version of the IPC went public in 1968, and after the Strasbourg Agreement, which came into force in 1975, it became the common classification system for patent documents of inventions including patent applications, utility models and utility certificates.(WIPO 2015)

The IPC has several functions (WIPO 2015):

1. The IPC is a powerful tool to determine the novelty and evaluate the inventive steps and non-obviousness that affect the patentability of each invention. Furthermore, it is considered as a powerful search tool for retrieval of the required scope of patent documents.
2. Facilitates access to the technology and legal information contained in patent documents.
3. A must-have tool for conducting a state-of-the-art search.
4. The basis for generating information-rich statistical analyses.

Structure of the IPC

In this part we present the structure of IPC as a complex and structured classification through the levels in its hierarchy (WIPO 2016b), (WIPO 2015).

IPC Sections, Class, Subclass and Groups

The IPC is divided into eight sections that represent the highest level of a hierarchy. A section is put together with **a)** the Section symbol thought the letters A – H, **b)** the Section titles which determines the broadest field of inventions, and **c)** subsections, which are the inner parts of sections without a specific code:

- **A** HUMAN NECESSITIES
- **B** PERFORMING OPERATIONS; TRANSPORTING
- **C** CHEMISTRY; METALLURGY
- **D** TEXTILES; PAPER
- **E** FIXED CONSTRUCTIONS
- **F** MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING
- **G** PHYSICS
- **H** ELECTRICITY

Subsection examples can be shown inside the B section:

B PERFORMING OPERATIONS; TRANSPORTING

- SEPARATING; MIXING
- SHAPING
- PRINTING
- TRANSPORTING
- MICRO-STRUCTURAL TECHNOLOGY; NANO-TECHNOLOGY

The sections are further divided into Classes, and these are represented by a two-digit code and Class titles that show the scope, and in some cases by a Class Index that contains information about the update.

B60 VEHICLES IN GENERAL

The subclasses represent a further structured class and they are the first specifying element of a given technical field.

B60R VEHICLES, VEHICLE FITTINGS, OR VEHICLE PARTS, NOT OTHERWISE PROVIDED FOR (fire prevention, containment or extinguishing specially adapted for vehicles A62C 3/07)

The IPC Groups works as the extension of the Subclass and can be distinguished as

a) Main group with a one to three-digit number, stroke and the number 00 together with title of the Main Group. For example:
B60R 22/00 Safety belts or body harnesses in vehicles

b) The IPC Subgroup is represented by two or three digit numbers together with the Subgroup title. For example:

B60R 22/00		Safety belts or body harnesses in vehicles
B60R 22/34	·	Belt retractors, e.g. reels
B60R 22/343	· ·	with electrically actuated locking means [2006.01]
B60R 22/347	· ·	with means for permanently locking the retractor during the wearing of the belt
B60R 22/35	· · ·	the locking means being automatically actuated [2006.01]
B60R 22/353	· · · ·	in response to belt movement when a wearer applies the belt [2006.01]
B60R 22/357	· · · ·	in response to fastening of the belt buckle [2006.01]

We also need to discuss the special hierarchy dot system in each of the Subgroups that precedes the Title of the Subgroup. The Subgroup with one dot is the head of following subgroups with two to maximum six dots. Consider the Subgroup **B60R 22/34 (Belt retractors, e.g. reels)** and the Subgroup 22/343 **with electrically actuated locking means**. The second Subgroup here should be read as **Belt retractors with electrically**

actuated locking means. The important fact is that we must consider determining a hierarchy level by dots, not by Subgroup number.

IPC complete notation looks as follows:

B Section

60 Class

R Subclass

22/00 Main Group

22/34 Subgroup

IPC Search Tools

The new IPC publication platform containing the IPC (2017.01) is still under development. Some of the functionalities, for example: Definitions, Catchword Index and advanced search for Terms and Cross references[1] will be implemented in the new publication platform during 2017. For these reasons we have decided to describe all the key functionalities in the former platform which is still fully accessible on (WIPO 2016a).

The IPC main framework (See Fig 1) provides many options for collecting information about invention and innovation developments, and for using it for further research in different patent search systems (WIPO 2016a).

Figure 1 The IPC main framework

The framework consists of:

- Current Symbol Search
- Terms Search
- Cross-references search
- Stats assistance
- Text categorization search (IPCCAT)
- Scheme Browsing
- Compilation of Amendments
- Catchwords Browsing
- IPC Green Inventory

Current Symbol Search

If we know any part of the IPC notation we can use the symbol search. The system will give us a highlighted result in an operational window. The input permits the following structure:

B60R 21/232 or B60R21/232 (no space)

[1] These tools and their functions are explained later.

Terms search

We can perform a basic term search in a different part of the IPC. On the one hand, there are several syntax-building limitations that a searcher faces during a search process in the IPC Term search. On the other hand, this framework provides a unique possibility to find the connections between IPC codes and therefore to narrow a search.(WIPO 2016b)

If we use IPC Terms search we must consider that we will be able to tune the search in any patent system we choose.

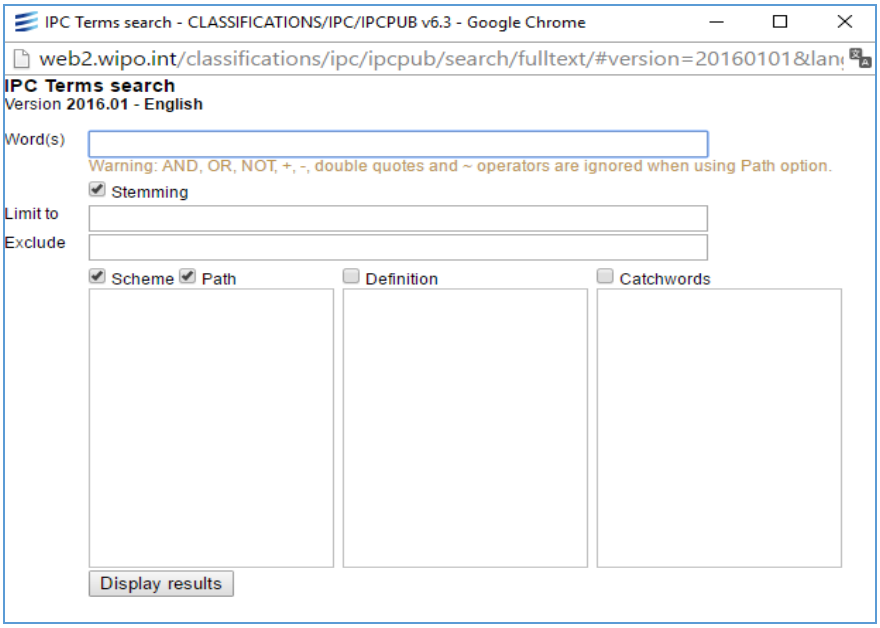


Figure 2 IPC Terms search framework

The search frameworks consists of (WIPO 2016b) :

- Search fields
  - Word(s)
    This search field has specific search features that differ slightly from any other secondary search system.
    - Special characters (e.g. %, /, (), [],…) are ignored during searches. The same applies for stopwords (and, or, not, such,... etc.). When using these words or characters the system gives us the same result set as without them.
    - Wildcards can be used when the stemming function is switched off. An asterisk (*) replaces any number or no characters. The question mark (?) is used when one or no character needs to be replaced. We can use more than one of these characters.
    - Double quotes represent a phrase (e.g. "safety belt"). Terms found can be separated by special characters or stopwords.
    - A tilde "~" implies the distance between two terms, e.g. "safety belt"~10. In this case the terms “safety” is being searched by 10 words from the term belt.
    - The terms that are preceded by the character "+" must be included in the result text. For example +safety +vehicle.
    - Terms that are preceded by the character "-" must be excluded from the result text. For example +safety +vehicle -belt.
    - Boolean operators are also possible to use, however they must be written all caps (AND, OR, NOT) otherwise they are considered as stopwords. When the Path option is switched on, these operators are ignored, except for wildcards and fuzzy search cases.
    - Fuzzy search uses tilde with a numeric value between 0.0.-1.0 to find similar words. It works with Levenshtein Distance, or Edit Distance algorithm.
  - Limit to
    When using this field, the terms will be searched only in the specific places of IPC. This is an important tool for narrowing the IPC Terms search.
  - Exclude
    In contrast to the Limit to field, we use this field to exclude IPC entities.

- Result windows
  - Scheme
    The scheme shows us the IPC places that contain the requested terms.
  - Path
    This option helps us to identify the hierarchical level of results that contain the terms. The first is the highest IPC entity connected to the requested terms and it is further structured in lower levels. However, the Path option does not work with single terms and searching operators.
  - Definition
    Very helpful tool that describes given IPC codes with their scope and relevant references to other IPC places.
  - Catchwords
    The main descriptive words with the references to relevant IPC codes.

Our information need for this paper is to define the trends in driver, passenger and pedestrian safety for the automotive industry purposes.

To define the automotive industry we have chosen the IPC B section that covers the transport industry and, in addition, class B60 that consists of vehicles in general. We can limit IPC search terms in this class and then conduct the search with the keywords safety, protection or prevention with the Path option. As a result we get the IPC main group B60R 21/00 that is described as *Arrangements or fittings on vehicles for protecting or preventing injuries to occupants or pedestrians in case of accidents or other traffic risks*.

Now we have a basic IPC code that we can follow to research our topic in different patent systems and we can consider it as the IPC code that should be taken as the main descriptive code on patent applications.

Cross-reference Search

Another strong search tool for the IPC framework that provides a unique function to make a cross analysis of the main IPC code for our information needs (WIPO 2016b) (see Fig 3).

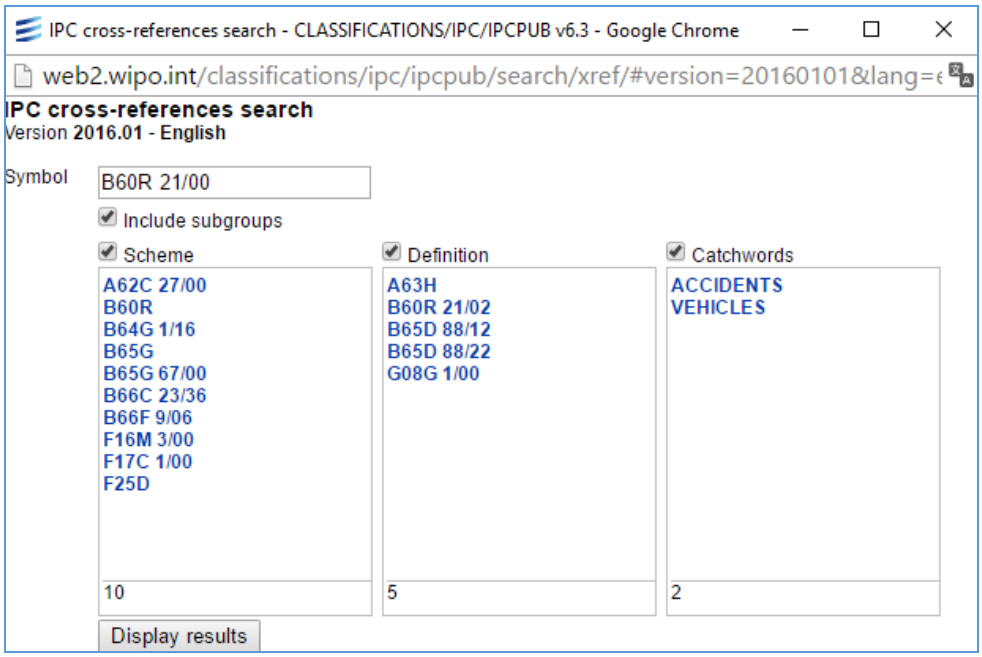


Figure 3 IPC cross-references search framework

We put the symbol B60R 21/00 to get information about references to our main topic IPC symbol. This search leads to these codes:

- A62C 27/00
- B60R
- B64G 1/16
- B65G
- B65G 67/00
- B66C 23/36
- B66F 9/06

- F16M 3/00
- F17C 1/00
- F25D

The code set should be used in a cross-search in a given patent system together with the keywords relating to the main topic of this paper.

STATS Assistance

STATS Assistance servers as the controlling search framework based on the same search features as the IPC terms search. If we put the specific terms in the Word(s) field we get the relevancy rate for the given IPC code. Moreover, it provides other suggestions for codes and we can check the relevancy and choose whether to consist them into our patent search.(WIPO 2016b)

Text categorization

This option is determined for inventors, patent agents, but also for searchers. The first two groups of users use it for application filling purposes to get the most relevant IPC code. Searchers use it for narrowing the initial search. For example, when they study a specific competitor's application, they can use the abstract, claim or description text and get other relevant IPC codes for the cross search (WIPO 2016b) (see Fig 4).

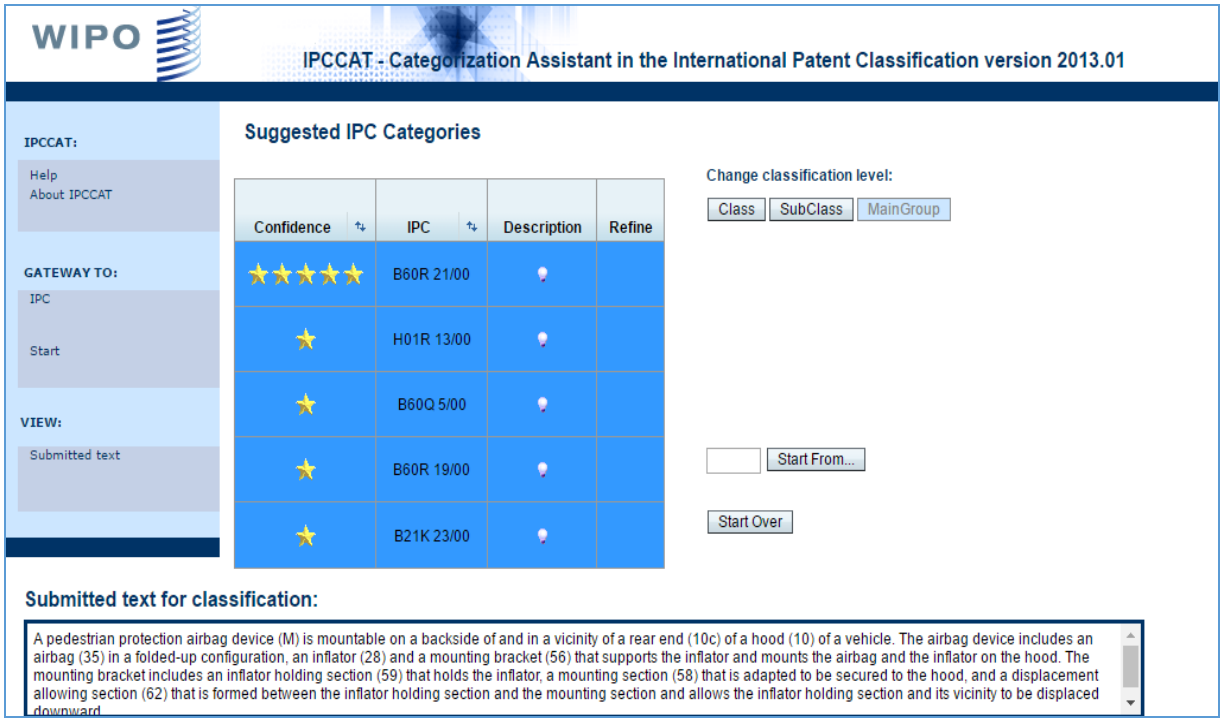


Figure 4 Text categorization framework. A pedestrian protection airbag device example.

Revised Concordance List

The IPC provides information about the changing features of different technological scopes through the Revised Concordance List (RCL). We monitor the change of the IPC Symbol from previous to the actual version of the IPC of specific scope.(WIPO 2016b, 2013)

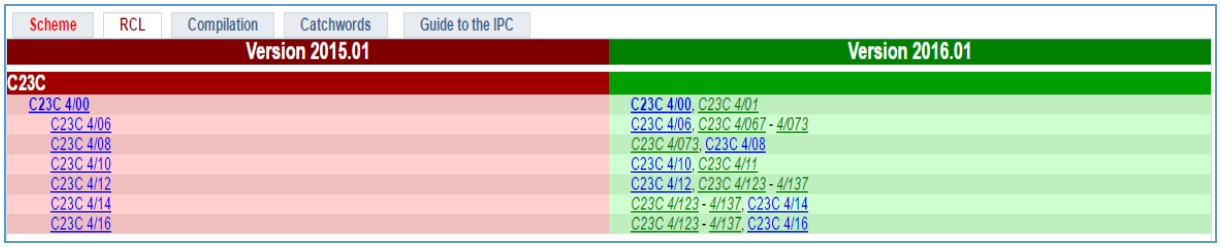
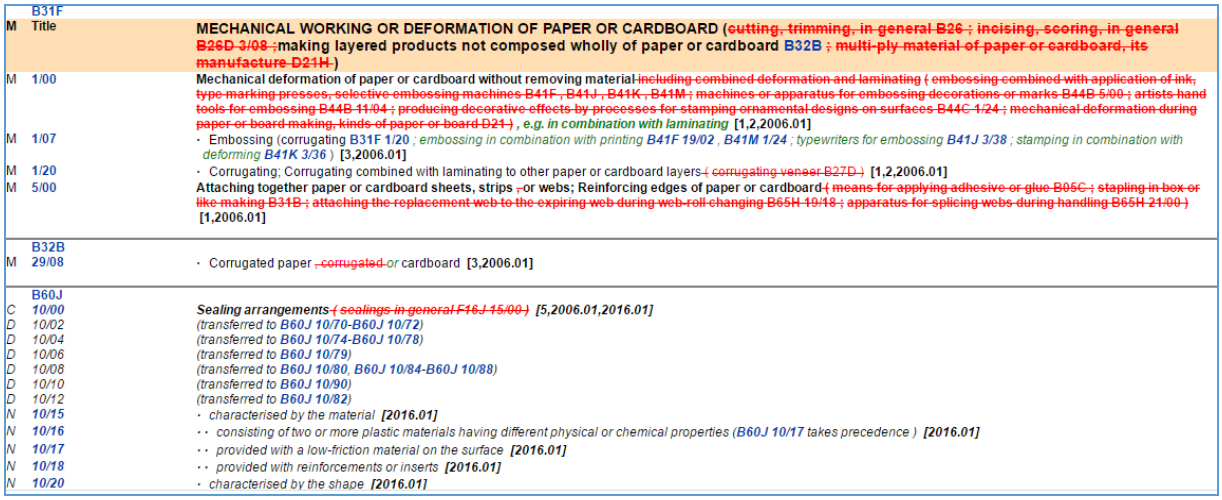


Figure 5 Revised Concordance List framework

Compilation of amendments

We can monitor potentially new emerging technologies through the Compilation of amendments section.



	B31F	
M	Title	MECHANICAL WORKING OR DEFORMATION OF PAPER OR CARDBOARD (~~cutting, trimming, in general B26 ; incising, scoring, in general B26D 3/08 ;~~making layered products not composed wholly of paper or cardboard B32B ; ~~multi-ply material of paper or cardboard, its manufacture D21H~~)
M	1/00	Mechanical deformation of paper or cardboard without removing material ~~including combined deformation and laminating (embossing combined with application of ink, type marking presses, selective embossing machines B41F , B41J , B41K , B41M ; machines or apparatus for embossing decorations or marks B44B 5/00 ; artists hand tools for embossing B44B 11/04 ; producing decorative effects by processes for stamping ornamental designs on surfaces B44C 1/24 ; mechanical deformation during paper or board making, kinds of paper or board D21) , e.g. in combination with laminating~~ [1,2,2006.01]
M	1/07	· Embossing (corrugating B31F 1/20 ; embossing in combination with printing B41F 19/02 , B41M 1/24 ; typewriters for embossing B41J 3/38 ; stamping in combination with deforming B41K 3/36) [3,2006.01]
M	1/20	· Corrugating; Corrugating combined with laminating to other paper or cardboard layers ~~(corrugating veneer B27D)~~ [1,2,2006.01]
M	5/00	Attaching together paper or cardboard sheets, strips ~~,~~ or webs; Reinforcing edges of paper or cardboard ~~(means for applying adhesive or glue B05C ; stapling in box or like making B31B ; attaching the replacement web to the expiring web during web-roll changing B65H 19/18 ; apparatus for splicing webs during handling B65H 21/00)~~ [1,2,2006.01]
	B32B	
M	29/08	· Corrugated paper ~~, corrugated~~ or cardboard [3,2006.01]
	B60J	
C	10/00	Sealing arrangements ~~(sealings in general F16J 15/00)~~ [5,2006.01,2016.01]
D	10/02	(transferred to B60J 10/70-B60J 10/72)
D	10/04	(transferred to B60J 10/74-B60J 10/78)
D	10/06	(transferred to B60J 10/79)
D	10/08	(transferred to B60J 10/80, B60J 10/84-B60J 10/88)
D	10/10	(transferred to B60J 10/90)
D	10/12	(transferred to B60J 10/82)
N	10/15	· characterised by the material [2016.01]
N	10/16	· · consisting of two or more plastic materials having different physical or chemical properties (B60J 10/17 takes precedence) [2016.01]
N	10/17	· · provided with a low-friction material on the surface [2016.01]
N	10/18	· · provided with reinforcements or inserts [2016.01]
N	10/20	· characterised by the shape [2016.01]

Figure 6 Compilation of amendments framework example

Compilation covers all the changes between old and updated version of IPC. Each of the modification has been published with following features (WIPO 2016b):

1. Type of modification represented by letter code:
   - ✓ **N** New entry (terms describing the novelty)
   - ✓ **M** Modification (text or data change, but not the technological scope)
   - ✓ **D** Deleted (deletion of the entry)
   - ✓ **C** File scope that has changed
2. **U** Unchanged entries
   Location in the IPC
3. Modification of the previous version of the IPC

Using Patent Systems for Key Player and Industry Trends Search

According to the World Health Organization (WHO) the number of traffic accidents have stagnated since 2007 relative to a 4% increase in world population and a 16% increase in motorization. In 2013 there were 1.2 million traffic accidents with over 50 million injuries. Moreover, traffic accidents play a major role in the mortality rate of young people aged between 15-29 years.(WHO 2015)

Nevertheless, the plateau in the number of accidents shows that the adopted measures in the developed countries have saved lives and the process of new measures could be considered to be efficient, e.g. Europe has the lowest mortality rate – 9.3. per 100 000 population (year). By contrast, the poorest statistics are found in the African region, where the mortality rate is 26.6 per 100 000 population. (WHO 2015)

The situation is even worse when we analyze pedestrian, cyclist and motorcyclist death rates. These make up nearly half of all deaths on the roads in the African region. Another negative number comes from the Americas, where motorcycle accidents rose from 15% to 20% of total road traffic deaths if we compare 2010 and 2013.(WHO 2015) The United Nations has brought a new initiative called the 2030 Agenda for Sustainable Development (United Nations 2015) and its two parts relates to improvement of road safety.

Two targets in the 2030 Agenda are:

1) **In part 3.6.** By 2020, halve the number of global deaths and injuries from road traffic accidents.
2) **In part 11.2.** By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations: women, children the disabled and the elderly.

Safe vehicles play a critical role in reducing the death and injury rate in terms of road safety and in past years the situation has led to implementing more safe elements into cars as the consumer demand for safer cars has increased. Moreover, the process of considering these elements as obligatory parts in vehicles has only just begun in some countries. Nevertheless, vehicles sold in 80% of countries do not meet the safety standards.(WHO 2015)

In view of these circumstances we can expect that the competitive environment in the field of vehicle and road safety elements will become even stronger and the demand for patent protection services and patent information services will be even stronger.

The purpose of this chapter is to design a patent search syntax with IPC codes relating to road safety and to reveal the latest technological developments in this field.

Before we build a state-of-the-art search syntax with the IPC codes related to traffic safety elements we also need to establish a keyword set describing a given search problem (see Fig. 7).

Entity	**Global Patent Index**
Safety field	prevention, protection, injury, death, safety, road safety, guard, accident
Traffic participants	vehicle, car, motorcycle, motorcyclist, motorbike, motorbiker pedestrian, passenger, driver, truck, lorry, cyclist, biker
IPC codes	B60R 21/00, A62C 27/00, B64G 1/16, B65G, B65G 67/00, B66C 23/36, B66F 9/06, F16M 3/00, F17C 1/00, F25D4.

Figure 7 A content analysis of the search problem

We have chosen the Global Patent Index (GPI) database (EPO 2014) for our example where we will perform a search with a specific syntax. The GPI as a bibliographic database provides unique insights into the more than 92 million patent documents.

The main characteristics of this source are:

Entity	**Global Patent Index**
Data Coverage	Only bibliographic records
Number of documents	101 034 000[2]
Updates	Weekly
Geographical Data Coverage	Countries under PCT (148) + Euroasian Patent Organization + European Patent Office + African Regional Intellectual Property Organization + 92 national offices[3]
Search modes	Expert
Advanced search syntax tools	Boolean, Proximity, Wildcards
Classification systems	IPC, CPC, US Patent Classification System, JP Classification (FI), JP Classification (F-Terms)
Legal status	INPADOC legal status codes
Patent family information	Simple patent family, INPADOC patent family
Data analysis outputs	• **Simple statistics** IPC, CPC, FI, F-Terms, Applicant, Cited Applicant, Inventor, Publishing Office • **Cross-reference** IPC, CPC, Inventor, Applicant, Date of Filling, Date of Publication, Date of Priority
Data visualization	Simple statistics graphs, two dimensional graphs
Data export	Full records in RTF, PDF, or XML format, Result lists in PDF, XML, XLS or HTML, Statistics and Cross-reference graphs in JSON, CSV or PDF

Figure 8. The Global Patent Index facts

As we have compared the whole functionality of the GPI framework to the WIPO Patentscope in another paper, here we will only specify the query syntax possibilities. The GPI works with these search tools:

- Boolean operators AND, OR, NOT, ANDNOT
- The new operator WITH brings more accurate results when using classification, legal status, applicant, inventor, priority and citation fields.
- The proximity operator /Xw implies the distance between two terms regardless of the order where X is the number. If use +Xw entry, the two terms must be in the same order within a given distance.
- Arithmetic operators
    - = equal to value in the field code
    - >=, <=, <, > are mainly used in date field codes.

[2] By 27th of January 2017
[3] By 27th of January 2017

- Wildcards
  - * (asterisk) implies zero or more characters and can be used in any part of the term.
  - # (hash) implies for one mandatory character
  - ? (question mark) implies for one or zero character
  - Wildcards cannot be used in phrase (double quotes)
- Brackets [] are used for a date range.
- Parentheses () affect the order of how the logic sets will be searched.

Latest developments search in the road safety field

Firstly, we perform a search with all the keywords separated into the sets given above in Fig. 7 We suggest this syntax:

WORD = ((vehicle? OR car? OR motorcyc* OR motorbike* OR pedestrian? OR passenger? OR driver? OR truck? OR lorr??? OR cyclist? OR biker?) AND (prevent* OR protect* OR injur* OR death? OR dead?? OR safe?? OR "road safety" OR guard? OR accident?))

We have over 600k results.

We must definitely narrow the search with a date limit. Searching for the latest developments mean setting the date range of patent application publications between 2014 and 2016.

WORD = ((vehicle? OR car? OR motorcyc* OR motorbike* OR pedestrian? OR passenger? OR driver? OR truck? OR lorr??? OR cyclist? OR biker?) AND (prevent* OR protect* OR injur* OR death? OR dead?? OR safe?? OR "road safety" OR guard? OR accident?)) AND PUD [2014,2016]

The result set now contains over 141k documents

The importance of IPC usage comes at this moment when we add it to the existing search syntax as a significant specifying factor.

WORD = ((vehicle? OR car? OR motorcyc* OR motorbike* OR pedestrian? OR passenger? OR driver? OR truck? OR lorr??? OR cyclist? OR biker?) AND (prevent* OR protect* OR injur* OR death? OR dead?? OR safe?? OR "road safety" OR guard? OR accident?)) AND PUD [2014,2016] AND IPC = ("B60R 21/" OR "A62C 27/" OR "B64G 1/16" OR "B65G" OR "B65G 67/" OR "B66C 23/36" OR "B66F 9/06" OR "F16M 3/" OR "F17C 1/" OR "F25D")

This set with IPC codes consists of 7 364 key patent families. We have chosen the first ten companies and will try to use different forms of their names in another search. The reason of this content analysis step is to get the most accurate number of patent applications for a particular company.

Results

The following list (See Fig. 9) represents the main innovative companies in the field of road safety between the years 2014 and 2016 (April) based on our searches.

Company Name	Number of patent applications (2014 – 2016)
Toyota	244
Hyundai	227
Autoliv	185
Bosch	120
Denso	120
Daimler	84
Toyoda Gosei	84
Ford	83
Fuji	78
Honda	67
Baic	60

Figure 9 The key innovating companies in the field of road safety

We can go further and analyze what kind of specific technology a particular company has been developing. The GPI data based upon our last search gave us the IPC codes analyzed in the following table (See Fig. 10).

IPC Code	Number of documents	Description of invention
B60R 21/02	569	Occupant safety arrangements or fitting
B60R 21/01	332	Electrical circuits for triggering safety arrangements in case of vehicle accidents or impending vehicle accident.
G08G 1/16	277	Anti-collision system
B60R 21/34	251	Inventions protecting non-occupants of a vehicle, e.g. pedestrians
B60R 21/0134	200	Arrangements or fittings on vehicles for protecting or preventing injuries to occupants or pedestrians in case of accidents or other traffic risks responsive to imminent contact with an obstacle
B60R 21/013	198	Electrical circuits for triggering safety arrangements in case of vehicle accidents or impending vehicle accidents including means for detecting collisions, impending collisions or roll-over.
B60R 21/015	171	Electrical circuits for triggering safety arrangements in case of vehicle accidents or impending vehicle accidents including means for detecting the presence or position of passengers, passenger seats or child seats, e.g. for disabling triggering.
B60R 21/207	152	Arrangements for storing inflatable members in their non-use or deflated condition; Arrangement or mounting of air bag modules or components in vehicle seats
B60R 21/0136	149	Electrical circuits for triggering safety arrangements in case of vehicle accidents or impending vehicle accidents including means for detecting collisions, impending collisions or roll-over responsive to actual contact with an obstacle
B60R21/231	122	Inflatable members characterised by their shape, construction or spatial configuration.

Figure 10 Industry trends technology analysis through the IPC codes

Subsequently, the GPI provides us with the possibility of characterizing the specific inventions of each searched company according to a given IPC code. We have chosen Toyota and the IPC code B60R 21/02. Our syntax for this example should be built as follows:

WORD = ((vehicle OR car OR motorcyc* OR motorbike* OR pedestrian? OR passenger? OR driver? OR truck? OR lorr??? OR cyclist? OR biker?) AND (prevent* OR protect* OR injur* OR death? OR dead?? OR safe?? OR "road safety" OR guard? OR accident?)) AND PUD [2014,2016] AND IPC = ("B60R 21/02") AND APP = (toyota*)

Seven documents from Toyota describe the company's latest developments in occupant safety arrangements or fitting (B60R 21/02). These are:

1. *Rear Seat Occupant Protection Device Of Vehicle*
2. *Vehicle Body Structure*
3. *Occupant Protection Device For Case Of Vehicle Side Collision*
4. *Automobile Front Pillar Bottom Structure*
5. *Right-Foot Protection Structure*
6. *Impact Absorption Structure*
7. *Glove Compartment Box With Tear Seam*
8. *Vehicle Side Door Structure*
9. *Vehicle Seat*

Conclusions

The International Patent Classification System is a global, regularly updated scheme that provides unique insight into different technology fields. In the first part of this paper, we discuss the structure, function and usage of this system and its framework to together with search possibilities of particular IPC codes for specific technology as the initial preparation for patent searches.

The second part describes an example of patent search for Competitive Technical Intelligence secondary information research activity purposes. This part also analyzes the search process for the given topic of forms of protection against, or prevention of injuries to drivers, passengers or pedestrians in the event of accidents or other traffic hazards for automotive industry purposes. We have focused on key player and industry trend analysis.

In this paper we chose the Global Patent Index database as a testing environment for two reasons. The first is that the GPI provides a significant number of efficient search and analytical tools. The second reason is the global regularly updated coverage of patent applications together with abstracts.

The results represent the importance of classification systems usage during different patent searches as a tool for narrowing the search set in the context of relevance or recall.

Our future work is directed toward to comparison of International Patent Classification and Cooperative Patent Classification and their differences in the patent search process.

Literature

COBURN, Mathhias M., 1999. *Competitive Technical Intelligence: A Guide to Design, Analysis and Action*. B.m.: American Chemical Society. ISBN 978-0841235151.

EPO, 2014. *Global Patent Index - GPI* [online]. Muenchen: European Patent Office [vid. 2016-04-13]. Dostupné z: http://documents.epo.org/projects/babylon/eponet.nsf/0/16B7F77528515906C1257C04003AB2FA/$File/GPI_UM_V23_EN.pdf

PIČMAN, Dobroslav, 2009. *Patentové a známkové rešerše*. Praha: Metropolitní univerzita Praha. ISBN 978-80-86855-44-8.

PORTER, L, DJ SCHOENECK, PR FREY, Diana M. HICKS a Dirk P. LIBAERS, 2007. Mining the Internet for competitive technical intelligence. *Competitive Intelligence Magazine* [online]. **10**(5), 24–28 [vid. 2016-05-20]. Dostupné z: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.73.3031

UNITED NATIONS, 2015. *Transforming our world: the 2030 Agenda for Sustainable Development* [online]. 2015. B.m.: United Nations. Dostupné z: http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E

WHO, 2015. *Global status report on road safety 2015* [online]. Geneva: WHO. ISBN 9789241564564. Dostupné z: doi:10.1136/injuryprev-2013-040775

WIPO, 2013. *Glossary of Terms Concerning Industrial Property Information And Documentation* [online]. 2013. ISSN 01722190. Dostupné z: doi:10.1016/0172-2190(93)90063-3

WIPO, 2015. *International Patent Classification Guide* [online] [vid. 2016-04-14]. Dostupné z: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf

WIPO, 2016a. International Patent Classification (IPC) Official Publication. *IP Services* [online] [vid. 2016-04-01]. Dostupné z: http://web2.wipo.int/classifications/ipc/ipcpub/

WIPO, 2016b. *IPC Internet Publication Help* [online] [vid. 2016-05-07]. Dostupné z: http://web2.wipo.int/classifications/ipc/ipcpub/static/htm/help_EN.htm#_Toc421895445

JEL Classification: O32

Shrnutí

Využití patentového třídění IPC pro oblast Competitive Technical Intelligence

Patentové informace představují významný zdroj pro oblasti Competitive Technical Intelligence a to v ohledu analýzy technologických aktivit v rámci konkurenčního boje. Patentové třídění posléze umožňuje průzkum trendů v určitém průmyslovém odvětví a rozklad aktivit klíčových hráčů na konkrétním technologickém poli. Tento článek pojednává o struktuře Mezinárodního patentového třídění (IPC) spolu s rešeršními metodami vedoucí k požadovaným analytickým výstupům. Autor tohoto článku využil pro demonstraci těchto metod Globální patentový index (GPI) a možnosti zjištění aktuálních technologických trendů a inovací v oblasti bezpečnostních prvků v automobilové dopravě pro ochranu před zraněními řidičů, pasažérů a chodců v případě nehody nebo jiných života ohrožujících situací se zaměřením na osobní vozidla, motocykly a nákladní vozy.

Klíčový slova: Competitive Technical Intelligence, Mezinárodní patentové třídění, IPC, analýza konkurence, bezpečnost v dopravě.

Aktuální otázky řízení eGovernmentu

Jana Fortinová
Jana.fortinova@vse.cz
Doktorand oboru Aplikovaná informatika
Školitel: doc. Ing. Jan Pour, CSc. (jan.pour@vse.cz)

Abstrakt: Cílem příspěvku je prezentovat návrh modelu řízení výkonnosti eGovernmentu se zohledněním potřeb organizací veřejné správy v ČR tak, jak vyplývají z analýz a průzkumů realizovaných na pracovišti autorky a z tuzemské i zahraniční literatury. Rozsah výzkumu se aktuálně soustřeďuje na řešení výkonnosti českých orgánů veřejné správy, avšak založeného i na zhodnocení publikovaných zahraničních přístupů.

V analytické části výzkumu byly využity statistické metody pro vyhodnocení údajů, kterými disponují mezinárodní orgány, zejména EUROSTAT. Aplikovanými metodikami jsou ve výzkumu převážně metodiky orientované na řízení organizací a zejména řízení jejich výkonnosti (CPM, Corporate Performance Management).

Návrh modelu řízení výkonnosti eGovernmentu je založen na uvedených obecných metodikách, ale je orientovaný na specifické potřeby veřejné správy. Součástí modelu je klasifikace IT služeb a úloh řízení IT řešených v rámci eGovernmentu a definování a kategorizace klíčových faktorů ovlivňujících výkonnost eGovernmentu. Detailní návrh modelu řízení výkonnosti eGovernmentu bude následně zahrnovat podrobnou specifikaci úloh řízení, manažerských a analytických metod, soustavu metrik, analytických a plánovacích aplikací a dalších objektů řízení eGovernmentu a jejich vzájemných vazeb. Navržený model bude průběžně verifikován ve vybraných orgánech veřejné moci a rozvíjen podle zde získaných poznatků.

Klíčová slova: Veřejná správa, řízení výkonnosti eGovernmentu, služby eGovernmentu, aplikace eGovernmentu

Úvod

Příspěvek se orientuje na otázky spojené s řízením eGovernmentu, řízením IT služeb, které poskytuje, a v návaznosti na to na možnosti zvyšování výkonnosti eGovernmentu na základě obecných principů a modelů řízení výkonnosti s respektováním zvláštních potřeb a možností, které jsou vlastní prostředí veřejné správy a orgánům veřejné moci (OVM). První část příspěvku je založena na aktuálních dostupných publikovaných zdrojích (MATES, SMEJKAL, 2012, BACAL, 2012, BOYNE, 2006, PALADINO, 2011, SHARK, TOPORKOFF, S., 2010, JEŠUTA, 2012, TAJTL, 2012, NOVOTNÝ a kol., 2010, ŠPAČEK, 2012) a zejména na výstupech ze seminářů České společnosti pro systémovou integraci (ČSSI) věnovaných veřejné správě v létech 2015 a 2016. Druhá část obsahuje návrhy autorky na řešení základního konceptu řízení výkonosti eGovernmentu a vybrané jeho části, tj. systému metrik, který by se měl postupně rozvíjet a představovat vstup pro řešení analytických a plánovacích úloh a aplikací, odpovídajících základním principům řízení výkonnosti (BOYNE, 2006, FIBÍROVÁ, ŠOLJAKOVÁ, 2005).

Metody výzkumu

V analytické části výzkumu byly využity statistické metody pro vyhodnocení údajů, kterými disponují mezinárodní orgány, zejména EUROSTAT. Aplikovanými metodikami jsou ve výzkumu převážně metodiky orientované na řízení organizací a zejména výkonnost a řízení jejich výkonnosti (CPM, Corpporate Performance Management).

Řízení výkonnosti je kombinace managementu, metodik a metrik podporovaná aplikacemi, nástroji a infrastrukturou, která umožňuje uživatelům definovat, monitorovat a optimalizovat výsledky a výstupy tak, aby bylo dosaženo cílů osobních či cílů organizační jednotky v souladu se strategickými cíli stanovenými na různých úrovních řízení podniku (osobní, procesní, skupinové a korporátní cíle podniku nebo podnikatelského ekosystému).

Z principů řízení výkonnosti je pak odvozen koncept řízení podnikové výkonnosti (Corporate Performance Management, CPM), který představuje holistický přístup k implementaci a monitoringu byznys strategie, kombinující dle (Coveney, 2003):

- metodiky – mezi které se zařazují metodiky podporující efektivní řízení podniku (např. Balanced Scorecard). Současně lze do této skupiny zařadit i technologické metodiky pro implementaci CPM systémů,
- metriky – které jsou v rámci implementace těchto metodik v podniku definovány,
- procesy – které používá organizace k implementaci a monitoringu řízení výkonnosti,
- aplikace a technologie – informační systémy pro podporu řízení výkonnosti na všech úrovních organizace, podporující dané metodiky, metriky a procesy.

CPM tak v sobě kombinuje nejrůznější moderní technologie a praktiky či postupy řízení podniků takovým způsobem, aby byla co nejvíce usnadněna formulace a vlastní uskutečňování podnikové strategie.

Současné otázky řízení eGovernmentu

Specifikace aktuálních otázek a problémů řízení eGovernmentu vychází z výše uvedených zdrojů, zejména materiálů publikovaných na stránkách ČSSI (http://www.cssi.cz). Problematika eGovernmentu je nesmírně široká a není tedy možné ji v jednom příspěvku postihnout. Proto se příspěvek zaměřuje pouze na některé oblasti, které mají k výkonnosti významný vztah.

Řízení IT služeb ve veřejné správě a eGovernmentu

Problematikou řízení IT služeb se zabývá řada publikací (VOŘÍŠEK, 2015, VOŘÍŠEK, POUR, 2012, NOVOTNÝ a kol., 2010) a řada metodik a modelů (ITIL, CobiT, CMMI). Aktuální pohled na IT služby ve veřejné správě prezentoval Prof. Voříšek ve své přednášce na semináři ČSSI v únoru 2016 (viz dále).

Vláda ČR připravila a schválila v listopadu 2015 návrh Strategie rozvoje IT služeb ve veřejné správě, v níž bylo identifikováno 16 klíčových nedostatků, a na jejich základě byly definovány cíle a odpovídající opatření. Další přehled uvádí alespoň vybrané nedostatky:

- eGovernment v ČR neřeší na potřebné úrovni vazby mezi základními službami poskytovanými orány veřejné moci a IT službami, resp. neřeší podporu základních služeb informatikou. To pak vede k tomu, že investice do IT nemohou být kvalitně posuzovány podle efektů, které IT do veřejné správy přináší, resp. by měly přinést.
- Neexistovala jednotná a přesná pravidla pro schvalování investičních záměrů do IT, schvalování plánovaných projektů, pravidla nákupu produktů a služeb v oblasti ICT. V současné době jsou taková pravidla připravena, pokud jde o projekt, a postupně se ověřují v praxi.
- Neexistují jednotná pravidla sledování a analýzy nákladů (investičních a provozních), výnosů a kvality služeb veřejné správy a IT služeb, což např. i znamená, že poskytovatele služeb nelze odměňovat na základě objemu a kvality poskytnutých služeb.
- I přes existenci základních registrů, informační systémy OVM nejsou dostatečně propojeny a již jednou získaná data nejsou důsledně sdílena, takže se stále nedaří minimalizovat ztráty podnikatelských subjektů i občanů s obíháním úřadů se stejnými daty.
- Velmi málo se využívají sdílené IT služby (zejména na bázi cloud computingu a jeho služeb), kdy jeden poskytovatel službu zajišťuje více OVM. Velmi intenzivně se diskutuje koncepce tzv. G-Cloudu, tedy cloudu pro Government, který by měl přinést významné úspory nákladů, pracnosti.

Významným problémem je v souvislosti s řízením IT služeb obrovský počet aplikací a informačních systémů pro OVM, jimž by měl charakter a funkcionalita zejména aplikačních IT služeb odpovídat. Těchto aplikací je aktuálně téměř 7 000, různě se překrývají, nejsou kompatibilní, často nejsou přesně známé ani náklady na jejich rozvoj a provoz, atd. To je i výraznou překážkou koncipování zmíněného G-Cloudu.

V každém případě máme-li smysluplně navrhovat a implementovat systém pro řízení výkonnosti eGovernmentu, pak jedním z jeho podstatných předpokladů je kvalitní systém, nebo alespoň katalog IT služeb. Prof. Voříšek uvádí v souvislosti s vytvářením katalogu např. následující otázky, které by bylo možné řešit:“

- jaké jsou duplicity v provozovaných službách
- jaké jsou celkové náklady ČR na IT služby (dle služeb, dle institucí, dle poskytovatelů,…),
- jaké jsou cenové rozdíly téže služby u různých institucí,
- jaké služby užívá daná instituce,
- kdo je provozovatelem těchto služeb,
- a další.

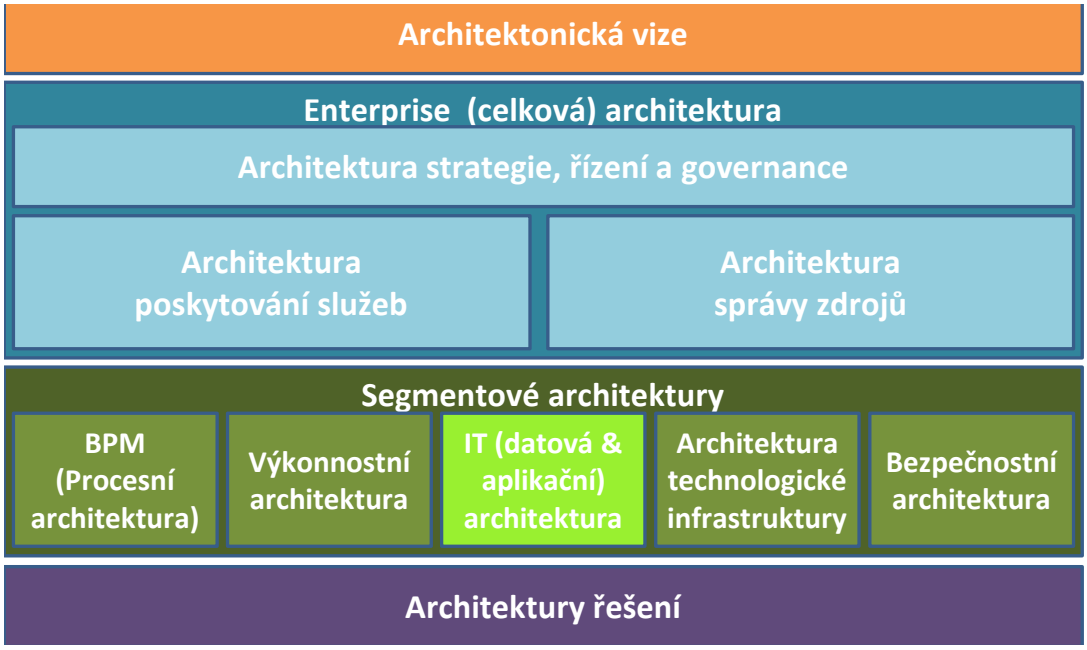
Architektury eGovernmentu – jejich principy a řešení

Význam architektur pro řízení podniků i orgánů veřejné moci (OVM) a jejich IT je při současných nárocích na kvalitu a výkonnost řízení zcela evidentní. To platí pro celý komplex architektur (podnikovou, aplikační, technologickou, datovou atd.), včetně tzv. architektury služeb, resp. IT služeb, které se věnuje detailněji část tohoto příspěvku. Na druhé straně jsou předmětem mnoha publikací, např. (Novotný a kol, 2010), otázky výkonnosti organizací jak v komerční, tak veřejné sféře. Řízení výkonnosti (performance management, PM) má svá pravidla a přístupy a samozřejmě celou škálu faktorů, které skutečně dosahovanou výkonnost pozitivně, či negativně ovlivňují. Jedním z těchto faktorů je nebo může být podniková architektura a z pohledu IT i architektura IT služeb.

Specifickými otázkami řešení architektur ve veřejné správě se dlouhodobě zabývá P. Hrabě (HRABĚ, 2014, HRABĚ, 2013) a naposledy formuloval klíčové principy architektur VS v příspěvku „Standardizace agend v kontextu NA VS v ČR“ na semináři ČSSI v roce 2016. K nim patří zejména:

- Architektura orgánu VS se chápe jako manažerská metoda, prostředek celostního poznávání organizace na podporu rozhodování a plánování strategických změn,
- Metodiky, rámce, nástroje, které se využívají k řešení a rozvoji architektur ve VS, jsou:
    - TOGAF - mezinárodně uznávaný rámec pro řízení tvorby podnikové architektury tam, kde se využívá IT: http://pubs.opengroup.org/architecture/togaf9-doc/arch/
    - ArchiMate® - nezávislý grafický modelovací jazyk, správa - Open Group, standard pro popis podnikové architektury: http://pubs.opengroup.org/architecture/archimate2-doc/
- Národní architektura (NA) je uplatnění metod podnikové architektury ve veřejné správě ČR, souhrn lokálních architektur OVM a centrálních architektur eGovernmentu.
- Národní architektonický rámec (NAR) je koncept, metodika postupu, standardy, návody pro tvorbu a údržbu NA a NAP
- Národní architektonický plán (NAP) je kromě jiného soubor architektonických dat (modelů) a diagramů, udržovaných společně Odborem hlavního architekta (OHA) a jednotlivými OVM členěný na: architektury úřadů a architektury sdílených řešení.

Uvedené principy dokresluje následující obrázek:



Obrázek 1: Návrh vrstev modelu architektury organizace (a státu)
Zdroj: (Hrabě, P: Disertační práce, 2015)

Důležitost kvalitně řešené architektury, resp. jednotlivých dílčích architektur jako vstupu pro řešení řízení výkonnosti eGovernmentu je zcela zřejmý. Problém je v tom, že zatímco na centrální úrovni je této oblasti věnována velmi silná pozornost, a to je výrazný posun, pak na jednotlivých orgánech je tato pozornost, již i z kapacitních důvodů, podstatně slabší. V každém případě ale řešení úloh a modelů výkonnosti se musí koncipovat v rámci zmíněných architektur.

Řízení výkonnosti eGovernmentu

Problematika řízení výkonnosti je zatím ve výzkumu a v literatuře orientována převážně na komerční sféru – CPM - a jeho různé modifikace - IT PM, Sales PM, CC PM atd. (Novotný a kol., 2010). Řízení výkonnosti v oblasti veřejné správy a eGovernmentu je předmětem zkoumání v podstatně menší míře. Pokud vezeme v úvahu i současné práce zaměřené na eGPM, pak i zde je velký prostor pro zkoumání různých variant a jejich účelnosti pro praktické užití, např. ve vymezení jednotlivých objektů řízení, jejich strukturalizace, vzájemných vazeb atd.

Některé problémy v oblasti řízení výkonnosti eGovernmentu a veřejné správy jsou v dalším přehledu:

- problémy a rezervy v řízení výkonnosti eGovernmentu mají zásadní dopad na řízení výkonnosti a kvality celé veřejné správy (UNITED NATIONS, 2014, BACAL, 2012, NOVOTNÝ a kol., 2010, PALADINO, 2011 atd.),
- řízení výkonnosti eGovernmentu (eGPM) je v současných publikacích, zejména zahraničních, prezentováno dosud na koncepční, hrubé úrovni, nebo na úrovni statistických přehledů a analýz, mezinárodních srovnání, globálních vývojových trendů apod., zatím ale chybí jeho dotažení do dílčích charakteristik jednotlivých objektů řízení a přesná specifikace jejich vazeb (VOŘÍŠEK, POUR, 2012),
- řízení výkonnosti eGovernmentu předpokládá propracovaný systém metrik, který musí mít celou řadu specifik vzhledem k podmínkám veřejné správy, např. v motivaci měření, dalšího rozvoje IT služeb, v řízení rizik projektů a využití aplikací, v kvalifikačních programech, v řízení nákladů a efektů IT, atd. (MATES, SMEJKAL, 2012, HRABĚ, 2013). Tyto systémy metrik jsou v dostupné literatuře prezentovány, ale je vhodné hledat nový koncept jejich vyjádření, včetně zohlednění možností analytických nástrojů pro práci s nimi,
- chyby v řízení výkonnosti eGovernmentu mají, zejména u aplikací celostátního významu, silný dopad nejen na efektivní využívání zdrojů a konkurenceschopnost jednotlivých podniků, ale celé národní ekonomiky, jak dokumentují srovnání mezinárodních statistik (EUROSTAT, OECD, ŠPAČEK, 2012), zejména z pohledu pokrytí a využívání IT služeb eGovernmentu.

Návrh modelu řízení výkonnosti eGovernmentu

Návrh modelu řízení výkonnosti eGovernmentu bude vycházet z principů CPM, IT PM a obdobných dalších (viz přehled zdrojů, např. Voříšek, 2015), ale bude orientovaný na specifické potřeby, možnosti a omezení služeb, aplikací, principů řízení v oblasti veřejné správy. Návrh konceptu modelu řízení výkonnosti eGovernmentu vychází z principů řízení výkonnosti a speciálně principů CPM, z modelu MBI (VOŘÍŠEK, POUR, 2012) a postupně i z dalších publikovaných zdrojů (HOMBURG, 2008, PALADINO, 2011 a dalších).

Další část příspěvku vychází z vlastních konstrukcí autorky publikovaných v (FORTINOVÁ, 2015). Návrh celkového konceptu řízení výkonnosti eGovernmentu je obsahem obrázku 2.

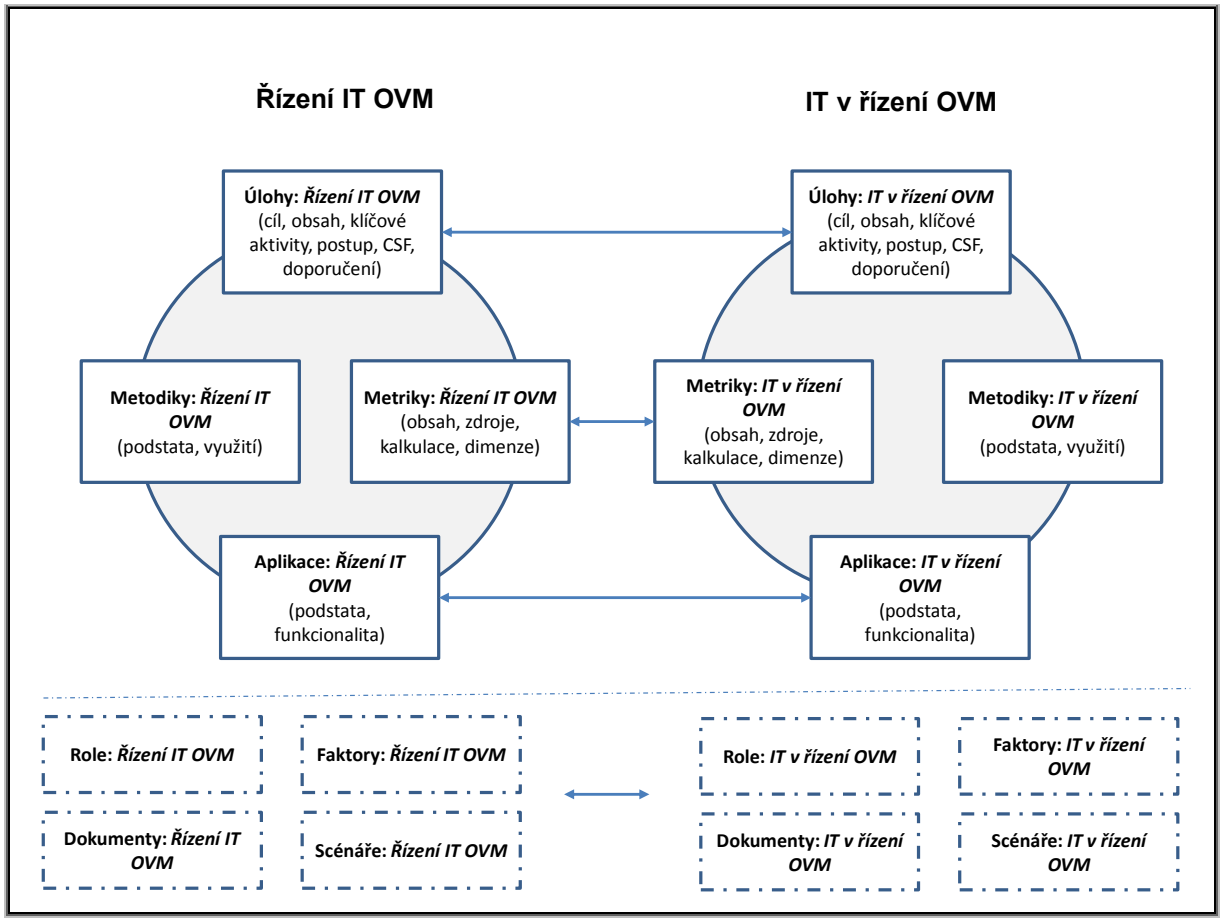
K návrhu je třeba připojit několik dalších poznámek:

- jednotlivé části – obdélníky na schématu představují objekty řízení výkonnosti (podrobněji VOŘÍŠEK, POUR, 2012),
- v horní části schématu jsou uvedeny standardní objekty řízení výkonnosti ve vzájemných vazbách:
  - úlohy (procesy a další charakteristiky),
  - metodiky, metody,
  - metriky,
  - aplikace, nástroje – zejména analytické a plánovací.
- uvedené základní objekty eGPM jsou rozděleny do dvou skupin, resp. samostatných schémat:
  - objekty řízení IT v OVM, tedy představující systém řízení IT v daném prostředí,
  - objekty řízení OVM s podporou IT (IT v řízení OVM), čímž se myslí základní oblasti řízení OVM (HR, majetek, sklady, poskytování služeb apod.), kde je významná podpora IT (a aplikací eGovernmentu),
  - důvod tohoto rozdělení je v tom, že obdobně, jako v komerční sféře se řeší otázky tzv. business – IT alignment, tak totéž je nutné řešit i ve veřejné správě a nejlépe na specifikaci vztahů zejména úloh a metrik řízení OVM a řízení IT v jeho rámci,
- ve spodní části schématu jsou uvedeny objekty (ve stejném rozdělení na 2 skupiny), které nejsou součástí základního konceptu CPM, ale byly nadefinovány a použity v modelu MBI (VOŘÍŠEK, POUR, 2012) a ukázaly se jako velmi efektivní. Proto i v návrhu eGPM se počítá s jejich využitím. Kromě těchto objektů zde budou využity i další součásti MBI (vlastnosti IS, předměty řízení a případně další.

Řešení modelu musí zahrnovat následující obsahové části:

- klasifikace IT služeb a IT úloh poskytovaných a řešených v rámci IT Governmentu – půjde o vlastní pracovní návrh klasifikace IT služeb a úloh, který bude odpovídat předpokládané koncepci modelu řízení výkonnosti eGovernmentu a bude postupně zpřesňován podle vývoje částí řešení práce,
- vymezení a klasifikace klíčových faktorů ovlivňujících výkonnost eGovernmentu a následně i jeho modelové řešení, vycházející zejména z formulace cílů rozvoje řízení informatiky ve veřejné správě, analýzy možností, omezení a rizik poskytovaných IT služeb a aplikací v rámci eGovernmentu (např. aktuálně specifikovaných dílčích aplikací eGovernmentu) atd. - faktory budou formalizovány tak aby bylo možné na nich realizovat i nezbytné analýzy, případně predikce,
- detailní návrh modelu řízení výkonnosti eGovernmentu, který bude zahrnovat zejména specifikaci úloh řízení, pracovních manažerských a analytických metod, soustavu metrik, aplikací a nástrojů a dalších objektů řízení a jejich vzájemných vazeb, včetně jejich kvalifikace a detailních hodnocení,

Specifickou částí celého řešení bude systém metrik a jejich hodnocení, proto je mu v další části věnována zvláštní pozornost.



Obrázek 2: Návrh konceptu řízení výkonnosti eGovernmentu (Zdroj: vlastní konstrukce)

Klíčové metriky hodnocení výkonnosti eGovernmentu

Pro objektivní a reálné hodnocení výkonnosti eGovernmentu bude třeba definovat soustavu klíčových charakteristik, které mohou mít charakter jak metrik, tak případně textového vyjádření (JEŠUTA, 2012, TAJTL, 2012). Pro takové účely je i potřebné navrhnout jejich základní klasifikaci, resp. hierarchickou strukturu. V daném případě návrh takové struktury člení metriky do těchto skupin:

- metriky cílového užití,
- metriky ekonomické,
- metriky provozní a technologické,
- ostatní metriky.

Další kapitoly textu obsahují pouze příklady zmíněných metrik s tím, že většina z nich bude následně doplňována a dopracována a bude tvořit v disertační práci navržený systém metrik jako součást koncepce eGPM a součást celého systému charakteristik (metrik i textových).

Metriky cílového užití eGovernmentu

Metriky cílového užití zahrnují portfolio poskytovaných služeb, jejich funkcionalitu a úroveň využití. Do této skupiny patří:

- rozsah poskytovaných a podporovaných IT služeb:
  - v současné době nejčastější charakteristika pro sledování úrovně eGovernmentu v rámci států EU a založená na standardu 20 služeb EUROSTATu),
  - pro detailnější analýzy je tato charakteristika velmi hrubá,
  - metrika – počet služeb poskytovaných v daném období, regionu apod.
- kvalita poskytovaných IT služeb:
  - je daná zejména úrovní poskytované funkcionality, kvalitou uživatelského interface apod.,
  - v mezinárodních statistikách se nevyužívá,
  - metrika – může být v tomto případě vyjádřená bodovou škálou
- úroveň využití poskytovaných služeb komerčními subjekty a občany,
  - používá se velmi často v mezinárodních statistikách v rozlišení na podniky a občany, s jejich přesným vymezením,
  - metrika – obvykle v procentuálním vyjádření – podle typů podniků, podle odvětví, podle kategorií občanů apod.
- skutečný počet uživatelů v rámci určitého období,
  - období musí být jasně definováno, je důležitá pro sledování vývoje ve využití služeb,
  - obvyklým obdobím v mezinárodních statistikách je 1 rok,
  - metrika – obvykle v absolutním nebo procentuálním vyjádření – podle obdobných dimenzí, jako v předchozím případě,

Metriky ekonomické

Metriky ekonomické představují:

- náklady na poskytované IT služby:
  - náklady na IT služby v členění podle kategorií IT služeb, poskytovatelů, OVM, nákladových druhů, případně dalších dimenzí, za definované období
  - metrika – náklady v tis. Kč,
- ekonomické efekty z IT služeb pro komerční subjekty:
  - představují efekty vyjádřitelné finančně, nebo efekty, které lze bez výrazného zkreslení na finanční vyjádření převést, příkladem je absolutní nebo relativní snížení nákladů na využití požadovaných funkcí VS, snížení časové náročnosti na jejich zajištění s převodem na náklady apod.
  - metrika – ekonomické efekty v tis. Kč,
- ekonomické efekty pro občany:
  - představují efekty obdobné, jako v předchozím případě, ale se zúžením na jednoho občana, tedy většinou snížení časové náročnosti s případným přepočtem, např. na ušlou mzdu apod.
  - metrika – ekonomické efekty v tis. Kč,

Metriky provozní a technologické

Metriky provozní a technologické zahrnují především úroveň aplikačních IT služeb a IT infrastruktury a do této skupiny patří:

- dostupnost IT služeb:
  - reprezentuje standardní hodnoty dostupnosti IT služby v čase,
  - metrika – procentuální vyjádření dostupnosti služby na celkové době provozu IT,
- chybovost IT služeb:
  - představuje počet chyb za jednotku času, která musí být jasně nadefinována, např. den, měsíc apod., podle typu chyby, služby
  - metrika – počet chyb v absolutním vyjádření
- počet a objem reklamací na poskytované IT služby:
  - obdobně jako v předchozím případě - počet reklamací za jednotku času podle typu reklamace, služby a autora reklamace,
- úroveň zajištění bezpečnosti služeb a další:
  - celková úroveň bezpečnosti IT služeb,

- v tomto případě půjde o kombinaci textu a metrik (např. počet služeb se zajištěnou bezpečností) – dle principů a standardů řízení bezpečnosti IT

Z navrženého přehledu vyplývá, že z pohledu možností vyjádření, měření a analýz výše navrhovaných metrik, půjde většinou o obě možnosti, tj. exaktního vyjádření a následně i textového. Další podstatnou částí řešení bude specifikace analytických dimenzí ve vztahu k metrikám, a to na základě realizovaných průzkumů. Ty se pak stanou i vstupem pro formulaci celého rámce analytických a plánovacích úloh řízení výkonnosti eGovernmentu.

Závěr

Návrh modelu řízení výkonnosti eGovernmentu musí vycházet z principů CPM, IT PM a obdobných dalších modelů (VOŘÍŠEK, 2015), ale bude orientovaný na specifické potřeby, možnosti a omezení služeb, aplikací, principů řízení v oblasti veřejné správy. Návrh bude zahrnovat tyto dílčí části:

- jako předpoklad řešení bude nutné provést klasifikaci IT služeb a IT úloh poskytovaných a řešených v rámci eGovernmentu - ve vazbě na připravovaný katalog IT služeb ve VS,
- vymezení a klasifikace klíčových faktorů ovlivňujících výkonnost eGovernmentu a následně i jeho modelové řešení, vycházející zejména z formulace cílů rozvoje řízení informatiky ve veřejné správě, analýzy možností, omezení a rizik poskytovaných IT služeb a aplikací v rámci eGovernmentu (např. aktuálně specifikovaných dílčích aplikací eGovernmentu) atd. - faktory budou formalizovány tak aby bylo možné na nich realizovat i nezbytné analýzy, případně predikce,
- vytvoření základní koncepce modelu řízení výkonnosti eGovernmentu vycházející ze struktury výše vymezených objektů řízení a zejména jejich vazeb, včetně jejich atributů a možností jejich analýz,
- detailní návrh modelu řízení výkonnosti eGovernmentu, který bude zahrnovat zejména specifikaci úloh řízení, pracovních manažerských a analytických metod, soustavu metrik, aplikací a nástrojů a dalších objektů řízení a jejich vzájemných vazeb, včetně jejich kvalifikace a detailních hodnocení,
- navržený model bude pak verifikován ve vybraných OVM a následně upravován podle zde získaných informací.

Literatura

BACAL, R.: Manager´s Guide to Performance Management. New York, McGraw-Hill 2012. ISBN 978-0-07-177225-9.

BOYNE, G.A. a další: Public service performance: Perspectives on measurement and management, Cambridge, 2006. ISBN 9780521859912

FIBÍROVÁ, J., ŠOLJAKOVÁ, L.: Hodnotové nástroje řízení a měření výkonosti podniku. Praha: ASPI, 2005. 264 s. ISBN 80-7357-084-X.

FORTINOVÁ, J.: Řízení výkonnosti eGovernmentu – problémy a perspektivy, Systémová integrace, 3/2015.

HRABĚ, P.: Zpráva o případové studii ke Koncepci Národní architektury veřejné správy ČR. Systémová integrace [online]. 2014, roč. 21, č. 1–2, s. 85–103. ISSN 1210-9479. Dostupné z: http://www.cssi.cz/cssi/zprava-o-pripadove-studii-ke-koncepci-narodni-architektury-verejne-spravy-cr.

HRABĚ, P.: Koncepce Národní architektury veřejné správy ČR. Systémová integrace [online]. 2013, roč. 20, č. 4, s. 7–21. ISSN 1210-9479. Dostupné z: http://www.cssi.cz/cssi/koncepce-narodni-architektury-verejne-spravy-cr.

HRABĚ, P.: Koncepce řízení výkonnosti a zodpovědnosti pomocí KPI v české veřejné správě. Systémová integrace [online]. 2013, roč. 20, č. 1, ISSN 1210-9479.

HOMBURG, V.: Understanding E-Government: Information Systems in Public Administration. London, Routledge, 2008

JEŠUTA, M.: Analýza efektivity e-služeb veřejné správy ČR, DP VŠE, Praha, 2012.

MATES, P., SMEJKAL, V.: E-government v České republice. Právní a technologické aspekty. Praha, Leges. 2012. ISBN 978-80-87576-36-6.

NOVOTNÝ, O., POUR, J., BASL, J., MARYŠKA, M.: Řízení výkonnosti podnikové informatiky. Professional Publishing, Praha, 2010. ISBN 978-80-7431-040-9.

OECD: Public sector innovation and e-government http://www.oecd.org/gov/public-innovation/ , OECD, 2014.

PALADINO, B.: Innovative Corporate Performance Management: Five Key Principles to Accelerate Results. Indianopolis, Wiley Publishing, 2011. ISBN: 978-0-470-62773-0.

PRAGR, A.: eGovernment a analýza Portálu veřejné správy, DP VŠE, Praha, 2015.

SHARK, A.R., TOPORKOFF, S.: Beyond eGovernment: Measuring Performance - a Global Perspective, Washington, DC, Public Technology Institute and ITEMS International, 2010.

ŠPAČEK, D.: eGovernment: cíle, trendy a přístupy k jeho hodnocení. C.H.Beck, 2012. ISBN 978-80-7400-261-8.

TAJTL, M.: Měření efektivity elektronické státní správy, DP VŠE, Praha, 2012.

UNITED NATIONS: United Nations E-Government Survey 2014, E-Government for the Future We Want. United Nations, 2014

VOŘÍŠEK, J., POUR, J. a kol.: Management podnikové informatiky, Professional Publishing, 2012, ISBN 978-80-7431-102-4

VOŘÍŠEK, J. a kol.: Principy a modely řízení podnikové informatiky. Praha, Oeconomia 2015. ISBN: 978-80-245-1440-6.

JEL Classification: M10, C88

Summary

Actual problems of eGovernment management

The aim of this article is to present the draft of a model for eGovernment performance management with regards to the needs of public administration organizations in the Czech Republic as they ensue from analyses and the research realized at the author's workplace and from the domestic as well as foreign literature. The research extent is currently focused at the performance of the Czech public administration authorities and is based also on the assessment of published foreign approaches.

The statistical data evaluating methods that are at the disposal for international bodies, especially EUROSTAT, were used in the analytic part of the research. The methodologies oriented to the management of corporations and above all to their performance management (CPM, Corporate Performance Management) were mostly used as applied methodologies.

This draft of the model of eGovernment performance management is based on quoted common methodologies but is oriented to specific needs of public administration. The classification of IT service and IT control tasks solved within the eGovernment frame as well as the definition and categorization of key factors influencing the eGovernment performance is part of the model. The detailed draft of the eGovernment performance management model brings an elaborated specification of management tasks, managerial and analytic methods, a system of metrics, analytic and planning applications and other eGovernment management objects and their mutual links. The proposed model will be currently verified in selected bodies of public administration and further developed according to received findings.

Keywords: Public administration, eGovernment performance management, eGovernment services, eGovernment applications.

Ergonomie software – identifikace a původ vzorů uživatelských rozhraní

Jiří Hradil

jiri.hradil@vse.cz

Doktorand oboru Aplikovaná informatika

Školitel: doc. Ing. Vilém Sklenák, CSc., (sklenak@vse.cz)

Abstrakt: Článek se zabývá identifikací vzorů v uživatelských rozhraních softwarových aplikací a snaží se vyvodit jejich původ a význam na základě asociací s fenomény okolního světa. Autor nejdříve zvolí čtyři základní archetypy (*Čas, Gravitace, Prostor a Podobnost*) a přiřadí k nim vzory uživatelských rozhraní. Každý vzor rozhraní je popsán, vysvětlen a zařazen do správného kontextu použití. Autor věří, že pokud se podaří vzory takto identifikovat a kategorizovat, lze je pak vylepšovat rychleji napříč různými doménami a aplikacemi. Uživatelská rozhraní a popsané vzory jsou součástí širšího výzkumu ergonomie software, jehož hlavním cílem je tvořit přístupnější a jednodušší software, který bude lidem sloužit lépe.

Klíčová slova: ergonomie software, uživatelské rozhraní, vzor uživatelského rozhraní, archetyp, layout.

Úvod

Ergonomie software je doménou výzkumu nahlížející na software pohledem člověka, jejíž hlavním cílem je harmonizace komunikace mezi člověkem a strojem. Člověk je v ergonomii software dominantním prvkem, software ho zohledňuje od svého prvopočátečního návrhu; člověk je v samotné DNA software.

Vědu o člověku nelze obsáhnout v jedné kategorii lidského poznání; i vědecké kategorie či obory byly odděleny umělými hranicemi, protože není v silách jednoho obsáhnout všechny obory najednou. Poznání a věda jsou komplexní a navzájem propojené, stejně jako celistvé vědění o čemkoli. Nahlížejíc moderním pohledem bychom mohli uvést, že věda o člověku je vědou multioborovou, protože přesahuje více těchto uměle vytyčených, ovšem všeobecně uznávaných a formálně definovaných hranic. Pokud je člověk v ergonomii software dominantní částí systému a ergonomie je vztažena hlavně ke člověku, pak i ergonomie software musí být nutně multioborová.

Tento článek je pouze jednou z částí výzkumu autora, zabývajícího se ergonomií software. Identifikace a původ vzorů uživatelských rozhraní a jejich archetypů má přesah do psychologie, fyziky či sociologie. Na popsané vzory je třeba nahlížet z pohledu tzv. „*měkkých*“ systémů. I když mají vzory odraz v okolním světě, jsou vyhodnocovány člověkem a pro člověka, a jejich interpretace může být do jisté míry subjektivní. Přesto se autor pokouší tyto vzory identifikovat, kategorizovat a popsat univerzálně tak, aby byly v aplikacích používány ve správném kontextu. Pokud budou uživatelská rozhraní generalizována na několik opakujících se typů, lze pak jejich ergonomii vylepšovat rychleji a s větším dopadem. Nové aplikace, následující popsané principy mohou být přístupnější, pohodlnější, jednodušší a ve všech směrech lepší pro své uživatele.

Současné vědecké přístupy a publikace řeší ergonomii software pouze zprostředkovaně či nepřímo, např. výzkumem psychologie barev (Birren, 2013) a (Löffler, 2015), úlohou člověka při výzkumu designu a ergonomie tvarů (Chapanis, 1999) či popisem návrhových vzorů (design patterns) při vývoji software (Freeman, Robson, Bates a Sierra, 2004). Konkrétní doporučení pro tvoru komponent návrhu uživatelských rozhraní popisuje (Krug, 2014), případně jsou součástí doporučení pro tvorbu komplexních aplikací v operačních systémech (Apple 2017a; Google 2017a; Microsoft, 2017a). Vzory rozhraní a komponent však postrádají odkaz na svůj původ a asociaci s okolním světem a tím se obtížně predikuje jejich chování.

Autor předpokládá následující:

- Výběrem archetypů a jejich odkazováním určíme hranice pro uživatelská rozhraní, která poté budou navrhována v mezích známého okolního světa a jejich ovládání bude predikovatelné.
- Uživatelská rozhraní budou používána ve správném kontextu.
- Software bude v důsledku jednodušší pro své uživatele.

Základní archetypy

Jako archetypy či pravzory pro uživatelská rozhraní byly zvoleny následující fenomény.

Čas

Čas představuje změnu a růst, neustálý pohyb, vyjadřuje závislost člověka na okolním prostředí, reflektuje interakci a propojuje akci s reakcí. Vnímáním času lze odkazovat základní časové rámce – minulost, současnost a budoucnost.

Jak uvádí (Thomson, 2014), z pohledu mentální časové osy je minulost v prostoru vnímána „vlevo“ a budoucnost „vpravo“. Test proběhl v Evropě, a tudíž nereflektuje východní kultury (psaní zprava-doleva) či starou čínštinu (shora-dolů). Lze tedy předpokládat, že v těchto kulturách by byly závěry testu rozdílné a uživatelská rozhraní mohou být chápána v jiném kontextu.

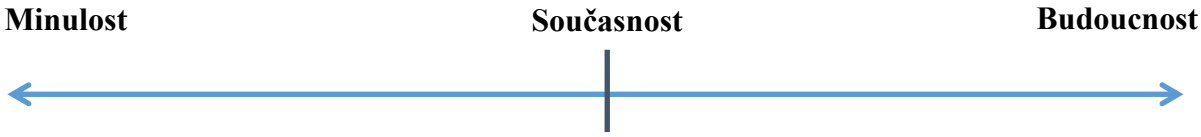
Způsob vnímání času zleva-doprava na mentální časové ose reflektují uživatelská rozhraní horizontálním layoutem, rozbalováním informace zleva-doprava (procházení adresářů – adresář vlevo, soubory v něm vpravo) a označováním prioritních komponent (vpravo je novější a tedy důležitější). Prioritu komponent je však třeba chápat ve správném kontextu. Komponenty vlevo „začínají“ komunikaci („akce“) a komponenty vpravo akci ukončují („reakce“). Jsou mezi sebou propojeny. V tomto kontextu nemusí vždy platit, že vlevo je „nedůležité“, protože bez levé části by se uživatel k pravé části nedostal. Tuto připomínku zohledňuje uživatelské rozhraní *Menu vlevo*.

Vnímání současnosti nelze zcela „odmyslet“. Nelze myslet „teď není“, protože i o „teď není“ myslíme zrovna teď, takže bychom nemysleli, čímž vzniká paradox. Neustálá přítomnost současnosti se nutně projeví i v návrhu uživatelského rozhraní. Pokud máme k dispozici pouze jednu pracovní plochu, tato vyjadřuje „teď“, čili současnost (Obrázek 1).

Obrázek 1: Vnímání mentální časové osy, 1 časový rámec Zdroj: (autor, 2017)

V případě 2 pracovních ploch (vzor *2-sloupcový horizontální layout*) lze vyjádřit minulost a současnost (Obrázek 2); v případě 3 ploch (*3-sloupcový horizontální layout*) lze projektovat minulost, současnost a budoucnost (Obrázek 3). Projektovat pouze současnost a budoucnost bez minulosti shledává autor minoritním, ale lze jej v případě potřeby použít také.

Obrázek 2: Vnímání mentální časové osy, 2 časové rámce



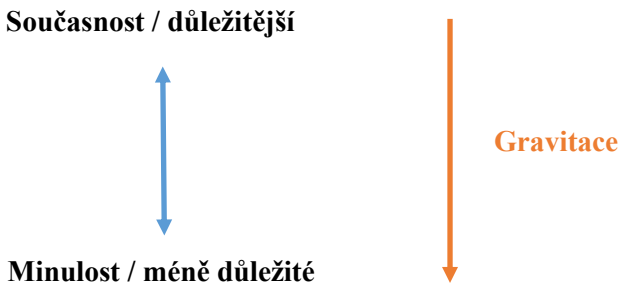
Obrázek 3: Vnímání mentální časové osy, 3 časové rámce Zdroj: (Thomson, 2014; autor 2017)

Archetyp *Času* lze rozpoznat v uživatelských rozhraních *2-sloupcový horizontální layout, 3-sloupcový horizontální layout, Časová osa, Menu vlevo a Menu vpravo.*

Gravitace

Vnímání gravitace, jakožto přírodní zákona a jejího vlivu na člověka se odráží v návrhu další typů uživatelských rozhraní, respektující princip „shora-dolů“. Nahoře je více důležité a dole méně důležité. Gravitaci je také možno chápat jako vertikálně položenou časovou osu. Projektovat budoucnost je v archetypu *Gravitace* obtížné, ne-li

nemožné. Ve vertikálním prostoru lze vyjádřit budoucnost nahoře, ovšem uživatelská rozhraní zobrazují data či ovládací prvky v úrovni očí uživatele, což implikuje současnost, nikoli budoucnost. Z tohoto důvodu je vhodné redukovat časové úseky v archetypu *Gravitace* pouze na současnost a minulost.



Obrázek 4: Princip archetypu Gravitace. Zdroj: autor, 2017

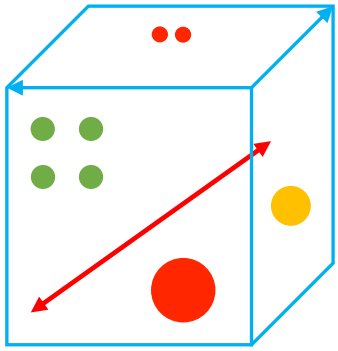
Archetyp *Gravitace* je základem pro uživatelské rozhraní *Seznam* a obecně pro rozhraní s vertikálním layoutem (ovládání shora-dolů), o jejichž popis bude rozšířena finální disertační práce autora (Hradil, 2017).

Prostor

Archetyp *Prostoru* je odvozen vnímáním světa člověkem. Předpokládejme, že aktér prostor vnímá (ví, že v něm je), uvědomuje si hranice prostoru (a tím i jeho velikost), rozpoznává objekty v něm a zná svou pozici. Pak lze uvést, že vnímání prostoru a objektů v něm je závislé na velikosti samotného prostoru a na vzdálenosti každého prvku od prvků ostatních. Prvky umístěné blízko sebe chápeme jako skupinu a hodnotíme je jako celek. Předpokládáme, že k sobě patří. Pokud existují skupiny prvků a některý prvek ve skupině není, získává tím samostatnost a exkluzivitu – něčím se liší od ostatních, seskupených prvků. Takový prvek má tedy zřejmě jiné, speciální vlastnosti (neposuzujeme, zda dobré či špatné, prostě není s ostatními prvky z nějakého důvodu).

Vnímání *Prostoru* zakládá vztahy *vzdálenosti, velikosti* a tím nepřímo i *důležitosti* v uživatelských rozhraních. Obecně platí, že pokud usilujeme o jednoduché rozhraní, mělo by v něm být co nejméně prvků. Každý prvek okupuje zorné pole uživatele, svou prezencí se dožaduje pozornosti a zvyšuje počet možných interakcí mezi ním a ostatními prvky či počet interakcí prvku a uživatele.

Velikost prostoru, velikost aktéra a jeho pozici v prostoru je základem pro vzor *Pracovní plocha uprostřed*. Počet prvků v prostoru, jejich umístění a vzdálenosti mezi nimi odráží vzor *Skupina*. Samotná velikost prvků zakládá vzor *Důležitější je větší.*



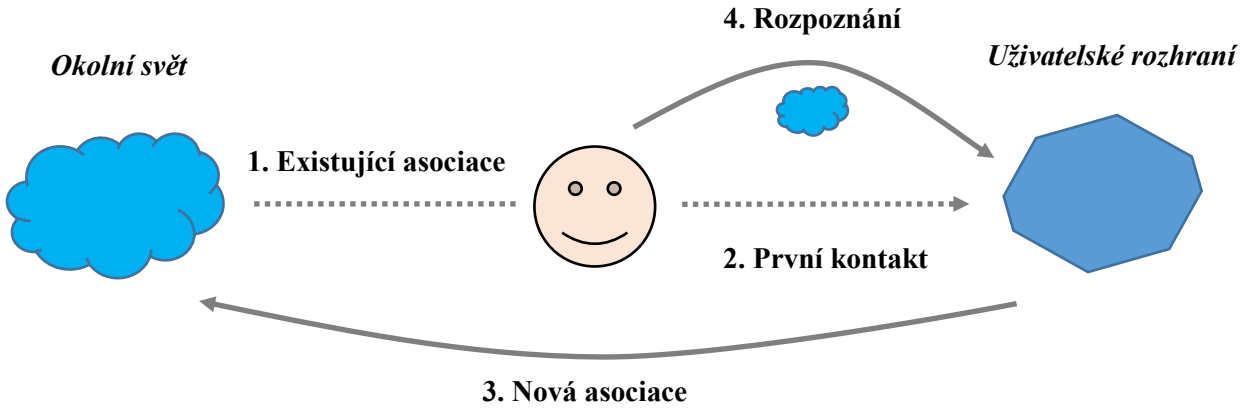
Obrázek 5: Princip archetypu Prostor (vzdálenost, velikost a důležitost) Zdroj: autor, 2017

Podobnost

Archetyp *Podobnost* znamená asociaci uživatelského rozhraní či jeho komponent s konkrétními objekty a jejich atributy z reálného světa. Výhodou je již existující mentální asociace a následná identifikace komponenty s jejím vzorem v reálném světě, a tedy i predikce jejího smyslu a funkce bez nutnosti učení. Asociace snižuje entropii. Podobnost, a tedy i asociace může být samozřejmě pouze přibližná na základě shody některých atributů či projevů původního objektu (zejména tvar, barva, chování) s komponentou. Obecně se jedná o atributy zachytitelné zrakem či sluchem (nověji i čichem i hmatem např. u rozhraní pro virtuální realitu).

Princip archetypu lze popsat následovně (Obrázek 6):

1. Uživatel má asociaci z okolního světa na základě předchozího poznání (*Existující asociace*).
2. Uživatel poprvé uvidí v rozhraní komponentu, kterou dosud neznal (*První kontakt*).
3. Uživatel si připodobní novou komponentu s objektem z okolního světa, který už zná; vznikne nová asociace (*Nová asociace*).
4. Uživatel od tohoto okamžiku rozpozná komponentu automaticky (*Rozpoznání*).



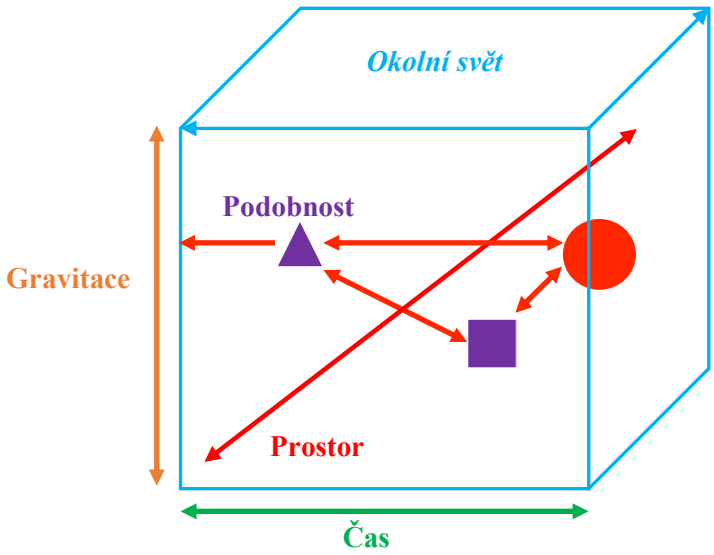
Obrázek 6: Princip archetypu Podobnost. Zdroj: autor, 2017

Z archetypu *Podobnost* vychází uživatelská rozhraní *Seznam, Pracovní plocha uprostřed, Grafika místo textu, Text místo grafiky, Grafika a text a Dlaždice*.

Samostatnou skupinou podobnosti je obor psychologie barev a asociace konkrétních barev dle jejich působení na člověka. Tato doména je natolik obsáhlá, že jí je věnována samostatná kapitola v rozpracované disertační práci autora (Hradil, 2017), nicméně již nyní lze doporučit publikaci na téma *Mapování barev na typické případy užití ve webové aplikaci* (Hradil a Sklenák, 2015).

Vztahy mezi archetypy

Vzájemné vztahy mezi základními archetypy *Čas*, *Gravitace*, *Prostor a Podobnost* lze graficky znázornit jako (Obrázek 7). Jak je zřejmé, archetypy navzájem koexistují, stejně jako jejich odrazy v reálném světě; gravitace je spojena s časem a prostorem a podobnost nastává v prostoru. Archetypy také působí na sebe navzájem. Při mapování vzorů je tedy možné, že jedno uživatelských rozhraní lze přiřadit k více archetypům najednou; jedná se pak o tzv. kompozitní (složené) rozhraní.



Obrázek 7: Vztahy mezi archetypy Zdroj: autor, 2017

Identifikace vzorů uživatelských rozhraní

Na základě rešerší systémů, publikací a doporučení (Apple, 2017a; Apple, 2017b; Ghisler Software GmbH, 2016; Google, 2017a; Google, 2017b; Google 2017c; Krug, 2014; Microsoft 2017a; Microsoft 2017b; Peter Norton Computing, 1986) byly identifikovány typické vzory uživatelského rozhraní a bylo provedeno jejich mapování na uvedené základní archetypy.

Tabulka 1: Přehled vzorů uživatelských rozhraní Zdroj: autor, 2017

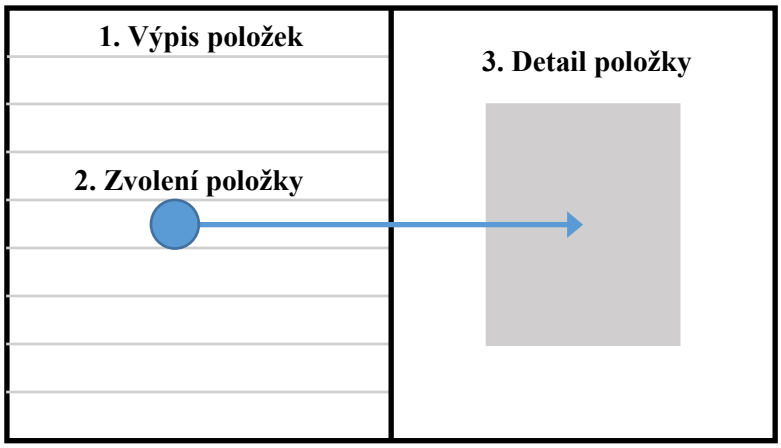
Název	***Popis***	***Archetyp***
2-sloupcový horizontální layout	Obrazovka je rozdělena na 2 části horizontálně.	*Čas.*
3-sloupcový horizontální layout	Obrazovka je rozdělena na 3 části horizontálně.	*Čas.*
Časová osa	Grafické zobrazení času zleva-doprava. Vlevo jsou starší entity, vpravo novější.	*Čas.*
Seznam	Výpis položek shora-dolů. Podobnost s papírovým seznamem úkolů.	*Gravitace, Podobnost.*
Pracovní plocha uprostřed	Pracovní část je umístěna uprostřed obrazovky. Podobnost s pracovním stolem.	*Prostor. Podobnost.*
Důležitější je větší	Důležitější, často používané nebo kritické prvky jsou větší než ostatní.	*Prostor.*
Skupina	Důležitější, často používané nebo kritické prvky jsou umístěny ve skupině o velikosti 1...N dále od ostatních prvků.	*Prostor.*
Grafika místo textu	Grafické vyjádření má přednost před textovým.	*Podobnost.*
Text místo grafiky	Textové vyjádření má přednost před grafickým.	*Podobnost.*
Grafika a text	Kombinace grafického a textového vyjádření.	*Prostor. Podobnost.*
Menu vlevo	Ovládací prvky jsou umístěny nalevo od pracovní plochy, hlavní ovládací prvky.	*Čas.*
Menu vpravo	Menu je doplňkovou částí aplikace / ovládá pravou část aplikace.	*Čas.*
Dlaždice	Zobrazení v mřížce. Drží tvar, jeví se uspořádaně.	*Podobnost.*

2-sloupcový horizontální layout

V tomto vzoru je obrazovka rozdělena na 2 části, které jsou umístěny vedle sebe. Zobrazení následuje archetyp *Času* – pak levá část zobrazuje data starší (méně důležitá) a pravá novější (více důležitá), nebo jsou části v důležitosti ekvivalentní; pak jde typicky o srovnávání dat či jejich duplicitní zobrazení.

Příklady použití:

- Vlevo výpis položek, vpravo detail vybrané položky (Obrázek 8).
- Vlevo minulost, vpravo současnost.
- Vlevo hrubý filtr, vpravo výpis dle zvoleného filtru.
- Zobrazení 2 položek či seznamů vedle pro srovnání – revize 1 dokumentu, 2 rozdílné dokumenty, popř. duplicitní zobrazení týž dat, např. *Norton Commander* (Peter Norton Computing, 1986) a *Total Commander* (Ghisler Software GmbH, 2016)
- Typicky procházení adresáře, zobrazení náhledu obrázku.



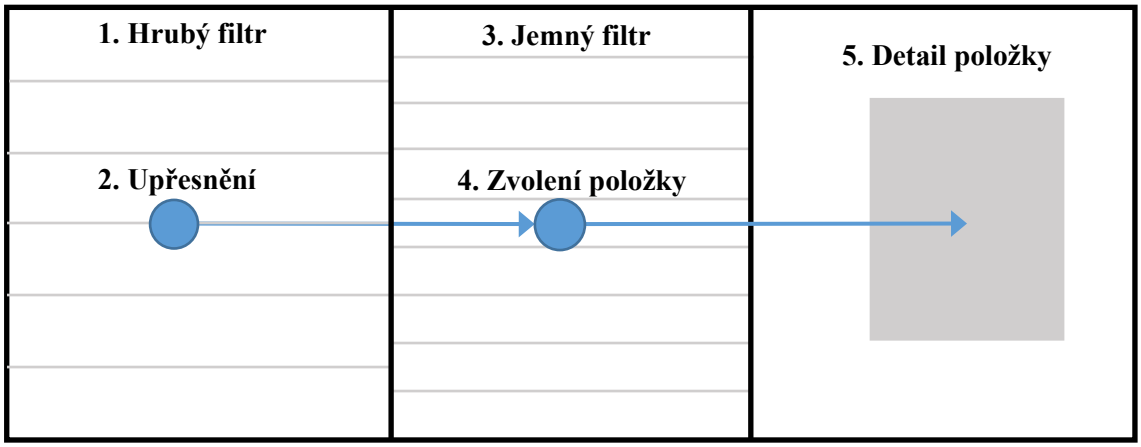
Obrázek 8: Příklad vzoru 2-sloupcový horizontální layout Zdroj: autor, 2017

3-sloupcový horizontální layout

Tento vzor je rozšířením 2-sloupcového horizontálního layoutu a přidává další úroveň podrobnosti. Zobrazení je ve formě 3 sloupců vedle sebe. Pokud zobrazení následuje archetyp *Času,* pak levá část zobrazuje minulost, střední část současnost a pravá budoucnost. Lze také následovat zobrazení z pohledu granularity (zrnitosti) – vlevo jsou nejhrubší (*coarse-grained*) data, uprostřed data v běžném rozlišení a vpravo data jemná (*fine-grained*).

Příklady použití:

- Vlevo hrubý filtr, uprostřed výpis položek, vpravo detail položky (Obrázek 9).
- Vlevo minulost, uprostřed současnost, vpravo budoucnost.
- Procházení adresáře ve 3 či více úrovních, např. *Apple Finder* (Apple, 2017b).



Obrázek 9: Příklad vzoru 3-sloupcový horizontální layout Zdroj: autor, 2017

Časová osa

Vzor *Časové osy* lze rozpoznat všude tam, kdy je třeba rychlá navigace v čase, zejména pro rychlý návrat do historie (tlačítko Zpět ve webových prohlížečích) či stále viditelnou stopu pozice v čase, ať už v procházení pomocí drobečkové navigace (Krug, 2014) či ve změnách provedených v čase (např. viditelnost změn dokumentu v revizích).

Příklady použití:

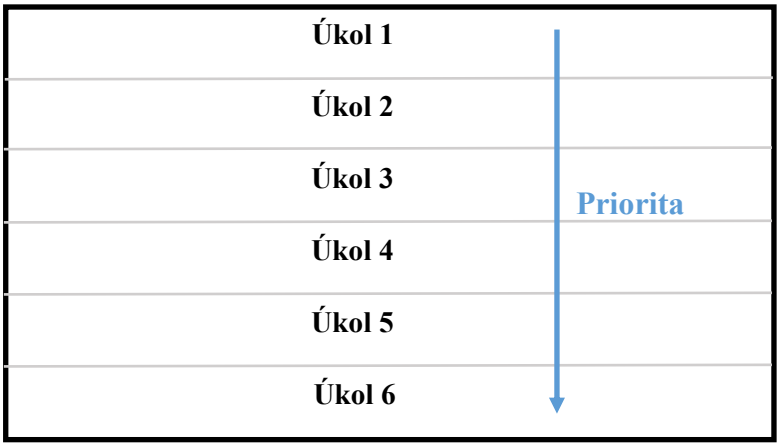
- Drobečková navigace.
- Revize dokumentu.
- Zálohy v čase.
- Průvodce v aplikaci (wizard).
- Rozdělení formuláře na více kroků, kdy je zřejmý časový postup a informace kolik ještě zbývá vyplnit.

Seznam

Vzor *Seznam* má charakteristiky archetypu *Gravitace* a je reprezentován formou vertikálního výpisu položek, typicky dle priority (nejdůležitější či nejnovější položka nahoře). Jedna položka seznamu obsahuje jednu zapouzdřenou informaci (např. email, dokument). Pokud má seznam více položek, je třeba posouvat jej dolů (scrolling) či stránkovat. Položky seznam je vhodné oddělovat mezerou či horizontální čárou, např. ve spojení se vzorem *Skupina*.

Příklady použití:

- Seznam úkolů – úkoly s největší prioritou nahoře (Obrázek 10).
- Seznam e-mailů – nejnovější nahoře
- Adresář kontaktů – seřazené dle abecedy.



Obrázek 10: Příklad vzoru Seznam Zdroj: autor, 2017

Pracovní plocha uprostřed

Vzor *Pracovní plocha uprostřed* je kompozitním vzorem odvozeným od archetypů *Prostor a Podobnost.* Typická reprezentace vzoru je maximalizované okno, jehož střední část je dominantní a uvozuje hlavní pracovní prostor. Zobrazení není rozděleno na více částí a hlavní pozornost je zaměřena na střed okna. Střed je nejbezpečnější část, protože má největší vzdálenost od okrajů (objekty ve středu nemůžeme „ztratit“, protože nespadnou mimo pracovní plochu). Podobnost s pracovním stolem je zřejmá.

Příklady použití:

- Jednoduché aplikace s minimalistickým ovládáním.
- Prohlížení fotografií, map, dokumentů.
- Grafické a textové editory (Obrázek 11).
- Kritické aplikace pro jeden konkrétní případ užití (např. vypnutí všech systémů najednou v případě nebezpečí), kde nelze rozptylovat pozornost jinými ovládacími prvky.

Obrázek 11: Příklad vzoru Pracovní plocha uprostřed Zdroj: autor, 2017

Důležitější je větší

Velikostí lze vyjádřit důležitost. Větší objekt strhává více pozornosti, protože okupuje větší plochu zorného pole, která je omezená. Větším pokrytím zorného pole lze vyjádřit převahu nad objekty menšími. Vzor *Důležitější je větší* lze kombinovat s *Pracovní plochou uprostřed,* protože větší objekty zabírají větší plochu. Dle (Javadi, Gebauer a Mahoney, 2013), zvýšení viditelnosti (prvků) zvětšuje kognitivní zátěž (cognitive load) a míru aktivace znalostí uživatele (knowledge activation). Prvek se větší mírou viditelnosti stává důležitějším.

Velikost lze ovlivnit také barvami. Jak uvádí (Birren, 2013), bílé a světlé objekty se jeví větší, než ve skutečnosti jsou, naopak objekty s tmavými barvami jsou menší. Pokud tedy existuje více stejně velkých objektů, lze jejich důležitost rozlišit také barvami. Vnímání barev je také podřízeno sociálnímu, náboženskému či geografickému kontextu (Löffler, 2015).

Příklady použití:

- Jakékoli rozhraní, kde je třeba upozornit na ovládací prvek.

Skupina

Skupina prvků obsahuje prvky s podobnými či unikátními vlastnostmi, ať to jsou jejich atributy, význam či četnost použití. Skupina vymezuje své členy oproti prvkům ostatním. Důležitost prvků ve skupině je vykoupena prostorem kolem této skupiny, který nelze využít jinak. Důležitost prvků, jejich unikátnost a vztah k prvkům ve stejné skupině je jedním z případů užití tohoto vzoru. Skupinu lze také rozdělit na podskupiny (a tím vytvořit skupiny nové). Například ve vzoru *Seznam* lze opticky zmenšit počet prvků v seznamu přidáním mezery mezi každý pátý prvek. Volné místo mezi prvky seznam odlehčí a manipulace s ním bude zjednoduší; uživatel již nemyslí jednotlivé prvky, ale skupiny, kterých je méně, než celkový počet prvků. Jak uvádí (Anderson, 2005), člověk si dokáže okamžitě vybavit, zpaměti zopakovat, a tím i uvažovat kolem 7 prvků najednou. Velikost skupiny by tedy neměla překročit tuto mez.

Příklady použití:

- Jakékoli rozhraní s více ovládacími prvky.
- Rozdělení položek seznamu do skupin maximálně 7 položek.
- Zvýraznění důležitosti prvků.
- Zpřehlednění a „odlehčení“ rozhraní.

Grafika místo textu

Vzor *Grafika místo textu* znamená, že grafické vyjádření má přednost před vyjádřením textovým. Výhodou použití grafiky (symbolu, obrázku, ikony) oproti textu je, že vnímání grafiky člověkem je rychlejší, protože okolní svět vnímáme také obrazově; navíc lze odkázat na archetyp *Podobnosti* a využít existující asociace použité grafiky k objektům okolního světa. Obrázek lépe a rychleji vystihuje podstatu než text. Text je třeba skládat dle písmen do slabik a slov a teprve potom použít asociaci dle naučeného slovníku čteného jazyka. Navíc při použití obrázku nevzniká problém s vícejazyčným prostředím oproti textu, který se musí překládat. Pokud se rozhodneme použít grafiku pro ovládací prvky, je třeba, aby grafika byla jednoznačná. Zatímco text je čten pořád stejně, obrázek může být pochopen rozdílně. Je tedy vhodné držet se zavedených vzorů, např. pro ikonu kopírování použít známé 2 stránky vedle sebe, pro nový dokument ikonu prázdné stránky apod.

Pro inspiraci je vhodné shlédnout více typů aplikací, které jsou stejného zaměření jako nově tvořená aplikace a pokud se jedná o známou aplikaci v doméně řešení, např. *Microsoft Word* (Microsoft, 2017b) v případě textových editorů, pak použít její ikony jako výchozí inspiraci pro ikony nové. Samozřejmě není nutné použít naprosto stejné ikony (i kvůli licenčním omezením), ale symboly by měly být alespoň podobné. Tímto způsobem bude nová aplikace uživatelům povědomá a proces adaptace a přijetí bude rychlejší a vřelejší. Dojde k asociaci dle archetypu *Podobnost*.

Nevýhodou použití vzoru *Grafika místo textu* je možná nejednoznačnost grafického vyjádření v porovnání s textovým vyjádření. Grafické symboly mohou být v různých kulturách, typech aplikací či kontextech vnímání odlišně. Překážkou také může být strojové ovládání uživatelského rozhraní, protože grafika postrádá sémantický význam a jeho detekce je náročnější, než u prostého textu (bylo by třeba zapojit metody strojového učení pro rozpoznávání grafiky apod.).

Použití vzoru:

- Aplikace poskytované pro vícejazyčná prostředí (není třeba překládat text).
- Aplikace pro děti, hry.
- Designové aplikace, u kterých je klíčové grafické vyjádření.
- Obecně grafické vyjádření, které může být konkurenční výhodou.
- Aplikace jako umělecké dílo.

Text místo grafiky

Vzor *Text místo grafiky* upřednostňuje textovou reprezentaci před grafickou. Používá se v případě, kdy je nutná naprostá jednoznačnost prvků a přesné předání informace nebo v případě, kdy je funkce ovládacího prvku tak specifická, že ji nelze graficky jednoznačně vyjádřit. Roli zde může hrát také přenosová kapacita – textové vyjádření je datově úspornější než grafické vyjádření. Text používá font (typ písma) a font samotný je grafickým vyjádřením. Proto je třeba použít takový font, který bude dobře čitelný a rozlišit, zda bude patkový či bezpatkový. Pro delší text či dokument je třeba použít patkový font, protože horní a spodní patka (malá horizontální čára) opticky tvoří řádek a čtení je tímto rychlejší a pohodlnější. Naopak v nadpisech či krátkých slovech patky ruší – jsou něco „navíc“ a je vhodné je eliminovat.

Nevýhodou použití vzoru *Text místo grafiky* je grafická podobnost textové reprezentace, a tudíž obtížné rozlišení textových ovládacích mezi sebou. Stejně jako u vzoru *Důležitější je větší* lze pro rozlišení použít barvy. Kritické ovládací prvky či důležité komponenty mohou být červené, protože červená barva je dle (Elliot a Maier, 2014) vnímána jako nebezpečí, a tedy upoutá pozornost dříve než ostatní barvy. Červený text bude vidět dříve než text černý.

Použití vzoru:

- Jednoduché aplikace.
- Aplikace nepoužívající barvy či grafiku.
- Aplikace, ve kterých musí být vyjádření jednoznačné.

Grafika a text

Vzor *Grafika a text* je kompozitním vzorem *Grafika místo textu* a *Text místo grafiky*. Obsahuje ideální kombinaci – používá grafické symboly pro rychlou orientaci a text jako vysvětlující či potvrzující doplnění. Kombinace je vhodná pro nové uživatele, u kterých adaptace na nový systém proběhne rychleji než v případě použití vzoru *Text místo grafiky* (grafické vyjádření je vnímáno rychleji než text) nebo vzoru *Grafika místo textu* (textové doplnění přidá ujištění o významu ovládacího prvku). Nevýhodou vzoru *Grafika a text* je větší prostorová náročnost a možné rušivé vnímání textu v případě typických případů užití používaných napříč různými typy aplikací (např. ikona tiskárny pro tisk).

Použití vzoru:

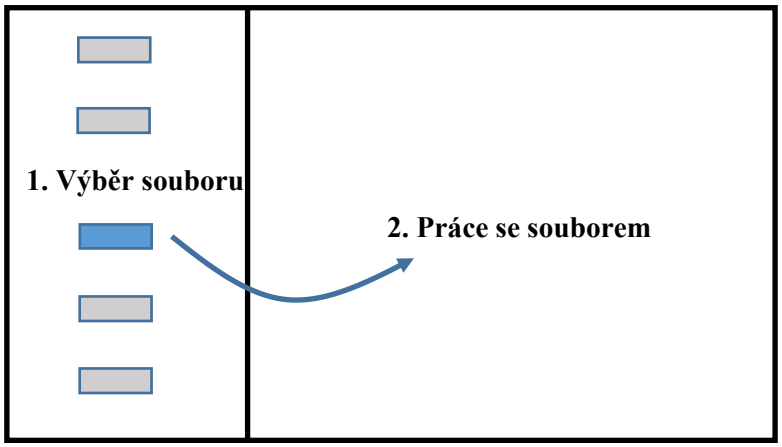
- Jakákoli aplikace, která může používat grafiku a má dostatek prostor pro přidání dodatečného textu.

Menu vlevo

Vzor *Menu vlevo* je vyvozen z archetypu *Čas* a je speciální variantou vzoru *2-sloupcový horizontální layout.* Menu aplikace je umístěno vlevo. Předpokládá se tedy, že je dominantní a bude použito dříve, než hlavní (středová) část aplikace.

Použití vzoru:

- Menu je startovním bodem práce v aplikaci.
- Menu je často využíváno a musí být viditelné a k dispozici.
- Základní ovládací prvky.
- Primární akce – např. výběr souboru a práce s ním (Obrázek 12).



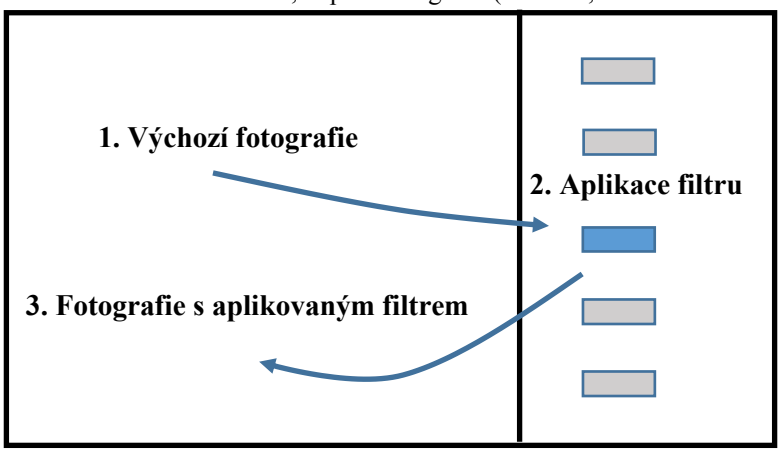
Obrázek 12: Příklad vzoru Menu vlevo Zdroj: autor, 2017

Menu vpravo

Vzor *Menu vpravo* také následuje archetyp *Čas*, ale menu je zde pouze doplňkovou částí aplikace. Pravá část je dle vnímání času rozpoznána později než část levá. Menu proto obsahuje komponenty, které jsou využívány s menší četností než ve vzoru *Menu vlevo*.

Použití vzoru:

- Rozšiřující ovládací prvky.
- Sekundární, doplňující akce, např. zobrazení fotografie a následná aplikace filtru (Obrázek 13).
- Pokračování levé/středové části.
- Meta-informace k zobrazenému detailu, např. k fotografii (citlivost, ohnisková vzdálenost atd.).



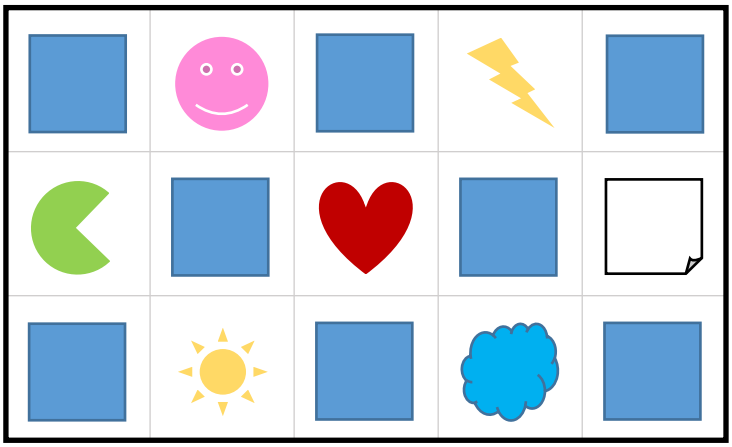
Obrázek 13: Příklad vzoru Menu vpravo. Zdroj: autor, 2017

Dlaždice

Vzor *Dlaždice* dle archetypu *Podobnost* umísťuje komponenty uživatelského rozhraní do mřížky. Název evokuje podobnost s dlaždicemi na chodníku. Efektem mřížky je uspořádanost, pokud jsou alespoň v jednom směru (horizontálním či vertikálním) stejné rozestupy mezi komponentami.

Příklady použití:

- Seznam obrázků, např. *Google Images* (Google, 2017b) nebo (Obrázek 14).
- Práce s aplikacemi v mřížce, např. *Microsoft Design Language* (Microsoft, 2017b), dříve označované jako *Metro*.
- Textové vyjádření dat v tabulce.



Obrázek 14: Příklad vzoru Dlaždice. Zdroj: autor, 2017

Závěr

V článku byly vybrány čtyři základní archetypy (fenomény) pro identifikaci uživatelských rozhraní a následně popsáno třináct vzorů rozhraní včetně doporučení případů užití. Smyslem článku je ukázat propojenost software s okolním světem prostřednictvím uživatelského rozhraní, které se fenomény okolního světa inspiruje. Lidé znají popsané archetypy velmi dobře a chápou minimálně jejich základní vztahy. Pokud tuto naučenou zkušenost myšlenkově přeneseme do uživatelských rozhraní při tvorbě software, zkrátíme dobu adaptace uživatelů na nový software, jeho ovládaní bude jednodušší a v konečném důsledku bude software sloužit lidem lépe.

Vybrané archetypy a identifikované vzory uživatelských rozhraní v práci autora nelze považovat za kompletní, spíše za prvotní a inspirující. Úroveň jejich podrobnosti či abstrakce lze kdykoli změnit dle aktuální potřeby. Pokud se však autoři rozhraní naučí tvořit či vylepšovat uživatelská rozhraní dle vzorů okolního světa, budou rozhraní v důsledku lepší, protože hranice vzorů již byly vytyčeny a jsou všeobecně známé. Člověk při používání takového odvozeného rozhraní nevstupuje do nového světa, ale do světa důvěrně známého.

Cílem dalšího výzkumu je pokračovat v hledání archetypů (či fenoménů), mapovat je na existující uživatelská rozhraní, snažit se vylepšovat návrh a používání rozhraní a tím finálně zlepšovat kvalitu software.

Literatura

ANDERSON, J.R, 2005. *Cognitive psychology and its implications*. New York, NY: W. H. Freeman. 6. vydání. ISBN 978-1429219488.

APPLE, 2017a. *iOS Human Interface Guideline* [online]. [přístup 2017-01-12]. Dostupné z: https://developer.apple.com/ios/human-interface-guidelines/overview/design-principles/

APPLE, 2017b. *Finder 10.12.3, macOS Sierra* [software]. [online]. [přístup 2017-01-12]. Dostupné z: http://www.apple.com/cz/macos/sierra/

BIRREN, Faber. *Color Psychology and Color Therapy*. Martino Fine Books (November 4, 2013). [cit. 2017-01-12]. ISBN 978-1-61427-513-8.

CHAPANIS, Alphonse, 1999. *The Chapanis Chronicles: 50 Years of Human Factors Research, Education, and Design*. Aegean Pub Co. 1999. ISBN: 978-0963617897.

ELLIOT, A. J. a M. A. MAIER, 2014. Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans. Annual Review of Psychology 65 (2014). [cit. 2017-01-13]. Dostupné z: http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-010213-115035

FREEMAN, Eric, Elisabeth ROBSON, Bert BATES, Kathy SIERRA, 2004. *Head First Design Patterns*. O'Reilly Media. 2004. ISBN 978-0-596-00712-6.

GHISLER SOFTWARE GMBH, 2016. *Total Commander 9.0a* [software]. [online]. [cit. 2017-01-12]. Dostupný z: http://www.ghisler.com

GOOGLE, 2017a. *Google Design* [online]. [přístup 2017-01-12]. Dostupné z https://design.google.com

GOOGLE, 2017b. *Google Images* [software]. [online]. [přístup 2017-01-12]. Dostupné z: https://images.google.com/

HRADIL, Jiří, 2017. *Ergonomie software a trendy dlouhodobé udržitelnosti uživatelského rozhraní.* Disertační práce. Fakulta informatiky a statistiky Vysoké škola ekonomická v Praze, Katedra informačního a znalostního inženýrství. Vedoucí disertační práce Vilém Sklenák. ***Rozpracováno***.

HRADIL Jiří a Vilém SKLENÁK, 2015. *Mapování barev na typické případy užití ve webové aplikaci* [online]. Acta Informatica Pragensia. 2015. sv. 4, č. 2, s. 154-173. [přístup 2017-01-12]. ISSN 1805-4951. Dostupné z: http://aip.vse.cz/index.php/aip/article/view/105/82.

JAVADI, Elahe, Judith GEBAUER a Joseph MAHONEY. The Impact of User Interface Design on Idea Integration in Electronic Brainstorming: An Attention-Based View. Journal of the Association for Information Systems [online]. 2013, vol. 14, no. 1, s. 1-21. ISSN 15369323.

KRUG, Steve, 2014. Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability (3rd Edition) (Voices That Matter). New Riders. 3. vydání. 2014. ISBN 978-0321965516.

LÖFFLER, Diana, 2014. *Happy is pink: designing for intuitive use with color-to-abstract mappings*. In: CHI '14 Extended Abstracts on Human Factors in Computing Systems (CHI EA '14). ACM, New

York, NY, USA, s. 323-326. [cit. 2017-01-05]. Dostupné z: http://doi.acm.org/10.1145/2559206.2559959

MICROSOFT, 2017a. *Microsoft Design Language* [software]. [přístup 2017-01-12]. Dostupné z: https://developer.microsoft.com/en-us/windows/apps/design

MICROSOFT, 2017b. *Microsoft Word. Office 365* [software]. [přístup 2017-01-01]. Dostupné z: https://products.office.com/en-us/business/office-365-business

PETER NORTON COMPUTING, 1986. *Norton Commander* [software]. [cit. 2017-01-12].

THOMSON, Helena, 2014, *Past is a blur if the right side of your brain is faulty*, New Scientist, 221, 2950, s. 13, Academic Search Complete, EBSCOhost, [přístup 2017-01-09]. ISSN 0262-4079.

JEL Classification: C88, C90, O33, L86

Summary

Software Ergonomics – Identification and Origin of User Interface Patterns

This paper identifies typical patterns of user interfaces in software applications and try to conclude their origin based on association with phenomena of the surrounding world. First, four basic archetypes are chosen (*Time, Gravity, Space and Similarity*) and then related user interfaces are assigned to them. Each user interface is described and put into the right usability context. Author believes that when we identify and understand these connections between archetypes and user interfaces, and follow the concept of inspiration from the outside world, we will be able to create more usable software which will serve people better. The user interface patterns research is just a part of author's broader research in software ergonomics field.

Keywords: software ergonomics, user interface, user interface pattern, archetype, layout.

Text mining na sociálních sítích: analýza politické události

Ivan Jelínek
Ivan.jelinek@vse.cz
Doktorand oboru aplikovaná informatika
Školitel: prof. Ing. Jaroslav Jandoš, CSc., jandos@vse.cz

Abstrakt: Nástroje pro zpracování dat ze sociálních sítí lze aplikovat v různých oblastech. Od sledování veřejného mínění, přes měření úspěšnosti marketingových kampaní po rychlou reakci na aktuální trendy mezi zákazníky. Tyto aplikace jsou kritické jak pro vysoký management podniků, tak o pro vrcholové vedení politických stran. Metoda analýzy dat ze sociálních sítí by tudíž pro oba sektory měla být v podstatě shodná.

Data dostupná z Facebooku a ostatních médií proto byla analyzována s cílem zjistit, jak jsou obě informační platformy propojeny. Obsah ze sociálních sítí byl zpracován nástrojem vyvíjeným na VŠE využívajícím open-source řešení a vlastních knihoven s využitím text miningových technik a srovnán s daty z ostatních médií.

Analýza prokázala silnou závislost mezi obsahem sociálních sítí a ostatních médií. Tato závislost kromě samotných četností byla prokázána i na úrovni obsahu příspěvků z Facebooku skrze shlukovou analýzu dat.

Odhadnout dopady politických událostí na uživatele je kritické pro osobnosti z řad politických stran. Metoda analýzy dat, která je připravovaná v rámci dizertační práce, byla proto uplatněna v této analýze.

Klíčová slova: sociální sítě, text mining, politická událost, publicistika.

Úvod

Internet a především sociální sítě se staly médiem pro sdílení informací široké části populace. Lidé si skrze měj vyměňují svá stanoviska, názory a dojmy. V první řadě se tak děje textem – příspěvky, které uživatelé publikují soukromě nebo veřejně – nebo jinými způsoby odvislými od dané sociální sítě (retweet na Twitteru, like na Facebooku, ...).

Pro analýzu mínění uživatelů je proto nutné zapojit více technik datové analýzy. K analýze textu použít nástroje a techniky text miningu, k mimotextovým komunikacím jiné. Nástroje, které by podobnou analýzu provedli jsou podstatné v různých situacích, kdy se sociální sítě tvoří platformu pro komunikaci sledovaných lidí.

V případě marketingových kampaní lze tak sledovat dopad na zákazníky firmy, na to co o firmě sdílejí, jak je kampaň živá a úspěšná. Je možné následně reagovat na vzniklou situaci, upravit podnikové procesy, zvyšovat prodej, podpořit firemní CSR, apod.

Sociální sítě ale představují odlišnou realitu od té fyzické. Kromě této je vhodné sledovat i odraz v ostatních médiích. Cílem nemusí být jen analýza marketingové kampaně, ale podobným způsobem je možné analyzovat i jiné oblasti sociálních sítí a veřejného života – politickou scénu.

Motivace a popis řešeného problému

Ve veřejném prostoru představují média zdroj informací pro společnost. Jsou prostředkem, který do jisté míry determinuje společenský diskurz k různým otázkám. V politické oblasti, která je vděčným objektem ostatních médií, dominují osobnosti a politické události. Proto pro aktéry v politickém prostoru je podstatné sledovat aktuální obsah médií k odhadu dopadu politické události na společnost.

Během několika posledních let došlo k prudkému rozvoji sociálních sítí, které se staly cenným zdrojem informací o názorech obyvatel, voličů. Existuje-li vztah mezi informacemi na sociálních sítích a v ostatních médiích, můžeme z toho vyvozovat, že oba tyto světy do jisté míry odráží aktuální společenskou debatu. Bude tedy možné sledovat jak společnost reaguje na aktuální události, o kterých média referují.

Ačkoli již existují metody pro průzkum veřejného mínění spadající do oblasti sociologie, představují sociální sítě další podstatný zdroj dat. Metody zpracování textových dat nebo analýzy vazeb mezi uživateli proto mohou přinést informace jiným způsobem nezjistitelné a podložené daty.

Pokud tedy existuje kauzální vztah mezi oběma světy, bude možné sledovat dopad politické události a podle potřeby reagovat na aktuální vývoj rychleji a přesněji než na podkladě sociologických šetření.

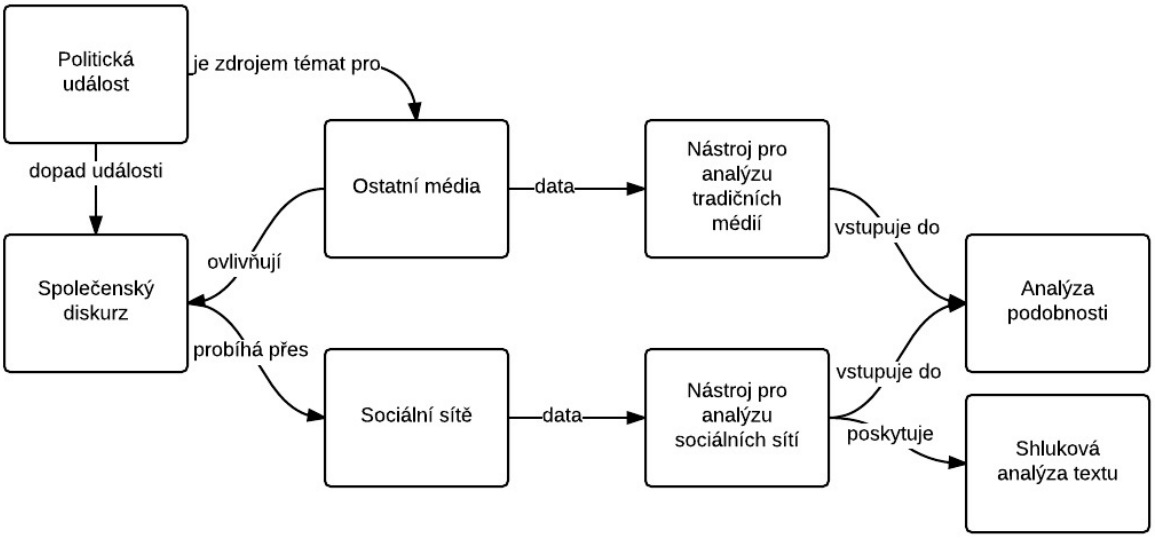
Výzkumné otázky

- RQ1: Jak se projevuje vazba mezi obsahem v ostatních médiích a na sociálních sítích?
  - Analýza hledající odpovědi k této výzkumné otázce leží především ve srovnávací frekvenční analýze obou zdrojů a shlukovací analýze dat ze sociálních sítí.
- RQ2: Jaké metriky je možné použít pro sledování dopadu politické události na sociálních sítích?
  - Dopad politické události je sledován pomocí dvou klíčových metrik, které sociální sítě nabízejí. Jejich použití ilustruje na sledované události jak reaguje svět sociálních sítí. Shlukovací analýzou dat je možné dále odhadnout co je obsahem diskurzu sociálních sítí.

Výzkumný model

Smysl použitých analytických metod je blíže popsán v politická událost. Události jsou klíčovými vstupy pro ostatní média, která o nich referují a utváří tak seznam témat pro veřejný diskurz. Ten je klíčový pro politické aktéry. Skrze ostatní média ale nemají žádné povědomí o tom, jak je veřejností události vnímána.

. Klíčovou událostí je zde politická událost. Události jsou klíčovými vstupy pro ostatní média, která o nich referují a utváří tak seznam témat pro veřejný diskurz. Ten je klíčový pro politické aktéry. Skrze ostatní média ale nemají žádné povědomí o tom, jak je veřejností události vnímána.



Obr. 1 Výzkumný model zdroj: autor

Společenský diskurz je do jisté míry přítomen i na sociálních sítích, které dovolují svým uživatelům zprostředkovat své názory širšímu publiku. Data z ostatních médií jsou zpracovávána nástrojem Anopress pro monitoring médií (blíže část Zpracování dat z ostatních médií). Sociální sítě jsou zpracovány nástrojem pro analýzu sociálních sítí (blíže část Zpracování dat ze sociálních sítí).

Výstupy z obou nástrojů jsou následně porovnány vůči sobě (blíže část Výsledky analýzy) pro zodpovězení RQ1. Pro lepší charakteristiku obsahu příspěvků na sociálních sítích je analýza doplněna o shlukovou analýzu textu (blíže tamtéž).

Pro zodpovězení RQ2 se použijí metriky, které sleduje nástroj pro analýzu sociálních sítí (blíže tamtéž).

Sledovaná událost

Pro odhalení vazby mezi sociálními sítěmi a ostatními médii je vhodné najít událost, která měla jasně časově ohraničený dopad. Jelikož média obvykle informují pouze o aktuálních událostech velkého dopadu, hledají v nich aktéry aby přinesly jejich stanoviska. Budou-li média referovat o aktérovi, který není často diskutovaný a stojí mimo hlavní politický proud a společenský diskurz, bude vazba mezi oběma informačními zdroji viditelnější.

Proto byla analyticky sledována událost vázající se k europoslanci Miloslavu Ransdorfovi, který byl 3. prosince 2015 zatčen v Kantonalbank Zürcher při předložení falešných dokumentů. Pan Ransdorf byl následně vyslýchán kriminální policií, zemřel v Praze 22. ledna 2016 (Kotalík 2015).

Sledovaná událost měla dopad především před koncem roku 2015. Analyzovaná data proto byla sbírána za celý rok 2015. Impakt do sociálních sítí a ostatních médií ležel především v prosinci 2015.

Sociální sítě

Dle definice Boyd a Ellison (Boyd & Ellison 2007) jsou stránky sociálních sítí (*social network sites*) definovány jako webové služby, které dovolují jedincům:

- vytvořit si veřejný nebo poloveřejný profil v rámci sítě,
- vytvořit seznam dalších uživatelů se kterými sdílím spojení a
- vidět a procházet jejich seznamy spojení.

Sociální sítě (*social network*) jako takový je ale pojem obecnější, jelikož jde jen o propojení lidí bez ohledu na použité médium. Pro účely této práce se uvažuje pod termínem sociální síť její webové podoba spíše odpovídající pojmu *social network site*, jelikož analyzovat je možné pouze zdroje jako: Facebook, Twitter, Sina Weibo, QQ, VKontakte, ...

V České republice je dle analýzy firmy Socialbakers (2017) ze sledovaných sociálních sítí v různých statistikách vedoucí sociální sítí Facebook. Další sociální sítě jako Twitter nebo YouTube nejsou v ČR tak populární. Z tohoto důvodu je vhodné pro analýzy sociálních sítí vybrat především Facebook, jelikož jeho dopad je nejvýraznější.

Relevantní text miningové úlohy

Zpracování přirozeného jazyka

Dalším dílčím úkonem při analýze nestrukturovaných dat ze sociálních sítí je zpracování přirozeného jazyka pro účely vyhledávání. Jedna z klíčových metod pro zpracování slov je *stemming*[1]. Stemováním českého jazyka se zabývá se své práci (Dolamic & Savoy 2009), kde navrhuje neagresivní verzi stemeru. Jejich navrhovaný stemingový algoritmus Dolamica a Savoye se zaměřuje pouze na přípony a koncovky českých slov – resp. některých tvarů slovních druhů. Tento přístup je sice v pořádku, není ale komplexní jako např. v Porter stemming algoritmu (Willett 2006). Chybí zde například rozpoznávání předpon, které navrhovaný stemmingový algoritmus opomíjí.

Jisté možnosti jak přistupovat k stemování nabízí ve své práci Strossa (2011), který kromě heuristického přístupu nabízí slovníkový, tezaurový přístup.

Stemming jako činnost zpracování dat je ve své podstatě jazykově závislá. Je jasné, že pravidla pro ořezávání slov se budou lišit podle jazyka. Jistou specifičnost v této oblasti mají čínština, japonština a korejština, které se svojí morfologií a lexikologií liší od indoevropských jazyků (i od sebe navzájem). Jelikož ale popisovaný datový korpus obsahuje pouze češtinu, výjimečně slovenštinu, je možné použít pro zpracování slov heuristický přístup, který oproti tezaurům nepotřebuje aktuální znalost analyzovaného prostředí.

Shluková analýza dat

Shlukovací analýza je technika analýzy dat, která si klade za cíl

- rozdělit pozorování (data) na základě jejich podobnosti do shluků,
- odvodit popis celého shluku na základě společné charakteristiky.

Toho je možné docílit různými algoritmy. Nejobvyklejší je varianta k-středového algoritmu, který seskupuje vektory n-dimenzionálního prostoru do skupin tak, aby součet čtverců jejich rozdílů byl nejmenší. Charakteristika skupiny je pak dáma průměrem jeho členů (Řezanková et al. 2009).

Kromě k-středového algoritmu je k dispozici i algoritmus STC (Branson & Greenberg 2002) nebo algoritmus LINGO (Stefanowski & Weiss 2003).

Výstupem všech těchto metod je ale v případě shlukové analýzy textů seznam skupin shluků s jejich charakteristikou. Tuto charakteristiku proto můžeme vnímat jako témata, která se nápadně často opakují v datech.

[1] Stemming je úloha zpracování slova, kdy podle definovaného algoritmu se odpojí koncovka slova, případně se upraví kořen slova. Tímto způsobem je možné slova v různých pádech, časech nebo číslech sjednotit pod jedem vyhledávaný termín.

Kategorizace textu

V některých situacích je samotnou analytickou činností vybudování systému pravidel, podle kterých dochází k automatické kategorizaci příchozích dokumentů (dat nebo textu) podle jejich obsahu. Druhy automatické klasifikace je možné rozdělit do tří skupin: na základě předdefinovaných pravidel, statistické, machine learningové (Keretna et al. 2013).

Typickým příkladem machine learningového automatizačního algoritmu je vector space model (VSM) využívající váhu slova v dokumentu – tfidf metriku (Minier et al. 2007). Dalším podobným přístupem může být Naive-Bayes algoritmus, který používá základní bayesův teorém:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

kde A a B jsou události, P(A) a P(B) jsou pravděpodobnosti, že nastane jev A nebo B. Dále P(A|B) je pravděpodobnost, že nastane jev A pokud jev B nastal. Podobně P(B|A) je pravděpodobnost, že nastane jev B pokud nastal jev A. To vše za předpokladu, že jev B má nenulovou pravděpodobnost.

Tohoto teorému lze využít v situaci, kdy je pravděpodobnostní rozložení sledovaného jevu známé (apriorní). Tedy je-li kategorizovaná entita na podkladě trénovacích dat známá. Z této apriorní pravděpodobnosti je možné dovodit výslednou (aposteriorní) pravděpodobnost, že kategorizovaná entita bude patřit do výsledné skupiny.

Sentiment analýza textu

Sentiment analýza je jedním ze způsobů analýzy dat. Jedná se o charakterizaci podle toho, je-li text psán pozitivně, neutrálně nebo negativně. Toho lze dostáhnout různými způsoby. Jednou z možností je vybírání pozitivních nebo negativních slov v textu po jejich zpracování stemmingem. Takový princip ve své práci používá např. Ward a kol. (2011), kteří využívají takto identifikovaný sentiment na úrovni fráze, resp. entity. Celkovou pozitivitu nebo negativitu příspěvku počítají jako:

$$s_i = \frac{pos-neg}{pos+neg},$$

kde s_i je celkový sentiment příspěvku, *pos* je celkový počet pozitivních a *neg* je celkový počet negativních slov.

Kromě tohoto slovníkového přístupu, zpracovávající sentiment na úroveň fráze (položky definované ve slovníku), lze dohledat v literatuře i přístupy zpracovávající sentiment na úroveň kontextu nebo celého příspěvku.

Práce (Kamath et al. 2013) sleduje sentiment na kontext v poměru k metrice *tf-idf*[2]:

$$tf.idf(w, Pi) = \frac{N(w,Pi)}{N(Pi)}.\log(\frac{N}{N(w)}),$$

kde *tf-idf* metrika pro váhu slova *w* v příspěvku *Pi* obohacenou o logaritmickou složku zohledňující výskyt posuzovaného slova v daném dokumentu.

Kromě těchto dvou základních přístupů můžeme pozorovat třetí možnost opírající se o již popsanou kategorizaci. Jelikož je pravděpodobné, že pozitivní příspěvky budou mít společná slova (konečný výčet pozitivní slov) a stejně tak negativní slova budou pro svoji negativitu budou používat stejná slova, je možné kategorizovat příspěvky například podle Naive-Bayes algoritmu.

- Obecný problém analýzy sentimentu se skrývá ve dvou rovinách:
- sarkasmu a ironie,
- slangu a žargonu.

Problém sarkasmu. V běžném prostředí na sociálních sítích se lze často setkat s příspěvky, které jsou do jisté míry sarkastické nebo ironické. Takové chování uživatelů narušuje původní předpoklad analýzy sentimentu – tj. rozřazení pozitivity nebo negativity příspěvku podle jeho pozitivních nebo negativních slov.

Data, která zde máme dostupná (řetězec písmen), která jsou spojována do slov oddělených konkrétními znaky detekovaná podle vybraného způsobu jako pozitivní ale s sebou nesou negativní informaci.[3] Ačkoli se aktuální výzkum zaměřuje na detekci sarkasmu na sociálních sítích (Bamman & Smith 2015; González-Ibánez et al. 2011; Rajadesingan et al. 2015; Davidov et al. 2010), jsou dostupné přístupy zatím nedostatečné.

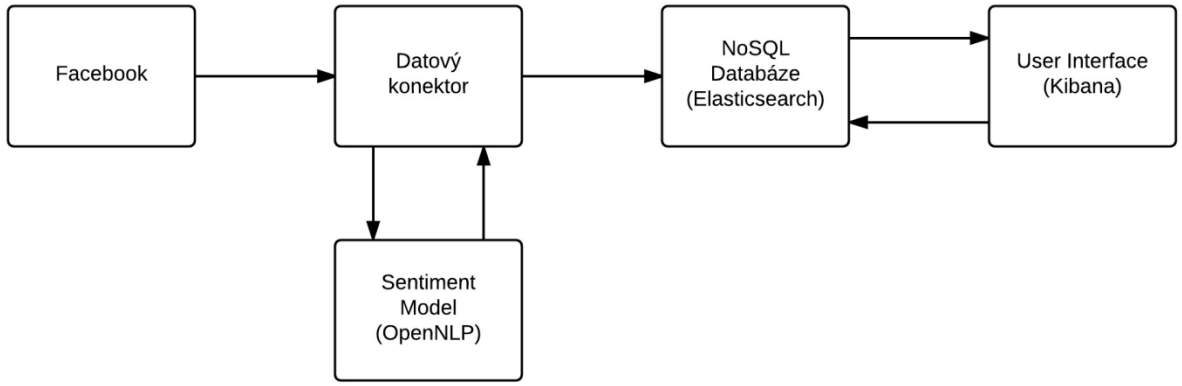
[2] Text frequency – inverse document frequency.
[3] Chápeme-li informaci jako význam, který přikládáme datům.

Problém slangu a žargonu. Slovník spisovné češtiny definuje žargon jako mluvu určité společenské skupiny, nejčastěji pracovního společenství či jako hantýrku nebo slang (Filipec 2010). Velmi specifický žargon, který mohou mít úzké skupiny obyvatelstva, může znesnadnit či úplně negovat snahu o sentiment analýzu příspěvku. Bez přímé znalosti kontextu této skupiny není možné upravit pravidla nebo slovníky pro analýzu sentimentu. U sentiment analýzy pomocí klasifikačních metod je nutné započítat tyto žargonové skupiny do trénovacích dat.

Zpracování dat ze sociálních sítí

Model technologie zpracování dat z Facebooku

Pro praktické ověření vztahu obsahu na sociálních sítí a ostatních médií je nutné připravit řešení, které bude zpracovávat data dostupná z Facebooku. Model navrženého řešení je uveden na Obr. 2.



Obr. 2 Architektura řešení zdroj: autor

Facebook - skrze standardní rozhraní Facebooku je možné přistupovat k datům, která jsou dostupná na veřejné části této sociální sítě. Do rozhraní je možné se autentizovat skrze přístupové tokeny vázající se k uživatelskému účtu. Ten definuje jaká data jsou dostupná. Pro účely této analýzy byly sbírány pouze veřejně dostupná data.

Datový konektor – aplikace založená na platformě Java používající open source knihovny. Datový konektor byl vytvořen autorem za přispění studentů magisterského a bakalářského studia informatiky VŠE. Na základě seznamu sledovaných facebookových stránek jsou data odebírána a dále zpracována komponentou OpenNLP. Po jejím zpracování jsou data uložena do NoSQL databáze Elasticsearch.

Sentiment model – byl implementován na řešení Apache OpenNLP jako kategorizační model obecných textů za použití modelu, který vznikl z trénovacích dat. Výstupem komponenty je ohodnocení textu jako pozitivního, neutrálního nebo negativního (nebo-li hodnot -1, 0, 1). Tato metrika je přejmuta datovým konektorem a data jsou jím obohacena. Sentiment je vyhodnocován a úroveň celého příspěvku.

NoSQL Databáze (Elasticsearch) – pro ukládání dat a zpracování textových položek bylo použito open source řešení Elasticsearch postavené na základě knihoven Apache Lucene. Umožňuje vyhledávat v datech podle různých kritérií a počítat statistiky číselných položek. Nástroj vytváří strukturu invertovaného indexu nad uloženými daty. Součástí nástroje je i komponenta pro shlukovací analýzu textových dat carrot2 (Stefanowski & Weiss 2003).

User Interface (Kibana) – je standardní řešení dodávané spolu s Elasticsearch, které umožňuje rychle a jednoduše analyzovat velké objemy textových dat, vytvářet vizualizace a sestavovat analytické dashboady.

Zpracované datové zdroje z Facebooku

Výsledky analýzy sociálních sítí jsou krajně odvislé od vstupních dat. Jinak vybrané datové zdroje logicky povedou k jiným, možná i zavádějícím výsledkům. Pro účely této analýzy byly vydefinovány stránky, které byly analyzovány. Jednalo o veřejné stránky politických stran, které jsou zastoupeny v Poslanecké sněmovně Parlamentu ČR a získaly více než 2,5% hlasů a stránky hlavních českých médií.

V případě, že politická strana má více různých stránek (např. pro celostátní a regionální, typicky municipální úrovně), byla analyzována pouze celostátní verze stránky. Na municipální úrovni není obvyklé projevovat názory, které se vážou k celostátní politice a sledované události.

Kromě stránek politických stran byly do analýzy zahrnuty veřejně dostupné stránky velkých českých médií: *iDnes.cz, iHned.cz, novinky.cz, Nova.cz, Ceskatelevize, ct24.cz, TV Prima*.

Tyto zdrojové stránky byly sledovány v průběhu roku 2015 aby bylo možné potvrdit, že sledovaná osoba (Miloslav Ransdorf) nebyl častým aktérem politického diskurzu ve společnosti. Celkem bylo zpracováno 1 214 190 příspěvků z těchto stránek od 192 154 jedinečných uživatelů s průměrným sentimentem 0,108.

V tomto datovém zdroji byly vyhledány příspěvky, které se vážou k sledované události. Jelikož média sledují především aktéry, byla analýza zaměřena na sledovanou osobnost. Jelikož je jméno *Ransdorf* specifické, je velmi obtížné si představit jiný kontext, kde by bylo toto slovo použito, než v odkazu na tuto událost. Proto u obou zdrojů bylo sledováno slovo *Ransdorf*.

Zpracování dat z ostatních médií

Model technologie zpracování dat z ostatních médií

Jako ostatní média je možné obecně vnímat tisk, televizi, rádio a internet (mimo sociální sítě, tedy v zásadě internetové verze novin nebo televize. Obvykle tito vydavatelé mají i své stránky na sociálních sítí, které zde nejsou uvažovány). Pro účely této analýzy byla použita databáze Anopress IT, která obsahuje zdroje českého periodického tisku, pořady českého televizního a rozhlasového zpravodajství, publicistiky, českých internetových serverů a vybraných slovenských a anglických zdrojů (Anopress IT 2017).

V této databázi bylo vyhledáváno klíčové slovo *Ransdorf* stejným způsobem jako v obsahu zpracované sociální sítě. Tyto články byly agregovány na měsíční bázi. Pro bližší srovnání vizte Tab. 1 v sekci Výsledky analýzy.

Zpracované datové zdroje z ostatních médií

Databáze Anopress IT obsahuje kromě regionálních a dalších zdrojů celostátní periodika jako jsou: *Hospodářské noviny, Lidové noviny, Blesk, Metro, Mladá Fronta, Právo, Sport*. Dále pořady *Českého Rozhlasu a České Televize.*

Sledované klíčové slovo v této situaci představuje odraz reality společenského diskurzu. Do prohledávaných dat byly zařazeny všechny dostupné datové zdroje. Úpravy ve vyhledávání, omezování zdrojů nebo jiná vyhledávací strategie nevedla v případě tohoto datové zdroje k výrazně odlišným výsledkům, proto lze konstatovat, že vyhledávání klíčového slova *Ransdorf* bylo dostačující jak pro svoji unikátnost, tak specifičnost. Vyhledané články v průběhu roku byly relevantní vzhledem ke sledované osobě a tématu.

Výsledky analýzy

Frekvenční analýza události

Pro výpočet měr podobnosti byla četnostní data agregovaná do Tab. 1.

Tab. 1 Srovnání měsíční frekvence události obou zdrojů zdroj: autor

Měsíc	Ostatní média	Sociální sítě
1	21	3
2	39	28
3	46	18
4	30	24
5	21	5
6	32	2
7	21	0
8	27	5
9	29	39
10	21	11
11	22	20
12	421	699
Celkem	730	854

Korelační míra představuje bezrozměrný koeficient r nacházející se v intervalu <-1;1>, kde -1 je silná negativní korelace a 1 je silná pozitivní korelace. V případech silné pozitivní korelace, kde hodnota korelačního koeficientu se přibližuje k 1 je možné stanovit, že přírůstek nebo úbytek jedné sledované veličiny velmi dobře vysvětluje přírůstek nebo úbytek druhé sledované veličiny.

Při frekvenční analýze četností textů na sociálních sítí a v ostatních médiích sledujeme silnou korelaci mezi oběma zdroji. Tato korelace dosahuje r = 0,9975.

Pro potvrzení vysoké míry podobnosti byl proveden výpočet Diceho míry podobnosti na měsíčních frekvencích uvedených v Tab. 1. Diceho míra je jedna z běžně používaných metrik pro výpočet podobnosti, která je definována jako:

$$S_D(x_i, x_j) = \frac{2\sum_{l=1}^{m} x_{il}x_{jl}}{\sum_{l=1}^{m} x_{il}^2+\sum_{l=1}^{m} x_{jl}^2},$$

kde x_i a x_j jsou souřadnice srovnávaných vektorů (resp. četností příspěvků v daném měsíci) a *l* je pořadí (v tomto případě pořadové číslo měsíce). Pro sledovanou událost je $S_D = 0,880$.

Obor hodnot Diceho podobnosti se pohybuje mezi <0; 1>, kde hodnoty blížící se k 1 udávají silnou podobnost obou sledovaných událostí.

Lze proto podloženě předpokládat, že událost, o které informují ostatní média rezonují i v obsahu sledovaných sociálních sítí ve stejně době.

Dále byly zaznamenány průměrné hodnoty sentimenty a počtů liků v jednotlivých měsících. Tyto výsledky jsou blíže prezentovány v Tab. 2.

Tab. 2 Sledované průměrné metriky v čase zdroj: autor

Měsíc	Sentiment (průměr)	Like (průměr)
1	-1	0
2	0,2	3,55
3	0	2,444
4	0	3,667
5	0	130
6	0	2,75
7	0	1
8	0	6
9	0,343	4,343
10	0,333	0,833
11	-0,2	3
12	-0,088	7,264

Z této tabulky můžeme proto dovozovat jaký byl dopad události do obecné nálady na Facebooku. Počet liků ale nelze interpretovat stejně jako sentiment.

Tlačítko „*Like*“ na Facebooku umožňuje jistým netextovým způsobem komunikovat s ostatními uživateli. Vyjadřuje tak možnost sdělit, že se zprávou souhlasím bez dalšího upřesnění.

Obecně lze přirovnat počet liků k počtům uživatelů, kteří byli natolik zasaženi zprávou, že se s ní ztotožňují natolik, aby svůj like přidali. Lze tak konstatovat, lze srovnat s mírou aktivity diskuze na Facebooku. Srovnáním s Tab. 1 lze dovozovat, že počty liků jsou ve vzájemném vztahu k obecnému počtu příspěvků a shodují se i s pokrytím ostatními médii.

Sentiment oproti tomu není ovlivněn aktuální situací ve společnosti, ostatními médii nebo počtem liků. Jelikož se jedná o metriku spočítanou při zpracovávání dat na základě podobnosti textu k srovnávacímu modelu, jde o vlastnost každého příspěvku, která se nemění. Celkový průměrný sentiment je tudíž odrazem pozitivity nebo negativity článků diskutujících aktuální události kolem sledované osoby.

Výsledky shlukovací analýzy

K zodpovězení otázky ohledně charakteru příspěvků z Facebooku, byla provedena shluková analýza textů obsahujících slovo *Ransdorf*. Výsledky shlukové analýzy pomocí algoritmu lingo jsou uvedeny v Tab. 3. Tyto shluky obsahují podstatně velká množství příspěvků, která jsou charakterizována těmito slovy.

Podle obsahu se obvykle nejedná o překvapivé příspěvky. Obsah se většinou koncentruje k vyhledávanému termínu. Můžeme ale sledovat i příbuzná slova, vyplývající z povahy události (T_6, T_8, T_9).

Z této povahy se vymyká T_7, která se váže k velké debatě vedené v pořadu Otázky Václava Moravce, kde vystupoval ministr obrany Stropnický a vyjadřoval se k této události.

Tab. 3 Tematické shluky podle algoritmu lingo zdroj: autor

	Popis shluku
T_1	Europoslanec Ransdorf
T_2	Soudruh Ransdorf
T_3	Miloslav Ransdorf
T_4	Chtěl Ransdorf
T_5	Komunista
T_6	Banky
T_7	Stropnický
T_8	Švýcarská policie
T_9	Švýcarská banka
T_{10}	Události Komentáře
T_{11}	Zeman
T_{12}	Evropský parlament

Závěr

Tato práce si kladla za cíl zodpovědět otázky vztahující se k vazbám mezi ostatními médii a sociálními sítěmi. Bylo potvrzeno, že sledovaná veřejná část sociální sítě Facebook je silně provázána s obsahem ostatních médií. Oba zdroje vykázali silnou provázanost podle korelační analýzy i podle Diceho míry.

Silná vazba ostatních a sociálních médií se projevuje i na obsahové úrovni. Ve sledovaných příspěvcích z Facebooku se neobjevily témata, která by nebyla pokryta v ostatních médiích. Některé shluky příspěvků dokonce odkazují na konkrétní pořady z ostatních médií.

Práce sledovala dvě metriky vázající se k Facebooku: sentiment a počet liků. Obě metriky lze využít ke kvantifikovanému sledování dopadu politické události na části voličů.

Vztah k dizertační práci

Připravovaná dizertační práce se zabývá metodami analýzy sociálních sítí. Cílem dizertace je vytvořit metodiku, která bude pokrývat všechny podstatné kroky zpracování a obohacení dat, jejich analýzou a interpretací.

Tato práce se zabývá vztahem mezi ostatními médii a sociálními sítěmi. Velké množství požadavků na analýzy sociálních sítí je zaměřeno na otázku jestli a jak sledovaná událost dopadá na společnost. Sledování politického světa na sociálních sítích a v ostatních médiích je jedním z možných uplatnění připravované metodiky.

Kromě sledování politického života se analýza sociálních sítí používá při zjišťování veřejného mínění, při detekci přírodních katastrof, pro marketingové účely. Ve všech situacích se předpokládá, že sociální sítě jsou do jisté míry odrazem společnosti. Ačkoli není možné dojít k naprosto dokonalému obrazu reality, nabízejí nám jiný, detailní, pohled na její část. Jde o relevantní zdroj, který je možné analyzovat postupy, které nejsou dostupné v dotazníkových šetřeních. *Dizertační práce se zabývá tímto problémem z obecnějšího pohledu a analýza politické události je jedna z jejích aplikačních oblastí.*

Zároveň připravovaná metodika ověřuje na sebraných datech jak zpracovat a obohatit textová data v českém jazyce. K tomuto účelu je použit korpus textových dat ze sociálních sítí, který je jedním z výsledných artefaktů dizertační práce.

Literatura

Kotalík, J., 2015. PŘEHLEDNĚ: Co vše víme o bizarní kauze europoslance Ransdorfa. *iDNES.cz*. Available at: http://zpravy.idnes.cz/prehled-o-kauze-ransdorf-08b-/domaci.aspx?c=A151208_191308_domaci_jkk [Accessed January 18, 2017].

Anopress IT, 2017. Vyhledávání v databázi - Anopress IT, a.s. - dodavatel profesionálního monitoringu médií. Available at: http://www.anopress.cz/vyhledavani-v-databazi [Accessed January 18, 2017].

Bamman, D. & Smith, N.A., 2015. Contextualized Sarcasm Detection on Twitter. In *Ninth International AAAI Conference on Web and Social Media*. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10538 [Accessed November 8, 2015].

Boyd, D.M. & Ellison, N.B., 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), pp.210–230.

Branson, S. & Greenberg, A., 2002. Clustering Web Search Results Using Suffix Tree Methods. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.692.5027&rep=rep1&type=pdf.

Davidov, D., Tsur, O. & Rappoport, A., 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 107–116. Available at: http://dl.acm.org/citation.cfm?id=1870582 [Accessed October 9, 2015].

Dolamic, L. & Savoy, J., 2009. Indexing and stemming approaches for the Czech language. *Information Processing & Management*, 45(6), pp.714–720.

Filipec, J., 2010. *Slovník spisovné češtiny pro školu a veřejnost: s dodatkem Ministerstva Školství, Mládeže a Tělovýchovy České republiky*, Prague: Academia.

González-Ibánez, R., Muresan, S. & Wacholder, N., 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pp. 581–586. Available at: http://dl.acm.org/citation.cfm?id=2002850 [Accessed October 9, 2015].

Kamath, S.S. et al., 2013. Sentiment Analysis Based Approaches for Understanding User Context in Web Content. In IEEE, pp. 607–611. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6524473 [Accessed March 18, 2014].

Keretna, S., Hossny, A. & Creighton, D., 2013. Recognising User Identity in Twitter Social Networks via Text Mining. In IEEE, pp. 3079–3082. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6722278 [Accessed October 24, 2014].

Minier, Z., Bodo, Z. & Csato, L., 2007. Wikipedia-Based Kernels for Text Categorization. In IEEE, pp. 157–164. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4438094 [Accessed November 24, 2014].

Rajadesingan, A., Zafarani, R. & Liu, H., 2015. Sarcasm Detection on Twitter: A Behavioral Modeling Approach. In ACM Press, pp. 97–106. Available at: http://dl.acm.org/citation.cfm?doid=2684822.2685316 [Accessed October 9, 2015].

Řezanková, H., Húsek, D. & Snášel, V., 2009. *Shluková analýza dat*, Praha: Professional Publishing.

Socialbakers, 2017. December 2016 Social Marketing Report Czech Republic | Socialbakers. Available at: https://www.socialbakers.com/resources/reports/czech-republic/2016/december/ [Accessed January 22, 2017].

Stefanowski, J. & Weiss, D., 2003. Carrot2 and language properties in web search results clustering. In *Advances in Web Intelligence*. Springer, pp. 240–249. Available at: http://link.springer.com/10.1007/3-540-44831-4_25 [Accessed September 18, 2015].

Strossa, P., 2011. *Počítačové zpracování přirozeného jazyka*, Praha: Oeconomica.

Ward, C.B. et al., 2011. Empath: A framework for evaluating entity-level sentiment analysis. In IEEE, pp. 1–6. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6135866 [Accessed November 6, 2014].

Willett, P., 2006. The Porter stemming algorithm: then and now. *Program: electronic library and information systems*, 40(3), pp.219–223.

JEL Classification: C38, H80.

Summary

Text Mining on Social Networks: Analysis of Political Event

Tools for social media data processing can be applied to various issues. From public opinion mining, to marketing campaign measurement, to quick reactions to actual trends among the customers. These applications are critical both for high management of companies as well as for the high management of political parties. Method of analysis of such data should be the same for both stakeholders.

Available data from Facebook and other media were analysed in order to find out how are both information platforms connected between each other. The content from social media were processed by the tool developed at UEP using open source solutions and self-developed libraries. These libraries are using text mining techniques and then compared with other media data.

Data gathered from the Facebook and other media showed us strong dependency between social and other media. This dependency was proved on the level of frequency analysis and on the level of content on the Facebook. This was done though the clustering analysis of the text.

To estimate an impact of political event into the society is crucial for the political parties‘ persons. Method of data analysis that was used in this analysis was developed in the disertation thesis.

Keywords: social media, text mining, political event., publicistic.

Data mining with explanations

Viktor Nekvapil
xnekv05@vse.cz
Doktorand oboru Aplikovaná informatika
Školitel: prof. RNDr. Jan Rauch, CSc. (rauch@vse.cz)

Abstract: New approach of using domain knowledge in the evaluation phase of data mining task is presented. In the approach, an 'explanation' is perceived as a piece of domain knowledge that could help the user of data mining system to better understand results of a data mining task. Architecture of the newly proposed system is presented and requirements on the components are stated. As a source of domain knowledge, data from trusted organisations such as statistical offices or ministries are used to avoid lengthy collection of domain knowledge from experts. The feasibility of the proposed approach is demonstrated in the proof of concept implementation based on the LISp-Miner system, Python and SQLite. Clients' data of a big financial institution is used for analysis and, as a source of domain knowledge, data from Czech statistical office is used. Enhancements of the FOFRADAR framework based on the newly defined 'trusted knowledge' is proposed in the final section.

Key words: data mining, domain knowledge, explanation, evaluation phase, LISp-Miner, Python

Introduction

As (Qiang, 2006) stated in the paper ‚10 Challenges of data mining research', ‚*there is a strong need to integrating data mining and knowledge inference*'. Although there have been some achievements since the paper was made public (se e.g. Mansingh, 2011; Rauch, 2015), data mining systems are still ‚*unable to relate the results of mining to the real-world decisions they affect'*, as they wrote.

(Qiang, 2006) further claims that ‚*Doing these inferences, and thus automating the whole data mining loop requires representing and using world knowledge within the system. One important application of the integration is to inject domain information and business knowledge into the knowledge discovery process*'.

The ‚*world knowledge*', or '*domain information'* as they used this terms in the paper, refers apparently to the general knowledge about some domain, for example about banking. Pursuing this example, the basic knowledge is that banks lend money (e.g. in form of mortgage) to customers which have to pay certain interest rates. Based on prices of real estates, the volume per mortgage usually differs across regions. One can assume, that in regions with higher real estate prices, the average volume of mortgage will be higher as well. For example, when using a data mining system on clients' data of a bank, the pattern found could look like this: If the customer is from region X, the volume of mortgage is high. When evaluating whether this rule is interesting and can be possibly used in decision making of the bank, domain knowledge of an expert in the finance domain will suggest that this rule is in conformance with the overall expectations because the prices of real estates are high in region X. Because domain experts are often not at disposal or are costly, the research aim is to replace domain expert to some extent with automate means – the system which will offer additional information to the user instead of domain expert. To summarise, the user evaluates whether the resulting pattern is *interesting* or not using knowledge and experience of himself/herself, or if he/she do not have enough knowledge in the particular domain, knowledge from domain expert.

The usual way how to express *interestingness* of a pattern is to use *objective* interest measures such as confidence in association rules mining. The main drawback of these methods is that the found patterns are only *statistically* interesting, but often not interesting to the user. Some approaches (e.g. Silberschatz, 1995; Padmanabhan, 1998, more recently in De Bie, 2013) introduce *subjective* measures of interestingness and try to overcome the drawback using a *belief system* of a user, which is the set of domain knowledge of the user. Then, a pattern is interesting if it changes the belief system of the user. The main drawback of this approach is that the belief system of the user has to be explicitly defined prior to the data analysis, which is very time consuming and often not practically applicable in day-to-day data mining, because it requires more effort than the possible benefits of the data analysis.

The approach presented in this article tries to incorporate domain knowledge in the evaluation phase of data mining but avoids lengthy and complex task of building a belief system of the user. The idea is to present additional relevant data to the user which will help him better evaluate the patterns found.

Interesting approach which could be possibly employed uses linked open data (LOD), which allows for publishing interlinked datasets using machine interpretable semantics. For example, (Paulheim, 2014) developed an extension for Rapid Miner, which can extend dataset with additional attributes drawn from Linked open data cloud. The possible downturn of this approach is the nature of the data – it comes from the 'community' and there is no guarantee that the data are correct – basically the same principles as for example on Wikipedia.

The origin of the data is very important property when performing a business task. Therefore, the newly proposed approach of *Explanation system based on domain knowledge* deploys data from trusted sources – state institutions such as statistical office or ministries.

The rest of the paper is organised as follows: Firstly, description of Explanation system is presented and implementation possibilities are discussed. Furthermore, the proof of concept is presented. Lastly, the comparison of the Explanation system with the FOFRADAR framework is sketched.

Explanation system based on domain knowledge

Although the term *explanation* is broadly used in relation to expert systems (see e.g. Buchanan, 1988), we use it here in connection with a data mining system. Here, we perceive *explanation* as a piece of domain knowledge that could help the user of data mining system to better understand results of a data mining task.

As mentioned in literature, knowledge acquisition is a sensible problem (see e.g. Danubianu, 2013). This means that obtaining domain knowledge (often from experts in a specific field) is very costly and time consuming activity. Yet there are easily available resources that could be used instead of knowledge obtained from experts. The data collected from trusted institutions such as statistical organisations, police etc. could be considered as objective and capturing reality. Moreover, this data could be used in more domains thanks to the geographical dimension that they usually include. For example, census demographic data collected from Czech statistical office (CSO, 2015) could be used to investigate clients' behaviour in the bank in different regions but could also be used in construction industry to anticipate the demographic evolution and thus the demand for new buildings in different regions.

Based on this simple idea, we can link the results of data mining to the data from those trusted institutions and thus enhancing the possibilities of the user to better understand the results and use them further in organisation. (Kopanas, 2002) calls this phase 'Fielding the knowledge base', where the learned knowledge is combined with other domain knowledge in order to become part of decision-making. The newly proposed approach follows this idea; the main contribution is to enhance the domain knowledge of the user with another sources of knowledge that are objective and believable. From the CRISP-DM[1] perspective, the domain knowledge is used in the evaluation phase of the data mining task.

Architecture of the newly proposed approach

Architecture of the newly proposed approach is very simple (see Figure 1). The first component constitutes the data mining system (DMS). Basically, any data mining system can be used. The second part is Domain knowledge repository (DKR), where domain knowledge in form of data from trusted sources and/or domain knowledge obtained from domain experts is included. Third part presents the Explanation module, which collects and presents explanations for the results obtained from DMS. It is also possible to see the context of the explanation there. Last part of the system constitutes user.

[1] **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining, see CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. 2000. *CRISP-DM 1.0. Step-by-step Data Mining Guide.* [electronic document]. Available from http://www.crisp-dm.org/CRISPWP-0800.pdf

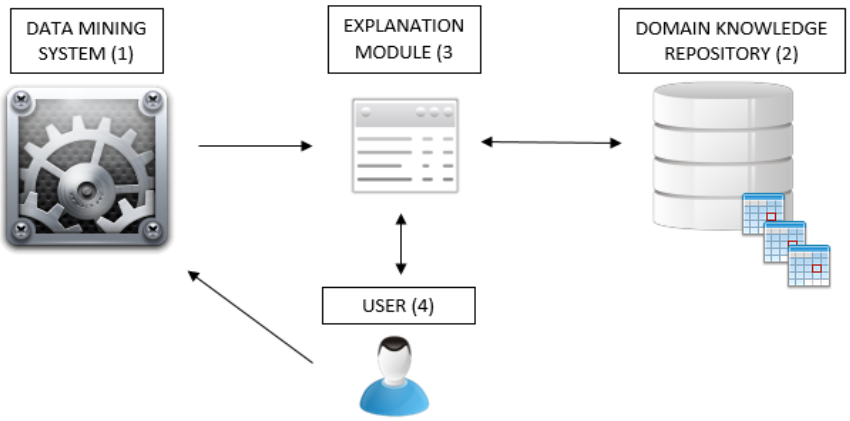


Figure 1: Architecture of the system - high level overview Source: Author

The key part of the proposed system is the Domain knowledge repository. Two possibilities have been identified how an explanation could be produced:

- DKR already have some pieces of knowledge that are relevant to the results of the data mining task (pre-calculated or manually inserted by domain expert)
- DKR needs to extract knowledge in its databases

To ensure desired behaviour, Domain knowledge repository and Explanation module should guarantee following properties:

1. The explanation must be relevant to the result
2. Number of explanations must be reasonable
3. The explanation must be understandable for the user
4. The explanation must be based on a trusted source of knowledge
5. The explanation must be formalised to ensure automation
6. The context of the explanation must be handed to user on request

1. The explanation must be relevant to the result

An explanation (a piece of domain knowledge) is relevant if it is connected to the results in some manner. For example, if the result is describing a specific region, then explanations concerning the specified region are conceived to be relevant.

2. Number of explanations must be reasonable

The number of explanations that is handed back to the user must be of a reasonable count to ensure that user is not overwhelmed by explanations that are of little use for him. Ideally, some sort of mechanism that picks up the most promising explanation should be involved.

3. The explanation must be understandable for the user

Although understandability is a rather subjective property, there are some basic guidelines of what is rather comprehensible and what is not. We assume that the user is familiar with the data which are subject of a data mining task but is not necessarily strong in maths or logic. Therefore, the syntax of an explanation should be intuitive and not far from a natural language.

4. The explanation must be based on a trusted source of knowledge

The first option of obtaining domain knowledge is to consult domain expert. As already stated, this way is costly and time-consuming. Besides this, there are various data publicly available that can be used as domain knowledge. Government institutions, EU institutions and statistical offices offer more and more data. This is boosted by the Open government data[2] initiatives, which offers catalogue of publicly available data sets. In the Czech Republic, the Open data[3] initiative offers a catalogue of data using the linked data paradigm which refers to the Czech Republic. The data from those organisations are generally considered to be trusted sources. Another source represents community encyclopaedic sources such as Wikipedia and its' linked data version DBPedia. Another linked data sources such as FreeBase can be used to obtain additional data as well. However, in our opinion, these community based sources lacks believability and objectivity.

[2] http://opengovernmentdata.org/
[3] http://www.opendata.cz/en

5. The explanation must be formalised to ensure automation
To formalise and represent domain knowledge, the common approaches include IF-THEN rules commonly used in expert systems, ontologies, hierarchies of attributes (Anand, 1995), mutual influence of attributes (Rauch, Šimůnek, 2014) and constraints (Han, 1999). Of course, some custom-made solution could be used as well if it suits better the aims of the framework.

6. The context of the explanation must be handed to user on request
We assume that in some cases it will be not clear to which degree the explanation is relevant or not. In this situation, the system should offer details and perhaps more knowledge that is related to the explanation. For example, when speaking about region as one dimension of explanations, then knowledge about surrounding regions could be offered to user.

Usage scenario

The usage scenario of the Explanation system is depicted in **Chyba! Nenalezen zdroj odkazů.**.

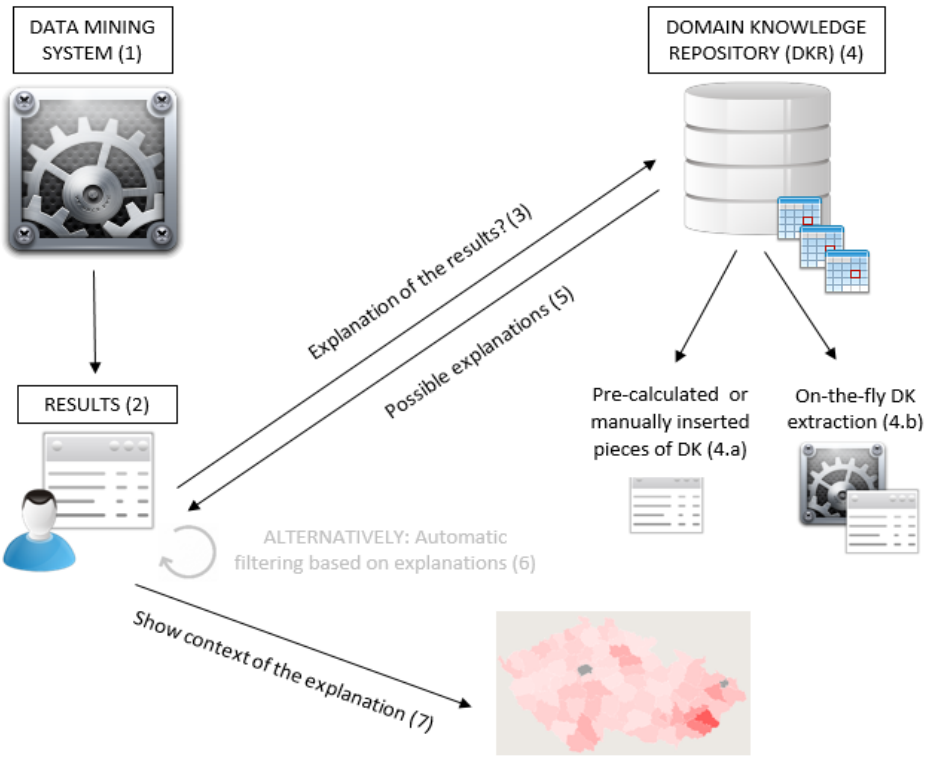


Figure 2: Usage scenario of the proposed approach Source: Author

The description of the phases is as follows:

1. Data mining task is executed using some data mining system
2. Results of the task are at disposal
3. Request for explanation of the results is sent to the DKR
4. DKR searches for possible explanations
5. Possible explanations are handed from the DKR. User can use the explanations to better understand the results
6. Alternatively, automatic filtering could take place (for example: hide results that are well explained)
7. User can request context for the explanation being handed to him. This could help him/her to assess the relevancy of the explanation.

Implementation possibilities

For each component of the model, there are several options how to implement it. In the following sections, these possibilities are discussed.

Data mining system

As already mentioned, basically any data mining system could be used in the system. For the purpose of simplicity, association rules will be used further in the course of experimentation because they are very intuitive to understand to user and well-studied by the data mining community.

Following systems for mining association rules are considered (not a full list of DM tools):

- RapidMiner[4] – basic version is free, custom extensions can be easily made
- SAS[5] – widely used in commercial domains (banking, insurance, pharma), very good possibilities of custom modifications due to programming language
- LISp-Miner[6] – free academic system, possible automation of DM task due to scripting language LMCL

Domain knowledge repository (DKR)

Following possibilities are considered as a *storage* for domain knowledge repository:

- Own relational database on a local computer
- Use of remote databases
- Web service based repositories
- Linked data paradigm

Following possibilities are considered as a *source* of domain knowledge

- Institutional organisations (statistical offices, EU offices, ministries, state and community organisations – often referred to as 'Open data')
- Community based sources (Wikipedia, DBpedia etc.)
- Knowledge elicited from domain experts and domain non-electronic sources (books etc.)
- Knowledge extracted by DKR from its databases

Following properties are considered when implementing DKR:

- Knowledge could be frequently updated,
- Knowledge must be machine readable,
- Knowledge bases (DKR) could be shared among different users

Explanation module

To the best knowledge of the author, there are no ready-made solutions that could act as an explanation module.

Requirements on the explanation module are as follows:

- displays results of the data mining task and assigns requirements on explanations to the DKR
- connects results of the data mining task and the explanation in an understandable manner
- offers context of the explanation on the user's request

The aim of the explanation module is to present additional information in form of domain knowledge which is related to the result of the data mining task and helps user to distinguish whether the result is interesting for him or not.

The user uses the explanation module to find out one of the following conclusions:

- The knowledge from data is in compliance with the domain knowledge
  - This situation assures user that his/her data are within expectations
- The knowledge from data is in contradiction to the domain knowledge
  - This situation is explained by the character of the data (for example, only VIP clients are present in the data, which influences the power of explanations) OR
  - The situation requires further investigation to clarify the contradiction
- The knowledge from data does not have a reasonable explanation based on domain knowledge
  - The DKR does not bring further benefits for the user. Possibly, there are no data about the relevant domain in DKR

[4] See https://rapidminer.com/ for more information on the system.
[5] See http://www.sas.com/en_us/software/analytics/enterprise-miner.html for more information on the system.
[6] See http://lispminer.vse.cz/ for more information on the system.

Proof of concept

Based on the requirements stated in chapter 'Architecture of the newly proposed approach', the proof of concept implementation verifies basic structure and interactions among the components of the newly proposed approach. It is a semi-automatic implementation based on LISp-Miner, SQLite and Python. The components used in the proof of concept implementation are stated in Table 1.

The LISp-Miner System has been chosen for its ability to automate the task through LMCL scripting language, which could be possibly used in future iterations. Simple database table is used as a storage of domain knowledge. SQLite database is used for this purpose. The DKR includes data about the average and median income in the Czech Republic in 2014, obtained from Czech Statistical Office. Python has very clear an easy to learn syntax, has a strong user community and suits for rapid prototyping.

Table 1: Components in the proof of concept

#	Component	Proof of concept solution
1	Data mining system	LISp-Miner system, 4ft-Miner procedure, AAD quantifier
2	Domain knowledge repository	SQLite database, one table, two types of explanations (pieces of domain knowledge) – Avg income and median income in CZ in 2014 per region
3	Explanation module	Python output to console

Requirements on explanations and their fulfilment in the proof of concept implementation are stated in Table 2.

Table 2: Requirements fulfilment in the proof of concept

#	Requirement	Proof of concept solution
1	Explanation is relevant	Same geographical dimension in result and in explanation, region in top 3 or bottom 3 based on the order of the measure
2	Reasonable number of explanations	Only two pieces of domain knowledge (explanations) in DKR
3	Understandable explanation	Syntax close to natural language
4	Based on trusted source	Czech statistical Office – data about income in 2014
5	Formalised explanation	Explanation formalised to following quadruplet: *region name (kraj), measure name, measure value, rank within measure*
6	Context of the explanation	Table with the measure in all the regions

Domain knowledge is formalised as follows:

- Each piece of domain knowledge is stored as per geographical unit
- Each geographical unit has its rank within the measure stored
- Each geographical unit has the absolute value of the measure stored
- Table in following structure – regional_unit (kraj), measure_name, measure_value, rank_within_measure, see Figure 3 for the structure.

kraj	measure_name	measure_value	rank_within_measure
Hlavní město Praha	income_avg	35115	1
Středočeský kraj	income_avg	27345	2

Figure 3: Structure of the DKR table

Process

Following steps were performed in the proof of concept implementation:

1. DKR is set up – data from Czech Statistical Office is inserted into SQLite database
2. Task is executed in the LISp-Miner (4ft-Miner procedure), attribute representing geographical unit is present
3. Results of the 4ft-Miner procedure are exported to XML file
4. Python script is used to
    - Parse the XML file and get the results of the task
    - Match the geographical units in the resulting rules and geographical units in DKR

- Fetch interesting explanations for each geographical unit. Relevancy criterion used: region is present in top 3 or bottom 3 rank based on the order of the measure
- Print the rules and possible explanations
- Retrieve context of the explanation on user's request

Case study

To validate the proposed framework on real data, following case study was performed. We are given a data set from a financial institution. There are data about clients, who were given a loan, including geographical and demographical client data, data from loan application, data concerning the agent who arranged the loan and so on. The task is to find some interesting facts about the geographical location of the client and the amount of loan taken. As a data mining system, we use the LISp-Miner system. We use the 4ft-Miner procedure and the Above average dependence (AAD) quantifier[7]. After several iterations, the task definition was set to Base=20 and AAD p=0.65 (see Figure 4).

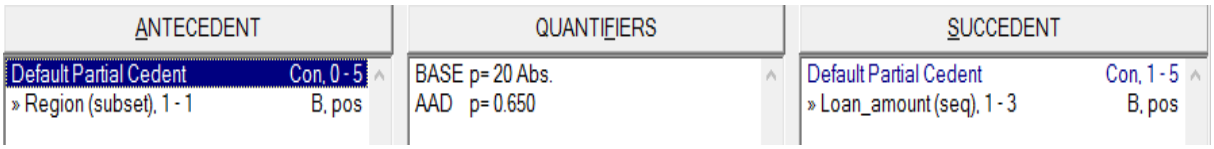


Figure 4: Task definition in LISp-Miner

As future work, the LMCL language could be used here to find the task settings which produce a reasonable number of rules. In the usage scenario (see Figure 2), this activity corresponds to number 1 (*Data mining system* and its usage).

Three rules were found which are relevant to the task definition (see Figure 5). This step corresponds to number *2 - Results* in the usage scenario (see Figure 2).

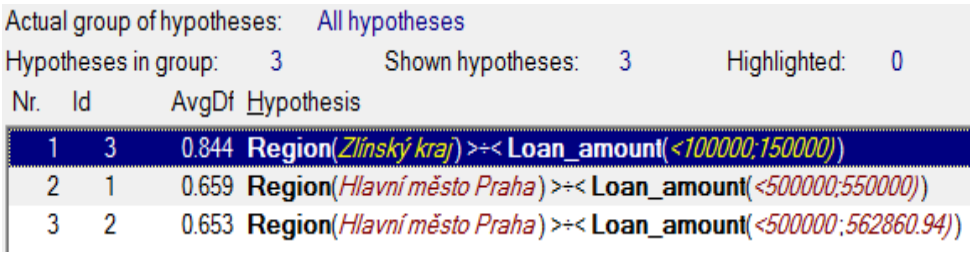


Figure 5: Results of the task

As the user knows from the distribution of the attribute Loan_amount, the Loan_amount <100000; 150000) is rather low (second lowest). Are there some explanations for this fact? This question corresponds to activity number 3 in the usage scenario (see Figure 2). The user exports the results of the task to XML file using the LISp-Miner functionality. In the Python script, this file is referenced and the data about the task results is integrated with the DKR based on the region presented in the rule.

From these data, those explanations which signal that a region presented in the rule is outstanding (unusual) regarding the selected measure will be handed back to the user. The unusualness will be considered as being top 3 or bottom 3 in the ranking according to the selected measure. An example output for the first rule is presented below (see Figure 6). Firstly, basic information about the rule is summarised. Then, all explanations found are presented (two for this example).

The user can conclude from the two explanations that income (both average and median) in 'Zlínský kraj' is rather low (second lowest – bottom 2). He/she can conclude that the clients' data he/she has at disposal in the data set are in conformance with his/her domain knowledge, because there is a direct proportion between amount of loan and income (which is part of the users' domain knowledge).

The reasoning used here is a combination of domain knowledge of the user and domain knowledge from a trusted source which brings the user to the conclusion that the rule is in conformance with his domain

[7] AAD is very similar to the commonly used LIFT measure, for more information on the 4ft-Miner procedure see RAUCH, Jan. *Observational Calculi and Association Rules* [online]. 1. ed. Berlin : Springer-Verlag, 2013. ISBN 978-3-642-11736-7. Available at: http://link.springer.com/book/10.1007/978-3-642-11737-4

knowledge. User knows that there is a direct proportion between amount of loan and income, but does not know if the income in the region ('Zlínský kraj') is rather low or high.

```
--------------------------------------------------
Rule ID: 34
Rule: Region(Zlínský kraj) >:< Loan_amount(<100000;150000))
Confidence: 0.178571
Support: 0.017606
Geo attribute found: Region
Coefficient 1 of geo attribute : Zlínský kraj
Explanations found:
--- Explanation 1 ---
Value of the geo attribute: Zlínský kraj
Measure: income_avg
Rank within measure: 13 (bottom 2)
Value of the measure: 23873
--- Explanation 2 ---
Value of the geo attribute: Zlínský kraj
Measure: income_median
Rank within measure: 13 (bottom 2)
Value of the measure: 21542
--------------------------------------------------
```

Figure 6: Output of the explanation module

At this stage, context of the explanation can be obtain in a very simple manner. User can request a context of the explanation by calling function context(name_of_the_measure), e.g. context(income_median). A table with the regions, values of the measure and order is retrieved.

Future development improvements

Following points could be implemented to enhance the robustness and easiness to use of the implementation of the Explanation system as shown in the proof of concept:

- user interface (web application)
  - to display rules and the explanations
  - to display context of the explanations in a more comfortable way (maps, tables)
  - to load new pieces of domain knowledge
  - to interact directly with LISp-Miner
- automation within LISp-Miner (using LMCL to adjust task settings if sufficient number of rules not found)
- better interaction with LISp-Miner (possibly no need to export xml file with results from LISp-Miner)
- domain knowledge repository enhancements – accommodate more types of geographical units (region, district, etc.) normalise database structure, include measurements units, upload new domain knowledge

Comparison with the FOFRADAR framework

- One of the possible solutions of the automatic formulation of conclusions using domain knowledge is presented in the FOFRADAR framework (Rauch, 2015). The framework is based on a logical calculus of association rules. The interpretation is based on mapping important items of knowledge to sets of association rules which can be considered as their consequences. Important items of knowledge are expressed using simple mutual influence among attributes. These are predefined relationships of attributes which are used to determine whether the association rule can be seen as a consequence of the item of knowledge or not. For example, simple mutual influence (SI-formula) Income ↑↑ Loan means: ‚*if Income increases, then Loan increases as well*'. The set of atomic consequences of this SI formula can be expressed for example by following union: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times Medium_{Loan} \cup High_{Income} \times High_{Loan}$, saying that '*if Income is high, than Loan is high or if Income is medium then Loan is medium or if Income is high then Loan is high*'. The definition of the *Income(low)*, *Income(medium)*, *Income(high)*, *Loan(low)*, *Loan(medium)*, *Loan(high*) levels is to be prepared in cooperation with domain expert. For example, each basic Boolean attribute *Loan(α)* satisfying condition $\alpha \subset Low_{Income}$ will be considered as saying '*Income is low*' if Low_{Income} = {(0;10000>,(10000;15000>,(15000;20000>}. The feature of this approach is that it is necessary to define such sets of categories in advance with the help of domain expert.

- Using the rankings of the measures from trusted organisations as shown in chapter Proof of concept, definitions of the sets of categories could be made automatically. In the context of FOFRADAR, we can refer to the data from trusted organisations as *trusted knowledge*. The ranking according to the specific measure (that is *Trusted knowledge*) could be used as a decision on which level of measure is considered to be *'high'* and which to be *'low'* without the need to consult domain expert. Consider this example. In total we have 14 districts (kraje) in the Czech Republic. If we want to take 5 levels of coefficient α into account (*very low, low, medium, high, very high*), we can assign the ranking according to selected measure (e.g. Average income) to the coefficients (α) as depicted in Figure 7.

Coefficient α	very low	low	medium	high	very high
Rank within measure	14,13,12	11,10,9	8,7	6,5,4	3,2,1

Figure 7: Coefficients and respective ranks

For example, if we have '*Zlinsky kraj*' ranked as 13th in the measure Average income, its coefficient will be '*very low'*. We can write: *Zlinsky kraj => Average income (very low)*, meaning that *in* '*Zlinsky kraj' district, Average income is very low*. This could be seen as a new type of *rank-based* items of domain knowledge which could be used to draw *rank-based* atomic consequences, agreed consequences and logical consequences as presented in (Rauch, 2015). This approach will be further developed in future research.

Conclusions

We presented an Explanation system which utilises domain knowledge from trusted organisations. The above presented implementation proves the feasibility of the approach. More work needs to be addressed regarding validation of the approach on different data and use case scenarios. The proof of concept implementation could be elaborated upon in several directions as presented in previous sections. Furthermore, the relevancy of the explanations handed back to the user presents challenging research topic. Moreover, further elaboration on enhancements of the FOFRADAR framework based on *trusted knowledge* and *rank-based* items of domain knowledge as proposed in this paper remains as future work.

References

ANAND, S., BELL, D., HUGHES, J., 1995. *The role of domain knowledge in data mining*. IN Proceedings of the fourth international conference on Information and knowledge management (CIKM '95), Niki Pissinou, Avi Silberschatz, E. K. Park, and Kia Makki (Eds.). ACM, New York, NY, USA, 37-43.

DE BIE, T., 2013. *Subjective interestingness in exploratory data mining*. In Advances in Intelligent Data Analysis XII: 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013.

BUCHANAN, B. G., SMITH, R. G., 1998. *Fundamentals of expert systems*. Annual review of computer science, 1988, 3.1: 23-58.

CZECH STATISTICAL OFFICE (CSO), 2015. Výsledky sčítání lidu, domů a bytů 2011 (in Czech) [online]. https://www.czso.cz/csu/czso/otevrena_data_pro_vysledky_scitani_lidu_domu_a_bytu_2011_-sldb_2011- Last modified on 14 th April 2015.

DANUBIANU, M., 2013. *Joining Data Mining with a Knowledge-Based System for Efficient Personalized Speech Therapy* In Computational Collective Intelligence. Technologies and Applications. Volume 8083 of the series Lecture Notes in Computer Science pp 245-254

HAN, J., LAKSHMANAN, L., NG, R.T., 1999. *Constraint-based, multidimensional data mining*. *In IEEE Computer*, vol.32, no.8, pp.46-50, Aug 1999.

KOPANAS I, AVOURIS N., DASKALAKI S., 2002. *The role of domain knowledge in a large scale data mining project.* In: Vlahavas IP and Spyropoulos CD (eds). Methods and Applications of Artificial Intelligence, Proceedings for the Second Hellenic Conference on AI,SETN 2002, Thessaloniki. Lecture Notes in Artificial Intelligence, Vol. 2308. Springer-Verlag: Berlin, pp 288–299.

Mansingh, G., Osei-Bryson, K.-M., Reichgelt. H.: Using ontologies to facilitate post-processing of association rules by domain experts, Information Sciences, 181(3), 2011, 419–434.

PADMANABHAN, B., TUZHILIN, A., 1998. *A belief-driven method for discovering unexpected patterns*. In Proc. of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 94-100, 1998.

PAULHEIM, H., RISTOSKI, P., MITICHKIN, E., BIZER, C., 2014. *Data Mining with Background Knowledge from the Web*. RapidMiner World, At Boston, USA. August 2014

RAUCH, J., ŠIMŮNEK, M., 2014. *Learning Association Rules from Data through Domain Knowledge and Automation*. In: Rules on the Web (RuleML 2014). Springer LNCS, 2014.

RAUCH, J., 2015. *Formal Framework for Data Mining with Association Rules and Domain Knowledge – Overview of an Approach*. Fundamenta Informaticae, 137 No 2, pp. 1–47

SILBERSCHATZ, A., TUZHILIN, A., 1995. *On subjective measures of interestingness in knowledge discovery*. In Proc. of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 275-281, 1995.

QIANG, Y., XINDONG, W., 2006. *10 Challenging Problems in Data Mining Research*, International Journal of Information Technology & Decision Making, Vol. 5, No. 4, 2006, 597-604.

JEL Classification: C88

Shrnutí

Data mining s vysvětleními

V článku je popsán nový způsob využití doménové znalosti ve fázi vyhodnocení výsledků data miningové úlohy. „Vysvětlení“ je chápáno jako doménová znalost, která pomůže uživateli data miningového systému lépe porozumět výsledkům úlohy. Je popsána architektura nově navrženého systému a požadavky na jednotlivé komponenty. Jako zdroj doménové znalosti se používají data z důvěryhodných organizací, jako jsou statistické úřady nebo ministerstva. Důvodem je snaha vyhnout se zdlouhavému sběru doménové znalosti od expertů. Využitelnost systému je demonstrována v proof of concept implementaci využívající systém LISp-Miner, Python a SQLite. Jsou použita klientská data z velké finanční instituce a jako zdroj doménové znalosti data z Českého statistického úřadu. V závěru je navrženo rozšíření FOFRADAR frameworku s využitím nově definovaného pojmu „trusted knowledge“.

Klíčová slova: data mining, doménová znalost, vysvětlení, vyhodnocení výsledků, LISp-Miner, Python.

Big data from the perspective of data sources and legal regulation

Richard Novák

xnovr900@vse.cz

Ph.D. student of applied computer science

Thesis Advisor: doc. Ing. Vlasta Svatá CSc., (vlasta.svata@vse.cz)

Abstract. The paper deals with the Big Data in relation to a research questions: What are the available data types designated as "Big data"? and What are their specifics? It provides an overview of available data sources and a suggestion of new data categorization regarding Big Data specifics is presented. The paper selects Twitter as one specific data source from the area of social networks and focuses on it as a model case.

The Big Data phenomenon and Twitter model case is a good example of how technologies advanced legal regulation that usually follows with some delay. The conclusions shows that different ways of Big data categorization could be defined and all categories can contain the personal data that are subject of legal regulation in current Data Protection Act.

The article also describes current principles of Data Protection Act that will be followed in 2018 by even stronger regulation with coming General Data Protection Regulation.

Keywords: Big Data, data categorization, data collection, Data Protection Act, legislation, personal data, social networks, Twitter.

Introduction

Before we approach the Big data categorization and collection topic we should describe some of the existing definitions of Big Data and its attributes as it seems to be a good start for the following discussion.

For definition of Big Data we have selected definitions provided by the two consulting companies, McKinsey and Gartner:

"Big Data is a term very often used only in relation to high data volume, which is not right. It does not refer to the specific volume of stored data as TeraBytes or PetaBytes but to the volume that is exceeding the level typical for the different industries and the ICT environment they were designed for"(Manyika et al. 2011).

Definition of Big data is indeed much broader than just a reference to the volume of stored data that is, at least for some of the scientists, less important in comparison to other attributes as visible in following definition of Gartner:

"Big data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Gartner 2013).

The Gartner´s definition summarizes so called 3V that is an abbreviation for terms marking data with high: Volume, Variety and Velocity.

Volume means that thanks to the new data sources we have available much bigger volume than we are ready to deal with in existing ICT environment.

Variety means that data structure vary from standard structured data stored in enterprise databases to unstructured or semi-structured data as images, videos, systems logs, plain text and some others more deeply discussed in the next chapters are.

Velocity means either that data are changing very fast or that the need for speed of data processing is very high as we expect results of analysis almost in real time.

Veracity is the last and only occasionally used attribute meaning that is easy to jump to either wrong or none conclusion since the 3V data characteristics above makes data analysis very complex for later judgment (Chen et al. 2012).

The article focuses on the area of Big data and how to deal with this new type of data. The structure of the article follows main research questions:

What are the available data types designated as "Big data"? What are their specifics? And How we can collect them?

The issue of Big data is very important today. It is new ICT innovation tool (for more about ICT innovation see (Hanclova & Ministr 2013) and has the potential to enhance user satisfaction (more about user satisfaction is described in (Nemec & Zapletal 2012) and (Sudzina et al. 2011). Big data can extensively enhance the scope of Artificial Intelligence, since the more data means broader base for decision-making (Mikulecky & Tucnik 2013) as well as knowledge sharing (Oskrdal 2009). The Big data open also new questions related to ethics and privacy (Sigmund 2014). For above mentioned reasons we believe, that the topic of Big Data deserves to be analyzed in detail.

Overview of available data sources

Approach to data collection

When comes to the area of Big Data collection we should be aware of the purpose of such collection. In that aspect the Big data approach is partially similar to discipline of Competitive Intelligence (CI) as described for example in book Competitive Intelligence by Z. Molnár (Molnár 2012). One of many CI definitions is provided by Wikipedia:

"Competitive Intelligence is the action of defining, gathering, analyzing, and distributing intelligence about products, customers, competitors, and any aspect of the environment needed to support executives and managers making strategic decisions for an organization".

Competitive Intelligence defines in its investigation phase the Key Intelligence Topics (KIT) and Key Intelligence Questions (KIQ) that are further investigated in so called Intelligence cycle or process. The Intelligence process consists of several repeated phases: planning and directions, collections, processing and exploitation, analysis and production, dissemination and integration.

The Intelligence process was originally designed for the purpose of Intelligence services as CIA[1] and it is further described for example by author P. Zeman (Zeman 2010). Phases of Intelligence process are shown below in the Figure 1.

On the contrary, the Big Data approach has defined an area of interest that could be seen as similar to the KIT at Competitive Intelligence sector but the KIQ are replaced by the observation of the Big Data pool with a goal to find a new value or Competitive advantage in such data that was not known at the beginning of data collection.

To provide a short summarization, while within the Competitive Intelligence approach it is essential to define KIT and KIQ precisely, within Big Data the observation phase is essential as there the exact KIQ are not known at the beginning of data collection and observation.

Figure 1: Intelligence proces Source: Zeman, 2010

[1] (CIA) Central Intelligence Agency is one of the principal intelligence-gathering agencies of the United States federal government.

Data sources

When speaking about data collection and data categorization or classification we should primarily discuss data sources and data types in general and then focus on Big Data specifics.

Data sources in general may be called **white**, **grey**, **not published** and **black**. For deeper insight and further description we recommend special literature because for the purpose of this paper we will summarize it only shortly while using the description done by Z. Molnár in (2012).

White Data sources are official published data such as press releases, annual company reports, newspapers, magazines, official statistics and others. **Grey Data** sources are data that were not originally intended to be published but could be found in specific institutions or databases as SIGLE (System for Information on Grey Literature in Europe). Special category of sources is **Not Published Data** that are mainly known to its owner only and can be harvested by primary research as interviews and questionnaires.

There also exist **Black Resources** that are outside of public zone and are considered as closed data sources and secret information, e.g. medical, financial, and public authorities' secret data. Harvesting such data is considered unethical and could violate the law as well.

Speaking about White Data sources we should not forget the worldwide initiatives called **Open Data** that is defined as follows:

"Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copy write, patents or other mechanisms of control"(Aberer 2007). *The term "open data" itself is recent, gaining popularity with the rise of the Internet and World Wide Web and, especially, with the launch of open-data government initiatives such as http://www.data.gov/ or in the Czech Republic http://opendata.cz/"*

The Open Data national initiative exists also in the Czech Republic and is supported mainly by academics from University of Economics in Prague and from Faculty of Mathematics and Physics of Charles University in Prague.

Data location and structure

Based on its location data can be divided to internal and external. Internal data means the internal data of the organization that is stored, managed and often also produced by internal processes, systems and people.

External data is located outside the organization. Organizations often monitor external data sources and try to import them if they find them important for its function and if they are able to manage its import to internal systems. External data is very important for Competitive Intelligence and also for Big Data.

Another view of data types stems from the distinction whether the data are structured or unstructured. Structured data have defined format and can be stored to the specific location (table, row and column) of databases. Above structured data is possible to perform certain set of standard procedures, such as calculation, queries, reports and others.

Unstructured data is without given format: such as plain text or documents of special format (images, videos, and CADs). The simple examples of unstructured data are documents in MS Office / Open Office. The unstructured data are usually handled on the level of files or objects and are not manageable by standard Relational Database Systems as for example MS SQL, MySQL, and Oracle.

Table 1 summarizes it, however such standard view of data from the point of enterprises and its IT is not reflecting all the new Big Data challenges that we will focus on in next chapter.

Table 1: Standard categorization of data types Source: Molnár, 2012

Data Type	Internal source	External source
Structured	Enterprise databases and information systems as ERP, CRM etc.	Specific databases, catalogs, price lists etc.
Unstructured	Text files, emails, meeting minutes, special format report etc.	Web pages, social networks, emails etc.

Big data and different categorizations

Variety of data structure is one of the main attributes in Big Data definition. Big Data may be structured, unstructured but also semi-structured data that could be understood as special sub-category in Table 1 above.

Buneman's (1997) definition of semi-structured data: *"A form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data."*

As an example of semi-structured data we can name human readable formats which are generally all texts in ASCII format or special formats as XML (eXtensible Markup Language) or JSON (JavaScript Object Notation).

They are special databases as Hadoop or noSQL database engines that are designed to work with semi-structured and unstructured Big Data. For detailed description of these technologies see literature related to Big Data technologies e.g. (Manyika et al. 2011).

The opposite of the human readable formats is machine readable data that is data in binary formats or more complex structures that is readable by computers and machines in general.

From the point of view of Big Data categorization, which we are going to suggest, it is more important to know how the data is generated than if it is structured. For example spatial and transactional data are typically produced by machines and equipment. The machines that produce data are linked to the billing systems, telematics systems or other systems that can store and manage the transactional data without any human interaction.

We may recognize the need for a new more scalable categorization of Big Data based on the criterions that combine innovaBig Data source and original purpose of its generation.

One possible view on Big Data categorization (classification) is based on terminology of Homogenous and Heterogeneous Networks as proposed by Jiawei Han who is considered to be one of leaders of Knowledge and Data Discovery discipline (Han et al. 2012).

Homogenous networks, based on prof. Han´s classification, connect subjects and objects with the same class of data as people with people in social networks or machine with machine in telematics applications so it is a projection of real life to the data networks that are easier for humans to work with. On the other hand, the heterogeneous networks are more close to the real life as it is connecting people with their activities, peers, devices, generally everybody with everything.

For the purpose of this paper we have decided to use our own categorization (classification) specific for Big Data. Before we approach to that categorization we should summarize who or what generates the data or, in other words, what is the driver for data explosion in Big Data environment. Generally, data can be generated by people, machines or digital processes and data manipulation. New factors in Big Data approach represent for example smart devices in the area of machines or social media and Internet in general in the area of human activities. Hence, thanks to new applications we face new ways of people interactions and communication. In the area of data processing and manipulation we are currently in the situation where costs of storages are so low that we simply may digitally store everything about our personal or business activities without having to erase anything. It means that the more data we have, the more we produce by its manipulation. All these new factors interact together and this causes the data explosion marked as Big Data phenomenon. The simple summary of description above is shown in Table 2 below.

Table 2: Data generators and new Big Data factors Source: Author, 2015

Data Drivers Who/What generates data?	**Description of new Big Data factors**
People	Applications as social media and other in internet and Web 2.0 environment make people to communicate more, to interact and share rich content
Machines	Penetration of smart phones and smart devices increased significantly in last decade and these machines produced almost constantly high volume of data with variable data structure
Digital processes & Data manipulation	Costs of storages is so low that we store digitally everything about our activities and erase nothing so the more data we have, the more we produce by data manipulation

The more detailed categorization that we are going to propose is clustering the variable Big Data generators to the logical categories based on similarity of original data source. Suggested categorization has some intersections

between categories but it was created with the purpose to group generators of data to the logical categories with respect to Big Data specifics and also with respect to possible further data processing or commercial usage. Suggestion of such Big Data categorization may be found in Table 3 below.

Table 3: Big Data categorization suggestion Source: Author, 2015

ID	Data sources	Examples	Description and Purpose
1	**Smart devices**	Phones, TVs, Fridges	Devices that are online and can generate data and interact with other users and devices
2	**Social networks**	Facebook, Twitter, LinkedIn, YouTube, Instagram	Connect people with other people to entertain and exchange the information and content
3	**Spatial and transactional systems**	GPS, Payment systems, Telemetric systems	Locate subjects and objects, inform about systems and machines statuses, generate the transactions
4	**Corporate systems**	ERP, CRM, CMS, DW	Collecting business data about customers, employees, products, technologies and its attributes
5	**Systems with special securing efforts**	Public authorities, healthcare systems, security systems	Collecting private and sensitive data about persons, assets, finance and other data that are stored and managed with special security efforts
6	**World Wide Web - Internet**	Content of Internet designed for easy access of its users	System of interlinked hypertext documents accessed via the Internet. With internet tools as search engines it is very easy to consume the content but also to contribute to it
7	**Media production tools and equipment**	Cameras, Dictaphones	Video, audio, voice records creation, production and manipulation for private and business purposes
8	**Special sources**	CAD, emails, OCR	Special digital activities as for example Computer Aided Design (CAD) that thanks to humans or machines generate data that can be shared immediately (online) or stored offline and shared later

For legal reasons is essential to know what the content of such data sources is and whether any personal data in mentioned Big Data categories can be found.

Personal data definition is in the Czech Republic following the EU directives and is defined in the Data Protection Act. Under definition of this Act, a name and surname, birth identification number, address or a telephone number of a natural person fall within the scope of personal data and therefore shall receive the protection from the law.

Table 4 below summarizes the Big Data sources and through main drivers for data generation reflects also the possibility of such categories to contain personal data.

Table 4: Big Data categorization related to its drivers and personal data Source: Author, 2015

ID	Data sources	Main driver for Data	Can contain personal data?
1	**Smart devices**	Machines	Yes
2	**Social networks**	People	Yes
3	**Spatial and transactional systems**	Machines	Yes
4	**Corporate systems**	Digital processes and Data manipulation	Yes
5	**Systems with special securing efforts**	All	Yes
6	**World Wide Web - Internet**	People	Yes
7	**Media production tools and equipment**	Machines	Yes
8	**Special sources**	All	Yes

As visible from Table 4 above **all data categories** can contain personal data either directly or indirectly through the context or multiple data sources combination. This is an important finding for the legal consequences it might have.

Practical example of data obtained from social networks

When focusing on Big Data collection and analysis, we have decided to choose the social networks and Twitter as an example for the so called social mining done within this network (similar approach was used and described in (Smutny et al. 2013)).

Social networks and their further mining, analysis of the data and possible related ICT human capital is a topic that is well described by several authors (Boyd & Ellison 2007), (Malinova et al. 2012) or (Doucek 2009).

The data collection from Twitter is possible for developers through defined API (Application Programming Interface), through the use of programming rules that are officially published and described by Twitter at https://dev.twitter.com/docs/api/1.1.

We have defined the key words for data mining with the purpose of gathering all data from Twitter users whose tweets contain following key words. “GTS virtuální operator”, “virtuální operátor GTS”, “GTS MVNO”, “MVNO GTS”, “mobilní služby pro firmy”, “GTS mobilní služby”, “mobilní služby GTS”, “GTS Czech”, “GTS mobil”, “mobil GTS”.

When defining the key words we had in mind a telecommunication company GTS Czech, where the author works and which in 2013 launched the services of Mobile Virtual Network Operator (MVNO) that we try to monitor via Twitter.The purpose of such data collection could be Sentiment Analysis of Twitter users.

The result of such data collection done by authors with help of external programmers in 2013 is shown in table below:

Table 5: Result of data collection from Twitter done in 2013 Source: Author, 2013

ID	Key words	Tweet Message	User login	User name	Time
1	GTS virtuální operátor	Zajímavá nabídka ... Virtuál *GTS* startuje. Nejlevnější tarif stojí 35 korun bez daně - iDNES.cz - http://t.co/jZRR1kjHQs	http://twitter.com/xelagem	Michal Novak	14.5.2013 15:21
2	GTS virtuální operátor	Virtuál *GTS* startuje. Nabízí neomezené tarify pro firmy a živnostníky. http://t.co/pCkDyXqpIq	http://twitter.com/emanprague	eMan	14.5.2013 11:06
3	GTS virtuální operátor	Virtuál *GTS* startuje. Nabízí neomezené tarify pro firmy a živnostníky. http://t.co/sgz1X6uCFr	http://twitter.com/TomasCermak	Tomáš Čermák	14.5.2013 11:06
4	mobilní služby pro firmy	Operátor GTS Czech rozšiřuje své konvergentní portfolio *pro firmy* , spouští *mobilní služby* *** http://t.co/SUAlgeoLSx	http://twitter.com/feeditcz	FeedIT.cz	14.5.2013 17:18
5	GTS mobilní služby	pc-politika.cz - Články - technologie: Operátor *GTS* Czech spouští *mobilní služby* - http://t.co/lEjxvZwiLt	http://twitter.com/szabog66	Gabriel Szabó	14.5.2013 19:23
6	GTS mobilní služby	Operátor *GTS* Czech rozšiřuje své konvergentní portfolio pro firmy, spouští *mobilní služby* *** http://t.co/SUAlgeoLSx	http://twitter.com/feeditcz	FeedIT.cz	14.5.2013 17:18
7	GTS Czech	RT @*GTS _Czech* : @*GTS _Czech* : Dnes jsme na tiskove konferenci představili konkretni konvergentni mobilní nabídku pro firmy - viz.16:54 http://â€¦	http://twitter.com/LadVach	Ladislav Vachutka	14.5.2013 19:27
8	GTS Czech	pc-politika.cz - Články - technologie: Operátor *GTS Czech* spouští mobilní služby - http://t.co/lEjxvZwiLt	http://twitter.com/szabog66	Gabriel Szabó	14.5.2013 19:23
9	GTS Czech	@*GTS _Czech* : Dnes jsme na tiskove konferenci představili konkretni konvergentni mobilní nabídku pro firmy - viz.16:54 http://t.co/3w1pL3fkPZ	http://twitter.com/GTS_Czech	GTS Czech	14.5.2013 19:04
10	GTS Czech	Operátor *GTS Czech* rozšiřuje své konvergentní portfolio pro firmy, spouští mobilní služby *** http://t.co/SUAlgeoLSx	http://twitter.com/feeditcz	FeedIT.cz	14.5.2013 17:18
...	...	...	...	...	...

The experiment have proved, that such social mining was possible, later, when the results were analyzed, it brought interesting and novelty results.

Big data and legal regulation

As shown in Table 5, we are technically able to collect data from Twitter containing the key words but also the whole message and other attributes such as the names of Twitter users or time of placing the tweet. This brings us to the question of whether such collection and later analysis of users' messages is legal.

The European Union became aware of the dangers hidden in unregulated data gathering much earlier and in 1995 adopted the European Data Protection Directive (Data Protection Directive, 1995). The Directive came into effect on October 25, 1998 and the Czech Republic has implemented it in the Data Protection Act. The protection under the Data Protection Directive is again given to personal data only but it is broadly complemented by the practice of the European Court of Justice (ECJ). Currently there is novelty of Data Protection Act named as e General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) that was adopted on 27 April 2016. It enters into application 25 May 2018 after a two-year transition period and, unlike a directive, it does not require any enabling legislation to be passed by national governments. When the GDPR takes effect it will replace the data protection directive (officially Directive 95/46/EC) from 1995.

Basically, it can be said that as far as we gather and store data that are anonymous, we are, from the Czech legal perspective, safe. The problem however arises when we collect data that can be attributable to a particular

natural person and when such data are able to give us a means of identifying the person (e.g. first name and surname of the person). Such information as personal data is subject to the Data Protection Act and therefore must be secured within the rules given by this Act.

According to the Data Protection Act an entity that would like to gather, store or preserve data that may contain personal data has to register itself and the intended aim of gathering within the Office for Personal Data Protection unless such gathering, storing or preserving is done under special legislation that imposes such an obligation. With respect to a general requirement that any kind of handling with personal data may be done with consent of the recipient only, the Data Protection Act also requires for data controllers to notify the recipient of personal data about the extent and aim of data gathering and about its further handling unless the recipient already knows such information.

Personal data may be collected by a data controller (or by a data processor who is however subordinated to a data controller) for specific, explicit and legitimate purposes only and for those purposes also processed. Further processing is not forbidden but it must be done in a way compatible with the original purposes in only such a way that is perceived as being in line with the law. That means that a data controller is allowed to further process personal data for a purpose different from the original purpose for which it was gathered but such processing needs to fall within the scale compatible with the original purpose which was disclosed to the recipient during the gathering.

Conclusion

The article proposed new categorization (classification) specific for Big Data. It took into consideration source of data (Data Drivers / Who / What generates data? – on the scale People /Machines / Digital processes & Data manipulation) and new Big Data factors – see Table 2.

We also presented more detailed categorization of clustering the variable Big Data generators to the logical categories based on similarity of original data source (Smart devices, Social networks, Spatial and transactional systems, Corporate systems, Systems with special securing efforts, World Wide Web – Internet, Media production tools and equipment, Special sources). It was created with the purpose to group generators of data to the logical categories with respect to Big Data specifics and also with respect to possible further data processing or commercial usage – see Table 3.

We also mapped possible collisions with the Personal Data Act – see Table 4 and illustrated on practical example, how big data can be harvester from social network site Twitter.

The next steps in PhD thesis will extend the existing research and this paper about the area of Ethics related to Big data phenomenon that logically follows the existing publication of the author focused on Big data from the perspective of data sources and legal regulation.

References

Aberer, K., 2007. The semantic Web. In 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 ASWC 2007. Busan, Korea.

Act No. 101/2000 Coll., on the Protection of Personal Data and on Amendment to Some Acts (Data Protection Act), as amended.

Boyd, D.M. & Ellison, N.B., 2007. Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication, 13(1), pp.210–230. DOI: 10.1111/j.1083-6101.2007.00393.x

Buneman, P., 1997. Semistructured Data. Tutorial. In PODS '97, Symposium on Principles of Database Systems.

Doucek, P., 2009. ICT HUMAN CAPITAL - RESEARCH AND DEVELOPMENT WORK IN ICT. IDIMT-2009: SYSTEM AND HUMANS, A COMPLEX RELATIONSHIP, 29, pp.83–94. WOS:000273266300009, ISBN:978-3-85499-624-8

Gartner, 2013. Big Data Analytics - Free Gartner Research. Gartner IT glossary. Available at: http://www.gartner.com/it-glossary/big-data/ [Accessed January 18, 2015].

Han, J., Kamber, M. & Pei, J., 2012. Data mining: concepts and techniques 3rd ed., Waltham: Morgan Kaufmann.

Hanclova, J. & Ministr, J., 2013. THE USE OF ICT TOOLS AS INNOVATIONS IN SMES IN THE CZECH-POLISH BORDER AREAS. IDIMT-2013: INFORMATION TECHNOLOGY HUMAN VALUES, INNOVATION AND ECONOMY, 42, pp.129–136 WOS:000333272200015, ISBN:978-3-99033-083-8

Hortonworks &Teradata, 2013. Hadoop & the Enterprise Data Warehouse: When to Use Which. Webinars.

Chen, H., Chiang, R. & Storey., V., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), p.25.

Malinova, L., Pavlicek, A. & Rosicky, A., 2012. COMPARATIVE SURVEY OF STUDENTS' BEHAVIOR ON SOCIAL NETWORKS (IN CZECH PERSPECTIVE). 38, pp.373–375. WOS:000315973000042, ISBN:978-3-99033-022-7

Manyika, J. et al., 2011. Big data: The next frontier for innovation, competition, and productivity, Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation [Accessed January 18, 2015].

Mikulecky, P. & Tucnik, P., 2013. Beyond Artificial Intelligence J. Kelemen, J. Romportl, & E. Zackova, eds., Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-34422-0_12, WOS:000313206700012, ISBN:978-3-642-34422-0, ISSN: 2193-9411

Molnár, Z., 2012. Competitive intelligence, aneb, jak získat konkurenční výhodu, Praha: Oeconomica, ISBN: 978-80-245-1908-1.

Nemec, R. & Zapletal, F., 2012. THE PERCEPTION OF USER SATISFACTION IN CONTEXT OF BUSINESS INTELLIGENCE SYSTEMS' SUCCESS ASSESSMENT. IDIMT-2012: ICT SUPPORT FOR COMPLEX SYSTEMS, 38, pp.203–211 WOS:000315973000022, ISBN:978-3-99033-022-7

Oskrdal, V., 2009. SHARING KNOWLEDGE: USING OPEN-SOURCE METHODOLOGIES IN IT PROJECTS. IDIMT-2009: SYSTEM AND HUMANS, A COMPLEX RELATIONSHIP, 29, pp.399–408. WOS:000273266300040, ISBN:978-3-85499-624-8

Sigmund, T., 2014. PRIVACY IN THE INFORMATION SOCIETY: HOW TO DEAL WITH ITS AMBIGUITY? IDIMT-2014: NETWORKING SOCIETIES - COOPERATION AND CONFLICT, 43, pp.191–201. WOS:000346148300021, ISBN:978-3-99033-340-2

Smutny, Z., Reznicek, V. & Pavlicek, A., 2013. MEASURING THE EFFECTS OF USING SOCIAL MEDIA IN THE MARKETING COMMUNICATIONS OF THE COMPANY: PRESENTATION OF RESEARCH RESULTS. , 42, pp.175–178 WOS:000333272200020, ISBN:978-3-99033-083-8

Sudzina, F., Pucihar, A. & Lenart, G., 2011. ON THE RELATIONSHIP BETWEEN PERCEIVED IMPORTANCE OF ERP SYSTEMS SELECTION CHARACTERISTICS AND SATISFACTION WITH SELECTED ERP SYSTEMS. IDIMT-2011: INTERDISCIPLINARITY IN COMPLEX SYSTEMS, 36, pp.59–69. WOS:000299185000006, ISBN:978-3-85499-873-0

Zeman, P., 2010. Zpravodajský cyklus – klišé nebo nosný koncept? Obrana a strategie, 10(1), pp.45–64. Available at: http://www.defenceandstrategy.eu/cs/aktualni-cislo-1-2010/clanky/zpravodajsky-cyklus-8211-klise-nebo-nosny-koncept.html#.VOSDPvmG8Uc [Accessed January 18, 2015].

JEL Classification: M150

Shrnutí

Big data z perspektivy datových zdrojů a právní regulace

Článek se zabývá pojmem Big data ve vztahu k výzkumné otázce: Jaké jsou dostupné datové typy a zdroje pro které používáme v současnosti pojem Big data? a Jaká jsou jejich specifika? Článek poskytuje přehled dostupných datových zdrojů a představuje několik vlastních návrhů pro kategorizaci Big dat na základě jejich specifických odlišností. Článek si vybírá Twitter jako jeden konkrétní datový zdroj z oblasti sociálních sítí a používá ho jako modelový případ.

Fenomén Big data a modelový příklad Twitteru je dobrou ukázkou toho jak technologie předchází právní regulaci, která obvykle reaguje až s určitým zpožděním. V závěru článku ukazujeme, že ačkoliv můžeme navrhnout odlišné kategorie v oblasti Big data, ve všech se mohou vyskytovat osobní údaje, které jsou předmětem právní regulace.

Článek se také zabývá principy současné právní normy označované jako Zákon o ochraně osobních údajů a jeho plánovanou novelou označovanou v evropském kontextu jako General Data Protection Regulation.

Klíčová slova: Big Data, kategorizace dat, legislativa, osobní data, sběr dat, sociální sítě, Twitter, Zákon o ochraně osobních údajů.

Koncepce měření výkonnosti veřejné správy

Miroslav Řezáč
miroslav.rezac@vse.cz
Doktorand oboru Aplikovaná informatika
Školitel: doc. Ing. Ota Novotný, Ph.D. (ota.novotny@vse.cz)

Abstrakt: Měření výkonnosti patří k neodmyslitelným součástem soukromého sektoru. V oblasti veřejné správy, nicméně tyto postupy a praktiky natolik rozšířené nejsou, zejména pokud se zaměříme na Českou republiku.

Článek se zabývá návrhem přístupu k měření výkonnosti ve vybraných oblastech veřejné správy za použití definovaných KPI indikátorů. Následně prezentuje, na základě výpočtu v oblasti nemovitostí, potenciál úspor, pokud by došlo v souvislosti s výsledky měření i k přijmutí několika rozpočtových opatření.

Klíčová slova: měření výkonnosti, KPI, klíčový indikátor výkonnosti, správa nemovitostí, benchmarking, státní správa, úspory

Úvod

Reporting KPI, měření a následné řízení výkonnosti na jejich základě, by v dnešní době mělo být součástí každé velké společnosti soukromého sektoru. Úkolem manažerů těchto společností je data sbírat, analyzovat, převést na informace a na jejich základě přijímat rozhodnutí, která společnost řídí. Tyto aktivity probíhají jak na úrovni řízení samotného podniku (pobočky, dceřiné společnosti), tak zejména na centrální úrovni podnikového holdingu (matky), kde jejich význam dramaticky stoupá.

Hlavní podstatou řízení výkonnosti je definování klíčových parametrů výkonnosti (KPI) pro každé oddělení nebo oblast. Mezi klíčové KPI patří např. výkonnost, kvalita a efektivnost. Každé KPI má také nastavenu svoji váhu, na jejímž základě se následně dá spočítat hodnocení zaměstnance, výkonnost oddělení či oblasti a celková pozice firmy v rámci konkurenčního prostředí. Z tohoto důvodu je nutné mít vytvořený systém reportingu, který všechna tato data obsahuje a umožňuje díky nim zpracovávat i další analýzy a modely. Měření výkonnosti je tedy nezbytné pro efektivní řízení organizace [Wagner, 2009].

KPI se obvykle používají v oblastech, jako je měření výkonu infrastruktury a operací, měření parametrů aplikací, lidské zdroje, ve finančním managementu a podobně. KPI definují standardy, které mají zajistit, že se výkon pohybuje v žádaném rozmezí, a že procesy fungují tak, jak mají. Výkyvy jsou pak indikátorem různých problémů, které je potřeba řešit.

Ve společnostech soukromého sektoru se ale také často stává, že dat se sbírá spousta bez ohledu na jejich potřebu a nikdo s nimi dále nepracuje. Na úrovni veřejného sektoruje situace velice obdobná. V současné době proto vznikají aktivity, které umožňují monitorovat a měřit výkonnost jednotlivých organizačních složek státu (benchmarking), na jejich základě však dále nejsou přijímána rozhodnutí.

KPI a benchmarking ve VS

Metriky KPI v oblasti VS (veřejné správy) vychází z dlouholetých zkušeností zemí jako je např. Velká Británie ([Cabinet Office, 2010]), Kanada ([OAGC 1]) nebo Nový Zéland ([NZ Treasury, 2011]). Ukazatele výkonnosti a efektivnosti veřejné správy jsou v ní rozděleny na ty, které měří tzv. „front office“ tj. služby veřejné správy poskytované obyvatelstvu a firmám, „middle office“, tj. podpůrné funkce státu a ty, které měří tzv. „back office“, tj. část správy zdrojů státu, konkrétně interní služby a provozní aktivity OVM (orgánů veřejné moci), které jsou nezbytné pro poskytování jejich “front office“ služeb.

Měření KPI a řízení výkonnosti interních služeb, tedy „back office“, je obvykle prvním stupněm přechodu k efektivnímu řízení veřejné správy. Tato oblast navíc umožnuje plošné srovnávání orgánů veřejné moci mezi sebou. Na něj následně navazuje měření a řízení výkonnosti externí služeb „front office“, již však resortně a agendově neporovnatelných.

Impulsem pro vznik funkčního systému měření výkonnosti, potažmo jakýchkoli manažerských informací, není dokonalý logický model indikátorů a jejich měření, ani technicky dokonalý IT systém pro sběr informací, ale poptávka po těchto informacích, typicky ze strany seniorních finančních manažerů ([NERV KPI, 2013]). Jedině jsou-li informace využívány, existuje vůle poskytovat informace užitečné a reálné, a jedině skrz opakované

používání a analýzu dat lze zjistit, jaká zlepšení jsou třeba a jak systém skrze používání postupně zlepšovat. V opačném případě se sběr dat stane byrokratickým cvičením.

Např. v Austrálii, na Novém Zélandě a v Kanadě jsou za benchmarking zodpovědná ministerstva, která zároveň mají na starosti výdajovou stránku rozpočtu a současně mají povinnost zaručit funkčnost a efektivitu státní správy jako celku (NZ: Treasury; Kanada: Treasury Board Secretariat; Austrálie: Department of Finance).

Přístupy aplikované v zahraničí (VB, NZ)

Stejně jako v podnicích je kromě měření a porovnávání KPI je ve veřejné správě třeba ještě udělat další krok k využití výsledků k vlastnímu řízení veřejné správy - například pro nastavení limitů rozpočtových kapitol nebo odměňování vrcholových manažerů.

Ze srovnávacího výzkumu (VB, NZ, Austrálie, USA) ale nevyplynul případ, kde by se podařilo informace tohoto typu zapojit do rozpočtování nebo strategických a policy rozhodnutí. Finanční informace se daří do rozhodování a rozpočtování zapojovat tehdy, jsou-li podstatné z pohledu konkrétních služeb nebo úspěšnosti iniciativ či politik ([Bouchal, 2013]).

Komplexní systémy řízení výkonnosti jsou pak úspěšné především tehdy, jsou-li navázány na zájem a priority premiéra (typicky ve VB Blair a systém PSAs; v Austrálii i na Novém Zélandu je pozornost napříč státní správou věnována především těm aspektům systému měření výkonnosti, o něž se zajímá premiér, popř. je využívá šéf úřadu vlády, který je typicky nadřízeným jednotlivých manažerských šéfů ministerstev —Austrálie, Nový Zéland, Kanada).

New Public Management

Snahu měřit a řídit agendu veřejné moci v zahraničí nejsilněji odráží koncept nazývaný New Public Management (NPM). Původně existující systém veřejné správy byl hodnocen jako vysoce byrokratický, a proto bylo třeba vytvořit koncept nového řízení organizací veřejné správy, který umožní odbourat tyto negativa a zavede nové změny řízení ve veřejné správě, čímž dojde ke zdokonalení o tržní principy a prvky ustálených pravidel řešení veřejných potřeb.

Reformy New Public Management vytvářejí také novou administrativní kulturu, která se odlišuje od byrokratických hodnot a bere v potaz faktor lidské motivace. Tím nahrazuje řízení podle pravidel řízením podle cílů a výsledků, čímž se zvyšuje účinnost, efektivita a transparentnost. Již v 80. letech 20. století se začala projevovat důležitost zvýšení efektivnosti ve veřejném sektoru a v této době jsou také viditelné snahy využívat jako vzor prvky řízení, které byly do té doby běžné pouze v soukromém sektoru. [MALÝ, 2011]

Vznik konceptu New Public Management je datován k roku 1968 na konferenci v Minnowbrooku a rozvíjel se právě v 80. a 90. letech 20. století. K prvnímu využívání principu NPM došlo na Novém Zélandu a rozšířil se více ve Velké Británii a USA než v kontinentální Evropě. [SIEGEL, M.] Dále se rozvíjel i v dalších zemích, například v Nizozemí, Německu, Švédsku, Dánsku, Finsku, Švýcarsku, Kanadě, Austrálii, ve Francii, na Islandu a v posledním desetiletí můžeme snahy o zavedení NPM pozorovat v některých zemích Afriky, Střední Ameriky a také v zemích střední a východní Evropy. V závislosti na zemích, kde se NPM prosazoval, se rozvinulo několik manažerských přístupů k řízení ve veřejném sektoru, které se více či méně odlišovaly:

- Britský přístup se vyvíjel přibližně od šedesátých let po reformách Margaret Thatcherové, v počátcích se zaměřoval především na zvyšování kvalifikace státních zaměstnanců, v současnosti se prosazuje v dobře manažersky zajištěném poskytování veřejných služeb občanům.
- Americký přístup byl charakteristický spoluprací více vědních oborů při realizaci reforem ve veřejné správě a zkušenosti o moderních metodách managementu přebíral z úspěšných soukromých podniků.
- Evropský kontinentální přístup zase začal přizpůsobovat vzájemně právo a právní normy s obsahem funkcí a chování pracovníků veřejné správy. [HRABALOVÁ, 2006]

Tabulka 3: NPM vs. původní koncept řízení VS [OCHRANA, 2012]

Hodnotící pohled	„Předreformní“ způsob řízení	Způsob řešení navrhovaný konceptem NPM
Podstata řízení	Hierarchické řízení, striktně byrokratické	Decentralizované řízení
Organizační aspekt	Veřejná správa je hierarchicky učleněná, má striktně vymezenou organizační strukturu	Decentralizovaná struktura. Řízení je realizováno na základě výsledků.
Role zaměstnance veřejné správy	Byrokrat jako monopolista, „vládce“ nad občanem	Byrokrat jako „služebník“ občana (orientace na zákazníka)

Hodnotící pohled	**„Předreformní“ způsob řízení**	**Způsob řešení navrhovaný konceptem NPM**
Motivy k činnosti	Systém nepobízí k výkonu a službě občanovi	Iniciuje kvalitní výkon, sleduje jej v indikátorech a má nástroje k jeho dalšímu podněcování.
Poskytování veřejných služeb a statků	Je obvykle poskytován univerzální veřejný statek či služba.	Prvky flexibilního poskytování veřejných služeb ve formě individualizovaných produktů.
Způsob řízení veřejných zdrojů	Přísně byrokratický (pyramidový) systém řízení (přidělování) zdrojů	Zdroje jsou alokovány s ohledem na stanovené cíle a se zřetelem na omezující podmínky
Forma řízení veřejných zdrojů s ohledem na výstupy	Procesně orientované řízení alokačních aktivit. Daná rozpočtová jednotka plní centrálně stanovené úkoly.	Řízení orientované na výsledek, monitorované v kontrole ex ante a kontrole ex post na bázi výkonových ukazatelů.

Případová studie - pilotní ověření v rámci NERV KPI

V rámci mého pilotního výzkumu bylo nutné, zejména s ohledem na dostupnost dat, zúžit zkoumanou oblast na problematiku řízení hospodaření/správy nemovitostí.

Obsahem správy nemovitého majetku je optimalizace řízení správy majetku a podpůrných procesů s tím spojených. Hlavním cílem jsou celkové úspory v podpoře a správě majetku, rychlejší vyřízení požadavků, efektivnější využívání majetků a podpora všech uživatelů nemovitostí respektive zaměstnanců společností.

Způsob měření a benchmarkingu v oblasti správy nemovitostí

Se zohledněním výše zmíněných zkušeností ze zahraničních států pro oblast nemovitostí by správně nadefinované metriky KPI měly poskytovat, v souladu s [Vyskočil, 2007] a [Hrabě, 2013], odpovědi na následující otázky:

- Jaká je nákladová efektivita funkcí správy nemovitostí?
- Absolutně /meziročně a v poměru k velikosti organizace (počtu zaměstnanců)
- Zajišťuje správa nemovitostí efektivní (úsporné) využití kancelářské plochy, resp. využívá organizace kancelářské plochy úsporně (m^2/zam.)?
- Jak se zlepšuje profesionalita správy nemovitostí?
- Jaká je kvalita poskytovaných služeb správy nemovitostí?

Na výše uvedené otázky reaguje v Evropské unii nový standard pojmů a funkcí v oblasti správy nemovitostí stanovuje (sada norem označovaná EN 15221, 1-7, [EN 15221, 2007]).

Text normy předpokládá, že ukazatele výkonnosti v oblasti správy nemovitostí budou rozděleny do následujících šesti kategorií:

Finanční ukazatele

- Celkové náklady správy nemovitostí, členěné podle jednotlivých skupin výstupů a přepočtené na FTE, pracoviště a m^2.

Prostorové ukazatele

- Výměra čisté podlahové plochy (NFA) přepočtená na FTE, jmenovité zaměstnance a pracoviště
- Sekundární poměry ploch:
- NFA / TLA (Net Floor Area / Total Level Area) (%)
- IFA / TLA (Internal Floor Area / Total Level Area) (%)
- GFA / TLA (Gross Floor Area / Total Level Area) (%)

Ukazatele životního prostředí

- Celková produkce CO_2 (tzv. uhlíková stopa), v tunách ročně, celkově a přepočtená na FTE a m^2 výměry NFA.
- Celková spotřeba všech energií (kWh ročně), celkově a přepočtená na FTE a m^2 výměry NFA.
- Celková spotřeba vody (m^3 ročně), celkově a přepočtená na FTE a m^2 výměry NFA.
- Celková produkce odpadu (v tunách ročně), celkově a přepočtená na FTE a m^2 výměry NFA.
- Další ukazatele životního prostředí, členěné dle kategorií výstupů správy nemovitostí, viz. detaily v normě EN 15221

Ukazatele kvality služeb

- Ukazatel kvality služeb dle porovnání se standardy, celkově a členěné podle vybraných kategorií výstupů

Ukazatele spokojenosti

- Ukazatel spokojenosti se správou nemovitostí dle průzkumu spokojenosti uživatelů, celkově a členěné podle vybraných kategorií výstupů.

Ukazatele produktivity (vnitřní, back-office ukazatele organizace správce nemovitostí)

- Ukazatele správy vnitřních zdrojů útvarů a organizací správy nemovitostí, jako například:
- Fluktuace zaměstnanců
- Nemocnost
- Docházka apod.

Pro dosažení celkových nákladů facility managementu, EN 15221-7 předepisuje přidat přepočtené kapitálové výdaje k ročním provozním výdajů, ale odečíst roční příjmy. Po sběru a analýzy dat, mohou finanční kritéria být vyjádřena v nákladech na FTE, na pracovišti nebo na čtvereční metr.

Doporučení jak řídit oblast správy nemovitostí, popsané v další kapitole, si logicky nedává za cíl suplovat roli vlády a rozhodovat o přístupu k nemovitostem ve vlastnictví státu, ale poskytuje pouze návody/rady, jak se dají sledovat, vyhodnocovat a kontrolovat případné neefektivnosti a jaká rozhodnutí by mohla vláda na základě získaných dat přijmout.

Tedy konkrétně, pokud se vláda rozhodne, že bude tzv. bydlet primárně ve vlastním, že bude co nejlépe využívat svých budov (tržní nebo vyšší nájemné ve svých budovách, tržní nebo nižší smlouvy o pronájmu u třetích stran, efektivně využité prostory apod.)

Návrh KPI pro měření oblasti správy nemovitostí v ČR

Pro měření výkonnosti v oblasti správy nemovitostí v ČR byly v rámci pilotního projektu doporučeny následující KPI:

$$KPIvarA = \frac{a + b + c - d}{\frac{e}{f}} \qquad KPIvarB = \frac{a + b + c}{\frac{e}{f}}$$

kde

a jsou výdaje na provoz a údržbu všech budov vlastněných institucí,
b je nájemné plochy najaté institucí,
c jsou výdaje za energie a vodu najatých nemovitostí,
d jsou příjmy z pronájmu budov spravovaných,
e je počet m^2 plochy,
f je počet FTE (full time employees).

Vláda však může učinit i rozhodnutí, že je cílem maximálně využít státních budov a sestěhovat všechny instituce z pronájmu do volných státních budov. Potom lze zvažovat naznačenou variantu KPIvarB.

Tyto ukazatele by měly být primárními zdroji pro regulaci nákladů na budovy státu. Pro detailní rozhodování manažerů jednotlivých státních institucí a pro vyvrácení některých mýtů, které může nastolit pouhý jeden ukazatel, doporučuji využít následující sadu doplňkových ukazatelů, které již poskytují přesnější odraz studované skutečnosti a eliminují většinu možných protiargumentů neefektivity.

Sada dílčích ukazatelů (pro management jednotlivých institucí)

- počet m^2 / počet FTE
- odchylky cen při nájmu prostor
- odchylky cen při pronájmu prostor

Návrh způsobu řízení výkonnosti v oblasti správy nemovitostí v ČR

Metriky a KPI jsou základem efektivního řízení, nikoliv však jeho cílem. Pro oblast nemovitostí mezi vytyčené cíle řízení patří:

- Optimálně vybavené pracoviště zaměstnance veřejné správy
- Trvale udržitelný rozvoj objektů a vybavení VS
- Minimalizace nákladů na správu nemovitostí při zachování ostatních cílů

- Minimalizace spotřeby médií a energií

Na základě sledovaných hodnot KPI lze dle mého návrhu oblast nemovitostí řídit dvěma způsoby:

- strop výdajů na správu a nájem (v rozlišení dle kategorie organizace)
- maximální plocha na 1 FTE

Po zvážení obou alternativ bylo zvoleno zastropování výdajů v rozpočtu udělat na úrovni KPI(A) nebo KPI(B) spočítaných přes všechny instituce dané kategorie = KPI(median). Strop stanovit na KPI(median) + 10%.

Data

Pro pilotní ověření metodiky byla zvažová data ze dvou zdrojů – systému CRAB a databáze VDK. Prvním z nich byl systém CRAB (Centrální Registr Administrativních Budov). Jedná se o jednotnou databázi evidující administrativní budovy na celorepublikové úrovni, která je spravována ÚZSVM (Úřad pro zastupování státu ve věcech majetkových). Z úvodní analýzy tohoto zdroje vyplynulo, že proběhla iniciální pasportizace budov a registr byl době analýzy naplněn daty 14ti organizací (z celkového počtu cca 670 státních institucí, byť se jednalo o klíčové, silové resorty), neb evidence dat k tomuto datu probíhala na principu dobrovolnosti. Z důvodu neúplnosti obsahu dat v tomto systému, k datu analýzy, bylo od tohoto zdroje upuštěno a pro pilotní ověření byly zvoleny podklady z druhého zdroje, databáze VDK (Vládní dislokační komise).

Obsah této databáze tvoří zejména data k užívaným objektům pocházející z dotazníkového šetření napříč státními institucemi z roku 2006. Z informací ÚZSVM vyplývá, že tato data představují informace o cca 60% objektů, které jsou v užívání státu. Jedná se o stavová data, která byla v průběhu let 2007 a 2008 u vybraných objektů ještě aktualizována, s cílem jejich zpřesnění.

Aplikovatelnost KPI

Z analýzy dostupných dat vyplývá, že metriku KPI(A) nelze nad dostupnými pilotními daty sestavit, protože neobsahují údaj o nájemném, které je organizacím placeno. Obsahují však nájemné, které organizace platí jinde, včetně ostatních atributů potřebných pro sestavení metriky KPI(B). Tato je tudíž na dostupná data aplikovatelná a lze ji v pilotu využít.

Ke každému objektu proto byla dopočítána hodnota KPI(B) podle vzorce

KPI(B) = (výdaje na provoz a údržbu objektu (včetně energie, vody, plynu) + nájemné placené v objektu / počet m^2 plochy / počet FTE

Pro každou organizaci (tzn. OSO) byl následně ze všech objektů stanoven medián KPI(B), a její jednotlivé objekty následně byly podrobněji srovnány mezi sebou (peer group). Plošná aplikace jednoho KPI (nastavení jeho cílové hodnoty) pro celou veřejnou správu by totiž mohly vést ke špatným rozhodnutím.

Jako příklad lze uvést ukazatele ze správy nemovitostí, kde např. soudy mají ze své podstaty jiné požadavky na velikost kancelářské plochy na zaměstnance než např. hygienické stanice. Na druhou stranu porovnávání tohoto KPI mezi jednotlivými soudy navzájem má velký význam.

Doplňkově byl spočítán též medián KPI(B) v rámci jedné organizace (OSO) v konkrétním kraji, který zohledňuje v případě objektů organizace i různé nájemné v rámci krajů. Zbylé objekty organizace pak opět byly s tímto výsledkem srovnány.

Představený systém měření a porovnávání výkonnosti prostřednictvím KPI může kromě podpory vlastního řízení z pozice vlády dále sloužit různým účelům - od měření hospodárnosti práce se zdroji a účinnosti jejich využití pro výstupy, přes vyhodnocování investic ve VS (včetně IT investic), dále přes motivaci a odměňování manažerů a zaměstnanců VS až k výkonovému a cílově i k výsledkovému rozpočtování.

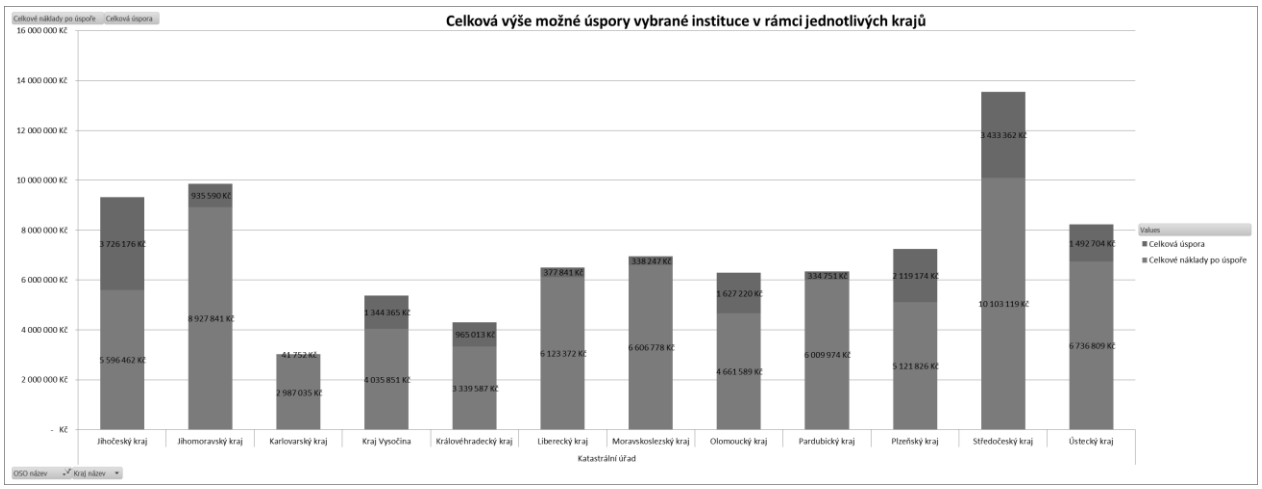
V rámci pilotního ověření byl spočítán i jeden z dílčích ukazatelů - počet m^2 / počet FTE. Opět na úrovni každé organizace a doplňkově na úrovni organizace v rámci kraje.

Pilotní výstupy

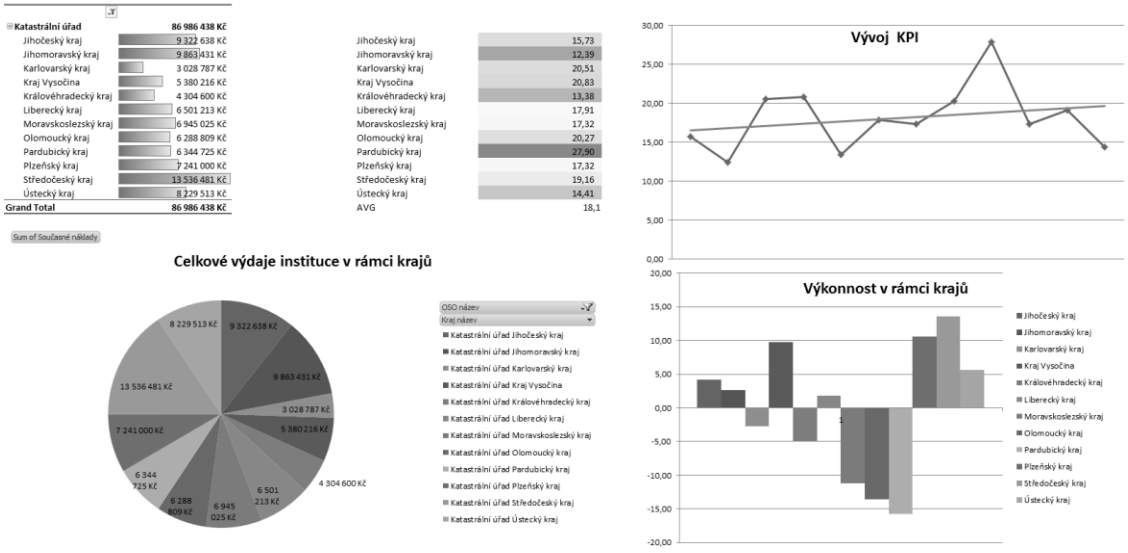
Výpočet vybraného KPI i dílčího ukazatele, které jsou popsány v předchozí kapitole se úspěšně nad pilotními daty povedlo realizovat. Nejdříve byl proveden nápočet všech relevantních atributů na úrovni objektů. Tyto pak následně byly sdruženy do skupin podle příslušnosti k OSO a v těchto skupinách stanoven medián ukazatele KPI(B) i dílčí ukazatel m2/FTE, variantně i podle OSO v jednotlivých krajích.

V oblasti řízení hospodaření s nemovitostmi byly aplikovány obě navržené metody. Byl stanoven strop výdajů na správu a nájem v rozsahu – medián skupiny + 10%, stejně tak byla stanovena pro každou skupinu i maximální plocha je jednoho zaměstnance.

Pro demonstraci efektů, které pramení z tohoto manažerského rozhodnutí, byla na základě stanovených stropů v pilotním vzorku dat spočítána celková úspora nákladů. Z výsledků vyplývá, že při stanovení stropů na základě organizace/organizační složky lze dosáhnout úspory až přibližně 870 mil. Kč. V případě zastropování výdajů na organizaci v konkrétním kraji se pak dostáváme k 777 mil. Kč.



Obrázek 2: Příklad potenciální úspory dle jednotlivých krajů



Obrázek 3: Příklad zpracovaného dashboardu zobrazujícího KPI

Aktualizace dat

Souhrnné výsledky předchozího výzkumu (Procházka, Voříšek, Novotný, 2013) byly presentovány tehdejšímu premiérovi ČR (Ing. Nečas), který následně zadal tuto oblast rozvíjet ministerstvu financí. Na tuto aktivitu následně navázal 1. místopředseda vlády a ministr financí Ing. Babiš oficiálním dopisem se žádostí o poskytnutí maximální součinnosti ÚZSVM v oblasti CRAB. Proto jím byli osloveni zástupci ústředních orgánů státní správy s výčtem objektů, které dosud nejsou zaevidovány v CRAB s povinností zaevidovat chybějící objekty do CRAB do 31. 8. 2014. Ke konci září 2014 byl registr z 95 procent naplněn, proto bylo možné realizovat další, přesnější výpočty.

Z výsledků je zřejmé, že

- stát nyní v průměru inkasuje za pronájem svých budov 1071 korun za m2. Zároveň ale státní úřady v soukromých pronajatých budovách platí v průměru 1437 korun za m2.

Pokud by dále byly aplikovány principy zastropování kancelářské plochy na jednoho zaměstnance, lze realizovat následující úsporu :

Výpočet úspory zastropování plochy na 12m^2:

- rozdíl mezi Průměrnou plochou kanceláří na 1 zaměstnance a stanoveným stropem – číslicí 12, násobený počtem Dislokovaných zaměstnanců
- Možná úspora výdajů za nájemné při snížení užívané kancl. plochy (výpočet: Možná úspora při respektování kancelářské plochy max. 12m^2 na osobu * Nájem na m^2 a rok),
- Možná úspora výdajů za provoz a údržbu při snížení užívané kancl. plochy (výpočet: Možná úspora při respektování kancelářské plochy max. 12m^2 na osobu * Výdaje za provoz a údržbu na m^2 a rok),
- Možná úspora výdajů za nájemné při snížení výše nájmu na průměrné nájemné v dané lokalitě (výpočet: rozdíl mezi Nájmem na m^2 a rok a Průměrným nájemným v dané lokalitě násobený užívanou plochou celkem)

Závěr

Při zmenšení kancelářské plochy připadající na jednoho zaměstnance na průměrných 12 m^2 by v celé České republice ubylo 358 tisíc m^2 kancelářské plochy obsazené státními zaměstnanci ve státních objektech. U komerčních objektů by díky tomuto snížení celkové hrazené nájemné mohlo klesnout až o 126,5 milionu korun. Dalších zhruba 48 milionů korun by se ušetřilo za výdaje na provoz a údržbu těchto objektů. Dohromady by tedy standardizace průměrné plochy kanceláří na 12 m^2 v komerčních objektech mohla přinést úspory ve výši přibližně 174,5 milionu korun.

Ušetřit nemalé prostředky by bylo možné také redukcí jednotkového nájemného, které státní instituce hradí za užívání kancelářských ploch v komerčních objektech. Pokud by se podařilo tyto výdaje za nájem snížit na průměr zjištěný z CRAB v jednotlivých obcích, znamenalo by to úsporu zhruba 187 milionů korun.

Literatura

[Burton, 2007] BURTON, B. Developing Performance Management Programs. In: Gartner Sympozium Cannes. Cannes: Gartner, 2007.

[Fibírová, 2005] FIBÍROVÁ, J., ŠOLJAKOVÁ, L. Hodnotné nástroje řízení a měření výkonnosti podniku. Praha ASPI, 2005. 264 s. ISBN 80-7357-084-X.

[Hrabalová, 2005] HRABALOVÁ S., V. KLÍMOVÁ a S. NUNVÁŘOVÁ. Metody a nástroje řízení ve veřejné správě. Brno: Masarykova univerzita, 2005. 130 s. ISBN 80-210-3679-6.

[Kaplan, 2000] KAPLAN, R.S., NORTON, D.P. Balanced Scorecard – Strategický systém měření výkonnosti podniku. Praha: Management press, 2000. 267 s. ISBN 80-7261-032-5.

[Koudelka, 2007] KOUDELKA, Z. Samospráva. Praha: Linde, 2007. 399 s. ISBN 978-80-7201-665-5. s. 20.

[Linhart, 1996] LINHART, J., Petrusek, M., Vodáková, A., Maříková, Hana. 1996. Velký sociologický slovník. Praha: Karolinum. ISBN 80-7184-310-5.

[Malý, 2011] MALÝ, Ivan a Juraj NEMEC. Možnosti zvyšování efektivnosti veřejného sektoru v podmínkách krize veřejných financí. Brno: Masarykova univerzita, 2011. 222 s. ISBN 978- 80-210-5668-8.

[Mates, 2001] MATES P., WOKOUN R. A KOL. Malá encyklopedie regionalistiky a veřejné správy, 2001. 1.vydání, Praha:: Prospektrum, ISBN 80-7175-100-6.

[Ochrana, 2010] OCHRANA, F., J. PAVEL a L. VÍTEK. Veřejný sektor a veřejné finance: Financování nepodnikatelských a podnikatelských aktivit. Praha: Graga Publishing, 2010. 261 s. ISBN 978-80-247-3228-2.

[Parmenter, 2007] PARMENTER, D. Key Performance Indicators (KPI): Developing, Implementing, and Using Winnig KPIs. Willey, John & Sons, Incorporated, 2007. 256 s. ISBN 0-470-09588-1.

[Průcha, 2002] PRŮCHA, P., POMAHAČ, R. Lexikon, správní právo. 1. vyd. Ostrava: Sagit, 2002. 686 s. ISBN 80-7208-314-7. s. 574 – 575.

[Siegel, 2011] SIEGEL, M., J. STEJSKAL a P. STRÁNSKÁ KOŤÁTKOVÁ. Management veřejného sektoru, 1. díl. Pardubice: Univerzita Pardubice, 2011. 114 s. ISBN 978-80-7395-415-4.

[Šulák, 2004] ŠULÁK, M., VACÍK, E. Měření výkonnosti firem. Plzeň: ZČU Plzeň 2004. 138 s. ISBN 80-7043-258-6

[Učeň, 2008] UČEŇ, P. Zvyšování výkonnosti firmy na bázi potenciálu zlepšení. Grada Publishing as, 2008.

[Voříšek, 2008] VOŘÍŠEK, J., aj. Principy a modely řízení podnikové informatiky, Praha, Oeconomica, 2008, 446 s. ISBN 978-80-245-1440-6

[Vyskočil, 2007] VYSKOČIL, V., ŠTRUP, O., PAVLÍK, M.: Facility Management a Public Private Partnership, 1. vyd. Kamil Mařík – Professional Publishing, 2007. ISBN: 80-86946-34-4

[Wagner, 2009] WAGNER, J. Měření výkonnosti: jak měřit, vyhodnocovat a využívat informace o podnikové výkonnosti. 1. vyd. Praha: Grada, 2009. 248 s."

Internetové zdroje:

[Bundesanstalt 1] Bundesanstalt für Immobilienaufgaben [cit. 2014-12-29] Bundesanstalt für Immobilienaufgaben. <http://www.bundesimmobilien.de/>

[Cabinet Office, 2010] Cabinet Office, 2010. [cit. 2014-12-29] Back Office Benchmark Information 2009/2010. <http://www.cabinetoffice.gov.uk/resource-library/back-office-benchmark-information-200910>

[Cabinet Office, 2012] Cabinet Office, 2012. [cit. 2014-12-29] State of the Estate in 2011. <http://www.cabinetoffice.gov.uk/resource-library/state-estate-2011>

[France Domaine 1] Le portail de l'Économie, des Finances, 2012. [cit. 2014-12-29] Cessions immobilières de l'Etat. <http://www.economie.gouv.fr/cessions/>

[National Audit Office, 2011] National Audit Office, 2011. [cit. 2014-12-29] Public Audit Forum performance indicators. <http://www.nao.org.uk/help_for_public_services/ performance_measurement/paf_performance_indicators.aspx>

[NZ Treasury, 2011] New Zealand Treasury, 2011. [cit. 2014-12-29] Administrative and Support Services Benchmarking Report. <http://www.treasury.govt.nz/statesector/ performance/bass/benchmarking/2008-10>

[OAGC 1] Office of the Auditor General of Canada. [cit. 2014-12-29] Performance Audit Manual. <http://www.oag-bvg.gc.ca/internet/docs/pam_e.pdf>

Normy:

[EN 15221, 2007] ČSN EN 15221-1 (762001) A Facility Management. Část 1, Termíny a definice = Facility Management. Part 1, Terms and definitions

Pracovní dokumenty skupiny NERV KPI:

[NERV KPI, 2013] NERV KPI, Výstupní dokument pracovní skupiny NERV KPI. Praha, 2013.

[Hrabě, 2013] HRABĚ, P., Koncepce KPI pro pilotní ověření v oblasti správy nemovitostí. Praha, 2013.

[Bouchal, 2013] BOUCHAL, P., Benchmarking: čerstvé zkušenosti z Velké Británie a poznatky z dalších zemí. Praha, 2013

Summary

KPI Based public sector management

Corporate performance measurement and management is nowadays generally widespread discipline in the private sector. In the public sector, however, these activities are still not applied in majority of countries including the Czech Republic.

This paper presents approach for managing selected areas of public administration with using defined key performance indicators (KPI) connected to governmental budgeting processes. Paper also describes how this approach was applied and tested in the area of state real estate management and budgeting.

Keywords: performance management, efficiency, KPI, key process indicators, property management, real estate management, benchmarking, public administration.

Metody pro rating hráčů v kompetitivním prostředí

Marek Dvořák

marek.dvorak@vse.cz

Doktorand oboru ekonometrie a operační výzkum

Školitel: prof. RNDr. Ing. Petr Fiala, MBA, CSc., (petr.fiala@vse.cz)

Abstrakt: V tomto článku se zabývám ohodnocením hráčů v kompetitivním prostředí. Pokud existuje množina hráčů střetávající se ve hře, která dovolí určit jednoznačného vítěze, pak předpokládám, že schopnosti každého z nich lze popsat jednorozměrnou náhodnou veličinou. Jednotlivá hra pak spočívá v porovnání náhodného pozorování dvou těchto rozdělení reprezentující tyto hráče. Po odehrání turnaje lze pak na základě výsledků odhadnout tyto skryté veličiny pomocí nelineárního matematického programování.

Byly odhadnuty skryté hodnoty hráčů na základě simulace a také kvalita tohoto odhadu.

Klíčová slova: Saatyho matice, ranking hráčů, nelineární programování

Úvod

Odhadování skrytých vlastností jednotek je jedna z mnoha možných aplikací vícekriteriálního rozhodování a matematického programování. Jedním z úloh pojednává o správném ordinálním či kardinálním seřazení hráčů na základě jejich (mnoha) výher či proher s jinými hráči účastnících se turnaje. Pokud za jednotky považujeme nějaké hráče, potom tyto hráče lze popsat pomocí skrytých kvantitativních hodnot. Tyto hodnoty mohou vyjadřovat hráčovy schopnosti, či jiné veličiny obtížně měřitelné přímo. V této práci uvažuji pouze zjednodušený model, kdy schopnosti hráče jsou popsané jako jednorozměrná veličina.

Žádný hráč také není perfektně konzistentní. To znamená, že jeho výkon náhodně kolísá kolem nějaké očekávané hodnoty. Proto schopnosti hráče nepopisuji jednou hodnotou, ale pravděpodobnostním rozdělením se střední hodnotou a směrodatnou odchylkou. Čím menší směrodatná odchylka, tím více můžeme říci, že je hráč konzistentnější, a naopak.

Toto skryté rozdělení je možno odhadnout pro každého hráče na základě odehraných zápasů s jinými hráči. Množina zápasů mezi hráči je turnaj.

Existuje několik variant algoritmů, které se zaměřují na odhad skrytých parametrů hráče, jako příklad lze uvést například ELO [1] nebo GLICKO-2 [2]. Tyto algoritmy však staví na základě online stochastické optimalizace a gradientních metod a jsou závislé na správně zvolených interních parametrech, jako například délka kroku. Přístup, který zde představuji, je od tohoto oproštěn, protože odhaduje všechny hráče, účastnící se turnaje, najednou.

Hráči

Hráče u je možno popsat pomocí normálního rozdělení, resp. pomocí log-normálního rozdělení:

$$u \sim \mathrm{N}(\mu; \sigma)$$
$$R = e^{u},$$

kde μ je očekávaná schopnost hráče, σ je konzistence hráče, e je Eulerovo číslo a R je rank hráče. Náhodný pokus z tohoto rozdělení reprezentuje konkrétní realizaci míry schopnosti hráče.

Hra spočívá ve střetu dvou hráčů i a j, respektive realizaci hodnoty ze svého vlastního rozdělení, a hru vyhraje ten, jehož hodnota je větší:

$$u_i \sim \mathrm{N}(\mu_i; \sigma_i) \sim \mathrm{N}(\ln R_i; \sigma_i)$$
$$u_j \sim \mathrm{N}(\mu_j; \sigma_j) \sim \mathrm{N}(\ln R_j; \sigma_j)$$
$$g_{ij} = \begin{cases} 1; u_i \geq u_j \\ -1; u_i < u_j \end{cases}.$$

Pravděpodobnost výhry p_{ij} hráče i nad hráčem j lze pak spočítat analyticky jako:

$$p_{ij} = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}\right)$$

Turnaj

Jestliže máme n hráčů, jejichž parametry rozdělení μ a σ neznáme, ale známe výsledky turnaje (mnoho náhodně odehraných her mezi náhodně zvolenými hráči) jako matici t_{ij}, která značí, kolikrát vyhrál hráč i nad hráčem j, pak lze aproximovat očekávaný rank hráče R^*_j vynásobený konstantou c, jako geometrický průměr řádků Saatyho matice s_{ij} [4]:

$$s_{ij} = \frac{t_{ij}}{t_{ji}}$$

$$cR_j^* = \left(\prod_{i=1 \neq j}^{n} s_{ij}\right)^{\frac{1}{n}}$$

Konstanta c je zde proto, že na absolutní hodnotě ranku hráče nezáleží, na čem záleží je relativní poměr ranku dvou hráčů, a ten zůstává stejný i když jsou oba ranky násobené konstantou:

$$\frac{R_i}{R_j} = \frac{cR_i}{cR_j}$$

Odhad parametrů

Z matice turnaje t_{ij} můžeme odhadnout pravděpodobnost výhry hráče i nad j:

$$\hat{p}_{ij} = \frac{t_{ij}}{t_{ij} + t_{ji}}$$
$$\hat{p}_{ij}^* = \text{probit}(\hat{p}_{ij})$$

a následně můžeme odhadnout parametry $\hat{\mu}$ a $\hat{\sigma}$ pomocí úlohy nelineárního matematického programování:

$$z = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\hat{p}_{ij}^* - \frac{\hat{\mu}_i - \hat{\mu}_j}{\sqrt{\hat{\sigma}_i^2 + \hat{\sigma}_j^2}}\right) \rightarrow \min.$$

Simulace

Simulace jsem se rozhodl provádět v turnajích, ve kterém soutěžilo 15 hráčů. Parametry μ a σ jsem pro každý turnaj a hráče náhodně generoval z rozdělení:

$$\mu \sim N(\ln 1500); 1)$$
$$\sigma \sim \ln N\ (0.13; 0.3)$$

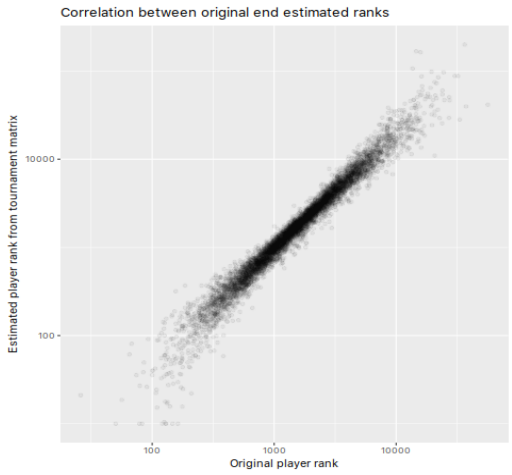
Skryté parametry hráčů náhodně generovaného 15 členného turnaje pak vypadají například takto:

Hráči náhodně generovaného turnaje zdroj: vlastní simulace, 2017

R_i	σ_i
957.345	1.011
2546.856	1.590
1850.927	1.340
1491.198	1.183
1355.182	1.530
1358.474	1.032
2825.487	0.829
406.680	1.405

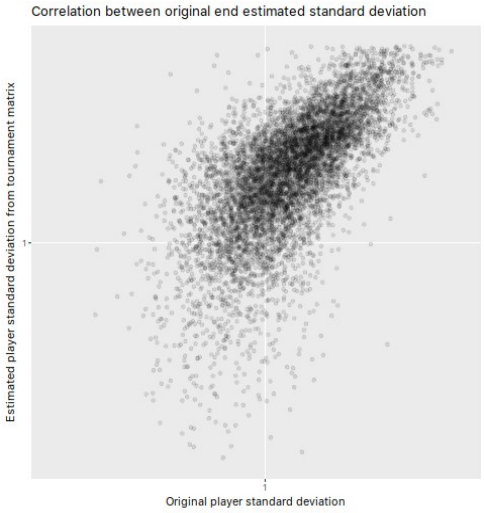
R_i	σ_i
4826.313	0.779
538.262	0.915
1014.723	1.195
603.942	0.451
1399.545	1.146
6125.788	1.701
2253.595	1.531

V každém turnaji bylo odehráno 10000 her (tzn. cca. 89 her mezi každým párem hráčů). Na základě matice t_{ij} pak byly odhadnuty parametry $\hat{\mu}$ (resp. $\hat{R}$) a $\hat{\sigma}$. Celá simulace se pak skládala z 500 odehraných turnajů, tzn., že dohromady bylo generováno 7500 náhodných hráčů. Pro matematickou optimalizaci jsem použil balík *nloptr* z repozitářů statistického softwaru R [5]. Jako algoritmus pro odhad byl použit BOBYQA [3] s počátečními hodnotami rovnající se R^* pro parametry středních hodnot a 2 jako parametry směrodatné odchylky. Následující grafy znázorňují vztah mezi původními skrytými hodnotami hráče a jejich odhadnutými hodnotami pomocí matematického modelu a výsledků turnaje:



Vztah mezi původním rankem hráče a odhadnutým rankem po odehrání turnaje

Zdroj: vlastní simulace, 2017

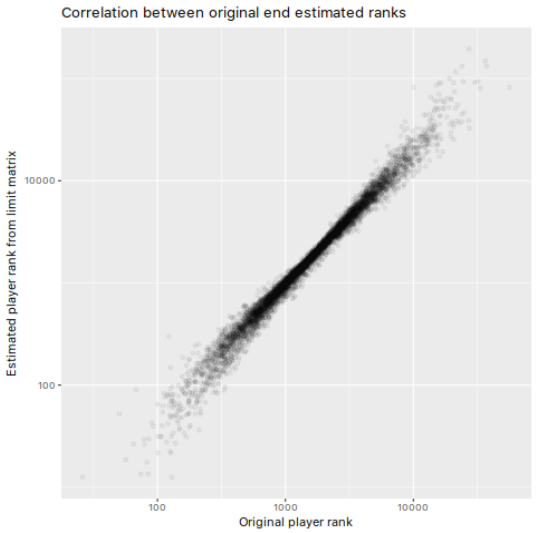


Vztah mezi původní směrodatnou odchylkou hráče a směrodatnou odchylkou odhadnutou po odehrání turnaje

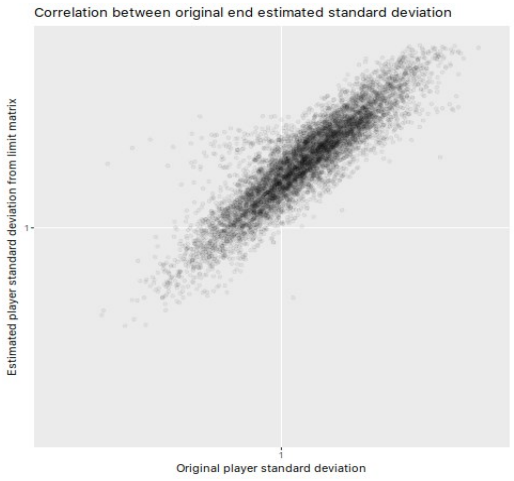
Zdroj: vlastní simulace, 2017

Je zajímavé všimnout si tvaru přesýpacích hodin rozdělení porovnání ranku hráče a jeho nejužšího místa, které je okolo ranku 1500. Toto je s největší pravděpodobností způsobené modem hustoty pravděpodobnosti rozdělení hráčů v generované simulaci. Množství těchto hráčů pak způsobilo zvýšenou přesnost odhadu okolo této hodnoty a výsledný zúžený graf. Zároveň je zajímavé povšimnout si odhadů směrodatné odchylky, který je poměrně nadhodnocen, což by bylo možné vysvětlit relativně nadhodnoceným odhadem výchozí hodnoty pro algoritmus nelineární optimalizace.

Následující grafy zobrazují vztah mezi původními skrytými hodnotami hráče a jejich odhadem z limitní pravděpodobnostní tabulky vypočítané ze vztahu pro p_{ij}:



Vztah mezi původním rankem hráče a odhadnutým rankem z limitní matice Zdroj: vlastní simulace, 2017



Vztah mezi původní směrodatnou odchylkou hráče a směrodatnou odchylkou odhadnutou z limitní matice Zdroj: vlastní simulace, 2017

Závěr

Navrhl jsem matematický aparát pro odhad střední hodnoty schopnosti (resp. ranku) a směrodatné odchylky hráčů účastnících se kompetitivního turnaje. Tento způsob holistického odhadu je v kontrastu s více tradičními metodami na základě stochastického sestupu využívanými známějšími metodami pro odhad ranku, jako je ELO nebo GLICKO.

Následně jsem vytvořil simulaci, kde jsem ukázal, že tato metoda je poměrně robustní a odhady ranků jsou relativně přesné.

Tento model by bylo možno dále vylepšit. Bylo by možno prozkoumat modálnost účelové funkce a věnovat větší pozornost lokálním extrémům. Dále by bylo možné zkoumat vliv kvality odhadu na zvolených hyperparametrech, jako počet hráčů, počet her v turnaji apod. Dále je možno zaměřit se na teoretický základ a

zkoumat vliv jiných omezujících podmínek, jako korelace mezi parametry, či existence nových parametrů a jejich odhad. A v neposlední řadě by bylo zajímavé zobecnění, jako například možnost remízy mezi hráči či existence týmové hry.

Literatura

[1] ELO, Arpard. *The rating of chessplayers, past and present*. Arco Pub., 1978

[2] GLICKMAN, Mark. The glicko system. *Boston University*., 1995

[3] POWELL, Michael. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge, 2009*

[4] SAATY, Thomas. A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, *15*(3), pp.234-281, 1977

[5] YPMA, Jelmer et al. R interface to NLopt. *R Foundation for Statistical Computing*. Vienna, 2016

JEL Classification: C51, C67, C87.

Dedikace: Příspěvek je zpracován jako součást výzkumného projektu IGA F4/54/2015 Interní grantové agentury, Fakulty Informatiky a Statistiky, Vysoké školy ekonomické v Praze.

Sumary

Methods for player ranking in competitive environment

I‘ve designed a method for rank estimation of players attending a competitive tournament. These holistic estimation method are in stark contrast to more traditional approaches based on stochastic descent used by more famous algorithms like ELO and GLICKO.

Afterwards, I‘ve simulated a test run where I‘ve showed a robustness and precision of my approach.

Keywords: Saaty matrix, player ranking, non-linear programming

Analýza šikmosti finančních aktiv

Lukáš Frýd

lukas.fryd@vse.cz

Doktorand oboru Ekonometrie a operační výzkum

Školitel: Prof. RNDr. Václava Pánková, CSc., (vaclava.pankova@vse.cz)

Abstrakt: Práce se zaobírá analýzou vztahu mezi výnosem akcie společnosti Goldman Sachs a druhým, třetím a čtvrtým centrálním momentem. Pro odhad centrálních momentů jsou sestaveny realizované momenty s využitím vysokofrekvenčních dat. Samotná analýza vztahů pak spočívá v aplikaci waveletové koherence. Výsledky naznačují, že existují vzájemné vztahy na vyšších škálách. Zároveň je možné sledovat proměnlivost těchto vztahů v čase. V dlouhodobém investičním horizontu v řádu měsíců byla jednoznačně prokázána pozitivní korelace mezi výnosem a realizovaným třetím momentem a dále negativní korelace mezi výnosem a čtvrtým centrálním momentem.

Klíčová slova: wavelet, waveletová koherence, šikmost, špičatost, vysokofrekvenční data, realizované momenty.

Úvod

Obecný předpoklad používaný ve finanční ekonometrii je, že logaritmické diference cen aktiv se řídí normálním rozdělením. Tento předpoklad se však jeví jako nereálný například díky existence tzv. těžkých chvostů León a kol. (2005). Přestože je snaha tento nedostatek řešit například využitím studentova t-rozdělení, nepracuje se s myšlenkou, že mohou existovat vyšší podmíněné momenty. Následně je využito nejčastější analýzy v podobě ARMA-GARCH modelů.

Z empirických zjištění je již známo, že kromě tlustých chvostů, vykazují výnosy aktiv i proměnlivou šikmost, případně i špičatost Leon, Rubio, Serna (2004). Nelze tak předpokládat symetrické rozdělení výnosů. Cenu šikmosti demonstrovali již v roce 1976 Kraus a Litzenberger, kteří rozšířili CAPM model Sharp (1964), o faktor šikmosti. Ten odhadli jako $Cov[(r_{Mt} - \bar{r}_M)^2(r_{it} - \bar{r}_i)]/(r_{Mt} - \bar{r}_M)^3$, kde r_{Mt} je výnos tržního portfolia v čase t a r_{it} je výnos aktiva i v čase t. Parametr řídící vliv faktoru šikmosti vyšel záporný a statisticky významný. Následující roky byly vyhrazeny zejména modelu GARCH Bollerslev (1986) a hypotéza dynamiky vyšších momentů nebyla podrobněji rozpracována.

Až v roce 1999 přišli Harvey a Siddique s myšlenkou využit metodologii GARCH modelu k odhadu i podmíněné šikmosti. Model dostal název GARCHS(1,1,1) a jeho popis je dán soustavou tří rovnic. Pro popis podmíněné střední hodnoty, rozptylu a šikmosti. Metodologii modelu GARCH využil i Brooks a kol. a v roce 2005 publikovali model popisující podmíněnou špičatost. S podobnou prací odhadu podmíněné šikmosti a špičatosti, ale s jiným předpokladem rozdělení náhodné složky přispěli i León a kolektiv (2005).

Vliv šikmosti na výnos aktiv testoval například Zhang (2006). Autor provedl analýzu na průřezových datech společností rozdělených podle sektorů, kdy jeho závěrem bylo, že mezi šikmostí a výnosem existuje negativní závislost. Ke stejném závěru ohledně vztahu výnosu a šikmosti dospěli i Mitton a Vorkink (2007). V roce 2015 publikovali Amaya a Christoffersen analýzu vlivu šikmosti a špičatosti na výnos aktiv pomocí Fama-French regrese. K odhadu jednotlivých momentů autoři využili vysokofrekvenční průřezová data společností. Jejich závěr ohledně vlivu šikmosti na výnos aktiv byl stejný jako u předešlých citovaných autorů. Naopak vztah mezi špičatostí na výnosem aktiv vyšel pozitivní.

Kromě Mitton a Vorkink (2007) všechny citované studie pracují s implicitním předpokladem homogenních investorů. Tento předpoklad se již dnes zdá jako příliš silný. Například ve fraktální hypotéze finančních trhů Peters (1994), se předpokládá, že existují různé skupiny investorů obchodujících v různých obdobích. Tuto hypotézu podpořila například studie od Krištoufek (2013), kde autor aplikoval waveletovou analýzu na vysokofrekvenční data světových indexů. Z tohoto důvodu bude tato práce zaměřena na testování vztahu mezi výnosem akcie společnosti Goldman Sachs a dalšími třemi momenty pomocí waveletové analýzy. Za předpokladu platnosti fraktální hypotézy finančních trhů tak mohou existovat odlišné vztahy mezi uvedenými momenty pro různé škály.

Metodologie

Realizované momenty

Předpokládejme, že proces logaritmické ceny aktiva p_t definovaný na pravděpodobnostním prostoru $(\Omega, \mathfrak{F}, P)$ je spojitý proces na [*0,T*], kdy *T* je konečné celé číslo. $\mathfrak{F}_t$ představuje $\sigma - algebru$ obsahující všechny známé informace v čase *t*, takové, že $\mathfrak{F}_s \subseteq \mathfrak{F}_t$ pro $0 \leq s \leq t \leq T$. Proces vývoje výnosu v interval [*t-h,h*], kdy $0 \leq h \leq t \leq T$ je pak dán jako $r_{t,h} = p_t - p_{t-h}$ Andersen a kol. (2003). Podle Protter (1992) lze okamžitý výnos rozložit na dvě části. První z nich představuje střední hodnotu výnosu a druhou je lokální martingal. Vývoj výnosu na interval [*t-h,h*] pak má podobu:

$$r_{t,h} = \int_{t-h}^{t} \mu_s ds + \int_{t-h}^{t} \sigma_s dW_s, \tag{1}$$

kde μ_s je integrovatelný a predikovatelný proces s konečným stochastickým rozptylem, σ_s je striktně pozitivní càdlàg stochastický proces Baruník, Vácha (2015) a W_t představuje Brownův pohyb.

Pro proces z rovnice (1), lze odvodit tzv. proces kvadratické variace QV_t Barndorff-Nielsen a Shephard (2002), který je možné popsat jako:

$$QV_t = r_t - r_{t-h} = \int_{t-h}^{t} \sigma_s ds, \tag{2}$$

Jelikož v reálném světě lze pozorovat pouze diskrétní vývoj výnosu, je možné provést odhad kvadratické variace pomocí realizované variace RV_t podle Barndorff-Nielsen a Shephard (2002):

$$RV_t = \sum_{i=1}^{N} r_{t,i}^2 \xrightarrow{p} QV_t, pro\ N \to \infty \tag{3}$$

V případě složitějších stochastických procesů, než popisuje rovnice (1), například modely se stochastickou volatilitou Neuberger (2012), je vhodné analyzovat i variace vyšších řádů. K odhadu kubické variace (RC) lze podle Amaya (2015) využít práce o kvadratické variaci a realizované variaci od Barndorff-Nielsen, Kinnebrock, Shephard (2010), nebo Jacod (2012). Odhadem kubické variance je realizovaný třetí moment (RC):

$$RC_t = \sum_{i=1}^{N} r_{t,i}^3 \xrightarrow{p} CV_t, pro\ N \to \infty \tag{4}$$

Stejně tak podle Barndorff-Nielsen a Shephard (2004) a Jacod (2012) lze dokázet, že realizovaný čtvrtý moment (RQ) konverguje v pravděpodobnosti ke kvarterní variaci (QUV):

$$RQ_t = \sum_{i=1}^{N} r_{t,i}^4 \xrightarrow{p} QUV_t, pro\ N \to \infty \tag{5}$$

Z výše uvedených rovnic vyplívá, že při dostatečně jemném dělení, je možné odhadnout charakteristiky stochastického procesu z rovnice pomocí sum mocnin výnosů.

Wavelety

Waveletová analýza patří do skupiny spektrálních metod, které umožňují analýzu časových řad i ve frekvenční doméně. Jednou z nejoblíbenějších spektrálních metod je Fourierova transformace. Jak uvádí například Baruník, Vácha (2015), zásadním nedostatek Fourierovy transformace v analýze ekonomických a finančních řad spočívá v odhadu spektra (bude vysvětleno dále) časové řady, které je spíše globální než lokální. Tím se vytrácí

informace v časové doméně, což znemožňuje například analýzu šoků, nebo strukturních zlomů Baruník, Vácha (2015). Naproti tomu waveletová transformace umožňuje analyzovat časovou řadu jak ve frekvenční doméně, tak v časové doméně. Základní myšlenkou je využití waveletové funkce jako filtru, který je transformována původní časová řada.

Waveletová funkce $\psi(t)$ musí splňovat následující dvě vlastnosti Percival (2000):

$$\int_{-\infty}^{\infty} \psi(t)\,dt = 0 \qquad (6)$$

$$\int_{-\infty}^{\infty} \psi^2(t)\,dt = 1$$

Následně je možné definovat wavelet jako:

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \qquad (7)$$

kde *s* je měřítko, kdy při jeho změně dochází ke smrštění vlnky a *u* udává polohu waveletové funkce. $\frac{1}{\sqrt{s}}$ je tzv. normalizační faktor.

V této práci bude využita pouze spojitá waveletová transformace, jejíž hodnoty $W_x(u,s)$ získáme jako následující konvoluci:

$$W_x(u,s) = \int_{-\infty}^{\infty} x(t)\frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)dt \qquad (8)$$

Potom $|W_n(s)|^2$ nazveme jako waveletové spektrum, které měří lokální rozptyl časové řady *x(t)* na škále *s*. V případě, že chceme analyzovat vzájemný vztah dvou časových řad pro různé škály *s*, je vhodným nástrojem krosswaveletové spektrum. Mějme následující rovnici:

$$W_{xy}(u,s) = W_x(u,s)W_y^*(u,s) \qquad (9)$$

kde $W_x(u,s)$ a $W_y(u,s)$ jsou spojité waveletové transformace časových řad *x(t)* a *y(t)* a hodnoty $W_y^*(u,s)$ jsou komplexně sdružené vzhledem k hodnotám $W_y(u,s)$ Bašta (2010). Absolutní hodnota $|W_{xy}(u,s)|$ představuje krosswaveletové spektrum, díky němuž je možné odhadnout absolutní hodnotu lokální kovariance mezi dvěma časovými řadami *x(t)* a *y(t)* v různých časových škálách. Jelikož cílem práce je odhadnout vztah mezi výnosem aktiva a jeho vyššími momenty, je vhodnějším nástrojem analýzy waveletová koherence Percival (2000), Baruník, Vácha (2012).

Tato metoda umožňuje odhadnout časové úseky, ve kterých obě časové řady vykazují společný pohyb. Waveletová koherence $R^2(u,s)$ je definována jako:

$$R^2(u,s) = \frac{\left|S\left(s^{-1}W_{xy}(u,s)\right)\right|^2}{S(s^{-1}|W_x(u,s)|^2)S\left(s^{-1}\left|W_y(u,s)\right|^2\right)} \qquad (10)$$

kde S představuje operátor vyhlazení Percival (2000). Waveletovou koherenci lze považovat za odhad lokální hodnoty kvadrátu korelačního koeficientu mezi *x(t)* a *y(t)*, což z definice korelace znamená, že musí platit $0 \leq R^2(u,s) \leq 1$. Bašta (2010)

V této práci bude využita komplexní waveletová funkce, konkrétně Morletův wavelet Percival (2000). Díky tomu je možné pomocí waveletové koherence získat informace o tzv. fázové diferenci $\phi_{xy}(u,s)$, která je definovaná jako:

$$\phi_{xy}(u,s) = \tan^{-1}\left[\frac{\Im\left\{S\left(s^{-1}W_{xy}(u,s)\right)\right\}}{\Re\left\{S\left(s^{-1}W_{xy}(u,s)\right)\right\}}\right] \tag{11}$$

kde $\Im\left\{S\left(s^{-1}W_{xy}(u,s)\right)\right\}$ představuje imaginární část a $\Re\left\{S\left(s^{-1}W_{xy}(u,s)\right)\right\}$ reálnou část waveletové koherence. Díky této dodatečné informaci, bude možné měřit nejen sílu vzájemného vztahu, ale i její směr, tedy zdali jsou časové řady kladně, či záporně korelovány.

$$\phi_{xy}(u,s) \in \begin{cases} \left(\frac{\pi}{2},\pi\right) - mimo\ fázi, y\ vede \\ \left(0,\frac{\pi}{2}\right) - ve\ fázi, x\ vede \\ \left(-\frac{\pi}{2},0\right) - ve\ fázi, y\ vede \\ \left(-\pi,\frac{\pi}{2}\right) - mimo\ fázi, x\ vede \end{cases} \tag{12}$$

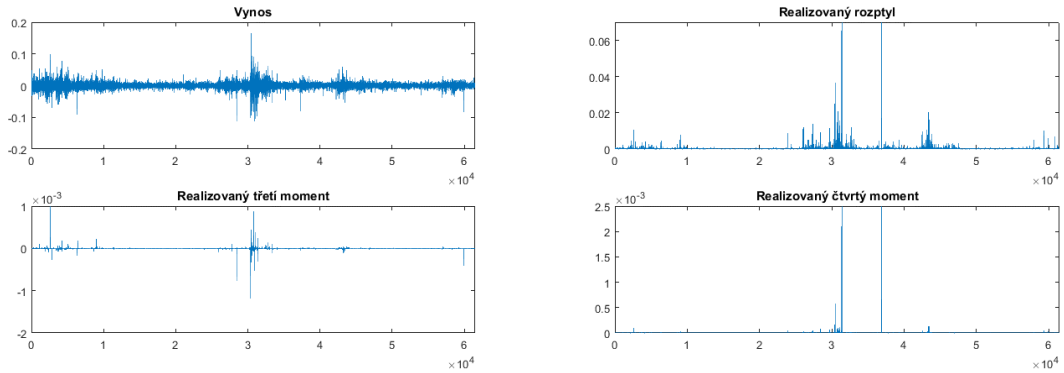
Ve (mimo) fázi zde bude znamenat pozitivně (negativně) korelované časové řady. Fázová diference zobrazená v grafu waveletové koherence pak bude mít podobu šipky. Šipka doprava (doleva) bude značit pozitivní (negativní) korelaci a její otočení vzhůru bude představovat, že první časová řada vede druhou o 90◦ a naopak. Ve většině případech bude graf obsahovat kombinaci poloh, kdy například směr šipky doprava s kladným sklonem bude znamenat pozitivní korelaci s tím, že první časová řada vede druhou.

Data

V práci byla použita minutová data společnosti Goldman Sachs za období 4.5.1999 až 29.11.2016. Pro následnou analýzu byly na základě rovnic (3,4,5) odhadnuty 30-minutové realizované momenty. Vývoj výnosu a realizovaných momentů je zobrazen na obrázku 1. Z obrázku je zřejmé, že k největším změnám došlo v roce 2008 během finanční krize. Dále je na grafu vidět například dopad internetové bubliny z roku 2001. Popisné charakteristiky jednotlivých časových řad jsou zobrazeny v tabulce 1.

Tabulka 4: Popisné charakteristiky dat

	Výnos	**RV**	**RC**	**RQ**
Průměr	0.00002	0.00007	0	0.
Směr. Od.	0.0065	0.00070	0.00001	0.00002
Šikmost	-0.0085	71.54490	-9.8339	127.47
Špičatos	36.55	6746	7843	16995
Min	-0.1127	0	-0.0012	0
Max	0.1636	0.0774	0.001	0.003



Obrázek 4: Vývoj realizovaných momentů

Empirická část

Waveletové koherence mezi výnosem akcií Goldman Sachs a jejich realizovaným druhým, třetím a čtvrtým momentem jsou zobrazeny na obrázku 2. Na ose x je vynesen čas pro období 4.5.1999 až 29.11.2016 po 30 minutových krocích a na ose y jednotlivé časové škály. Škála 1 představuje 30-minut a škála 16384 odpovídá přibližně 341 dním. Vzhledem k tomu, že jeden rok obsahuje zhruba 250 obchodních dní, tak se jedná o 1,361 obchodního roku. Síla lokální korelace určuje barevné spektrum, jehož škála je zobrazena na pravé straně obrázku 2. Kromě barevného spektra byla testována statistická významnost lokální korelace. Nulová hypotéza předpokládá, že obě časové řady jsou generovány bílým šumem. Oblasti ve kterých můžeme zamítnou nulovou hypotézu pro $\alpha = 5\%$, jsou hraničeny černou konturou. Jelikož jsou některé hodnoty waveletových koeficientů ovlivněny tzv. okrajovými podmínkami Percival (2000), zavádí se tzv. oblast vlivu. Ta je na obrázku 2 reprezentována oblastí ležící pod křivkou ve tvaru písmene U. Vzájemný vztah časových řad definovaný podle rovnice (11) je znázorněn pomocí šipek.

Vztah výnos a realizovaný rozptyl

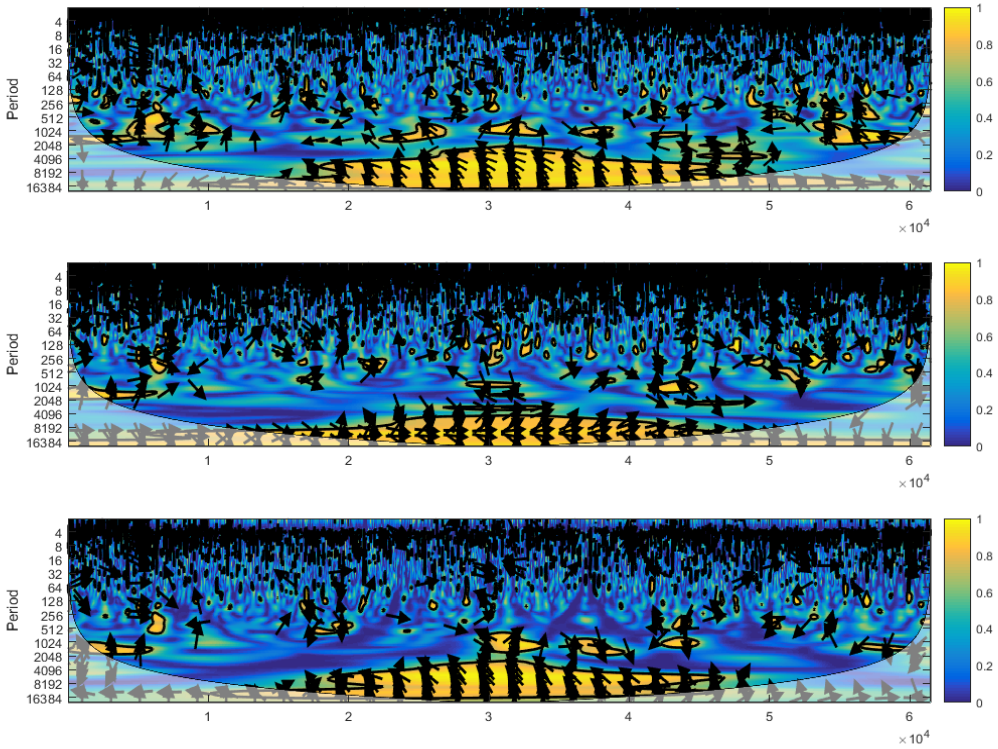
První graf z obrázku 2 zobrazuje vztah mezi výnosem a rozptylem časové řady. Z obrázku je zřejmé, že pro škály menší jak 128 (), nelze předpokládat existenci vzájemné korelace. Oblasti statistické významnosti se objevují od škály 512. Zřejmá statistická významnost lokální korelace je pak od škály 4096. V této oblasti je většina šipek otočena doleva nahoru což značí zápornou korelaci mezi výnosem a volatilitou s tím, že výnos řídí variaci.

Vztah mezi výnosem a realizovaným třetím momentem

Pro analýzu výnosu a realizovaného třetího momentu slouží prostřední graf z obrázku 2. Oblasti statistické významnosti jsou položeny přibližně na stejných škálách jako v předešlém případě. Zajímavým úkazem je proměnlivý vztah vůdčí časové řady pro škály v rozmezí 256 až 2048. V tomto rozmezí existují čtyři statisticky významné oblasti, ve kterých existuje pozitivní vztah mezi výnosem a RC. Pro počátek časové řady vychází, že výnos řídí RC, naopak během a po finanční krizi 2008 (hodnoty x>3) ukazují opačný směr. Pozitivní korelace mezi výnosem a RC dokládají i časové škály nad 4096 s tím rozdílem, že pro interval 4096-8192 řídí výnos RC a od 8192 naopak řídí RC výnos.

Vztah mezi výnosem a realizovaným čtvrtým momentem

Poslední třetí graf na obrázku 2 znázorňuje vztah mezi výnosem a čtvrtým momentem. Opět je zřejmé, že přibližně do škály 512 můžeme mluvit o statisticky významné lokální korelaci zcela výjimečně. Zajímavější jsou až škály nad hodnotu 512. Mezi 512 až 2048 existují oblasti statisticky významné korelace, které zejména během finanční krize dokládají proměnlivý vztah mezi oběma časovými řadami. Pro hodnotu x=3 existuje v daném rozmezí škál negativní vztah mezi výnosem a RQ, kdy výnos určuje druhou časovou řadu. V dalším průběhu pak naopak dochází k momentů pozitivní korelace s proměnlivou řídící časovou řadou. Nejzřetelnější vzájemná vazba nastává opět pro škály nad 2048. Na rozdíl od vztahu výnos vs. RC, zde existuje jednoznačná negativní korelace mezi výnosem a RQ, kdy RQ řídí výnos.



Obrázek 2: Waveletová koherence pro vývoj výnosu a realizované variace, výnosu a realizovaného třetího momentu, výnosu a realizovaného čtvrtého momentu.

Závěr

Cílem práce bylo analyzovat vztah mezi výnosem akcie společnosti Goldman Sachs a jeho vyššími momenty. V citované literatuře dospěli autoři k závěrům, že mezi výnosem a šikmostí existuje negativní vztah a naopak mezi výnosem a špičatostí kladný vztah. Pomocí vysokofrekvenčních dat byl odhadnut druhý, třetí a čtvrtý centrální moment výnosu akcie. Pro analýzu vzájemných vztahů byla využita metoda waveletové koherence, na jejímž základě bylo zjištěno, že vzájemná korelace mezi výnosem a uvedenými momenty se objevuje až pro časovou škálu 512-1024 a to převážně v obdobích zvýšené volatility. Zároveň tyto časové úseky vykazovali nestabilitu znaménka korelace a řídícího procesu. K jednoznačným závěrům ohledně vzájemného vztahu svědčí až škály nad 4096. Pro všechny tři analyzované vztahy existuje statisticky významná lokální korelace. Variace a čtvrtý centrální moment vykazují shodně negativní korelaci s výnosem, kdy výnos je řídící časová řada. Naopak analýza třetího realizovaného momentu dokládá kladnou korelaci s výnosem pro dané časové škály. Pro škály v rozmezí 4096 až 8192 je řídící časovou řadou výnos, naopak pro škály nad 8192 je naopak řídící realizovaný třetí moment. Analýza výnosu ve vtahu k třetímu a čtvrtému realizovanému momentu dokládají odlišné závěry než Zhang (2006), Mitton a Vorkink (2007) a Amaya a Christoffersen (2015). Přestože v této práci byla analyzována pouze akcie jedné firmy, tak je zřejmé, že záleží na časové škále na které je daný vztah zkoumán.

Literatura

AMAYA, D., CHRISTOFFERSEN P., JACOBS, K., and VASQUEZ, A. 2015. Does Realized Skewness Predict the Cross-Section of Equity Returns? Journal of Financial Economics; Issue: 118; 2015; Pages: 135-167

ANDERSON, T. G., BOLLERSLEV, T., DIEBOLD, F. X., & LABYS, P. 2003. Modeling and forecasting realized volatility. Econometrica, 71, 579–625.

BOLLERSLEV, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307-327. 1986.

BARNDORFF-NIELSEN, O.E., and SHEPHARD, N. 2002. Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models. Journal of the Royal Statistical Society 64, 253-280.

BARNDORFF-NIELSEN, O.E., KINNEBROCK, S., and SHEPHARD, N. 2010. Measuring downside risk realised semivariance, in Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle, ed. by T. Bollerslev, J. Russell, and M. Watson. Oxford University Press.

BAŠTA, M. 2010. Waveletová transformace a její aplikace při analýze ekonomických a finančních časových řad. Kvalifikační práce

BROOKS, C., BURKE, S. P., HERAVI, S., & PERSAND, G. 2005. Autoregressive conditional kurtosis. Journal of Financial Econometrics, 3, 399–421.

BARUNÍK, J, VÁCHA, L. 2012. Comovement of energy commodities revised: Evidence from wavelet coherence analysis. Energy Economics 34(1), pp. 241-24

BARUNÍK, J., VÁCHA, L. 2015. Realized Wavelet-Based Estimation of Integrated Variance and Jumps in the Presence of Noise. *Quantitative Finance*. 15, 8, s. 1347–1364. ISSN 1469-7688.

FAMA, E.F., FRENCH, K.R. 1992. The cross-section o expected stock returns. Journal of Finance 59, 427-465.

HARVEY, C.R., SIDDIQUE, A. 1999. Autoregressive conditional skewness. Journal of Financial and Quantitative Analysis 34 (4), 465-487.

JACOD, J. 2012. Statistical Methods for Stochastic Differential Equations, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Volume 124.

KRAUS, A, and LITZENBERGER, R. 1976. Skewness preference and the valuation of risk. Journal of Finance 31, 1085-1100.

KRIŠTOUFEK, L., 2012. Fractal markets hypothesis and the global financial crisis: Scaling,

investment horizons and liquidity. Adv. Complex Syst. 15, 1250065 (2012).

LEON, A, RUBIO, G, and SERNA, G. 2005. Autoregressive Conditional Volatility, Skewness And Kurtosis. The Quarterly Review of Economics and Finance, 45, 599 – 618.

MITTON, T., and K. VORKINK, 2007. Equilibrium Underdiversi.cation and the Preference for Skewness. Review of Financial Studies 20, 1255-1288.

NEUBERGER, A, 2012. Realized skewness. Review of Financial Studies 25(11), 3423-3455.

PERCIVAL, D. B. and A. T. WALDEN, A. T. (2000). Wavelet Methods for Time series Analysis. Cambridge University Press.

PETERS, E. 1994. Fractal Market Analysis – Applying Chaos Theory to Investment and Analysis. John Wiley & Sons, Inc., New York

PROTTER, P. E. 2005. Stochastic integration and differential equations. Springer.

SHARPE, W, 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance 19, 425-442.

ZHANG, Y. 2006. Individual Skewness and the Cross-Section of Average Stock Returns. Working Paper.

Dedikace: Příspěvek vznikl s podporou projektu IGA F4/73/2016 Interní grantové agentury Vysoké školy ekonomické v Praze.

Summary

Analysis of skewness in financial data

We analyse correlation between return of Goldman Sachs stock and higher central moments. The higher moments were estimated by the realized moments methods from high frequency data. We utilize wavelet transform analysis and obtain wavelet coherency spectra which give the information about the local correlation distribution across scales and its evolution in time. We show that in the short time investment horizons does not exist statistical significant correlation between return and realized moments. On the other hand for scale between 512-1024 we found several statistical significant areas. These areas were connected with highly volatile period.

The most interesting discovery was found on long time investment horizons. Between scale 4096 – 8192 we found statistical significant area of local correlation for all time series. For return and variance we found negative correlation relationships where return lead variance. The same conclusion is for correlation between return and realized fourth moment. The contrary conclusion is connected with correlation between return and realized third

moment. We found that between scale 4096 – 8192 there exist positive correlation. The RC is led by return. On the other hand for higher scales the correlation is positive too but return is led by RC. The findings about correlation between return and realized third, fourth moments is not in hand with assertions of Zhang (2006), Mitton a Vorkink (2007) a Amaya a Christoffersen (2015). We found that scale is crucial for relationship determination.

Keywords: wavelet, wavelet coherence, skewness, kurtosis, high-frequency data, realized moments.

Market Microstructure Noise

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: This article reviews a wide range of literature about market microstructure noise contained in high-frequency financial data. We present general framework for efficient prices, jumps, market microstructure noise, quadratic variation, integrated variance and realized variance. At first, possible causes of market microstructure noise are described. Next, the effect of the noise on realized variance is explored. Finally, various methods for robust estimation of integrated variance and forecasting models are presented.

Keywords: High-Frequency Data, Efficient Price, Market Microstructure Noise, Integrated Variance, Realized Variance.

Introduction

Engle (2000) coined a term *ultra-high-frequency data* or simply *high-frequency data* referring to data that consist of observations recorded within seconds or even fractions of seconds. The availability of these high-frequency financial data in the past 20 years allows econometricians to construct more precise estimates, models and forecasts while facing some new challenges.

A risk of financial assets such as stocks or currencies can be measured by integrated variance. However, using high-frequency data to estimate this measure is not as straightforward as in the case of lower frequencies. Observed prices of financial assets are contaminated by so-called market microstructure noise, which at higher frequencies can significantly bias realized variance, a standard estimator of integrated variance.

An overview of recent advances in high-frequency data analysis, volatility estimation and market microstructure noise can be found in the books by Hautsch (2011) and Aït-Sahalia and Jacod (2014).

The Causes

The goal of high-frequency data analysis is to examine logarithmic *efficient prices* that contain the information about prices of an asset. However, we do not observe this underlying latent price process. Instead, we observe logarithmic prices x_t contaminated by so-called *market microstructure noise* due to various microstructure effects (e.g. bid-ask spread, discreteness of prices). We formulate this problem in commonly used *additive noise setting* (Aït-Sahalia and Jacod, 2014)

$$x_t = p_t + e_t$$

where p_t is the efficient price process and e_t is the market microstructure noise. Furthermore, we denote

$$y_{s,t} := x_t - x_s, \quad r_{s,t} := p_t - p_s, \quad u_{s,t} := p_t - p_s.$$

Under the assumption of no arbitrage, the logarithmic efficient returns $r_{s,t}$ must follow a semi-martingale (Delbaen and Schachermayer, 1994). We consider two cases for the price process.

Assumption 1. Let returns $r_{s,t}$ follow a continuous semi-martingale

$$r_{s,t} = \int_s^t \mu(\tau)\,\mathrm{d}\tau + \int_s^t \sigma(\tau)\mathrm{d}W_\tau$$

where $\mu(\tau)$ is a finite variation càglàg drift process, $\sigma(\tau)$ is an adapted càdlàg volatility process and W_τ denotes a standard Wiener process (Hautsch, 2011).

Assumption 2. Let returns $r_{s,t}$ follow a semi-martingale with jumps

$$r_{s,t} = \int_s^t \mu(\tau)\,\mathrm{d}\tau + \int_s^t \sigma(\tau)\,\mathrm{d}W_\tau + \sum_{i=N_s+1}^{N_t} c_i$$

where N_t is a counting process and c_i are non-zero i.i.d. random variables (Barndorff-Nielsen and Shephard, 2006).

The noise is assumed to have zero expected value but besides that it can have various structures. In general, it can be dependent on efficient price and dependent in time (Hansen and Lunde, 2006). It can even be another semi-martingale. In that case, efficient price is not identifiable. However, the most basic model for market microstructure noise is the white noise assumption.

Assumption 3. Let the noise process e_t have following properties:

1. $\mathrm{E}[e_t] = 0$ for all t,
2. $\omega^2 := \mathrm{E}[e_t]^2$ for all t,
3. e_t is independent of p_s for all t and s,
4. e_t is independent of e_s for all $t \neq s$.

Empirical studies indeed show that market microstructure noise is present in high-frequency price processes. Formal methodology for testing for the presence of market microstructure noise was proposed by Awartani et al. (2009) and Aït-Sahalia and Xiu (2016).

The properties of the noise were among others analyzed by Hansen and Lunde (2006) and Bandi and Russell (2008). The variance of market microstructure noise was estimated by Bandi and Russell (2006), Aït-Sahalia and Yu (2009) and (Nolte and Voev, 2012). The correlation structure of market microstructure noise was examined by Aït-Sahalia et al. (2011) and Jacod et al. (2016) while Ubukata and Oya (2009) analyzed both the cross-dependence and the time-dependence.

The market microstructure noise can have various sources. The overview of possible market microstructure noise causes is presented in Table 1. Sources of market microstructure noise can be divided into three classes (Aït-Sahalia et al., 2011).

Table 1: Overview of market microstructure noise causes

Source of Market Microstructure Noise	Data
Frictions in Trading Process	
Bid-Ask Bounce	Transaction Data
Mid Price Interpolation	Quote Data
Discreteness of Price Changes	All Data
Rounding Error	All Data
Informational Effects	
Asymmetric Information	All Data
Partially Incorporated Information	All Data
Strategic Behavior	All Data
Trades on Different Markets	All Data
Gradual Response to a Block Trade	All Data
Inventory Control Effect	All Data
Difference in Trade Sizes	All Data
Price Pressure	All Data
Recording Errors	
Prices Recorded as Zeros	Low Quality Data
Misplaced Decimal Points	Low Quality Data
Missing Observations	Low Quality Data

The first class is comprised of frictions in trading process, i.e. market microstructure, a term coined by Garman (1976). The microstructure causes include bid-ask bounce present in transaction prices, approximation of actual

prices by mid prices interpolated from bid and ask prices, discreteness of transaction times and discreteness of price values (i.e. rounding of prices). Biais et al. (2005) surveyed extensive literature about market microstructure, price formation and trading process. The component of noise caused by bid-ask spreads and discrete price scales is analyzed by Taylor (2016).

The second class represents the informational effects, i.e. how information contained in prices affects behavior of economic agents and formation of new prices. Diebold and Strasser (2013) analyzed the impact of behavior of economic agents on cross-dependency of the market microstructure noise. Hendershott and Menkveld (2014) studied deviations from the efficient price caused by price pressures.

The third class consists of recording errors such as missing observations and incorrectly recorded prices. Such errors occur mostly in low quality data and their adjustment should be subject to preprocessing stage of analysis.

The Effects

Firstly, for prices sampled at times $a = t_0, t_1, \dots, t_n = b$ we define *quadratic variation*

$$QV_{a,b} \coloneqq \operatorname*{plim}_{\Delta^n \to 0} \sum_{i=1}^{n} r^2_{t_{i-1},t_i}$$

where **plim** denotes the limit in probability and $\Delta^n = \max\{t_1 - t_0, t_2 - t_1, \dots, t_n - t_{n-1}\}$ is the maximal lag between observations. As Δ^n goes to 0, n goes to infinity. Next, we define *integrated variance*

$$IV_{a,b} \coloneqq \int_a^b \sigma^2(\tau) d\tau.$$

Our main focus is the estimation of the integrated variance. Under Assumption 1 we have

$$QV_{a,b} = IV_{a,b}.$$

A natural estimator of quadratic variation (and integrated variance) is *realized variance*

$$RV^n_{a,b} \coloneqq \sum_{i=1}^{n} y^2_{t_{i-1},t_i}.$$

In the absence of the noise, this is a consistent estimator of quadratic variation with asymptotically normal distribution (Barndorff-Nielsen and Shephard, 2002).

In the presence of the noise realized variance can be decomposed to

$$RV^n_{a,b} \coloneqq \sum_{i=1}^{n} r^2_{t_{i-1},t_i} + 2\sum_{i=1}^{n} r_{t_{i-1},t_i} u_{t_{i-1},t_i} + \sum_{i=1}^{n} u^2_{t_{i-1},t_i}.$$

The first term corresponds to unbiased estimation of quadratic variation while the second and the third term represent the bias caused by market microstructure noise. Under Assumption 3 the expected value of realized variance is

$$\mathrm{E}[RV^n_{a,b}] = IV_{a,b} + 2n\omega^2,$$

which for $n \to \infty$ diverges linearly to infinity making the realized variance useless for quadratic variation estimation at higher frequencies. Under less restrictive assumptions about market microstructure noise, the bias of realized variance can have more complex structure. Figure 1 shows realized variances estimated from different frequencies (different periods). We can see nonlinear behavior of realized variance bias. The bias at lower frequencies can be caused by drift term in efficient price while at higher frequencies this bias diminishes. On the other hand, the bias at higher frequencies is caused by market microstructure noise. Negative bias at high frequencies suggests negative cross-correlation with the efficient price (Hansen and Lunde, 2006). Andersen et al. (2000) refer to these plots as volatility signature plots.

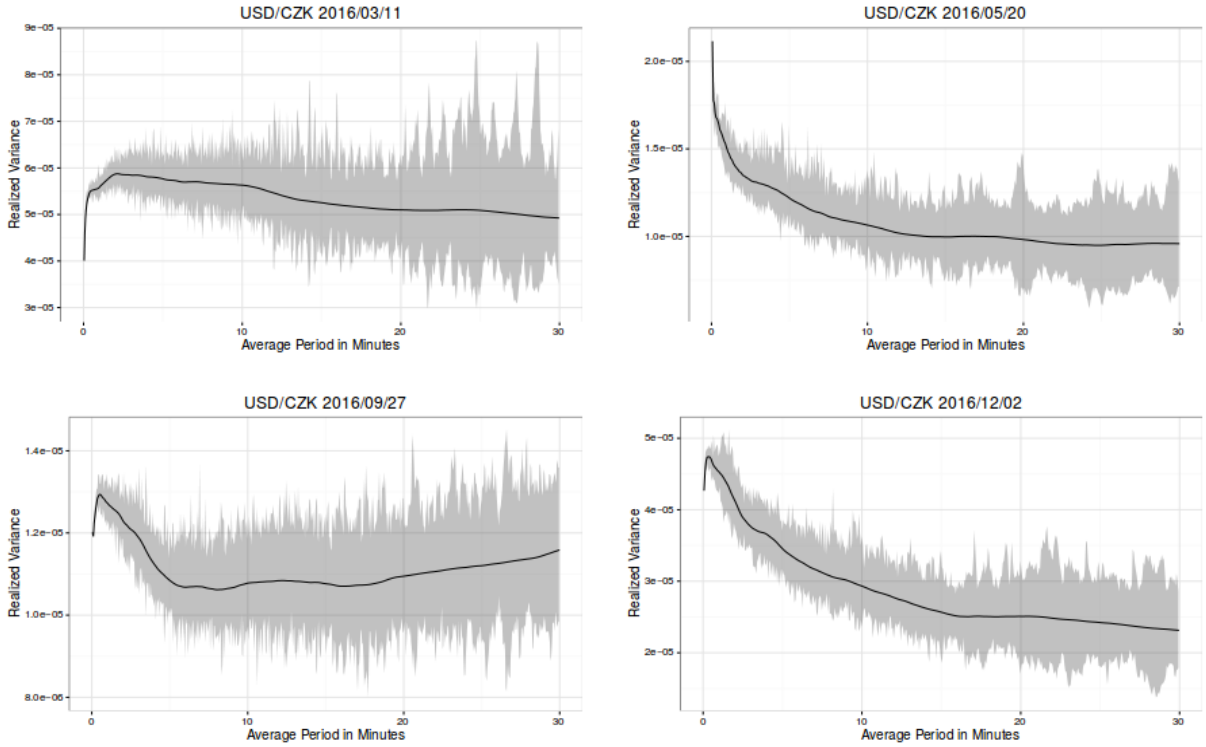


Figure 5: Bias of daily realized variances with 95% confidence intervals for USD/CZK pair.

Realized variance can also be affected by sampling schemes. Oomen (2005; 2006) addresses this issue and identifies four alternative sampling schemes: calendar time sampling, transaction time sampling, tick time sampling and business time sampling.

So far we have considered the efficient price not to contain jumps. Under Assumption 2 we have

$$QV_{a,b} = IV_{a,b} + JV_{a,b},$$

where

$$JV_{a,b} := \sum_{i=N_a+1}^{N_b} c_i^2.$$

In the absence of the noise realized variance is still consistent estimator of quadratic variation, but is biased estimator of integrated variance because of the jump component. Integrated variance can be estimated by *bipower variation* (Barndorff-Nielsen, 2004)

$$BV_{a,b}^n := \frac{\pi}{2} \sum_{i=2}^{n} |y_{t_{i-1},t_i}| |y_{t_{i-2},t_{i-1}}|.$$

The Solutions

One of the main problems surrounding market microstructure noise is to estimate unbiased integrated variance. When the noise is relatively small the bias of realized variance can be minimized by finding the optimal sampling frequency (Zhang et al., 2005; Bandi and Russell, 2008). However, bias correction and consistent robust estimation are more effective ways to deal with the noise of any size.

One of the early estimators robust to market microstructure noise was introduced by Zhou (1996). It was later studied by Hansen and Lunde (2006) and showed that this estimator is unbiased but inconsistent.

When assuming constant volatility in the efficient price process, the returns contaminated by white noise follow MA(1) process and the maximum likelihood estimator (Aït-Sahalia et al., 2005; Aït-Sahalia and Yu, 2009) can be used.

The first consistent estimator of integrated variance was *two-scale realized volatility* (Zhang et al., 2005). The idea behind this estimator is to combine realized variance estimated using the highest frequency and grid of realized variances estimated from lower frequency for bias adjustment. Later, it was extended to *multi-scale realized volatility* (Zhang, 2006), two-scale realized volatility robust to time series dependency (Aït-Sahalia et al., 2011) and *multi-scale discrete sine transform* (Curci and Corsi, 2012). A similar idea was employed by *least squares estimator* (Nolte and Voev, 2012) utilizing linear regression in which realized variances estimated from different data subsamples act as dependent variable while the number of observations act as explanatory variable.

Realized kernel estimator (Barndorff-Nielsen et al., 2008) captures serial correlations induced by microstructure noise with a kernel. It was later studied by Barndorff-Nielsen et al. (2009), Bandi and Russell (2011) and Barndorff-Nielsen et al. (2011).

Another approach is to remove noise by locally averaging returns before computing realized variance. The pre-averaging estimator was proposed by Jacod et al. (2009) with extensions by Hautsch and Podolskij (2013) and Jacod and Mykland (2015).

Other estimators include *Fourier series estimator* (Malliavin and Mancino, 2002; Mancino and Sanfelici, 2008), *quasi-maximum likelihood estimator* (Xiu, 2010), *alternation estimator* (Large, 2011), *wavelet estimator* (Høg and Lunde, 2003) and *maximum overlap discrete wavelet transform* (Baruník and Vácha, 2015). Nielsen and Frederiksen (2008) and Gatheral and Oomen (2010) compared finite sample accuracy of various estimators. Table 2 presents an overview of integrated variance estimators.

Table 2: Overview of integrated variance estimators

Estimator	Origin
Bias-Corrected Realized Volatility	Zhou (1996)
Fourier Series Estimator	Malliavin and Mancino (2002)
Wavelet Estimator	Høg and Lunde (2003)
Maximum Likelihood Estimator	Aït-Sahalia et al. (2005)
Two-Scale Realized Volatility	Zhang et al. (2005)
Multi-Scale Realized Volatility	Zhang (2006)
Realized Kernel	Barndorff-Nielsen et al. (2008)
Pre-Averaging Estimator	Jacod et al. (2009)
Quasi-Maximum Likelihood Estimator	Xiu (2010)
Alternation Estimator	Large (2011)
Multi-Scale Discrete Sine Transform	Curci and Corsi (2012)
Least Squares Estimator	Nolte and Voev (2012)
Maximum Overlap Discrete Wavelet Transform	Baruník and Vácha (2015)

Most of the estimators can be represented using *quadratic form* (Sun, 2006; Andersen et al., 2011)

$$QE_{a,b}^{n} := \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}\, y_{t_{i-1},t_i} y_{t_{j-1},t_j}$$

where q_{ij} are suitable weights. Weights $q_{ii} = 1$ and $q_{ij} = 1$ for $i \neq j$ correspond to realized variance. Another quadratic estimators are two-scale realized volatility, multi-scale realized volatility, least squares estimator, realized kernel, pre-averaging estimator, Fourier series estimator and quasi-maximum likelihood estimator.

Generalization to multivariate processes and estimation of realized covariances were among others analyzed by Bandi and Russell (2005), Sheppard (2006), Voev and Lunde (2007), Zhang (2011) Corsi and Audrino (2012) and Haugom et al. (2014).

All above mentioned estimators are model-free. However, to forecast future volatility, some sort of modelling is required. One of the first approaches to model realized variance is the HAR model (Corsi, 2009). It is an autoregressive-type model featuring realized variances over different time horizons. McAleer and Medeiros (2008) extended HAR to capture long memory and asymmetries while Busch et al., 2011 proposed vector HAR model. Aït-Sahalia and Mancini (2008) compared HAR model, HES model, log-volatility model and fractional Ornstein-Uhlenbeck model. Another suggested models are HEAVY models (Shephard and Sheppard, 2010), eigenfunction stochastic volatility models (Andersen et al., 2011) and MIDAS model (Ghysels and Sinko, 2011). More recent model is realized GARCH (Hansen et al., 2012) which is a modification of classical GARCH model where lagged realized variances are used instead of lagged errors.

Discussion

In the past 20 years a wide range of literature answered rising high-frequency data analysis problems of estimation, modelling and forecasting integrated variance as well as investigation the price process. However, some issues related to market microstructure noise still remain to be addressed.

The causes of the noise are often overlooked as the realized variance literature focuses mainly on statistical point of view. More studies connecting the economic theory with statistical properties of integrated variance estimation and market microstructure noise would bring more understanding to high-frequency price processes.

The effects of the noise have been thoroughly explored and the need for robust estimators has been advocated in a vast number of papers. Unfortunately, there is still a significant amount of practitioners who do not take them into account and rely on realized variance without any bias correction. Perhaps this is because the idea of bias caused by higher frequencies is counter-intuitive at first sight.

The solutions to the bias of the noise can be accomplished by varied methods for integrated variance estimation. However, it can be unclear which method to use, especially in a finite sample case. This can be improved by developing a unified framework for univariate as well as multivariate price processes under general assumption about market microstructure noise.

References

AÏT-SAHALIA, Y., JACOD, J. *High-Frequency Financial Econometrics.* Princeton University Press, 2014. 10.1017/CBO9781107415324.004. ISBN 9788578110796.

AÏT-SAHALIA, Y., MANCINI, L. Out of Sample Forecasts of Quadratic Variation. *Journal of Econometrics.* 2008, 147, 1, s. 17–33. ISSN 03044076.

AÏT-SAHALIA, Y., XIU, D. A Hausman Test for the Presence of Market Microstructure Noise in High Frequency Data. 2016. Working paper.

AÏT-SAHALIA, Y., YU, J. High Frequency Market Microstructure Noise Estimates and Liquidity Measures. *The Annals of Applied Statistics*. 2009, 3, 1, s. 422–457.

AÏT-SAHALIA, Y., MYKLAND, P. A., ZHANG, L. How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise. *Oxford University Press for Society for Financial Studies, Review of Financial Studies*. 2005, 18, 2, s. 351–416. ISSN 0893-9454.

AÏT-SAHALIA, Y., MYKLAND, P. A., ZHANG, L. Ultra High Frequency Volatility Estimation with Dependent Microstructure Noise. *Journal of Econometrics*. 2011, 160, 1, s. 160–175. ISSN 03044076.

ANDERSEN, T. G. et al. Great Realisations. *Risk.* 2000, 13, s. 105–108.

ANDERSEN, T. G., BOLLERSLEV, T., MEDDAHI, N. Realized Volatility Forecasting and Market Microstructure Noise. *Journal of Econometrics*. 2011, 160, 1, s. 220–234. ISSN 0304-4076.

AWARTANI, B., CORRADI, V., DISTASO, W. Assessing Market Microstructure Effects via Realized Volatility Measures with an Application to the Dow Jones Industrial Average Stocks. *Journal of Business & Economic Statistics*. 2009, 27, 2, s. 251–265. ISSN 0735-0015.

BANDI, F. M., RUSSELL, J. R. Realized Covariation, Realized Beta and Microstructure Noise. 2005. Working paper.

BANDI, F. M., RUSSELL, J. R. Separating Microstructure Noise from Volatility. *Journal of Financial Economics*. 2006, 79, 3, s. 655–692.

BANDI, F. M., RUSSELL, J. R. Microstructure Noise, Realized Variance, and Optimal Sampling. *Review of Economic Studies*. 2008, 75, 2, s. 339–369. ISSN 00346527.

BANDI, F. M., RUSSELL, J. R. Market Microstructure Noise, Integrated Variance Estimators, and the Accuracy of Asymptotic Approximations. *Journal of Econometrics*. 2011, 160, 1, s. 145–159. ISSN 03044076.

BARNDORFF-NIELSEN, O. E. Power and Bipower Variation with Stochastic Volatility and Jumps. *Journal of Financial Econometrics*. 2004, 2, 1, s. 1–37. ISSN 1479-8409.

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: Statistical Methodology*. 2002, 64, 2, s. 253–280. ISSN 13697412.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N. Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation. *Journal of Financial Econometrics*. 2006, 4, 1, s. 1–30.

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.

BARNDORFF-NIELSEN, O. E. et al. Realized Kernels in Practice: Trades and Quotes. *Econometrics Journal*. 2009, 12, 3, s. 1–32. ISSN 13684221.

BARNDORFF-NIELSEN, O. E. et al. Subsampling Realised Kernels. *Journal of Econometrics*. 2011, 160, 1, s. 204–219. ISSN 03044076.

BARUNÍK, J., VÁCHA, L. Realized Wavelet-Based Estimation of Integrated Variance and Jumps in the Presence of Noise. *Quantitative Finance*. 2015, 15, 8, s. 1347–1364. ISSN 1469-7688.

BIAIS, B., GLOSTEN, L., SPATT, C. Market Microstructure: A Survey of Microfoundations,Empirical Results, and Policy Implications. *Journal of Financial Markets*. 2005, 8, 2, s. 217–264.

BUSCH, T., CHRISTENSEN, B. J., NIELSEN, M. O. The Role of Implied Volatility in Forecasting Future Realized Volatility and Jumps in Foreign Exchange, Stock, and Bond Markets. *Journal of Econometrics*. 2011, 160, 1, s. 48–57. ISSN 03044076.

CORSI, F. A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics*. 2009, 7, 2, s. 174–196.

CORSI, F., AUDRINO, F. Realized Covariance Tick-by-Tick in Presence of Rounded Time Stamps and General Microstructure Effects. *Journal of Financial Econometrics*. 2012, 10, 4, s. 591–616.

CURCI, G., CORSI, F. Discrete Sine Transform for Multi-Scale Realized Volatility Measures. *Quantitative Finance*. 2012, 12, 2, s. 263–279.

DELBAEN, F., SCHACHERMAYER, W. A General Version of the Fundamental Theorem of Asset Pricing. *Mathematische Annalen*. 1994, 300, 1, s. 463–520.

DIEBOLD, F. X., STRASSER, G. On the Correlation Structure of Microstructure Noise: A Financial Economic Approach. *The Review of Economic Studies*. 2013, 80, 4, s. 1304–1337. ISSN 0034-6527.

ENGLE, R. F. The Econometrics of Ultra-High-Frequency Data. *Econometrica*. 2000, 68, 1, s. 1–22. ISSN 0012-9682.

GARMAN, M. B. Market Microstructure. *Journal of Financial Economics*. 1976, 3, 3, s. 257–275.

GATHERAL, J., OOMEN, R. C. A. Zero-Intelligence Realized Variance Estimation. *Finance and Stochastics*. 2010, 14, 2, s. 249–283. ISSN 09492984.

GHYSELS, E., SINKO, A. Volatility Forecasting and Microstructure Noise. *Journal of Econometrics*. 2011, 160, 1, s. 257–271.

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.

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.

HAUGOM, E. et al. Covariance Estimation Using High-Frequency Data: Sensitivities of Estimation Methods. *Economic Modelling*. 2014, 43, s. 416–425. ISSN 02649993.

HAUTSCH, N. *Econometrics of Financial High-Frequency Data*. Springer Science & Business Media, 2011. ISBN 9783642219245.

HAUTSCH, N., PODOLSKIJ, M. Preaveraging-Based Estimation of Quadratic Variation in the Presence of Noise and Jumps: Theory, Implementation, and Empirical Evidence. *Journal of Business & Economic Statistics*. 2013, 31, 2, s. 165–183. ISSN 0735-0015.

HENDERSHOTT, T., MENKVELD, A. J. Price Pressures. *Journal of Financial Economics*. 2014, 114, 3, s. 405–423.

HØG, E., LUNDE, A. Wavelet Estimation of Integrated Volatility. 2003. Working paper.

JACOD, J., MYKLAND, P. A. Microstructure Noise in the Continuous Case: Approximate Efficiency of the Adaptive Pre-Averaging Method. *Stochastic Processes and their Applications*. 2015, 125, 8, s. 2910–2936. ISSN 0304-4149.

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 03044149.

JACOD, J., LI, Y., ZHENG, X. Statistical Properties of Microstructure Noise. 2016. Working paper.

LARGE, J. Estimating Quadratic Variation When Quoted Prices Change by a Constant Increment. *Journal of Econometrics*. 2011, 160, 1, s. 2–11.

MALLIAVIN, P., MANCINO, M. E. Fourier Series Method for Measurement of Multivariate Volatilities. *Finance and Stochastics*. 2002, 6, 1, s. 49–61.

MANCINO, M. E., SANFELICI, S. Robustness of Fourier Estimator of Integrated Volatility in the Presence of Microstructure Noise. *Computational Statistics & Data Analysis*. 2008, 52, 6, s. 2966–2989.

MCALEER, M., MEDEIROS, M. C. A Multiple Regime Smooth Transition Heterogeneous Autoregressive Model for Long Memory and Asymmetries. *Journal of Econometrics*. 2008, 147, 1, s. 104–119.

NIELSEN, M. Ø., FREDERIKSEN, P. Finite Sample Accuracy and Choice of Sampling Frequency in Integrated Volatility Estimation. *Journal of Empirical Finance*. 2008, 15, 2, s. 265–286.

NOLTE, I., VOEV, V. Least Squares Inference on Integrated Volatility and the Relationship Between Efficient Prices and Noise. *Journal of Business & Economic Statistics*. 2012, 30, 1, s. 94–108.

OOMEN, R. C. A. Properties of Bias-Corrected Realized Variance Under Alternative Sampling Schemes. *Journal of Financial Econometrics*. 2005, 3, 4, s. 555–577.

OOMEN, R. C. A. Properties of Realized Variance Under Alternative Sampling Schemes. *Journal of Business and Economic Statistics*. 2006, 24, 2, s. 219–237.

SHEPHARD, N., SHEPPARD, K. Realising the Future: Forecasting with High-Frequency-Based Volatility (HEAVY) Models. *Journal of Applied Econometrics*. 2010, 47, 4, s. 36–37. ISSN 01451707.

SHEPPARD, K. Realized Covariance and Scrambling. 2006. Working paper.

SUN, Y. Best Quadratic Unbiased Estimators of Integrated Variance in the Presence of Market Microstructure Noise. 2006. Working paper.

TAYLOR, S. J. Microstructure Noise Components of the S&P 500 Index: Variation, Persistence and Distributions. 2016. Working paper.

UBUKATA, M., OYA, K. Estimation and Testing for Dependence in Market Microstructure Noise. *Journal of Financial Econometrics*. 2009, 7, 2, s. 106–151.

VOEV, V., LUNDE, A. Integrated Covariance Estimation Using High-frequency Data in the Presence of Noise. *Journal of Financial Econometrics*. 2007, 5, 1, s. 68–104.

XIU, D. Quasi-Maximum Likelihood Estimation of Volatility with High Frequency Data. *Journal of Econometrics*. 2010, 159, 1, s. 235–250. ISSN 03044076.

ZHANG, L. Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach. *Bernoulli*. 2006, 12, 6, s. 1019–1043.

ZHANG, L. Estimating Covariation: Epps Effect, Microstructure Noise. *Journal of Econometrics*. 2011, 160, 1, s. 33–47.

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 01621459.

ZHOU, B. High-Frequency Data and Volatility in Foreign-Exchange Rates. *Journal of Business & Economic Statistics*. 1996, 14, 1, s. 45–52. ISSN 07350015.

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. F4/63/2016.

Shrnutí

Mikrostrukturní šum

Tento článek se zabývá širokou literaturou zabývající se mikrostrukturním šumem obsaženým ve vysokofrekvenčních finančních datech. Představíme obecný rámec pro eficientní ceny, skoky, mikrostrukturní šum, kvadratickou variaci, integrovanou varianci a realizovanou varianci. Nejdříve popíšeme možné příčiny šumu. Dále prozkoumáme efekt šumu na realizovanou varianci. Nakonec představíme různorodé metody pro robustní odhady a předpovědi integrované variance.

Klíčová slova: Vysokofrekvenční data, Eficientní cena, Mikrostrukturní šum, Integrovaná variance, Realizovaná variance.

Aplikace základního gravitačního modelu pro odhad migrace mezi Evropskými státy

Tatiana Polonyankina

tatiana.polonyankina@vse.cz

Doktorand oboru Ekonometrie a operační výzkum

Školitel: doc. Ing. Tomáš Cahlík, CSc., (tomas.cahlik@vse.cz)

Abstrakt: Gravitační model je zajímavý svojí analogií s Newtonovým gravitačním zákonem, kdy se logika gravitace přeneseně používá pro popis prostorové interakce mezi ekonomickými jednotkami. Model stejně jako gravitační zákon předpokládá, že síla interakce mezi jednotkami je pozitivně ovlivněna jejich velikostí a negativně ovlivněna vzdálenosti mezi nimi. Příspěvek používá gravitační model pro odhad závislosti migrace na HDP a vzdálenosti pro Evropské státy. V úvodu je stručně shrnuta aplikace gravitačního modelu na různá ekonomická odvětví. V teoretické části je definován základní gravitační model a také jeho úprava pro analýzu migrace. V další části je model použit pro odhad závislosti migrace. Regresní analýzou na datech za rok 2012 bylo zjištěno, že odhady vlivu vzájemné vzdálenosti uvažovaných zemí Evropské unie a HDP cílové země na migraci odpovídají teorii gravitačního modelu. Parametr definující vliv HDP země původu na migraci byl analýzou odhadnut jako negativní, kdy tento výsledek neodpovídá teorii modelu, která předpokládá naopak pozitivní vliv. Výsledek však ukazuje platnost ekonomické teorie o faktorech ovlivňujících migraci, kdy při ekonomické migraci dochází k přesunu obyvatel z méně ekonomicky vyspělých zemí do ekonomicky silnějších států.

Klíčová slova: gravitační model, prostorová závislost, migrace, Evropská unie, výběr modelu.

Úvod

Prostorová interakce je termín, který se v odborné ekonomické literatuře používá k popisu pohybu v prostoru způsobeného nebo ovlivněného lidskou aktivitou. Klasickým příkladem je migrace neboli pohyb osob, a to v každé své podobě: od pracovníků každý den dojíždějících do práce po cizince, kteří se rozhodli žít na území daného státu. Dále se tento termín používá například v souvislosti s pohybem zboží, služeb, při studiích mezinárodního obchodu nebo při analýzách sdílení a šíření informací.

Gravitační model je jeden ze způsobů, jak vysvětlit prostorovou interakci mezi jednotkami, protože zohledňuje geografickou závislost jednotek. Ve své základní podobě jej mezi prvními použil Tinbergenem (1962) pro modelování obchodních toků mezi jednotkami. V dalších letech byl model široce používán právě pro analýzu přeshraničních toků (například Anderson (1979)).

Model získal své jméno a základní logiku z fyziky, díky analogii s Newtonovým gravitačním zákonem, který říká: gravitační síla, kterou na sebe působí dvě vesmírná tělesa, pozitivně závisí na jejich velikosti a negativně na jejich vzdálenosti. Gravitační model definuje interakci mezi dvěma jednotkami jako funkci jejich vzdálenosti a jejich velikosti. Velikost jednotek dle gravitačního modelu má pozitivní vliv na interakci mezi jednotkami a je možné ji definovat pomocí fyzikálních veličin (hmotnost, objem, obsah) nebo přeneseně například jako ekonomickou sílu nebo jinou formu „přitažlivosti“ zkoumané jednotky pro ostatní jednotky v souboru v závislosti na poli aplikace modelu. Při práci na makroekonomické úrovni si zde například můžeme představit HDP. Vzdálenost může být také chápána jako absolutní nebo přeneseně jako relativní. Absolutní vzdálenost je definovaná jednotkami míry (kilometry). Definice relativní vzdálenosti se liší v závislosti na tématu, kterým se analýza zabývá. Můžeme si jí představit jako přístup k vodě, vzdělání, potravinám nebo pracovním příležitostem. Dvě místa se stejnou absolutní vzdálenosti od bodu se mohou lišit v relativní vzdálenosti a naopak. Definici základního gravitačního modelu je věnovaná další kapitola.

Důležitost zejména relativního pojetí prostorové závislosti a velikosti jednotky je vidět při aplikaci gravitačního modelu v různých odvětvích. Kingsley (1984) se zmiňuje o použití modelu urbanisty, dopravními analytiky, investory do nákupních center, sociálních analytiků. Constantin (2004) definuje několik aplikací gravitačního modelu, a to verze gravitačního modelu pro: A) obchodní toky, kde je gravitační model aplikován v podobě Riellyho zákona. Zákon zjednodušeně říká, že retailový klient je přitahován obchodním centrem úměrně velikosti centra a nepřímo úměrně druhé mocnině vzdálenosti od obchodního centra. B) gravitačního model pro analýzu toku komodit, kde tok závisí přímo na poptávce po komoditě a nepřímo na transakčních nákladech. C)

gravitační model pro analýzu migrace, kdy je migrace přímo závislá na populaci nebo ekonomické síle centra a nepřímo na vzdálenosti od centra.

Konkrétním příkladem použití gravitačního modelu je práce Eggera (2004), který zkoumá, jak Evropská integrace v 90. letech ovlivnila objem zahraničních přímých investic na území Evropské unie. Zjistil, že realizace jednotného evropského trhu a rozšíření EU v roce 1995 mělo pozitivní efekt na přímé zahraniční investice zemí členských států. Braconier (2005) aplikoval gravitační model při analýze závislosti přímých zahraničních investic na mzdách kvalifikované a nekvalifikované pracovní síly. Jedním ze závěrů je, že zahraniční investice jsou směřovány do států s nižší cenou pracovní síly.

Použití gravitačního modelu při studiích vlivu migrace bylo mezi prvními provedeno Ravenstein (1889) na data pro UK. Metulini (2013) ve své práci při odhadu prostorové závislosti mezi OECD zeměmi srovnal metodu nejmenších čtverců (MNČ) a modely kombinujícími gravitační model s modely pracující s prostorovou autokorelací, kdy MNČ se ukázala jako méně vhodná.

Základní gravitační model

Tato kapitola se věnuje definici klasického gravitačního modelu, kde v základní fázi pracujeme s obecnou maticí vzdáleností a obecně zadanou velikostí regionu. Výše byly popsány možnosti přeneseného významu použitých charakteristik, které budou využity až v aplikační části příspěvku.

Modely zohledňující prostorovou závislost je vhodné použít v případě, že je logické předpokládat existenci vztahů mezi jednotkami (regiony), které jsou ovlivněny jejich geografickým rozmístěním. U těchto modelů je vždy vyžadovaná informace o geografické poloze jednotek modelu. Tato informace je definovaná pomocí matice. Pro gravitační model se tato matice nazývá ”origin-destination” matice vzdáleností (“origin – destination regions distance matrix”).

Základní gravitační model je definován vzorcem níže a říká, že interakce T_{ij} mezi regiony i a j je závislá na P_i velikosti regionu i, P_j velikosti regionu j, d_{ij} vzdálenosti mezi regiony i, j a konstantě G.

$$T_{ij} = G\frac{P_i^{\,a} P_j^{\,b}}{d_{ij}^{\,c}}$$

V další fázi následuje úprava modelu pro regresní analýzu, kdy je provedena logaritmická transformace modelu a přidána náhodná složka.

$$\log T_{ij} = \log G + a\log P_i + b\log P_j - c\log d_{ij} + \varepsilon_{ij}$$

Pro snadnější práci s modelem je provedena substituce modelu výše, výsledná rovnice je:

$$y_{ij} = \alpha + \beta_o x_{oi} + \beta_d x_{dj} + \gamma d_{ij} + \varepsilon_{ij}.$$

Model je dále vyjádřen pomocí matic pro n regionů. Prostorová interakce mezi regiony je definovaná jako $n \times n$ čtvercová matice, jejíž prvky jsou hodnoty d_{ij}. Tuto matici můžeme upravit na $n^2 \times 1$, tak že z řádků matice vytvoříme sloupce. Vzniklá matice je pojmenovaná D. Stejný postup je uplatněn i pro matice X_d a X_o. Základní gravitační model používá pro vyjádření velikosti jen jednu proměnou x, model je možné rozšířit přidáním dalších vysvětlujících proměnných, potom k je počet proměnných, které definují velikost regionů. Rozšíření modelu o další proměnné lze využít zejména při relativním vyjádření velikosti a aplikaci modelu ve specifických odvětvích. Po těchto úpravách maticový zápis klasického gravitačního modelu, kde závislou proměnnou y je logaritmus interakce mezi regiony T_{ij} , bude:

$$y = \alpha J + X_i\beta_o + X_j\beta_d + D\gamma + \varepsilon,$$

kde

y je $n^2 x\ 1$ vektor logaritmu vysvětlované proměnné vyjadřující interakci mezi regiony,

J je $n^2 x\ 1$ jednotkový vektor,

X_i je $n^2 x\ k$ matice logaritmu charakteristik vyjadřujících velikost regionu původu,

X_j je $n^2 x\ k$ matice logaritmu charakteristik vyjadřujících velikost cílového regionu,

D je $n^2 x\ 1$ vektor logaritmu vzdáleností mezi původním a cílovým regionem,

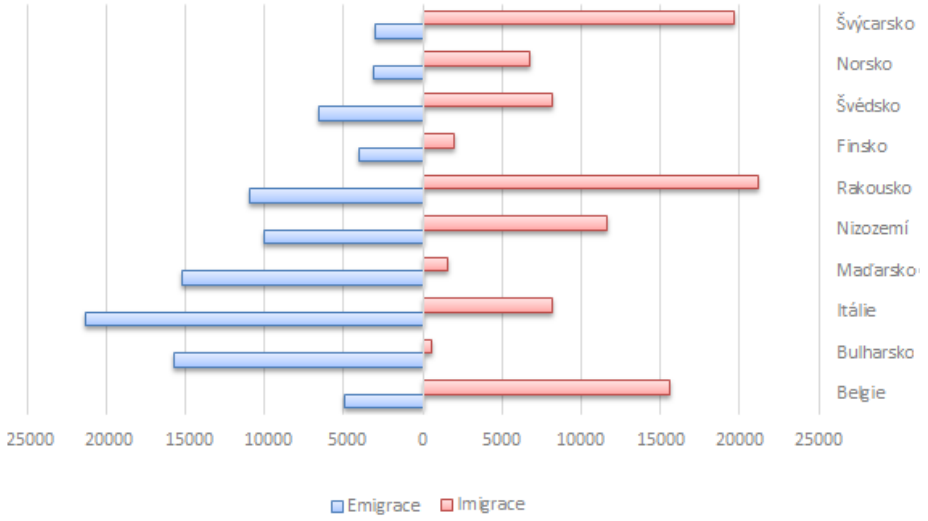
$\varepsilon \sim N(0, \sigma^2 I_{n^2})$,

$\alpha, \beta_d, \beta_o, \gamma$ jsou odhadované parametry modelu.

Aplikace modelu na reálná data

V aplikační části je model odvozený v předešlé kapitole použit pro odhad závislosti migrace mezi deseti Evropskými státy na HDP a vzájemné vzdálenosti těchto států. Zdrojem dat jsou webové stránky Eurostatu. Regresní analýza je provedena na datech pro rok 2012 a pro státy: Belgie, Bulharsko, Itálie, Maďarsko, Holandsko, Rakousko, Finsko, Švédsko, Norsko a Švýcarsko.

Jak vypadají počty migrujících obyvatel mezi těmito státy, je uvedeno v grafu 1 níže. Z grafu je patrné, že do Bulharsko přibylo ze zbylých devíti států nejméně obyvatel a naopak emigrace je jedna z nejvyšších ve zkoumaném vzorku. Opačné chování migrantů je patrné ve Švýcarsku, kde je jeden z nejvyšších přírůstků migrace ze zkoumaných států a nejnižší úbytek.



Obrázek 5: Migrace mezi státy analýzy Zdroj: Eurostat, 11/2016

Proměnné základního gravitačního modelu v předešlé kapitole jsou nahrazeny proměnnými, které se používají při aplikaci modelu na analýzu migrace: T_{ij} je migrace mezi státy *i* a *j*, která je závislá na P_i zde HDP státu i, P_j zde HDP státu j, d_{ij} vzdálenosti mezi státy i, j (měřená jako vzdálenost od středu) a konstantě *G*, tedy:

$$migrace_{ij} = G \frac{HDP_i{}^a HDP_j{}^b}{d_{ij}{}^c}.$$

HDP je měřené v eurech na obyvatele, aby byla zohledněna pouze ekonomická vyspělost státu a proměnná nebyla zkreslená velikostí státu. Po provedení úprav dle postupu v teoretické části je výsledný regresní model, kde $n = 10$ a $k = 1$.

$$y = \alpha J + X_i \beta_o + X_j \beta_d + D\gamma + \varepsilon,$$

kde

y je $100\ x\ 1$ vektor logaritmu vysvětlované proměnné vyjadřující migraci mezi regiony,

J je $100\ x\ 1$ jednotkový vektor,

X_j je $100\ x\ 1$ vektor logaritmu HDP cílové země (v eurech na obyvatele),

X_i je $100 \; x \; 1$ vektor logaritmu HDP země původu (v eurech na obyvatele),

D je $100 \; x \; 1$ vektor logaritmu vzdálenosti mezi zemí původu a cílové zemně (v km, měřeno od středu státu),

$\varepsilon \sim N(0, \sigma^2 I_{100})$,

$\alpha, \beta_d, \beta_o, \gamma$ odhadované parametry.

Cílem příspěvku je pomocí analýzy zjistit, zda pro migraci mezi státy Evropské unie v době hospodářské krize funguje princip základního gravitačního modelu. Dle ekonomické teorie a logiky gravitačního modelu je očekáváno, že HDP cílového státu bude mít pozitivní vliv na migraci a vzdálenost negativní vliv na migraci.

U vlivu HDP zdrojové země na migraci se ekonomická teorie a gravitační model rozcházejí. Dle gravitačního modelu HDP zdrojové země má pozitivní vliv na migraci, z ekonomického pohledu je ale možný negativní vliv. Ekonomická teorie o determinantech migrace říká, že migrace je často motivovaná nadějí na zlepšení ekonomické situace jedince. Obecně se předpokládá, že ekonomicky motivovaný jedinec se bude stěhovat z ekonomicky slabší země s nižším HDP do země s vyšším HDP. Empiricky zkoumáno například v studii Maydy (2003).

Odhad modelu je zapsán pomocí vzorce a také zaznamenán v tabulce 1. Koeficient determinace je 0,63. Všechny proměnné jsou statisticky významné při 1% hladině významnosti a testy na heteroskedasticitu a multikolinearitu jsou negativní. Model byl také testovaný na nezahrnutí významné proměnné do modelu dle Ramseyho testu, kdy test vyšel také negativní.

$$\hat{y}_{ij} = 5{,}29 - 0{,}61x_i + 0{,}97x_j - 0{,}28d_{ij}, \quad i, j = 1, \ldots 10$$

Tabulka 6: Výsledek regresní analýzy Zdroj: Eurostat, 11/2016

závislá proměnná	odhad parametru	st.chyba	t-hodnota	p-hodnota
konstanta	5,29	2,02	2,62	0,01
log HDP země původu	-0,61	0,14	-4,4	0,00
log HDP cílové země	0,97	0,14	7,00	0,00
log vzdálenosti	-0,28	0,03	-9,51	0,00

Odhady parametrů β_d, γ odpovídají očekávaným výsledkům a splňují ekonomickou verifikaci modelu i předpoklady gravitačního modelu. Regrese ukazuje, že vliv ekonomické vyspělosti na migraci je větší než vliv vzdálenosti. Menší vliv vzdálenosti souvisí s levnějším cestováním a volným pohybem osob v rámci Evropské unie. Odhad parametru β_d má opačné znaménko, než očekává gravitační model. Jak bylo zmíněno na začátku této kapitoly, vzhledem k ekonomické teorii výsledek naznačuje, že u migrace platí, že k emigraci dochází především u států s nižším HDP.

Závěr

Příspěvek se zabýval teorií a použitím gravitačního modelu. Tento model vychází z Newtonova fyzikálního zákona o gravitaci, který říká, že interakce mezi dvěma tělesy pozitivně závisí na jejich velikosti a negativně na jejich vzdálenosti. V úvodní fázi jsou shrnuty modifikace základního modelu při aplikaci na různá ekonomická odvětví. V teoretické části je definován základní gravitační model a v praktické části je model aplikován na reálná data.

Cílem bylo zjistit, zda princip základního gravitačního modelu funguje při jeho aplikaci na analýzu migrace mezi Evropskými státy v období hospodářské recese. Byl proveden odhad závislosti migrace ze zdrojového státu do cílového státu na HDP cílového státu, HDP zdrojového státu a jejich vzdálenosti. Výsledný odhad splňuje ekonomické předpoklady i předpoklady gravitačního modelu o statisticky významném pozitivním vlivu HDP cílové země a negativním statisticky významném vlivu vzdálenosti mezi zeměmi. Gravitační model předpokládá také pozitivní statisticky významný vliv HDP zdrojové země. Ekonomické teorie předpokládá vliv opačný, tedy emigrace je vyšší u státu s nižším HDP, kdy jedinec od změny země pobytu očekává také zlepšení ekonomické situace v ekonomicky silnější zemi s vyšším HDP. Výsledek regrese ukázal platnost ekonomické teorie a naznačuje, že jedinci průměrně odchází z ekonomicky slabších zemí do ekonomicky vyspělejších Evropských států. Dále regresní analýza ukázala, že vliv ekonomické vyspělosti na migraci je vyšší než vliv vzdálenosti mezi

státy. Tento výsledek se dá vysvětlit volným pohybem osob v rámci Evropské unie, rostoucí dostupností a zrychlování dopravy.

Literatura

ANDERSON J. A. A Theoretical Foundation for the Gravity Model. *The American economic Review.* 1979, 69 (1), 106-116.

METULINI R. Spatial gravity models for international trade: a panel analysis among OECD countries. *European Regional Science Association conference paper*. 2013, ersa13p522.

EGGER, P. a M. PFAFFERMAYR. Foreign direct investment and European integration in the 1990´s. *The World Economy*. 2004, 24(1), 99-110.

BRACONIER, H., P. J. NORBRACK a D. URBAN. Multinational enterprises and wage costs: Vertical FDI revisited. *Journal of International Economics.* 2005, 67(2), 446-470.

CONSTATNTIN, D. L. The use of Gravity Models for spatial interaction analysis. *Economy Informatics*. 2004, 1(4), 116-118.

HAYNES, Kingsley E. a A. Stewart. FOTHERINGHAM. *Gravity and spatial interaction models*. Beverly Hills: Sage Publications, c1984. ISBN 0803923260

LESAGE, James P. a R. Kelley. PACE. *Introduction to spatial econometrics*. Boca Raton: CRC Press, c2009. Statistics, textbooks and monographs. ISBN 142006424X.

MAYDA, A. International migration: A panel data of the determinants of bilateral flows. Quarterly Journal of Economics. 2007, 116(2), 549-599.

RAVENSTEIN, E. G. The laws of migration. *Journal of the royal statistical society*. 1889, 52(2), 241-305.

TINBERG, J. *Shaping the World Economy; Suggestions for an International Economic Policy*. 1962. Books (Jan Tinbergen). Twentieth Century Fund, New York.

JEL Classification: C52

Aplication of Basic Gravity Model on estimation of migration between European countries

Gravity model is interesting for its analogy with Newton's law of gravitation, when the logic of gravitation is used to describe the spatial interactions between economic units. The law and the model suppose that a force of interaction between units is positively influenced by the size of units and negatively by the distance between them.

The paper uses the gravity model to estimate a dependence of migration on the GDP and the distance for European Union countries. Introduction briefly summarizes the application of the model in different economic sectors. The basic gravity model and its modification for migration analysis are defined in the theoretical section. The model is applied on real data in the empirical part. Regression analysis was done for data of 2012 for ten European countries. The estimated impact of the distance and the GDP of a host country on migration fit the gravity model theory. The negative impact of GDP of an origin county on migration was estimated. This result does not correspond with the gravity model theory where a positive impact is expected, however this result is in line with the economic theory of migration factors. It is said that an economic migration usually force residents to move from economically less developed countries to countries with higher GDP.

Keywords: gravity model, spatial dependence, migration, European union, model selection.

Statistiky nad intervalovými daty

Ondřej Sokol

ondrej.sokol@vse.cz

Doktorand oboru Ekonometrie a operační výzkum

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

Abstrakt: Příspěvek se zabývá výpočtem těsné horní meze výběrového rozptylu v případě intervalových dat. Obecně je nalezení horní meze výběrového rozptylu ze znalosti pouze intervalových dat NP-těžký problém, ale při splnění určitých podmínek kladených na vstupní intervalová data lze použít některý z efektivních algoritmů. V příspěvku je analyzována průměrná výpočetní složitost zejména Fersonova algoritmu v případě náhodně generovaných dat. Exponenciální část složitosti tohoto algoritmu závisí na maximálním počtu překrytí zúžených intervalů, proto byla zkoumána střední hodnota maximálního počtu překrytí v případě, že středy a poloměry intervalů pocházejí z určitých, „běžných", rozdělení. Výsledky simulací naznačují, že v takovém případě roste maximální počet překrytí zúžených intervalů nanejvýš logaritmicky. Při předpokladu rovnoměrného rozdělení středů a konstantních poloměrech jsou odvozeny způsoby, jak lze střední hodnotu maximálního počtu překrytí nad náhodnými intervaly spočítat s využitím objemu řezů simplexu, nebo transformací na speciální instanci tzv. narozeninového problému.

Klíčová slova: částečná identifikace, intervalová data, výběrový rozptyl, narozeninový problém

Úvod

V mnoha oblastech vědy lze narazit na situaci, kdy data nelze přesně změřit či uložit, nebo jsou některá data z nějakého důvodu nedostupná. V takovém případě je možné se pokusit jejich distribuci *rozumně* odhadnout. Taková situace je obvykle v odborné literatuře nazývána částečná identifikace („partial identification" (Manski 2003)). Problematika je uvedena na následujících příkladech:

Ve městě, jehož demografické charakteristiky obyvatel jsou dostupné, se koná sčítání lidu. Zajímá nás přitom hlavně výše příjmu obyvatel. Část obyvatel svůj příjem poskytla, zbylá část se odmítla vyjádřit. U části obyvatel tedy jsou k dispozici údaje jak o demografické skupině (definované například pomocí instrumentálních proměnných vzešlých z demografických charakteristik), tak přesné údaje o příjmu, u druhé části pak pouze údaje o demografické skupině. Pro další práci si stanovíme následující předpoklad ohledně distribuce příjmů jednotlivých demografických skupin: obyvatelé, kteří odmítli odpovědět, mají příjem zcela jistě v intervalu minima a maxima z odpovědí obyvatel stejné demografické skupiny.

Nyní nás zajímají statistiky ohledně příjmu populace města či nějaké skupiny obyvatel. Jaký je průměr a medián příjmu obyvatel města? Jaký je rozptyl příjmů? Jelikož v předpokladu je používána intervalová distribuce, statistiky z nich vycházející budou taktéž intervaly. Hledají tedy jejich dolní a horní meze.

Pracovníci firmy spravující benzinové pumpy po celém světě dostali za úkol změřit opotřebení čerpacího zařízení. Měřící zařízení ale není stejné, každé má jinou, ale vždy známou, maximální odchylku. Údaje, které posílají na centrálu, pak mají podobu intervalů. Zajímají nás statistiky opotřebení celého souboru, případně jeho podmnožiny.

V této práci se budu zabývat specifickou situací, kdy odhadnutá distribuce nepozorovatelných dat je charakterizována intervalem, tedy pouze dolní a horní mezí. O distribuci na intervalu nejsou k dispozici žádné další informace. V následujících částech jsou popisovány způsoby práce s takto definovanými daty. Analyzovány jsou zejména vlastnosti Fersonova algoritmu pro výpočet horní meze výběrového rozptylu. Důraz je kladen na vlastnosti jevu překrytí zúžených intervalů, na kterém závisí exponenciální část výpočetní složitosti tohoto algoritmu.

Všechny zmíněné (i odkazované) algoritmy v tomto příspěvku fungují pro jakýkoliv podíl intervalových pozorování, kompletní pozorování zde vystupuje jako tzv. degenerovaný interval, kdy horní i dolní mez jsou stejné hodnoty. Zaměřovat se pak budu na nejobtížnější situaci, kdy všechna pozorování jsou nedegenerované a vzájemně unikátní intervaly.

Statistiky nad intervalovými daty

Předpokládejme, že jednorozměrná množina dat $x_1, \dots, x_n$ je nepozorovatelná, ale k dispozici jsou intervaly

$$x_1 = [\underline{x}_1, \overline{x}_1], \dots, x_n = [\underline{x}_n, \overline{x}_n],$$

pro které platí $\underline{x}_i \leq x_i \leq \overline{x}_i$ pro každý $i = 1, \dots, n$. V případě $\underline{x}_i = \overline{x}_i$ nazveme interval degenerovaný.

Pro interval $\boldsymbol{x} = [\underline{x}, \overline{x}]$ označme $x^C = \frac{1}{2}(\underline{x} + \overline{x})$ střed intervalu a $x^\Delta = \frac{1}{2}(\overline{x} - \underline{x})$ poloměr intervalu. Pro $\alpha > 0$ označme $\alpha\boldsymbol{x}$ jako α-zúžený interval $[x^C - \alpha x^\Delta, x^C + \alpha x^\Delta]$.

Mějme statistiku, tedy spojitou funkci dat, $S(x_1, x_2, \dots, x_n)$. Předpokládejme, že $x_1, x_2, \dots, x_n$ neznáme, ale máme k dispozici intervaly $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}$. Na $x_1, x_2, \dots, x_n$ lze pak nahlížet jako na realizace náhodné veličiny, jejíž rozdělení je dáno pouze mezemi intervalů. Pak i hodnota statistiky $S = S(x_1, x_2, \dots, x_n)$ je náhodnou veličinou. Jelikož jediná informace, co známe, jsou meze, kterých x_i může nabývat, lze tudíž určit pouze meze statistiky S, přičemž pro ně platí:

$$\underline{S} = \inf_{x \in R^n} \left\{ S(x_1, x_2, \dots, x_n) : \underline{x}_i \leq x_i \leq \overline{x}_i, i = 1, \dots, n \right\},$$

$$\overline{S} = \sup_{x \in R^n} \left\{ S(x_1, x_2, \dots, x_n) : \underline{x}_i \leq x_i \leq \overline{x}_i, i = 1, \dots, n \right\}.$$

Pak zřejmě platí $\underline{S} \leq S \leq \overline{S}$, přičemž meze statistik jsou vždy definované. Výpočet mezí statistik je ale v některých případech NP-obtížná úloha.

Jednoduchou úlohou je naproti tomu nalezení mezí střední hodnoty či kvantilů. V tomto případě lze využít stochastické dominance prvního řádu – dolní meze střední hodnoty i kvantilů lze vypočítat pouze z množiny dolních mezí jednotlivých intervalů. Analogicky horní meze lze vypočítat na množině horních mezí intervalů.

V této práci se zaměřím pouze na problematiku výběrového rozptylu nad intervalovými daty, což je v případě těsné horní meze obecně NP-obtížná úloha. Na druhou stranu pro tuto statistiku existuje několik specializovaných algoritmů, které za splnění určitých předpokladů pracují efektivně (Ferson et al. 2005; Dantsin et al. 2006; Xiang 2006).

Řada dalších statistik byla zkoumána jak z pohledu výpočetní složitosti, tak z pohledu existence specializovaných algoritmů pro specifické případy, například již zmíněný rozptyl (Ferson et al. 2002, 2005; Xiang 2006; Dantsin et al. 2006), entropie (Kreinovich 1996), vyšší momenty (Kreinovich et al. 2004) či t-statistika (Černý a Hladík 2014b). Ucelenější přehledy algoritmů pro výpočty mezí různých statistik lze nalézt v (Nguyen et al. 2012; Ferson et al. 2007).

Výpočet výběrového rozptylu nad intervalovými daty za známé střední hodnoty

V případě, že je známá střední hodnota ρ, ale jednotlivá pozorování jsou intervalová, je výpočet obou mezí rozptylu snadný. Taková situace může nastat, pokud jsou citlivá data dodatečně anonymizována – například známe průměr příjmu celého souboru klientů banky, ale data u jednotlivých klientů jsou uměle převedena do podoby intervalů za účelem ochrany osobních údajů.

Výpočet dalších statistik jako je například výběrový rozptyl či t-statistika, je pak možný v lineárním čase. Jedná se o řešení následujících optimalizačních úloh, kde ρ je známé:

$$\underline{\hat{\sigma}} = \min_{x \in R^n} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \rho)^2 : \underline{x} \leq x \leq \overline{x} \right\},$$

$$\overline{\hat{\sigma}} = \max_{x \in R^n} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \rho)^2 : \underline{x} \leq x \leq \overline{x} \right\}.$$

Jelikož střední hodnota ρ je známa, tak je možné minimalizovat každý člen sumy jednotlivě. Při minimalizaci vybíráme takové hodnoty x splňující $\underline{x} \leq x \leq \overline{x}$, které jsou nejblíže střední hodnotě ρ. Při maximalizaci výběrového rozptylu je postup analogický – pouze vybíráme hodnoty nejdále od střední hodnoty ρ.

Výpočet výběrového rozptylu nad intervalovými daty za neznámé střední hodnoty

Výpočet těsných mezí výběrového rozptylu v případě neznámé střední hodnoty není proveditelný v lineárním čase. Pro nalezení dolní a horní meze je třeba vyřešit následující optimalizační úlohy:

$$\underline{\hat{\sigma}} = \min_{x \in R^n} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 : \underline{x} \leq x \leq \overline{x} \right\},$$

$$\overline{\hat{\sigma}} = \max_{x \in R^n} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 : \underline{x} \leq x \leq \overline{x} \right\}.$$

Na rozdíl od předchozích úloh se zde vyskytuje tzv. problém závislosti, kdy při změně jedné hodnoty dochází k řetězení změn zasahujících do všech částí sumy.

Jedná se o optimalizaci konvexní účelové funkce nad konvexní množinou. V případě dolní meze se jedná o minimalizaci, v případě horní meze pak o maximalizaci. Pomocí obecných algoritmů je tedy nalezení dolní meze možné v polynomiálním čase (v případě použití specializovaných algoritmů pak $O(n \log n)$ (Xiang et al. 2007), případně i v lineárním čase, viz (Nguyen et al. 2012)). Na druhou stranu nalezení těsné horní meze výběrového rozptylu je výrazně obtížnější.

Jelikož se jedná o konvexní funkci nad konvexní množinou, optimální řešení se nachází v některém krajním bodě n-rozměrné krychle definované intervaly. Těchto vrcholů, které by bylo třeba prohledat, je 2^n. Obecně je nalezení těsné horní meze výběrového rozptylu nad intervalovými daty NP-obtížná úloha (Ferson et al. 2002; Černý a Hladík 2014a). V následujících částech se tedy budeme zabývat zejména hledáním těsné horní meze výběrového rozptylu.

Výpočet horní meze

Pro výpočet těsné horní meze výběrového rozptylu lze kromě obecných nelineárních solverů (které pracují obecně s exponenciální složitostí – jedná se o NP-obtížnou úlohu) použít některé ze specializovaných algoritmů. Tyto algoritmy za splnění určitých předpokladů dokáží spočítat horní mez v polynomiálním čase. V následujících odstavcích rozdělím tyto specializované algoritmy do dvou skupin dle předpokladů, za kterých pracují s polynomiální složitostí.

Do první skupiny lze zařadit Dantstinův (Dantsin et al. 2006) a Xiangův algoritmus (Xiang 2006). Tyto algoritmy vycházejí z podobné základní myšlenky. Jejich složitost je polynomiální, respektive loglineární, v případě, že žádný zúžený interval není vlastní podmnožinou jiného zúženého intervalu. V opačném případě je složitost těchto algoritmů exponenciální v počtu zúžených intervalů, které jsou vlastní podmnožinou jiného zúženého intervalu. Oba algoritmy lze upravit tak, že složitost bude exponenciální v počtu zúžených intervalů, které je třeba nahradit jednou hodnotou tak, že žádný zúžený interval nebude vlastní podmnožinou jiného zúženého intervalu. Pro ujasnění v prvním případě sčítáme zúžené intervaly, které jsou vlastní podmnožinou

jiného, v druhém pak ty, jež jsou nadmnožinou jiných zúžených intervalů. Vždy se ale jedná o součet. To vede k tomu, že v případě, kdy data – intervaly – jsou generované z nějakého běžného rozdělení, tak s rostoucím počtem intervalů roste lineárně součet zúžených intervalů, které jsou podmnožinou jiného. Důsledkem pak je, že výpočetní složitost je v průměru exponenciální (Sokol 2015).

Do druhé skupiny lze pak zařadit Fersonův algoritmus (Ferson et al. 2005). Ten pracuje následovně

1. Mějme n intervalů $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$. Spočtěme zúžené intervaly $\boldsymbol{x_1^{1/n}}, \boldsymbol{x_2^{1/n}}, \ldots, \boldsymbol{x_n^{1/n}}$.
2. Všech $2n$ krajních bodů zúžených intervalů $x_{(i)}$ seřadíme do posloupnosti dle velikosti. Definujme $x_{(0)} := -\infty$ a $x_{(2n+1)} := \infty$. Bez újmy na obecnosti $x_{(0)} \leq x_{(1)} \leq \cdots \leq x_{(2n+1)}$.
3. Nalezneme hodnoty $\overline{E}$ a $\underline{E}$ a vybereme všechny intervaly $\left[x_{(q)}, x_{(q+1)}\right]$ takové, že $\left[x_{(q)}, x_{(q+1)}\right] \cap \left[\underline{E}, \overline{E}\right] \neq \emptyset$.
4. Pro každý interval $\left[x_{(q)}, x_{(q+1)}\right]$ vybereme $x_1, x_2, \ldots, x_n$ dle následujících pravidel:
   a. pokud $x_{(q+1)} < \overline{x_i^{1/n}}$, pak $x_i := \overline{x_i}$,
   b. pokud $x_{(q)} > \underline{x_i^{1/n}}$, pak $x_i := \underline{x_i}$,
   c. v ostatních případech je třeba prozkoumat obě možnosti.

   Výsledkem bude n-tice $x_1, x_2, \ldots, x_n$, u které zkontrolujeme, zda její průměr leží v intervalu $\left[x_{(q)}, x_{(q+1)}\right]$. Pokud ano, pak z prvků n-tice spočítáme rozptyl a uložíme ho.
5. Ze všech hodnot uložených rozptylů vybereme nejvyšší.

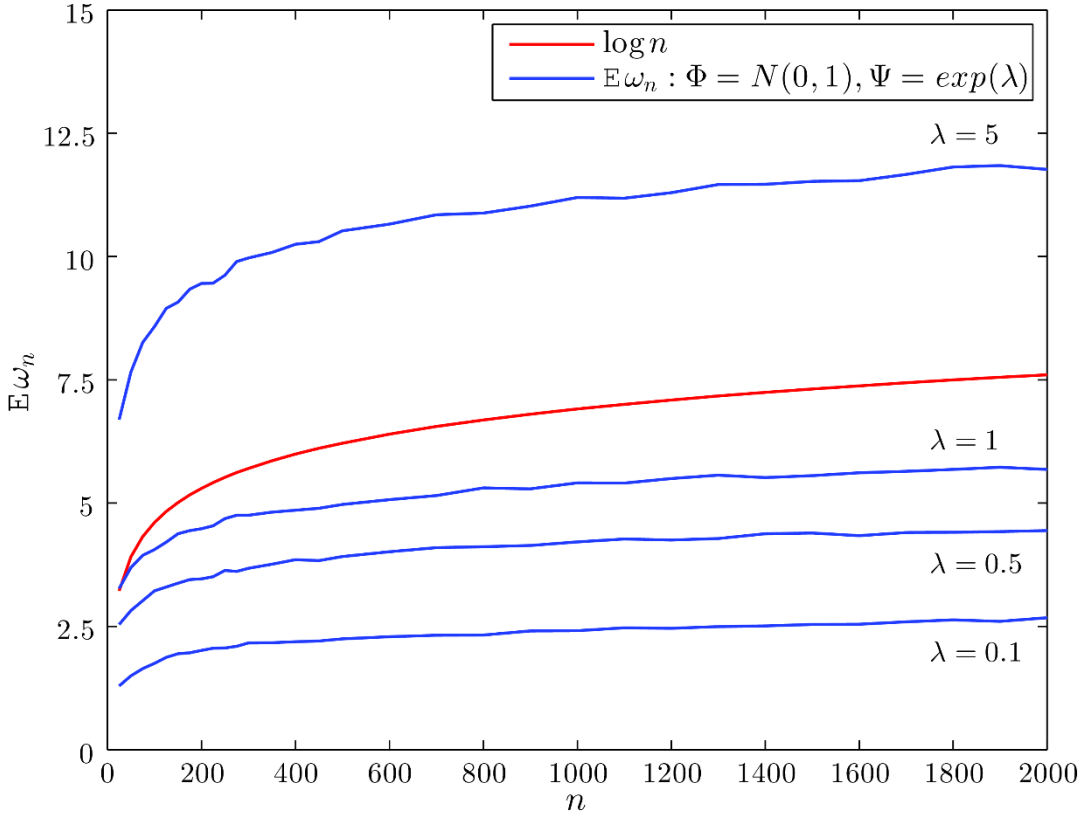
Výpočetní složitost je tedy polynomiální právě tehdy, pokud se žádné zúžené intervaly nepřekrývají. V případě, že k překrytí dochází, tak je složitost algoritmu exponenciální v maximálním počtu zúžených intervalů, které mají společný alespoň jeden bod. Označme tento počet ω_n:

$$\omega_n = \max\left(|K| : \left(\bigcap_{j \in K} \boldsymbol{x_j^{\frac{1}{n}}}\right) \cap \left[\underline{E}, \overline{E}\right] \neq \emptyset; K \subset \{1, 2, \ldots, n\}\right).$$

Počet maximálního překrytí intervalů lze reprezentovat pomocí grafu. V něm vrcholy značí jednotlivé zúžené intervaly a hrana mezi vrcholy značí, že zúžené intervaly mají společný alespoň jeden bod. Pak ω_n je maximální klikou tohoto grafu.

Úloha je obecně NP-obtížná, ale v případě, že data pocházející z běžných rozdělení, tak s velkou pravděpodobnostní je hodnota ω_n „malá“. Na základě simulačních experimentů (viz Obrázek 6) byla stanovena následující hypotéza (Černý a Sokol 2015):

Hypotéza. Nechť středy $x_1^C, \ldots, x_n^C$ jsou generované z rozdělení Φ, a poloměry $x_1^\Delta, \ldots, x_n^\Delta$ jsou generované z nezáporného rozdělení Ψ a výběry jsou nezávislé. Pokud je Φ spojité rozdělení s konečným prvním a druhým momentem a jeho funkce hustoty je omezena shora a Ψ má konečný první a druhý moment, pak $E(\omega_n) = O(\log n)$ a $Var(\omega_n) = O(1)$.



Obrázek 6: Střední hodnota maximálního počtu překrytých zúžených intervalů

Analýza složitosti

Hlavním cílem tohoto příspěvku je představit některé vlastnosti a možné interpretace $E(\omega_n)$. Pro zjednodušení budeme uvažovat $\Phi = \mathrm{U}(0,1)$ a $\Psi = 1$. Pravděpodobnostní rozdělení maxima překrytí zúžených intervalů lze za těchto podmínek interpretovat zajímavými způsoby. Dva z nich budou v následujících odstavcích představeny.

Využití řezů simplexu

Tento postup byl představen v (Sokol a Rada 2016). Mějme středy intervalů $x_1^c \leq x_2^c \leq \cdots \leq x_n^c$, zaveďme $x_0^c = 0$ a $x_{n+1}^c = 1$. Bez újmy na obecnosti můžeme předpokládat $0 = x_0^c \leq x_1^c \leq x_2^c \leq \cdots \leq x_n^c \leq x_{n+1}^c = 1$. Pak

$$d_i = x_i^c - x_{i-1}^c, i = 1, \dots, n+1,$$

přičemž platí $\sum_{i=1}^{n+1} d_i = 1$ a $d_i \geq 0$ pro všechna i.

V takovém případě má proměnná $d = (d_1, d_2, \dots, d_{n+1})$ Dirichletovo rozdělení s parametry $\alpha = (1,1, \dots, 1)$. Proměnná d má tak rovnoměrné rozdělení na $(n+1)$-dimenzionálním simplexu S_{n+1} s vrcholy $a_1 = (1,0, \dots, 0), \dots, a_{n+1} = (0, \dots, 0,1)$. Vzhledem k tomu, že $d_{n+1} = 1 - \sum_{i=1}^{n} d_i$ lze simplex bez ztráty informace převést na n-dimenzionální simplex

$$\Delta_n = \left\{ d \in R^n : d \geq 0, \sum_{i=1}^{n} d_i \leq 1 \right\}.$$

Překrytí zúžených intervalů pak lze vyjádřit s využitím geometrických vlastností Δ_n. Uvažujme množinu indexů $I \subseteq \{1, \dots, n\}$, přičemž $|I| = k$ pro dané k. Pak platí:

$$\bigcap_{i \in I} \boldsymbol{x}_i^{\frac{1}{n}} \neq \emptyset \Leftrightarrow \max_{i \in I} x_i^c - \min_{i \in I} x_i^c \leq \frac{2}{n} \Leftrightarrow \sum_{i=1+\min\{j \in I\}}^{\max\{j \in I\}} d_i \leq \frac{2}{n}.$$

Tedy k zúžených intervalů má společný alespoň jeden bod právě tehdy, když nejvzdálenější středy jsou dostatečně blízko sebe. Z toho důvodu se při zkoumání maximálního počtu překrytí můžeme omezit pouze na množiny sousedních středů, pro $|I| = k$ musí platit $\max\{i \in I\} - \min\{i \in I\} = k - 1$. Nazvěme tuto množinu sousední.

Můžeme definovat polorovinu, která bude využita pro výpočet pravděpodobnosti, že se překrývá k sousedních zúžených intervalů. Díky rovnoměrnému rozdělení středů na Δ_n ji lze vyjádřit následovně:

$$H_k^n := \left\{ d \in R^n : \sum_{j=2}^{k} d_j \leq \frac{2}{n} \right\}.$$

Takto definovaná polorovina H_k^n odřízne část Δ_n, přičemž platí, že čím větší objem je odříznut, tím větší je pravděpodobnost překrytí k zúžených sousedních intervalů. Označme $C(n,k) = H_k^n \cap \Delta_n$. Pak pravděpodobnost $P(n,k)$, že se překrývá k sousedních zúžených intervalů vyjádříme:

$$P(n,k) = \frac{Vol\big(C(n,k)\big)}{Vol(\Delta_n)} = n! Vol\big(C(n,k)\big).$$

Cílem je analyticky spočítat $E(\omega_n)$. K tomu lze využít standardní vzorec

$$E(\omega_n) = \sum_{k=1}^{n} k\big(P(\omega_n = k)\big),$$

kde využijeme $P(\omega_n = k) = P(\omega_n \geq k) - P(\omega_n \geq k+1)$. Pro $P(\omega_n \geq k)$ pak platí

$$P(\omega_n \geq k) = n! Vol\left(\bigcup_{I \subseteq [n], |I| = k} C(n,I) \right),$$

kde

$$C(n,I) = \sum_{i=1+\min\{j \in I\}}^{\max\{j \in I\}} d_i \leq \frac{2}{n}.$$

Je možné, že pro různé množiny I_j, takové, že $|I| = k$, se odřezané geometrické útvary $C(n,I_j)$ překrývají. Nelze tedy objemy $C(n,I_j)$ jednoduše sčítat a je třeba využít princip inkluze a exkluze. Další problém spočívá v enumeraci objemu řezu. Existuje sice několik zajímavých přístupů, například (Gerber 1981; Varsi 1973), které

využívají rekurzivní vzorec pro výpočet objemu řezu simplexu, ale bohužel nejsou vhodné pro důkaz limitního chování $E(\omega_n)$.

Transformace na narozeninový problém

Zaveďme indikátor

$$Z_i(t) = \begin{cases} 1, t \in \boldsymbol{x}_i^{\frac{1}{n}}, \\ 0, t \notin \boldsymbol{x}_i^{\frac{1}{n}}, \end{cases}$$

pak $P\left(Z_i(t) = 1, t \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]\right) = \frac{2}{n}$. Omezíme se zde pouze na interval $t \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$, jelikož pro krajní body intervalu je pravděpodobnost vždy menší. Tímto omezením bude výsledkem horní odhad $E(\omega_n)$. Pro $t, s \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$ a $i \neq j$ platí, že $Z_i(t)$ a $Z_j(t)$ jsou nezávislé, stejně tak $Z_i(t)$ a $Z_j(s)$.

Dále označme

$$A(t) = \sum_{i=1}^{n} Z_i(t).$$

Tato proměnná tedy značí počet překrytých zúžených intervalů v bodě t. Naším cílem je najít maximum těchto veličin na celém intervalu $[0,1]$. Platí, že $A(t)$ má binomické rozdělení $Bi(n, \frac{2}{n})$. Pro $t, s \in \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$ platí

$$|t - s| \leq \frac{2}{n}: \quad cov(A(s), A(t)) = n\left(\frac{2}{n} - |t - s|\right) - \frac{4}{n},$$
$$|t - s| \geq \frac{2}{n}: \quad cov(A(s), A(t)) = -\frac{4}{n}.$$

Tedy s rostoucím počtem pozorování v množině dat kovariance klesá.

Nechť T_n je množina $\left\lfloor \frac{n}{2} \right\rfloor + 1$ bodů ekvidistantně rozmístěných na intervalu $[0,1]$. Pak nutně musí platit, že každý zúžený interval zasahuje do právě jednoho bodu z množiny T_n. Kovariance pak v tomto případě popisuje situaci, kdy interval může zasáhnout do právě jednoho bodu. Při takovém rozmístění bodů pak platí

$$\omega_n = \max(A(t); t \in T_n).$$

Jedná se tedy o maximum $2n + 1$ binomických proměnných s parametry $Bi(n, \frac{2}{n})$ (lze aproximovat Poissonovým rozdělením $Pois(2)$), které jsou záporně korelované a korelace s rostoucím n jde limitně k nule.

Poznámka. Pokud by proměnné byly nezávislé, distribuci maxima lze aproximovat funkcí $O(\log n / \log \log n)$ (Kimber 1983).

Výsledek uvedený v poznámce by dokázal platnost hypotézy na tomto určitém nastavení rozdělení středů a poloměru.

Při použití rozdělení $\Phi = \mathrm{U}(0,1)$ a $\Psi = 1$ pro generování intervalů je řešený problém – nalezení maximálního počtu překrytí zúžených intervalů – speciální variací narozeninového problému. V teorii pravděpodobnosti je narozeninovým problémem (někdy je používán název narozeninový paradox) myšlena úloha spočívající v nalezení minimální velikosti skupiny náhodně vybraných lidí takové, že pravděpodobnost, že nějací dva lidé ze skupiny budou mít narozeniny ve stejný den, byla alespoň 50 %. Za autora této úloha je udáván Richard von

Mises. Za obvyklých předpokladů, že datum narození je rovnoměrně rozložené a že rok má 365 dní, je to skupina o velikosti 23 osob.

V této variaci je $2n + 1$ dnů a rozdělení data narození je rovnoměrné. Cílem je přitom nalezení střední hodnoty distribuce v závislosti na n. Přesná enumerace střední hodnoty v závislosti na n je výpočetně velmi náročná. Existují zajímavé aproximace této distribuce, např. (Matthias Arnold a Werner Glaß 2013), ale pro důkaz limitního chování $E(\omega_n)$ je nelze použít, protože ignorují nenulovou kovarianci $cov(A(s), A(t))$.

Závěr

V první části byly představeny specializované algoritmy pro výpočet těsné horní meze výběrového rozptylu v případě intervalových dat. Byla zkoumána jejich výpočetní složitost, přičemž důraz byl kladen především na situace, kdy intervaly pocházejí z „běžných“ rozdělení. V další části byl pak podrobně rozebrán Fersonův algoritmus spolu s analýzou složitosti. Exponenciální část složitosti tohoto algoritmu závisí na maximálním počtu překrytí zúžených intervalů, proto byla zkoumána střední hodnota maximálního počtu překrytí v případě, že středy a poloměry intervalů pocházejí z určitých, „běžných“, rozdělení.

Byly odvozeny způsoby, jak lze střední hodnotu maximálního počtu překrytí spočítat s využitím geometrických vlastností řezů simplexu, nebo transformací na speciální instanci tzv. narozeninového problému. Oba tyto způsoby jsou použitelné pro odhady jednotlivých hodnot, ale zatím se je nepovedlo upravit pro výpočet limitního chování maximálního počtu překrytí zúžených intervalů v závislosti na počtu intervalů.

Literatura

ČERNÝ, Michal a Milan HLADÍK, 2014a. Intervalová data a výpočet některých statistik. Robust 2014.

ČERNÝ, Michal a Milan HLADÍK, 2014b. The Complexity of Computation and Approximation of the t-ratio over One-dimensional Interval Data. *Computational Statistics & Data Analysis* [online]. 12., **80**, 26–43. ISSN 0167-9473. Dostupné z: doi:10.1016/j.csda.2014.06.007

ČERNÝ, Michal a Ondřej SOKOL, 2015. Interval data and sample variance: A study of an efficiently computable case. In: *Mathematical Methods in Economics 2015 (MME)*.

DANTSIN, Evgeny, Vladik KREINOVICH, Alexander WOLPERT a Gang XIANG, 2006. Population Variance under Interval Uncertainty: A New Algorithm. *Reliable Computing* [online]. 8., **12**(4), 273–280. ISSN 1573-1340. Dostupné z: doi:10.1007/s11155-006-9001-x

FERSON, Scott, Lev GINZBURG, Vladik KREINOVICH, Luc LONGPRÉ a Monica AVILES, 2002. Computing Variance for Interval Data is NP-hard. *ACM SIGACT News*. **33**(2), 108–118.

FERSON, Scott, Lev GINZBURG, Vladik KREINOVICH, Luc LONGPRÉ a Monica AVILES, 2005. Exact Bounds on Finite Populations of Interval Data. *Reliable Computing* [online]. 6., **11**(3), 207–233. ISSN 1573-1340. Dostupné z: doi:10.1007/s11155-005-3616-1

FERSON, Scott, Vladik KREINOVICH, Janos HAJAGOS, William OBERKAMPF a Lev GINZBURG, 2007. *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty* [online]. B.m.: Sandia National Laboratories. Dostupné z: http://prod.sandia.gov/techlib/access-control.cgi/2007/070939.pdf

GERBER, Leon, 1981. The Volume Cut Off a Simplex by a Half-space. *Pacific Journal of Mathematics* [online]. 1. 6., **94**(2), 311–314. ISSN 0030-8730, 0030-8730. Dostupné z: doi:10.2140/pjm.1981.94.311

KIMBER, A. C., 1983. A note on Poisson maxima. *Probability Theory and Related Fields* [online]. **63**(4), 551–552. ISSN 0044-3719, 1432-2064. Dostupné z: doi:10.1007/BF00533727

KREINOVICH, Vladik, 1996. Maximum Entropy and Interval Computations. *Reliable Computing*. **2**(1), 63–79.

KREINOVICH, Vladik, Luc LONGPRÉ, Scott FERSON a Lev GINZBURG, 2004. *Computing Higher Central Moments for Interval Data*. B.m.: University of Texas at El Paso.

MANSKI, Charles F, 2003. *Partial identification of probability distributions*. B.m.: Springer Science & Business Media.

MATTHIAS ARNOLD a ANDWERNER GLAß, 2013. Simple Approximation Formulas for the Birthday Problem. *The American Mathematical Monthly* [online]. **120**(7), 645. ISSN 00029890. Dostupné z: doi:10.4169/amer.math.monthly.120.07.645

NGUYEN, Hung T., Vladik KREINOVICH, Berlin WU a Gang XIANG, 2012. *Computing Statistics under Interval and Fuzzy Uncertainty* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg. Studies in Computational Intelligence [vid. 2016-01-22]. ISBN 978-3-642-24904-4. Dostupné z: http://link.springer.com/10.1007/978-3-642-24905-1

SOKOL, Ondřej, 2015. Výpočet horní meze výběrového rozptylu nad intervalovými daty. In: *Nové trendy v ekonometrii a operačním výzkumu*: *Mezinárodní vědecký seminář nové trendy v ekonometrii a operačním výzkumu*. Bratislava: Ekonóm. ISBN 978-80-225-4181-7.

SOKOL, Ondřej a Miroslav RADA, 2016. How to Prove Polynomiality of Computation of Upper Bound Considering Random Intervals? In: *Mathematical Methods in Economics (MME2016)*. Liberec: TU Liberec, s. 779--785. ISBN 978-80-7494-296-9.

VARSI, Giulio, 1973. The Multidimensional Content of the Frustum of the Simplex. *Pacific Journal of Mathematics* [online]. 1. 5., **46**(1), 303–314. ISSN 0030-8730, 0030-8730. Dostupné z: doi:10.2140/pjm.1973.46.303

XIANG, Gang, 2006. Fast Algorithm for Computing the Upper Endpoint of Sample Variance for Interval Data: Case of Sufficiently Accurate Measurements. *Reliable Computing* [online]. 2., **12**(1), 59–64. ISSN 1385-3139, 1573-1340. Dostupné z: doi:10.1007/s11155-006-2965-8

XIANG, Gang, Martine CEBERIO a Vladik KREINOVICH, 2007. Computing Population Variance and Entropy under Interval Uncertainty: Linear-time Algorithms. *Reliable computing*. **13**(6), 467–488.

JEL Classification: C44

Dedikace: Příspěvek vznikl s podporou projektu IGA F4/63/2016 Interní grantové agentury Vysoké školy ekonomické v Praze.

Summary

Statistics over interval data

This work deals mainly with the computation of the tight upper limit of the sample variance when the exact data are not known but intervals which certainly contain them are available. Generally, finding the upper limit of the sample variance knowing only interval data is an NP-hard problem, but under certain conditions imposed on the input data an appropriate efficient algorithm can be used. We focus mainly on the average computational complexity of Ferson's algorithm over randomly generated data.

Using Ferson's algorithm, the computation of the maximal possible variance over interval-valued dataset can be realized in polynomial time in the maximal number of narrowed intervals intersecting at one point; narrowed means that the intervals are shrinked proportionally to the size of the dataset.Considering random datasets, experiments allowed for conjecturing that the maximal number of narrowed intervals intersecting at one point is at most of logarithmic size for a reasonable choice of the data-generating process.

Assume uniform distribution of centers and constant radii. The computation of expected value of the maximal number of narrowed intervals intersecting at one point is reducible to the evaluation of the volume of some special simplicial cuts or to the special instance of birthday problem. Both cases are analyzed in this work.

Keywords: partial identification, interval data, sample variance, birthday problém.

Application of data envelopment analysis for study of efficiency of small and medium enterprises

Soldatyuk Nataliya

xsoln900@vse.cz

PhD student of Quantitative Methods in Economics

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

Abstract: The article describes results of an empirical research focused on measuring the impact of selected macroeconomic indicators on relative operational efficiency of the small and medium enterprises across countries of Eastern Europe. For these purposes Data Envelopment Analysis techniques were used to measure relative efficiency of companies across 14 countries of Eastern Europe. Than regression approach was used for estimation of relationship between macroeconomic indicators and DEA efficiency. Macroeconomic indicators were selected based on problem areas identified by companies as the major obstacles on their way to efficient operation. Data used for this analysis has been taken from the results of the enterprise survey for 2013 year conducted by the World Bank.

Keywords: relative efficiency, Data Envelopment Analysis, small and medium enterprises.

Introduction

The worldwide experience proves the special role and the importance of small and medium enterprises (SMEs) within national economies. For instance, according to the European Commission report[1], 99.8% in the non-financial business sector are small and medium enterprises. About 92% of the total business sector belong to micro businesses, which employ fewer than 10 workers. At the same time SMEs contribute 58 cents in every euro in economy. After the ascension of the last three decades, it is thought that the small and medium enterprises will be the main vector of the economic progress going further, both in the developed countries and the ones in transition.

From the employment point of view, the SME sector provides significant portion of jobs worldwide and it is continuously creating new job opportunities. The impact of SME varies depending on economy: it is stronger in emerging markets and weaker in developed economies. According to European commission report[1] in 2014 SMEs employed around 90 million people in EU, which is 67% of total employment. In conditions of continuously growing world population, creation of new working positions is critical. The World Bank[2] estimates need of 600 million jobs in the next 15 years to absorb a growing global workforce. Therefore, the SME sector is greatly responsible for the economic growth and therefore it is an important sector to focus on.

Despite of its important role in economy, small and medium enterprises face a lot of issues on their way to efficient operation. Among the main obstacles mentioned by companies during the World Bank survey were the access to finance, practice of the informal sector, tax rates, corruption and others.

One of the biggest obstacles is a limited access to finance. According to the World Bank[2] estimation, yet more than 50% of SME lack access to finance. The financing gap is even larger when informal enterprises are taken into account. For lending institutions SMEs represent a segment with high credit risks as they are less likely to be able to secure bank loans than large firms. Considering that SMEs belong to high risk portfolio of customers, often only loans with high value of collateral are available for them. This issue has been frequently discussed and different steps were taken by some European countries in order to simplify approval procedure for SMEs. Thus as one of macroeconomic indicators we included into the analysis is an average value of collateral needed for a loan (% of the loan amount). The other indicators are discussed in the further sections.

The main objectives of the study are the following:

- To measure DEA efficiently for the SME sector (aggregated by countries)
- To estimate relationship between selected macroeconomic indicators and DEA efficiency of small and medium enterprises in countries of Eastern Europe.

[1] http://ec.europa.eu/growth/toolsdatabases/newsroom

[2] http://www.worldbank.org/en/topic/financialsector/brief/smes-finance

Methodology

This section presents methodological steps adopted in this study, which is Data Envelopment Analysis and linear regression.

Data Envelopment Analysis is a commonly used optimization method for comparing the relative efficiency of given decision making units (DMUs) based on multiple inputs and outputs of those units. The efficiency then is measured in a bounded ratio scale by the fraction "weighted output" to "weighted input". The biggest advantage of this method is that it allows us to measure the efficiency taking into account inputs and outputs that are not easily comparable.

DEA approach has been widely applied in various industries, for example to measure efficiency of education, medical care, utilities sector, banks or insurance branches and ecological environment (Sherman and Gold, 1985) and (Grmanová and Jablonsky, 2009). There are more unusual examples of DEA application, where the efficiency of football teams was measured using DEA models (García-Sánchez, 2007 and Guzmán 2007).

It is also used to evaluate SME performance in different countries and industries. Various articles were published describing application of DEA approach to SME efficiency. For instance, Halkos and Tzeremes (2010) have used DEA models to show the effect of foreign ownership on SME performance in Greece.

In our study decision making units (DMUs) are small and medium enterprises. It was decided to apply Banker-Charnes-Cooper (BCC) model for efficiency measuring. BCC model is one of basic extensions of classical CCR (Charnes-Cooper-Rhodes) model, however it fully suits purposes of the study. Here it was assumed that all model inputs and outputs variables expressed by crisp numbers. Another assumption is variable returns-to-scale, which is also accommodated by BCC model.

Lets consider n DMUs, m inputs and s outputs. Denote vectors of inputs and outputs $x \in R^m$ and $y \in R^s$ respectively. Then input-oriented BCC problem is expressed with a real variable θ and a nonnegative vector $\lambda = (\lambda_1,...,\lambda_n)^T$ of variables as follows:

$$(BCC_o) \quad \min_{\theta,\lambda} \theta$$

$$subject\ to \quad \theta\, x_o - X\lambda \geq 0$$

$$Y\lambda \geq y_o$$

$$e^T\lambda = 1$$

$$\lambda \geq 0,$$

where the index o shows the DMU under evaluation. $X = (x_j) \in R^{m\,x n}$ and $Y = (y_j) \in R^{s\,x n}$ are matrices of inputs and outputs. Variable returns-to-scale is expressed by condition $e\lambda = 1$.Feasible values for θ are between 0 and 1, where 1 means 100% relative efficiency and 0 means 0% relative efficiency.

Regression approach was applied on the second stage of the study to measure effect of macroeconomic indicators on SME efficiency. All observations refer to 2013 year and thus simple linear regression was applied. The regression equation was used in following form:

$$\theta_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon_i, \quad i = 1,...,n,$$

where θ_i is DEA efficiency, x_1, x_2, x_3, x_4 are macroeconomic indicators discussed in the section model formulation, ε_i is identically distributed independent error. Parameter estimates were obtained by ordinary least squares method.

Model formulation

The study is divided into two phases. During the first phase the efficiency of enterprises is measured on company level. During the second phase the impact of macroeconomic indicators on operational efficiency of companies is investigated.

Construction of DEA model

The main challenge during the modelling DEA was selection of model inputs and outputs. The choice of inputs and outputs was driven not only by models commonly found in the literature review, but also by the availability of data. Thus the selection of variables for DEA model was inspired by article of Jablonsky and Grmanova (2009), where authors have applied two stage DEA model with interval data to measure efficiency of telecommunication operators in Czech Republic. The author identifies main inputs in company's operation as the following: operational costs, market potential and number of opening hours. In our model we used similar structure of inputs, however we have separated particular components of operational costs in order to conduct more detailed analysis. Operational costs now have three separate components: a) cost of labour, b) cost of services such as electricity, water and fuel and c) other costs including material costs and rental costs. We have also added a number of employees as a model input. We expressed market potential by inverse number of market competitors for each company, however usage of this characteristic led to its excessively high coefficients and it was decided to exclude it from the final model. Another model input is number of working hours per week.The argumentation for inclusion of this variable was that in case of SME there is often no standard length of working week and time in operations could variate from few days per week to 24/7 nonstop operations. Thus outputs of such operations cannot be treated equally. This approach commonly used in other works (for instance, Grmanova and Jablonsky, 2007). Output of SME operations is represented by annual revenue weighted by GDP per capita. As we want to compare working efficiency of enterprises operating in different countries, in the first phase of the study we want to eliminate economy impact to investigate it in the second phase of the study. Thus we weight total annual revenue by GDP per capita. The final version of DEA model included five inputs (cost of labour, cost of services, material and other costs, number of working hours per week and number of employees) and one output (revenue).

The outcome of DEA model is presented by several measurements calculated for each company:

- working efficiency score
- weight of each input
- level of additional improvement (decrease in inputs) needed for a unit to become efficient.

Several data sources were used for the study: the main data sample containing production inputs and outputs of SMEs was obtained from the World Bank database, additional data were retrieved from Yandex databases.

Main data sample for our analysis was obtained from the annual World Bank's Enterprise Survey. The enterprise survey collects data from key manufacturing and service sectors in every region of the world. One of the main objectives of collecting these data is to assess the constraints to private sector growth and enterprise performance. The data sample was created as a result of personal interviews with enterprises conducted by Word Bank employees. During the interview enterprises have been asked 125 questions covering various business aspects. Questions could be divided into 11 groups:

- Control Information, registration number, date of registration,
- General Information: ownership, start – up,
- Infrastructure and Services: power, water, transport and communication technologies,
- Sales and Supplies: import, export, supply and demand conditions,
- Degree and Competition: price and supply changes, competitors,
- Land: land ownership, land access issue,
- Crime: extent and losses due to crime,
- Business – Government Relations: quality of public services, consistency of policy,
- Investment Climate Constrains: evaluation of general obstacles.

The dataset contains valuable information about enterprises activity from very different perspectives. Datasets gathered from different countries have been standardized into one data sample, thus for all observations same set of variables is present and it's internationally comparable. However one of the main issue during the work with such sample was data quality. A lot of variables have missing values for the majority of observations. At the first stage before defining a structure of our model, data quality analysis and data cleaning have been performed.

The Enterprise Survey also includes indication of company's industry and main product, which allowed us to conduct study for each industry separately. For the study we have selected 5 most frequent industries: production

of food (specifically bakery only), metal production (basic metals only), construction services (only building construction), land passenger transportation and land cargo transportation (only track transportation).

The World Bank Enterprise Survey is focused on Small, Medium and Large enterprises. Under definition of Small enterprises fall companies which employ less than 20 employees, Medium enterprises employs up to 100 of employees. Companies with number of employees higher than 100 are considered to be large enterprises. We have excluded Large enterprises from our data sample as our research mainly focuses on Small and Medium enterprises.

Selection of repressors

Although there are many macroeconomic indicators which could be included into analysis, in this study we decided to focus of environmental factors, which companies themselves identify as major obstacles on their way to efficient operation. As part of the enterprise survey the companies have been asked to name three major obstacles they face. The figure below shows the most frequently named obstacles.

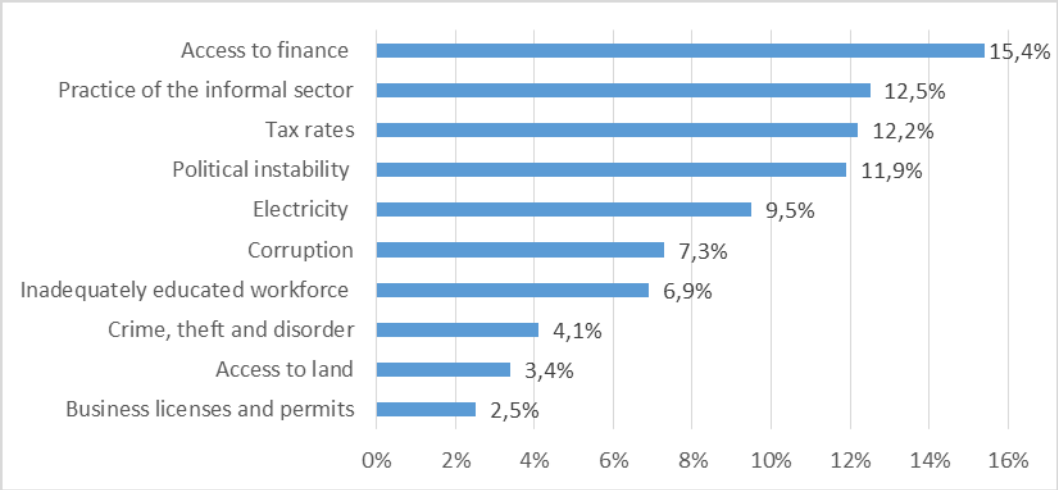


Figure 1: Most frequent obstacles indicated by SMEs Source: The World Bank official website

Thus we selected four macroeconomic indicators for regression analysis, each of them to reflect above mentioned obstacles. The indicators are following:

- Access to finance is represented by value of collateral needed for a loan (% of the loan amount) (variable Finance in the tables below). We would like to use other indicators such as loan approval rate, average interest rate or average credit limit, however these indicators were not available.
- Practice of the informal sector is represented by percentage of unregistered businesses (Inform)
- Tax rates are reflected by average tax to EBITDA ratio for SME sector (Tax)
- Electricity prices (Euro per 100 kWh) (Electr)
- Corruption is represented by percentage of firms experiencing at least one bribe payment request (Corrupt)
- Crime level (percentage measure of security provided by the World Bank)(Crime)

Results and discussion

In total 1134 observations from 14 European countries were analysed. Observation distribution by countries and industries is presented in the table below.

Table 1: Observations by countries and industries

Number of companies	**Bakery**	**Cargo Transportation**	**Building Construction**	**Metal Production**	**Passenger Transportation**
Albania	16	1	26	16	2
Bosnia and Herzegovina	7	7	24	27	9
Bulgaria	55	19	37	82	12
Czech Republic	2	6	18	27	2
Estonia	8	3	30	18	13

Number of companies	**Bakery**	**Cargo Transportation**	**Building Construction**	**Metal Production**	**Passenger Transportation**
Macedonia	27	15	34	25	4
Latvia	6	9	26	9	6
Lithuania	11	10	21	18	1
Montenegro	8	2	5	7	2
Poland	9	2	26	20	8
Romania	27	17	46	35	6
Serbia	9	8	35	31	10
Slovak Republic	4	6	20	14	2
Slovenia	6	5	32	41	2
Industry total	**195**	**110**	**380**	**370**	**79**

The DEA modelling was conducted separately for each industry. As output of DEA modelling efficiency scores were derived for each company. Due to high numbers of observations all DEA results cannot be displayed and thus table below presents aggregated DEA results in form of average efficiency scores. Although the results are not easily comparable in such format, but they still reflect the general trends.

Table 2: Average efficiency scores by countries and industries

	Bakery	**Cargo Transportation**	**Building Construction**	**Metal Production**	**Passenger Transportation**
Albania	0.79	0.08	0.22	0.74	0.47
Bosnia and Herzegovina	0.70	0.36	0.17	0.53	0.34
Bulgaria	0.58	0.43	0.26	0.62	0.32
Czech Republic	0.76	0.75	0.41	0.61	0.59
Estonia	0.54	0.60	0.39	0.67	0.48
Macedonia	0.65	0.47	0.32	0.64	0.29
Latvia	0.47	0.37	0.27	0.70	0.19
Lithuania	0.72	0.72	0.27	0.77	0.33
Montenegro	0.55	0.71	0.24	0.71	0.59
Poland	0.78	0.09	0.28	0.70	0.44
Romania	0.48	0.48	0.23	0.69	0.42
Serbia	0.72	0.50	0.22	0.55	0.30
Slovak Republic	0.60	0.49	0.19	0.65	0.14
Slovenia	0.52	0.59	0.40	0.61	0.76
Industry Average	0.61	0.49	0.28	0.64	0.38

It is interesting to notice that there is no stable trend across all 5 industries. Generally, we may conclude, that Lithuania, Czech Republic and Slovenia show relatively high efficiency scores across all industries, but for other countries efficiency trends are very various, which is expected. For instance, Albanian companies have high relative scores in bakery, but very low in cargo transportation.

Tables below shows coefficient estimations for 5 analysed industries using OLS for $\alpha < 0.05$:

Table 3: OLS estimations for bakery industry

Variable	Coefficient estimation	Std. Error	t - value	Pr (>|t|)
Intercept	15.326	26.338	0.582	0.561
Corrupt	0.072	0.449	0.160	0.872
Finance	0.028	0.063	0.436	0.664
Inform	0.383	0.265	1.444	0.150
Tax	2.162	1.045	2.068	0.040
Electr	-1.158	1.102	-1.050	0.295
Crime	0.145	0.169	0.856	0.393

Table 4: OLS estimations for land cargo transportation industry

Variable	Coefficient estimation	Std. Error	t - value	Pr (>|t|)
Intercept	-2.752	33.916	-0.081	0.935
Corrupt	-0.255	0.974	-0.262	0.793
Finance	-0.132	0.095	-1.375	0.172
Inform	0.584	0.345	1.691	0.093
Tax	-1.353	1.743	-0.776	0.439
Electr	3.856	1.805	2.136	0.035
Crime	0.635	0.283	2.241	0.027

Table 5: OLS estimations for building construction industry

Variable	Coefficient estimation	Std. Error	t - value	Pr (>|t|)
Intercept	19.477	13.836	1.408	0.160
Corrupt	-0.890	0.324	-2.745	0.006
Finance	0.025	0.040	0.645	0.519
Inform	0.050	0.149	0.338	0.735
Tax	-0.921	0.557	-1.651	0.099
Electr	1.426	0.677	2.106	0.035
Crime	0.086	0.122	0.709	0.478

Table 6: OLS estimations for metal production industry

Variable	Coefficient estimation	Std. Error	t - value	Pr (>|t|)
Intercept	35.002	13.405	2.611	0.009
Corrupt	0.076	0.328	1.519	0.129
Finance	0.076	0.042	1.821	0.069
Inform	0.033	0.149	0.224	0.822
Tax	0.169	0.545	0.310	0.756
Electr	1.162	0.615	1.887	0.060
Crime	-0.134	0.103	-1.302	0.193

Table 7: OLS estimations for land passenger transportation industry

Variable	Coefficient estimation	Std. Error	t - value	Pr (>|t|)
Intercept	32.472	46.891	0.693	0.490
Corrupt	0.937	1.107	0.847	0.400
Finance	-0.180	1.107	-1.506	-0.136
Inform	-0.080	0.530	-1.377	0.172
Tax	-1.692	1.229	-1.377	0.172
Electr	3.727	1.889	1.973	0.052
Crime	0.363	0.328	1.109	0.270

From table 3, for bakery industry only tax rates have significant impact, however positive value of coefficient means that growth of tax rates is associated with higher efficiency. From table 4, electricity prices and crime level have impact on efficiency of land cargo transportation industry, however it is difficult to interpret the sign of the coefficients. From table 5, corruption has an impact on building construction industry, increase of corruption level is associated with decrease of DEA efficiency. As it is shown in tables 6 and 7 there is no

significant dependency between analysed macroeconomic indicators and metal production and passenger transportation industries.

Conclusions

The study deals with efficiency analysis of SME sector in different countries of Eastern Europe. The procedure is based on DEA model, which evaluates efficiency of each enterprise. Data for the analysis was taken from results of the annual enterprise survey provided by the World Bank. Further linear regression approach is applied in order to measure the impact of selected macroeconomic indicators on relative operational efficiency. Macroeconomic indicators were selected based on problem arears identified by companies as major obstacles on their way to efficient operation. They include such obstacles like access to finance, presence on informal sector, corruption level, tax rate and others. Regression analysis has shown no relationship between DEA efficiency and variables presenting access to finance. On the other hand the regression analysis identified that growth of corruption and electricity prices have negative on building construction and bakery industries. However positive coefficients were estimated for electricity prices for other industries, which is difficult to interpret.

For the future study we will extend both selection of analysed industries and macroeconomic factors.

References

ANOUZE, A., OSMAN, I., EMROUZNEJAD, A. (2014) Handbook of Research on Strategic Performance Management and Measurement Using Data Envelopment Analysis, *IGI Global*., ISBN: 978-1-4666-4474-8.

AZIZI, H. (2013) A note on data envelopment analysis with missing values: an interval DEA approach. The International Journal of Advanced Manufacturing Technology, vol. 66, issue 9, pp 1817-1823.

BOUSSOFIANE A., DYSON R.G., THANASSOULIS E. (1991) Applied data envelopment analysis. *European Journal of Operational Research*, vol. 52, p.1-15.

CHARNES A., COOPER W.W., RHODES E. (1978) Measuring the efficiency of decision-making units. *European Journal of Operational Research*, vol. 2, p.249-444

COOPER W., SEIFORD M., TONE K. (2007). Data Envelopment Analysis. A Comprehensive Text with Models, Applications, References and DEA-Solver Software. Second edition. ISBN 978-0-387-45283-8.

GARCÍA-SÁNCHEZ I.M. (2007) Efficiency and effectiveness of Spanish football teams: a three-stage-DEA approach. *Cejor*, vol. 15, p.21–45.

GRMANOVA E., JABLONSKY J. (2009) Efficiency analysis of Slovak and Czech insurance companies using data envelopment analysis models. *Ekonomicky casopis*, vol. 57, p.857-869.

GUZMÁN I., MORROW S. (2007) Measuring efficiency and productivity in professional football teams: evidence from the English Premier League. *Cejor*, vol. 15, p.309–328.

HALKOS, G., TZEREMES N. (2010) The effect of foreign ownership on SMEs performance: An efficiency analysis perspective. *Journal of Productivity Analysis,* vol. 34, pp. 167-180.

KORHONEN P.J., LUPTACIK M. (2004) Eco-efficiency analysis of power plants: An extension of data envelopment analysis. *European Journal of Operational Research*, vol. 154, p.437-446.

MALEKOHAMMADI N., JAAFAR A.B., MONSI M. (2010) Setting Targets with Interval Data Envelopment Analysis Models via Wang Method. Sains Malaysiana 01/2010; 39:485-489.

SHERMAN H.D., GOLD F. (1985) Bank Branch Operating efficiency Evaluation with data envelopment analysis. *Journal of Banking and Finance*, vol. 9, p.297-315.

SMITH P. (1990) Data Envelopment Analysis Applied to Financial Statements. *Omega*, vol. 18, p.131-138.

SMIRLIS Y., MARAGOS E., DESPOTIS D. (2006). Data Envelopment Analysis with Missing Values: An Interval DEA Approach. Appl Math Comput 177(1): 1-10.

JEL Classification: C20, C61

Acknowledgements

This study was supported by the IGA VŠE/IG403036/63/2016.

Shrnutí

Použití analýzy obalu dat pro výzkum efektivity maloobchodních společností

Příspěvek shrnuje výsledky empirického studia zaměřeného na měření dopadu vybraných makroekonomických ukazatelů na relativní výkonnosti maloobchodních společností v zemích východní Evropy. Pro měření relativní výkonnosti firem byla použita analýza obalu dat. Následovně byla aplikovaná regresní analýza pro odhad vztahu mezi makroekonomickými ukazatele a DEA efektivitou. Makroekonomické ukazatele byly vybrány na základě problémových oblasti, které společností vidějí jako hlavní překážky efektivní činnosti. Data použita pro tuto analýzu byla převzata z výsledků průzkumu maloobchodních společností provedeného Světovou bankou.

Klíčová slova: efektivita**,** analýza obalu dat, maloobchodní společností.

Vliv profesní mobility na hodnocení pracovní spokojenosti

Michal Švarc

xsvam00@vse.cz

Doktorand oboru Ekonometrie a operační výzkum

Školitel: doc. Ing. Tomáš Cahlík, CSc., (tomas.cahlik@vse.cz)

Abstrakt: Hierarchická profesní mobilita, konkrétně pak vzestupná a sestupná profesní mobilita, a její vliv na hodnocení spokojenosti pracovního života u osob v předdůchodovém věku stojí ve středu tohoto článku. Odhad vlivu tranzice v rámci hierarchicky uspořádaných profesních tříd je proveden na datech z evropského výzkumu The Survey of Health, Ageing and Retirement in Europe (SHARE), britského průzkumu, English Longitudinal Study of Ageing (ELSA) a americké studie Health and Retirement Study (HRS), za pomoci modelů diskrétní volby s eliminací fixních efektů pomocí prvních diferencí. V případě sestupné profesní mobility se podařilo prokázat pozitivní vliv na redukci stresu z pracovního vytížení. U osob, které zažily vzestupnou mobilitu, bylo zjištěno zlepšení hodnocení pracovního prostředí ve většině sledovaných aspektů.

Klíčová slova: profesní mobilita, modely diskrétné volby, marginalizace

ÚVOD

Hlavním cílem této práce je identifikace vlivu sestupné profesní mobility (downward occupational mobility) na profesní spokojenost a hodnocení pracovního života u osob v předdůchodovém věku. První část tohoto textu je převážně sociologická a je věnována sociální stratifikaci a profesní mobilitě odvozené ze změny socio-ekonomické pozice jednotlivce. V této části jsou rovněž identifikována kritéria pro hodnocení profesního života, která jsou následně využita jako závislé proměnné v empirické analýze. Taktéž jsou zde uvedeny proměnné, které vstupují jako regresory v matematickém modelu. Následující pasáž textu popisuje datovou základnu, způsob jejího vzniku, včetně harmonizace proměnných napříč jednotlivými datovými zdroji. Poslední část tvoří jednotlivé modely pro odhad vlivu profesní mobility na pracovní spokojenost a hodnocení profesního života, přičemž jako vhodný aparát ekonometrické analýzy byl zvolen model diskrétní volby – ordinální logit – ve formě prvních diferencí pro minimalizaci fixních efektů jednotlivců.

Profesní mobilita

Sociální stratifikace je výrazem označujícím segmentaci společnosti dle socioekonomických charakteristik a je institucializovanou formou sociální nerovnosti. [ŠANDEROVÁ, 2000]. Zařazení do společenských segmentů je pak závislé na takzvané sociální pozici, která je jednotlivci přiřazena. Tyto pozice jsou odvozovány jak na základě objektivních kritérií, jako jsou:

- příslušnost k sociálním skupinám (etnická příslušnost, pohlaví, úroveň vzdělání),
- ekonomické charakteristiky jedince,
- mocensko-politická příslušnost
- kulturní kritéria
- rozsah neformálních sociálních sítí,

tak i na základě subjektivního vnímáni. Jedním z nejdůležitějších kritérií subjektivní percepce pro utváření sociální pozice je prestiž, a to jak prestiž zastávané pracovní pozice, tak i prestiž spojená s konkrétní osobou a jejím společenským uznáním. Tyto sociální pozice lze hierarchicky uspořádat, přičemž sociální pozice z vyšších hierarchických vrstev poskytují osobám větší životní šance, prostor pro seberealizaci a percepci větší respektovanosti okolím. [GIDDENS, 1999]

Sociální pozice odvozovaná na základě ekonomických charakteristik jedince je základem pro profesní mobilitu. Profesní mobilita může mít buď vzestupný či sestupný charakter. V případě vzestupné profesní mobility jedince dochází k tranzici napříč hierarchickým uspořádáním segmentované společnosti směrem vzhůru. Přesun z nižší vrstvy hierarchie do vyšších vrstev je většinou spjat s většími kvalifikačními požadavky na zastávanou pozici, vyšší úrovní zodpovědnosti a většími časovými nároky. Tyto zvyšující se nároky bývají spjaty i s rostoucím finančním ohodnocení jedince. [LAIN, 2012]

V současné době je používána pro klasifikaci socio-ekonomické pozice takzvaná syntetická pozice, která v sobě agreguje tři charakteristiky jedince, a to profesi, vzdělání a příjem. Nejhojněji užívaným měřícím nástrojem se

standardizovanou podobou je ISCO (International Standrard Classification of Occupations). Dalším klasickým systémem třídění je klasifikace SOC (Standard Occupational classification). Tato klasifikace je užívána zejména ve Spojených státech amerických a v Anglii.

Sestupná profesní mobilita a hodnocení pracovní spokojenosti

Obecně lze očekávat, že sestupná profesní mobilita je úzce spjata s nižší úrovní pracovní spokojenosti. [FELDMAN, 2010]. Na druhou stranu, sestupná mobilita může mít pozitivní vliv na hodnocení profesního života, a to zejména u pracovníků v důchodovém či předdůchodovém věku. U těchto osob dochází k proměně žebříčku životních hodnot, což se projevuje zejména poklesem kariérních aspirací, silnější preferencí nižšího pracovního vytížení a menšího pracovního stresu. Pro tento proces existují pojmy jako je „postupný odchod do důchodu“ či „bridge employment“ [DOERINGER, 1990] nebo „recareering“ [JOHNSON, 2009].

Jak již bylo dříve zmiňováno hlavním cílem této práce je odhalení vlivu sestupné mobility na hodnocení profesní spokojenosti u pracovníků v předdůchodovém věku, přičemž hodnocení profesní spokojenosti se odehrává na dvou úrovních. První úroveň je globální a vztahuje se k všeobecnému hodnocení profesní spokojenosti vzhledem k současnému zaměstnání. Druhá úroveň již jde do většího detailu a snaží se hodnotit profesní spokojenost v následujících aspektech, které se týkají:

- profesního uznání,
- míry samostatnosti či autonomie,
- intenzity pracovního vytížení.

Stranou ovšem není ponechána ani vzestupná mobilita a její vliv na hodnocení pracovní spokojenosti. Potřeba tohoto typu výzkumu byla identifikována již v práci autorů NG, SORENSEN, EBY, kteří se zabývali profesní mobilitou a jejími pozitivními a negativními důsledky [NG, 2007]. Další výzkumné otázky se tedy týkají vzestupné mobility a jejího vlivu na percepci profesního života a rozdílnosti vlivu profesní mobility na odlišné skupiny osob (třídění dle pohlaví a vzdělání).

Dostupná data

Oporou pro empirickou analýzu jsou tři datové zdroje, které jsou spojeny v jeden harmonizovaný dataset: Tyto datové zdroje pokrývají údaje o osobách v předdůchodovém věku ze Spojených států amerických, Velké Británie a Evropy. Zdrojem dat pro osoby z USA je výzkum Michiganské university s názvem Health and Retirement Study (HRS). Evropské země jsou zastoupeny díky výzkumu The Survey of Health, Ageing and Retirement in Europe (SHARE). Posledním zdrojem dat je britský datový archív a studie s názvem English Longitudinal Study of Ageing (ELSA).

Výše zmiňované datové zdroje jsou částečně harmonizovány díky The Program on Global Aging, Health, and Policy, avšak úplná harmonizace zejména proměnných týkajících se vzdělání, financí a profesního života není dostupná.

- The Survey of Health, Ageing and Retirement in Europe (SHARE)

  Tento výzkum obsahuje údaje přibližně o 123 tisících osobách z 20 evropských zemí a Izraele. Osoby jsou dotazovány v jednotlivých vlnách, přičemž první vlna byla realizována již v roce 2004. V současné době je k dispozici již vlna pátá, která byla ukončena koncem roku 2013.[SHARE, 2016]

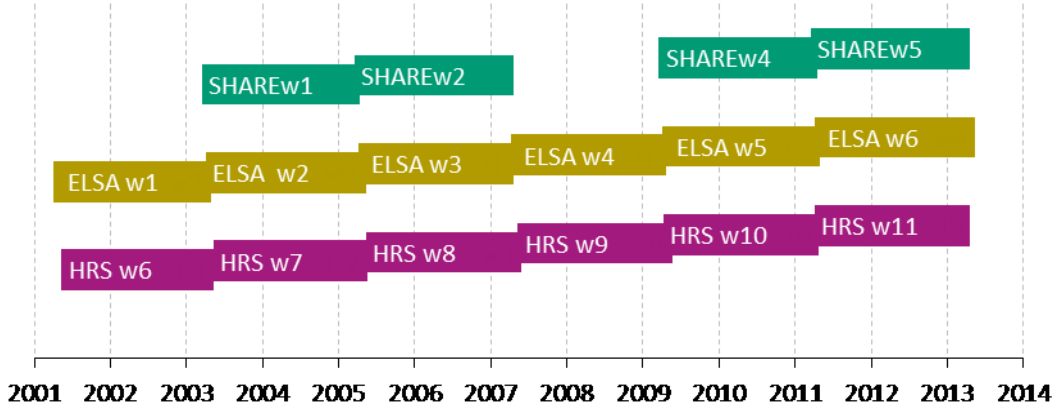
- English Longitudinal Study of Ageing (ELSA)

  Projekt ELSA sbírá údaje o osobách z Velké Británie již od roku 2001 a stejně jako SHARE i on se zaměřuje na osoby ve věku 50 a více let. V současnosti jsou k dispozici údaje o 20 tisících osobách za téměř 12 let.[ELSA, 2011]

- Health and Retirement Study (HRS)

  Sběr dat pro HRS probíhá již od roku 1992 pod záštitou National Institute on Aging. Jako oba výše zmiňované datové zdroje i HRS sbírá informace o přijmu, majetku, socio-ekonomických charakteristikách a zdravotním stavu osob starších 50 let.[HRS, 2016]

Pro zajištění srovnatelnosti údajů bylo pro empirickou analýzu vybráno posledních pět vln z každého datového zdroje. Graf č. 1 ukazuje jejich rozložení v čase.



Graf 1: Rozložení sběru dat v čase Zdroj: Vlastní

Vybrané proměnné a jejich harmonizace

Zvolené proměnné pro empirickou analýzu lze rozdělit do dvou skupin. První skupinu tvoří proměnné v čase neměnné, jako je například **pohlaví** a **země**, ve které jedinec žije. Jelikož se v drtivé většině jedná o osoby starší 50 let je do skupiny v čase neměnných proměnných zařazeno i **vzdělání**. Zatímco pro HRS a SHARE je dostupné vzdělání v počtu let, ELSA obsahuje pouze kategorie. Kvůli této skutečnosti byly vytvořeny následující kategorie, které umožnily harmonizaci a respektují logiku kumulativních kategorií ISCED: (0-10 let, 11 let, 12-15 let a 16 a více let). [ISCED, 2011]

Druhá skupina proměnných jsou proměnné v čase variabilní. Zde jsou zařazeny proměnné, jako jsou: **rodinný stav**, **ekonomicky status**, **změna zaměstnavatele**, **počet hodin odpracovaných za rok** a **hodinová mzda**. V případě hodinové mzdy je rovněž nutná harmonizace, a to zejména kvůli různým měnám v jednotlivých zemích. Z tohoto důvodu je zkonstruována nová proměnná, která je podílem hodinové mzdy daného člověka v dané vlně a mediánové mzdy v dané zemi a vlně. Takto zkonstruovaná proměnná, a zvláště její změna mezi vlnami, v sobě zahrnuje nejen změnu příjmu dané osoby, ale změnu příjmu všech osob v dané zemi. Z tohoto důvodu byla vytvořena ještě jedna proměnná určující procentuální změnu příjmu dané osoby ve dvou po sobě jdoucích vlnách.

Proměnné vztahující se k evaluaci profesní spokojenosti jsou taktéž součástí množiny v čase variabilních proměnných. Osoby byly nejprve dotazovány na jejich spokojenost se současným zaměstnáním a následně jim byly kladeny upřesňující otázky na spokojenost s finančním ohodnocením, pracovním uznání, možností dalšího rozvoje dovedností a kariérního růstu. Rovněž fyzická náročnost práce a pracovní vytížení byly předmětem zkoumání spokojenosti ekonomicky aktivních osob. Všechny otázky týkající se spokojenosti byly kladeny ve formě výroků, na které respondent mohl odpovídat tak, že vybral jednu možnost ze čtyřstupňové ordinální škály typu (určitě souhlasím, spíše souhlasím, spíše nesouhlasím a určitě nesouhlasím).

Poslední proměnnou, která patří do kategorie v čase variantních proměnných je **profesní status**. Zatímco SHARE poskytuje údaje o profesním statusu dle ISCO (International Standard Classification of Occupations), ELSA a HRS kódují profesní status dle SOC (Social Occupational Classification). Další komplikaci představuje skutečnost, že HRS reportuje profesní status pouze na nejvyšší, jednomístné, úrovni. Jelikož cílem práce je zkoumání vlivu profesní mobility a ne profesního statusu, vyvstává potřeba hierarchického uspořádání obou klasifikací (jak SOC 2000, tak ISCO). Pro data z ELSA a SHARE je možné využít k definici profesní mobility teoretického konceptu „bílých a modrých límečků (white & blue collars)“, který je používán například ve průzkumech pracovních podmínek v Evropě agentury Eurofound. [EUROFOUND, 2016]

Tento koncept definuje 4 skupiny, které mohou být hierarchicky uspořádatelé, a to takto:

1. High Skilled White Collar workers (HSWC),
2. Low Skilled White Collar workers (LSWC),
3. High Skilled Blue Collar workers (HSBC),
4. Low Skilled Blue Collar workers (LSBC).

Do těchto skupin je možné rozřadit jak ISCO-88 klasifikaci, tak i dvoumístnou klasifikaci SOC-2000. Toto rozřazení je k nalezení v PŘÍLOZE 1. Na základě definované kategorie dle konceptu bílých a modrých límečků je možné určit vzestupnou a sestupnou mobilitu. Klíč pro identifikaci profesní mobility je v Tabulce 1, přičemž jsou uvažovány pouze tranzice mezi po sobě jdoucími vlnami.

Tabulka 1: Definice vzestupné a sestupné mobility Zdroj: Vlastní

VZESTUPNÁ PROFESNÍ MOBILITA		SESTUPNÁ PROFESNÍ MOBILITA	
vlna $_t$	vlna $_{t+1}$	vlna $_t$	vlna $_{t+1}$
LSBC	HSWC	HSWC	LSWC
LSBC	LSWC	HSWC	HSBC
LSBC	HSBC	HSWC	LSBC
HSBC	HSWC	LSWC	HSBC
HSBC	LSWC	LSWC	LSBC
LSWC	HSWC	HSBC	LSBC

Bohužel v případě dat z HRS je k dispozici pouze jednomístná klasifikace SOC 2000, a tak není možno využít tohoto konceptu. Určitou možnost pro identifikaci profesní mobility v datovém souboru HRS představuje nalezení silně korelované proměnné se profesní mobilitou, která by posloužila jako proxy pro identifikaci profesní mobility. Na základě dostupné teorie se zdá být vhodnou proxy proměnnou k profesní mobilitě změna hodinové mzdy. Tato teorie se opírá o hypotézu, že pokud osoba zažije sestupnou mobilitu, pak i její hodinová mzda poklesne, a naopak, v případě vzestupné mobility by mělo docházet u osob k nárůstu hodinové mzdy. Tato hypotéza je testovatelná na datech z ELSA a SHARE. Jelikož změna hodinové mzdy i profesní mobilita jsou ordinální proměnné, je jejich závislost zjišťována pomocí měr asociace jako je Goodman&Kruskalova gamma a Kendallovo τb. [GOODMAN,1959]

Tabulka 2: Analýza závislosti profesní mobility a změny hodinové mzdy Zdroj:Vlastní

ELSA

PROFESNÍ MOBILITA	ZMĚNA HODINOVÉ MZDY			
	POKLES	BEZ ZMĚNY	VZESTUP	CELKEM
SESTUPNÁ MOBILITY	105	9	123	237
BEZ ZMĚNY	2,784	1,191	4,538	8,513
VZESTUPNÁ MOBILITA	82	12	106	200
CELKEM	2,971	1,212	4,767	8,950

Pearson chi2(4)	39.4733	Pr = 0.000
likelihood-ratio chi2(4)	47.2748	Pr = 0.000
Cramér's V	0.0470	
gamma	0.0278	ASE = 0.047
Kendall's tau-b	0.0064	ASE = 0.011
Fisher's exact		0.000

SHARE

PROFESNÍ MOBILITA	ZMĚNA HODINOVÉ MZDY			
	POKLES	BEZ ZMĚNY	VZESTUP	CELKEM
SESTUPNÁ MOBILITY	175	2	126	303
BEZ ZMĚNY	570	6	520	1,096
VZESTUPNÁ MOBILITA	118	3	121	242
CELKEM	863	11	767	1,641

Pearson chi2(4)	6.0909	Pr = 0.192
likelihood-ratio chi2(4)	5.8941	Pr = 0.207
Cramér's V	0.0431	
gamma	0.0991	ASE = 0.047
Kendall's tau-b	0.0498	ASE = 0.024
Fisher's exact		0.144

Ač dle výsledků z Tabulky 2 nelze na 5% hladině významnosti zamítnout hypotézu o nezávislosti změny hodinové mzdy a profesní mobility, dozajista hodnoty velmi blízké nule pro Goodman&Kruskalovu gammu a Kendallovo τb naznačují velmi slabou asociaci mezi těmito dvěma proměnnými.

Jelikož výsledky testů závislosti nejsou zcela jednoznačné, je zkonstruován ještě model, ve kterém je závislou proměnnou profesní mobilita a vysvětlujícími proměnnými jsou procentuální změna hodinové mzdy mezi po sobě jdoucími vlnami (**WoW_itearn_h**) a změna podílu hodinové mzdy a mediánu hodinové mzdy pro danou zemi a vlnu (**t_itearn_h_ind**). Model byl rovněž omezen pouze na osoby ekonomicky aktivní, které nebyly nezaměstnané ve dvou po sobě jdoucích vlnách, a u nichž bylo možno identifikovat profesní mobilitu. Výsledky

modelu jsou v Tabulce 3 a indikují, že ani jeden z regresorů nemá významnější vysvětlující význam pro model, a ani model jako celek není významný.

Tabulka 3: Výsledky regesního modelu Zdroj:Vlastní

```
Ordered logistic regression                     Number of obs   =       8485
                                                LR chi2(2)      =       1.08
                                                Prob > chi2     =     0.5818
Log likelihood = -3314.6801                     Pseudo R2       =     0.0002
------------------------------------------------------------------------------
PROFESNÍ
MOBILITA      |  Coef.     Std. Err.  z       P>|z|    [95% Conf. Interval]
--------------+---------------------------------------------------------------
t_itearn_h_ind|  .0041736  .0545912   0.08   0.939    -.1028231    .1111704
  WoW_itearn_h| -.0037999  .0032197  -1.18   0.238    -.0101103    .0025106
--------------+---------------------------------------------------------------
       /cut1  |-2.864845   .0481461                    -2.95921   -2.770481
       /cut2  | 3.057482   .0525497                    2.954486    3.160477
------------------------------------------------------------------------------
```

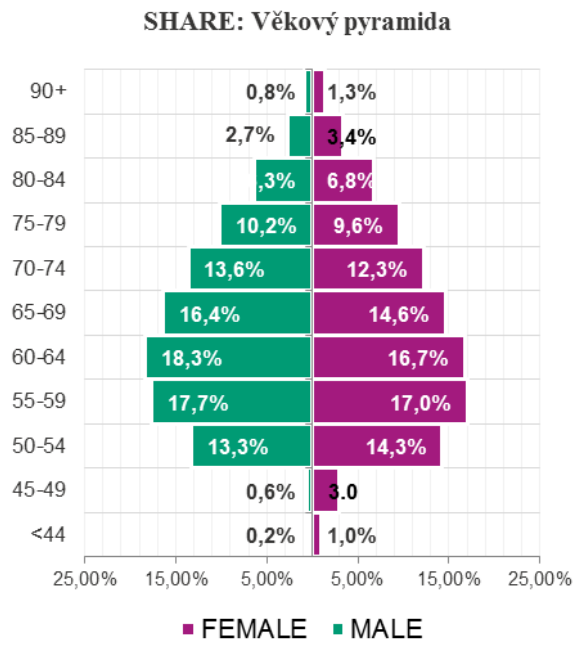
Na základě výše uvedených výsledků zkoumání vztahu mezi profesní mobilitou a změnou hodinové mzdy na datech z SHARE a ELSA není možné tvrdit, že by změna hodinové mzdy byla dobrou proxy proměnnou pro identifikaci profesní mobility. Z tohoto důvodu jsou data z HRS ponechána v další analýze stranou.

Po spojení ELSA a SHARE čítá datová soubor informace o cca 130 tisících osobách, přičemž poměr mužů a žen je přibližně 6:7. Drtivá většina osob je sezdána (65%). Dalšími dvěma největšími skupinami osob z hlediska rodinného stavu jsou osoby ovdovělé (15%) a osoby rozvedené či nikdy nesezdané (10%). Následující tabulka 4 ukazuje rozložení osob dle úrovně vzdělání. Z tabulky 4 je patrné, že největší skupinu osob tvoří lidé se středoškolským vzděláním.

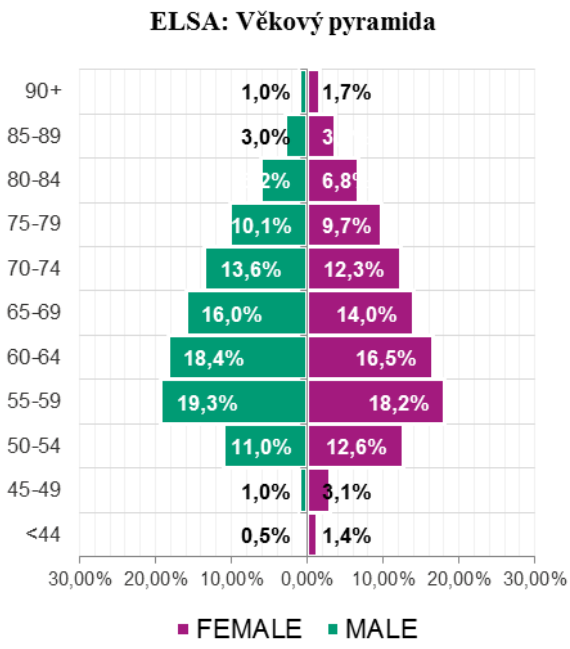
Tabulka 4: Struktura osob dle dosaženého vzdělání Zdroj:Vlastní

	Počet let vzdělávání	Proc.	Počet.
ELSA	0 - 10 (let)	28.2	5,072
	11 (let)	15.6	2,810
	12 - 15 (let)	28.9	5,195
	16 a více (let)	12.5	2,255
	chybějící údaje	14.7	2,649
	CELKEM	100.0	17,981

	Počet let vzdělávání	Proc.	Počet.
SHARE	0 - 10 (years)	44.6	48,600
	11 (years)	9.4	10,210
	12 - 15 (years)	30.7	33,506
	16 a více (years)	15.3	16,706
	chybějící údaje	0.0	23
	CELKEM	100.0	109,045



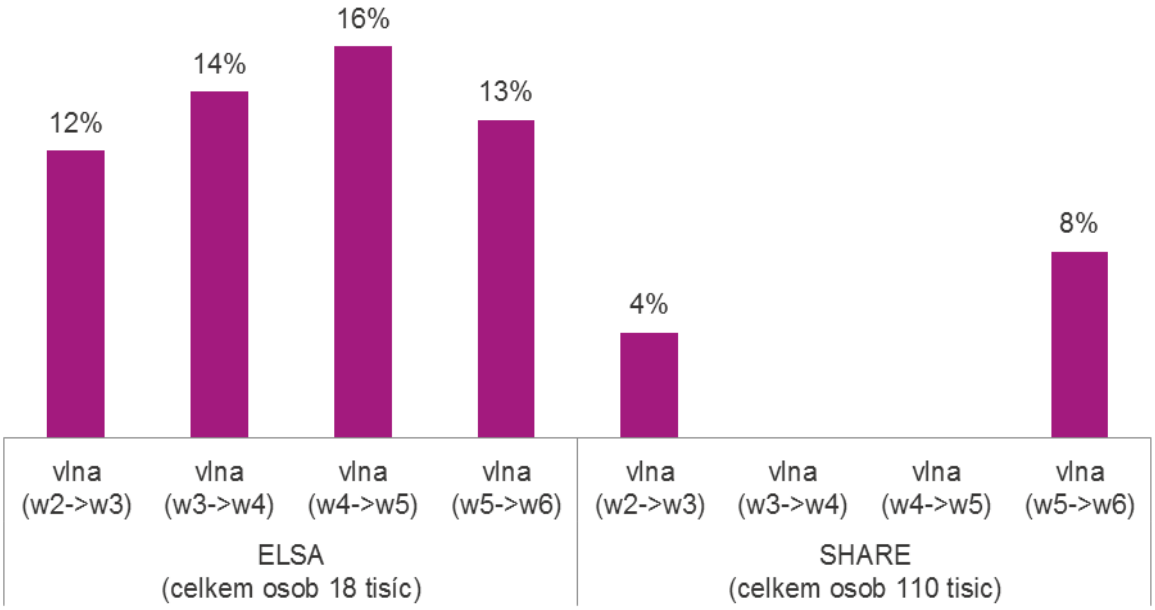
Graf 2: Věková pyramida Zdroj:Vlastní



Graf 3: Věková pyramida Zdroj:Vlastní

Jak již bylo patrné z počtu pozorování při ověřování vhodnosti změny hodinové mzdy jako proxy proměnné pro identifikaci profesní mobility, je počet ekonomicky aktivních osob relativně nízký, což je způsobeno zejména pokročilým věkem osob. Lepší představu o věkové struktuře osob poskytují následující věkové pyramidy.

Další prerekvizitou pro identifikaci profesní mobility je ekonomická aktivita osob. Rovněž u osob nezaměstnaných není možné určit profesní status a tudíž ani profesní mobilitu. Následující graf číslo 4 ukazuje procentuální zastoupení osob zaměstnaných a osob samostatně výdělečně činných napříč vlnami a datovými soubory.



Graf 2: Procentuální zastoupení ekonomicky aktivních osob Zdroj:Vlastní

Empirická analýza

V této předposlední části je nejdříve formulován model včetně vymezení všech proměnných a následně jsou zde k nalezení interpretované výsledky pro všechna hodnocení spokojenosti.

Specifikace modelu

Finální datový soubor byl po harmonizaci transformován do formy prvních diferencí, přičemž byly uvažovány pouze diference mezi dvěma po sobě jdoucími vlnami. Tento přístup umožnil identifikovat jak změnu ve spokojenosti osob, tak i profesní mobilitu. Navíc tento typ transformace je v souladu s principem eliminace fixních efektů jednotlivce pomocí metody first differences, která poskytuje konzistentní odhad při splnění požadavku striktní exogenity. [WOOLDRIDGE, 2010]

Dalším argumentem pro toto transformaci je skutečnost, že pro většinu osob je dostupná pouze jedna změna, a v takovém to případě transformace skrze first differences je totožná s modelem fixed effects (FE). Jelikož lze předpokládat, že nepozorovatelný individuální efekt jedince je korelovaný s použitými regresory, nelze předpokládat, že by RE model (random effects) poskytoval lepší odhady než model s fixními efekty. [WOOLDRIDGE, 2010]

Pro hodnocení změny spokojenosti osob je použita ordinální škála o třech hodnotách:

- (1) – zlepšení
- (0) – beze změny
- (-1) – zhoršení.

Pro modelování vlivu profesní mobility na změnu spokojenosti byl zvolen ordinální logit na transformovaných datech metodou prvních diferencí. Jako závislé proměnné byly popořadě užity následující změny hodnocení spokojenosti:

- *t_ep026*: změna celkové spokojenosti s prací,
- *t_ep028*: změna stresu způsobovaného prací,
- *t_ep029*: změna ve svobodě rozhodování,
- *t_ep032*: změna v uznání práce,

Sada vysvětlujících proměnných je pro všechny modelované změny totožná a obsahuje tyto proměnné:

- *DOM:* binární proměnná s hodnotou 1 pro případ sestupné mobility,
- *UOM:* binární proměnná s hodnotou 1 pro případ vzestupné mobility,
- *rT0agey*: věk osoby,
- *b_gender_M*: pohlaví – binární proměnná nabývající hodnoty 1 pro může,
- *b_edu_11, b_edu_12_15, b_edu_16* – sada binárních proměnných pro nominální proměnnou určující kategorii počtu let vzdělání,
- *t_jcten* – binární proměnná nabývající hodnoty 1 v případě změny zaměstnavatele,
- *t_itearn_h_ind* – standardizovaná změna hodinové mzdy (změna podílu hodinové mzdy a mediánu mzdy pro danou zemi mezi dvěma vlnami),
- *WoW_itearn_h* – procentuální změna příjmu mezi dvěma vlnami,
- *t_jhour_y_ind* – standardizovaná změna počtu odpracovaných hodin,
- *b_lbrf_emp_semp, b_lbrf_semp_semp, b_lbrf_semp_emp* – sada binárních proměnných pro změnu ekonomického statusu:
    - b_lbrf_emp_semp – nabývá hodnoty 1 pro osoby, jež se staly ze zaměstnanců osobami samostatně výdělečně činnými,
    - b_lbrf_semp_semp – nabývá hodnoty 1 pro osoby samostatně výdělečně činné,
    - _lbrf_semp_emp - nabývá hodnoty 1 pro osoby, jež se staly zaměstnanci a v předchozím období byly samostatně výdělečně činné.

Odhad modelů

Následující tabulka sumarizuje stěžejní výsledky z odhadnutých modelů vzhledem k vlivu profesní mobility na změnu hodnocení kvality profesního života. Znaménko plus obsažené v tabulce indikuje zlepšení hodnocené kategorie při výskytu sestupné či vzestupné mobility. Naopak znaménko mínus označuje zhoršení hodnocení. Přítomnost hvězdiček pak udává statistickou významnost.

Jako velmi významná vysvětlující proměnná se ukazuje i proměnná indikující změnu zaměstnavatele. Tato proměnná je signifikatní ve všech odhadnutých modelech, což poukazuje na pozitivní dopad změny pracovního prostředí na jeho hodnocení, a to i v tomto případě, kdy analýza zkoumá osoby v předdůchodovém věku.

Tabulka 5: Výsledky empirické analýzi **Zdroj:Vlastní**

	Změna celkové spokojenosti (H1)	Změna pracovního tlaku (H2)	Změna svobody rozhodování (H3)	Změna uznání odvedené práce (H4)
downward mobility	+	+ ***	+	- *
upward mobility	+ ***	+	+ ***	+ **
change of employer	+ ***	+ ***	+ **	+ ***

Hypotézu o pozitivním vlivu sestupné profesní mobility na celkovou spokojenost se bohužel nepodařilo na základě modelu prokázat. Rovněž statistický test závislosti dvou ordinálních proměnných není významný, což potvrzují i hodnoty koeficientů Goodman&Kruskalova gamma a Kendallovo τb blízké nule. Je-li odhlédnuto od statistické signifikance, lze usuzovat, že sestupná profesní mobility má spíše pozivitní vliv na hodnocení pracovního prostředí.

Za zmínění rovněž stojí i skutečnost, že vzestupná mobilita má signifikantně významný vliv na hodnocení profesní spokojenosti, a to ve všech zkoumaných aspektech, až na stres způsobovaný pracovním vytížením.

Dalším zajímavým zjištěním je skutečnost, že osoby starší mají větší tendenci hodnotit profesní život pozitivně.

Závěr

Pro analýzu sestupné profesní mobility a jejího vlivu na hodnocení profesního života a spokojenosti byla využita data z ELSA a SHARE. Bohužel datový zdroj HRS nemohl být použit, protože nebyla nalezena vhodná proxy proměnná, která by zastoupila profesní mobilitu.

Na základě výsledků ordinálního logitu lze odhadovat, že sestupná mobilita má pozitivní a signifikantní vliv pouze na na redukci stresu z práce. Na druhou stranu, osoby, které zažijí sestupnou mobilitu, ovšem reportují i nižší ocenění jejich práce.

Vzestupná mobilita jednoznačně přispívá k lepší celkové percepci práce. Jedinci se zkušeností se vzestupnou profesní mobilitou rovněž odpovídají, že roste jak jejich svoboda v rozhodování o jejich práci, tak i ocenění jejich práce.

Další významnou proměnnou, která přispívá k zlepšení hodnocení profesního života je zejména změna zaměstnavatele, takže i v případě osob předdůchodového věku lze hodnotit výměnu zaměstnavatele jako „změnu k lepšímu“. Proměnné pro pohlaví a úroveň vzdělání se ukázaly jako spíše nevýznamné regresory.

Literatura

GIDDENS, Anthony a Jan JAŘAB. *Sociologie*. Praha: Argo, 1999. ISBN 80-7203-124-4.

ŠANDEROVÁ, Jadwiga. *Sociální stratifikace: problém, vybrané teorie, výzkum*. Praha: Karolinum, 2000. ISBN 80-246-0025-0.

LAIN, D. Working past 65 in the UK and the USA: segregation into 'Lopaq' occupations? *Work, Employment & Society* [online]. 2012, 26(1), 78-94 [cit. 2016-05-23]. DOI: 10.1177/0950017011426312. ISSN 0950-0170. Dostupné z: http://wes.sagepub.com/cgi/doi/10.1177/0950017011426312

NG, THOMAS W. H. a DANIEL C. FELDMAN. THE RELATIONSHIPS OF AGE WITH JOB ATTITUDES: A META-ANALYSIS. *Personnel Psychology* [online]. 2010, 63(3), 677-718 [cit. 2016-05-28]. DOI: 10.1111/j.1744-6570.2010.01184.x. ISSN 00315826. Dostupné z: http://doi.wiley.com/10.1111/j.1744-6570.2010.01184.x

DOERINGER, Peter B. *Bridges to retirement: older workers in a changing labor market*. Ithaca, NY: ILR Press, School of Industrial and Labor Relations, Cornell University, c1990. ISBN 087546159X.

JOHNSON, Richard W, JEANETTE KAWACHI a ERIC K. LEWIS. (2009). Older Workers on the Move: Recareering in Later Life*, AARP Research Report*, pp. 1-55.

NG, Thomas W. H., Kelly L. SORENSEN, Lillian T. EBY a Daniel C. FELDMAN. Determinants of job mobility: A theoretical integration and extension. *Journal of Occupational and Organizational Psychology*[online]. 2007, 80(3), 363-386 [cit. 2016-05-28]. DOI: 10.1348/096317906X130582. ISSN 09631798. Dostupné z: http://doi.wiley.com/10.1348/096317906X130582

ELSA - English Longitudinal Study of Ageing [online]. London: The Institute for Fiscal Studies, 2011 [cit. 2016-05-29]. Dostupné z: http://www.elsa-project.ac.uk/

SHARE - Survey of Health, Ageing and Retirement in Europe [online]. Munich: Munich Center for the Economics of Aging (MEA), 2016 [cit. 2016-05-29]. Dostupné z: http://www.share-project.org/

HRS - Health and Retirement Study [online]. Ann Arbor: Survey Research Center, 2016 [cit. 2016-05-29]. Dostupné z: http://hrsonline.isr.umich.edu/

ISCED - UNESCO INSTITUTE FOR STATISTICS. *International standard classification of education*: ISCED 2011. Montreal, Quebec: UNESCO Institute for Statistics, 2012. ISBN 9789291891238.

EUROFOUND - European Foundation for the Improvement of Living and Working Conditions [online]. Dublin: Eurofound, 2016 [cit. 2016-05-30].

GOODMAN, Leo A. a William H. KRUSKAL. Measures of Association for Cross Classifications. II: Further Discussion and References. *Journal of the American Statistical Association* [online]. 1959, 54(285), 123-163 [cit. 2016-05-30]. DOI: 10.1080/01621459.1959.10501503. ISSN 0162-1459. Dostupné z: http://www.tandfonline.com/doi/abs/10.1080/01621459.1959.10501503

MCCULLAGH, Peter. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1980, **42**(2), 109-142.

OLOGIT - Ordered logistic regression. *MANUAL STATA* [online]. STATA, 2016 [cit. 2016-05-31]. Dostupné z: http://www.stata.com/manuals13/rologit.pdf

WOOLDRIDGE, Jeffrey M. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, MA: MIT Press, c2010. ISBN 978-0-262-23258-6.

JEL Klasifikace: J620, M51

Summary

Occupational mobility and its impact on work life satisfaction

Hierarchical occupational mobility, concretely downward (DOM) and upward (UOM) occupational mobility, and its impact on old workers and theirs job conditions and overall job satisfaction is the middle point of this paper. Estimation of the impact of hierarchical transition is done on dataset which was created with the help of merging and harmonizing three data source: The Survey of Health, Ageing and Retirement in Europe (SHARE), English Longitudinal Study of Ageing (ELSA), Health and Retirement Study (HRS). Discrete choice models with first differences correction for reducing fixed effects were used for impact estimation. Results show that person experienced with downward mobility reports significantly less work stress. On another hand, upward occupational mobility seems to have positive impact on the majority of working life aspects.

Keywords: occupational mobility, models of discrete choice, aging, marginalization.

Příloha 1

KATEGORIE	SOC 2000 - (2 digits)		ISCO-88 (1 digits)	
high skilled white collar workers	11	Corporate Managers	1	Legislators, senior officials, managers
	12	Managers and Proprietors in Agriculture and Services	2	Professionals
	21	Science and Technology Professionals	3	Technicians and associate professionals
	22	Health Professionals		
	23	Teaching and Research Professionals		
	24	Business and Public Service Professionals		
	31	Science and Technology Associate Professionals		
	32	Health and Social Welfare Associate Professionals		
	33	Protective Service Occupations		
	34	Culture, Media and Sports Occupations		
	35	Business and Public Service Associate Professionals		
low skilled white collar workers	41	Administrative Occupations	4	Clerks
	42	Secretarial and Related Occupations	5	Service workers and shop and market sales workers
	61	Caring Personal Service Occupations		
	62	Leisure and Other Personal Service Occupations		
	71	Sales Occupations		
	72	Customer Service Occupations		
high skilled blue collar workers	51	Skilled Agricultural Trades	6	Skilled agricultural and fishery workers
	52	Skilled Metal and Electrical Trades	7	Craft and related trades workers
	53	Skilled Construction and Building Trades		
	54	Textiles, Printing and Other Skilled Trades		
low skilled blue collar workers	81	Process, Plant and Machine Operatives	8	Plant and machine operators and assemblers
	82	Transport and Mobile Machine Drivers and Operatives	9	Elementary occupations
	91	Elementary Trades, Plant and Storage Related Occupations		
	92	Elementary Administration and Service Occupations		

Příloha 2

T_EP026: změna celkové spokojenosti s prací:

```
Ordered logistic regression                       Number of obs   =       5101
                                                  LR chi2(14)     =      33.38
                                                  Prob > chi2     =     0.0025
Log likelihood = -4657.9373                       Pseudo R2       =     0.0036
------------------------------------------------------------------------------
     t_ep026 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         DOM |   .1513926   .1224011     1.24   0.216    -.0885091    .3912943
         UOM |   .4022721   .1337518     3.01   0.003     .1401234    .6644208
     t_jcten |   .2781719   .0940567     2.96   0.003     .0938241    .4625197
     rT0agey |   .0071678   .0055439     1.29   0.196    -.0036981    .0180337
  b_gender_M |  -.0312615   .0577996    -0.54   0.589    -.1445466    .0820236
t_itea~h_ind |   .0281009   .0477251     0.59   0.556    -.0654385    .1216403
WoW_itearn_h |  -.0001465   .0033605    -0.04   0.965    -.0067328    .0064399
t_jhour_y~nd |  -.1507502   .1206911    -1.25   0.212    -.3873005       .0858
    b_edu_11 |  -.1381306   .1194227    -1.16   0.247    -.3721948    .0959337
 b_edu_12_15 |   .0375279   .0722558     0.52   0.603    -.1040909    .1791467
    b_edu_16 |   -.011733   .0839879    -0.14   0.889    -.1763462    .1528803
b_~_emp_semp |   .1910795   .2307527     0.83   0.408    -.2611875    .6433465
b_~semp_semp |   .1268327   .1608849     0.79   0.430    -.1884959    .4421612
b_l~semp_emp |   .0214204   .2699524     0.08   0.937    -.5076765    .5505174
-------------+----------------------------------------------------------------
       /cut1 |  -.9310947    .321018                     -1.560278    -.301911
       /cut2 |   2.045583   .3224277                      1.413637     2.67753
------------------------------------------------------------------------------
```

T_EP028: změna hodnocení pracovního tlaku

```
Ordered logistic regression                       Number of obs   =       5101
                                                  LR chi2(14)     =     128.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -5216.2169                       Pseudo R2       =     0.0122
------------------------------------------------------------------------------
     t_ep028 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         DOM |   .4466568   .1155191     3.87   0.000     .2202435      .67307
         UOM |   .1492493   .1262092     1.18   0.237    -.0981162    .3966147
     t_jcten |   .3451308   .0888891     3.88   0.000     .1709114    .5193503
     rT0agey |   .0114314   .0053055     2.15   0.031     .0010329    .0218299
  b_gender_M |   .0606815   .0543987     1.12   0.265    -.0459379     .167301
t_itea~h_ind |  -.0397319   .0445944    -0.89   0.373    -.1271354    .0476715
WoW_itearn_h |   .0006877   .0030629     0.22   0.822    -.0053155    .0066909
t_jhour_y~nd |  -.8997538   .1170184    -7.69   0.000    -1.129106   -.6704019
    b_edu_11 |   .0252861   .1149423     0.22   0.826    -.1999966    .2505689
 b_edu_12_15 |   .0059496   .0681008     0.09   0.930    -.1275256    .1394249
    b_edu_16 |   .0567161   .0790067     0.72   0.473    -.0981342    .2115664
b_~_emp_semp |   .2795332   .2246292     1.24   0.213    -.1607319    .7197984
b_~semp_semp |  -.0731727   .1597091    -0.46   0.647    -.3861967    .2398514
b_l~semp_emp |   .1683954   .2710581     0.62   0.534    -.3628688    .6996595
-------------+----------------------------------------------------------------
       /cut1 |  -.4261005   .3065176                     -1.026864     .174663
       /cut2 |  1. 846624   . 3077753                      1.243396    2.449853
```

T_EP029: změna hodnocení svobody rozhodovat si o práci samostatně

```
Ordered logistic regression                     Number of obs   =       5101
                                                LR chi2(14)     =      31.20
                                                Prob > chi2     =     0.0052
Log likelihood = -5116.3057                     Pseudo R2       =     0.0030
------------------------------------------------------------------------------
     t_ep029 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         DOM |   .0845531    .115601     0.73   0.465    -.1420206    .3111268
         UOM |   .3741605   .1295129     2.89   0.004       .12032     .628001
     t_jcten |   .1897422   .0899733     2.11   0.035     .0133977    .3660867
     rT0agey |   .0035703   .0053852     0.66   0.507    -.0069845    .0141252
  b_gender_M |   .0522393   .0550755     0.95   0.343    -.0557066    .1601852
t_itea~h_ind |   .0158719   .0457883     0.35   0.729    -.0738715    .1056152
WoW_itearn_h |   .0020155   .0031756     0.63   0.526    -.0042087    .0082396
t_jhour_y~nd |  -.0341432   .1137085    -0.30   0.764    -.2570078    .1887213
    b_edu_11 |   -.227454   .1168804    -1.95   0.052    -.4565354    .0016274
 b_edu_12_15 |  -.0324826   .0693235    -0.47   0.639    -.1683541    .1033889
    b_edu_16 |   .0582927   .0800559     0.73   0.467     -.098614    .2151994
b_~_emp_semp |   .1059267   .2183948     0.49   0.628    -.3221193    .5339728
b_~semp_semp |  -.0902711   .1575536    -0.57   0.567    -.3990704    .2185283
b_l~semp_emp |  -.7377424   .2585893    -2.85   0.004    -1.244568   -.2309168
-------------+----------------------------------------------------------------
       /cut1 |  -.9490644   .3118057                     -1.560192   -.3379365
       /cut2 |   1.497331   .3123409                      .8851543    2.109508
------------------------------------------------------------------------------
```

T_EP032: změna hodnocení uznání za odvedenou práci

```
Ordered logistic regression                     Number of obs   =       5079
                                                LR chi2(14)     =      41.00
                                                Prob > chi2     =     0.0002
Log likelihood = -4959.9789                     Pseudo R2       =     0.0041
------------------------------------------------------------------------------
     t_ep032 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         DOM |  -.2022366   .1187877    -1.70   0.089    -.4350561     .030583
         UOM |   .2739874   .1303363     2.10   0.036      .018533    .5294417
     t_jcten |   .4433762   .0922706     4.81   0.000     .2625292    .6242233
     rT0agey |   .0046487   .0054723     0.85   0.396    -.0060769    .0153742
  b_gender_M |   .0032896   .0559681     0.06   0.953    -.1064059    .1129851
t_itea~h_ind |   .0453922   .0462121     0.98   0.326    -.0451818    .1359662
WoW_itearn_h |   .0004926   .0031971     0.15   0.878    -.0057735    .0067587
t_jhour_y~nd |  -.1772344   .1193111    -1.49   0.137    -.4110799    .0566112
    b_edu_11 |  -.0656324   .1165542    -0.56   0.573    -.2940743    .1628095
 b_edu_12_15 |    .027412   .0703825     0.39   0.697    -.1105351    .1653591
    b_edu_16 |   .0552415   .0812637     0.68   0.497    -.1040324    .2145154
b_~_emp_semp |   .0421999   .2259337     0.19   0.852     -.400622    .4850218
b_~semp_semp |   .0283107   .1573612     0.18   0.857    -.2801115    .3367329
b_l~semp_emp |  -.2546449   .2683406    -0.95   0.343    -.7805828    .2712929
-------------+----------------------------------------------------------------
       /cut1 |  -.9746041   .3168818                     -1.595681   -.3535271
       /cut2 |   1.637361   .3175701                      1.014935    2.259787
------------------------------------------------------------------------------
```

Migrační modely

Daniela Krbcová

daniela.krbcova@vse.cz

Doktorand oboru statistika

Školitel: doc. Ing. Ladislav Průša, CSc., (ladislav.prusa@vups.cz)

Abstrakt: Článek si klade za cíl představení základních problémů modelování migrace, seznámení s problematikou nejistoty v datech a představení bayesovského přístupu v modelování migrace. V analytické části je potom nastíněn postup výpočtu intenzity emigrace a předpovědi tohoto ukazatele do budoucna.

Klíčová slova: migrace, modely, bayesovský přístup, emigrace

Úvod

Migrační problematika je velmi důležitá nejen pro politiku každého státu. Většina vyspělých zemí se potýká s problémy stárnutí populace, které jsou způsobeny v důsledku snižování plodnosti u žen v reprodukčním věku a snižování úmrtnosti v nejvyšších věkových skupinách. Do budoucna to znamená nejen úbytek obyvatelstva, ale hlavně problém s financováním důchodového systému, jelikož nemusí být dostatek ekonomicky aktivních obyvatel, kteří by odváděli peněžní prostředky ze svých výdělků. Jedním z možných řešení, může být nárůst počtu ekonomicky aktivních obyvatel doplněných z řad migrantů. Imigranti, kteří se do přijímající země stěhují, s cílem nalézt pracovní uplatnění, bývají lidé v produktivním věku, mladí a bez větších závazků. Průměrný věk migrantů bývá nižší než průměrný věk obyvatel v zemi, do které přichází (Josifidis a spol., 2013). Tím snižují průměrný věk obyvatel celé země a omlazují tak populaci. U ekonomických migrantů se však může vyskytovat jev, který není pro ekonomiku hostitelské země žádoucí. Setkáváme se u nich buď s prací na černo, čímž minimalizují odvody do daňového systému nebo odvádějí své příjmy v podobě remitencí (většinu z toho, co si v hostitelské zemi vydělají nebo jinak získají, posílají do své země rodinám). Často se také stává, pokud jsou přistěhovalci chudší než stávající obyvatelstvo, mohou chtít profitovat ze štědrého sociálního systému (Huber a spol., 2014). Takový sociální systém by pak v zemi měl být vhodně nastaven, aby nemohl být zneužíván. Dalším jevem, který bývá často pozorován u migrantů, je skutečnost, že si často uvědomují nové role a mění svoje individuální preference (například Mexičani v USA) a čím jsou v zemi déle, tím více se jejich návyky přibližují návykům v hostitelské zemi (úprava chování plodnosti apod.) (Battaglia, 2015).

Existuje mnoho teorií, které se zabývají migrací, každá má jiný pohled i přístup. Jedna zkoumá migraci z hlediska psychologického, sociologického nebo ekonomického. Jedna teorie se zabývá pohledem jedince (jeho přínosy, obětované příležitosti apod.), jiná teorie zkoumá přínosy pro společnost jako celek. Někteří autoři se pak snažili migraci jen popsat, jiní se pokoušeli vytvořit model, který by ji co nejlépe vystihoval. Ovšem modelování migrace má svá úskalí, protože data o počtu migrantů, která často do modelů vstupují, jsou nespolehlivá, nejednoznačná, nejasná a v čase se měnící definice toho, kdo je za migranta považován. To má za následky nestabilní odhady a těžko předvídatelné předpovědi s velkými zkresleními. Nepříznivé pro migrační odhady je také skutečnost, že závisí na množství velmi špatně předvídatelných faktorů (politické vlivy, sociologické faktory, ekonomická situace, životní prostředí atd.). Avšak i přesto, že se mnoho autorů pokusilo o kvantifikaci tohoto jevu, použití vhodných migračních modelů je velmi vzácné a postrádá komplexnost.

Ukazuje se, že jednoduché odhady nulové nebo konstantní migrace vedou k mnohem větším předpovědním chybám než sofistikovanější metody (Bijak, 2011). Mnozí autoři se tak pokoušejí modelovat migraci pomocí dekompozice (rozklad na základě nějakých konkrétních vlastností migrantů), častější však je, že jsou migrační toky modelovány a předpovídány na základě analýzy časových řad (ARIMA modely, VAR modely), např. Beer, Alho, Keilman (Bijak, 2011).

Abychom mohli získávat co nejméně zkreslené odhady, je důležité mít co nejspolehlivější data na vstupu, resp. co nejvhodněji naložit s jejich nepřesností a snažit se využít postupů, které by dávaly spolehlivější výsledky.

Bayesovský přístup k modelování migrace

Z důvodu stále se zvyšující poptávky nejrůznějších statistik demografických dat členěných na malé územní celky, např. tabulky života za malé územní celky, plodnosti etnických skupin, apod. čelí demografové neúplnosti, řídkosti, nespolehlivosti dat. I přesto, že se zdá bayesovská statistika jedním z nejvhodnějších nástrojů k řešení těchto problémů, statistické úřady, agentury se jí brání z důvodu určité míry subjektivity.

Bayesovský přístup je metoda, která umožňuje kombinovat subjektivní prvky (vlastní představy, expertní názory, zkušenosti, intuice) se vzorkem aktuálně dostupných dat (napozorovaných, naměřených). Nikoliv na celém možném výskytu událostí, které se mohly vyskytnout, ale nevyskytly jako je tomu v případě frekventistického přístupu (Bolstad, 2007). Tento přístup je proto vhodný zejména v sociálních vědách, kde není možné výsledky experimentu, pokusu opakovat.

Právě subjektivita tohoto přístupu je často mnohými odpůrci kritizována. Níže uvedený vzorec představuje základní myšlenku bayesovského přístupu.

$$p(\theta|x) = \frac{p(\theta) \times p(x|\theta)}{p(x)},$$

kde

$p(\theta|x)$ nazýváme posteriorní rozdělení, které je součinem pravděpodobnosti apriorního rozdělení $p(\theta)$ (představa o rozdělení před pořízením dat) a naměřených dat $p(x|\theta)$. Jmenovatel *p (x) je* tzv. vyrovnávací konstanta (suma nebo integrál přes všechny hodnoty, které mohou nastat) (Gelman, 2014), (Bolstad, 2007). Apriornímu rozdělení přikládáme vyšší váhu v případě, že máme data nespolehlivá nebo jich není dostatek. S narůstajícím počtem dat, klesá vliv prioru. V případě, že nemáme téměř žádnou prvotní představu o parametru a zvolíme např. rovnoměrné rozdělení nebo konstantu, pak hovoříme o tzv. neinformativním prioru. Výsledky jsou v tomto případě totožné jako z klasického přístupu statistiky.

Problematika nejistých odhadů

Rozlišuje se sedm důvodů, proč jsou výsledky populačních projekcí nespolehlivé (vykazují předpovědní chyby). Tři z nich jsou založené na měřících problémech, chybách (špatně rozpoznaný trend v datech, zaokrouhlovací chyby, chyby způsobené „skoky“ v datech). Dalším samostatným důvodem náhodnost parametrů prognostického modelu a polední tři jsou způsobené předpovědí exogenních veličin, náhlou a nečekanou změnou v parametru v budoucnu, která může způsobit diskontinuitu v trendu nebo špatně specifikovaným modelem. Možnosti, jak se s takovými chybami vypořádat, jsou následující (Bijak, 2011):

- Ignorovat nepřesnosti a modelovat jednoduchý deterministický model projekce
- Modelovat scénáře s více variantami projekce (vysoký, střední, nízký scénář) – často používané statistickými úřady
- Aplikovat stochastický přístupu a vyjádřit nepřesnost pomocí pravděpodobností.

Použitá metodologie

Uvažujeme intenzitu emigrace I_e (hrubá míra emigrace) jako podíl počtu vystěhovalých v roce *t* na tisíc obyvatel středního stavu v roce *t* (Roubíček, 1997).

$$I_e = \frac{E_t}{\bar{S}_t} \times 1000$$

kde

E_t je počet emigrantů v roce *t*,

$\bar{S}_t$ střední stav obyvatelstva v roce *t*.

Vzhledem k tomu, že časová řada je krátká (12 pozorování), nemá smysl použití složitějších modelů s více parametry (například ARIMA modelů), protože by neumožňovaly příliš smysluplné odhady. Podle Bijaka (Bijak, 2007) nejvyšší posteriorní pravděpodobnost dávají jednodušší modely a na základě Occamova pravidla (tzv. Occams´razor principle) je v případě podobných výsledků preferován model s nižším počtem parametrů. V tomto článku jsou na základě pravidla o nižším počtu parametrů vybrány dva modely: model (M_1) autoregresního procesu řádu 1 (AR 1) a model náhodné procházky s konstantou (model M_2). U chybové složky ε_t se předpokládá, že má charakter bílého šumu. Do obou modelů vstupuje logaritmus hodnot intenzity emigrace $y_t = \log(I_e)$.

$$M_1 : y_t = \beta_{10} + \beta_{11} y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid\ N(0, \tau_1^{-1}),$$

$M_2 : y_t = \beta_2 + y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid\ N(0, \tau_2^{-1}),$

kde sledovaný parametr $\boldsymbol{\theta_1}\,|M_1 = [\boldsymbol{\beta_1^T}, \tau_1]^T$ a složka $\boldsymbol{\beta_1} = [\beta_{10}, \beta_{11}]^T$.

Pro parametr druhého modelu $\boldsymbol{\theta_2}\,|M_2 = (\beta_2, \tau_2)$.

Oba modely mají věrohodnostní funkci *L* definovanou

$$L(\boldsymbol{\beta_1}, \boldsymbol{X_1}, \boldsymbol{y}|M_1) = (2\pi)^{-(N-1)/2}\ \tau_1^{(N-1)/2} \exp[-\tau_1/2(\boldsymbol{y} - \boldsymbol{X_1\beta_1})^T(\boldsymbol{y} - \boldsymbol{X_1\beta_1})],$$

$$L(\beta_2, \boldsymbol{x_2}, \boldsymbol{y}|M_2) = (2\pi)^{-(N-1)/2}\ \tau_2^{(N-1)/2} \exp[-\tau_2/2(\Delta\boldsymbol{y} - \boldsymbol{x_2}\beta_2)^T(\Delta\boldsymbol{y} - \boldsymbol{x_2}\beta_2)],$$

kde

$\boldsymbol{y}$ je datový vektor hodnot $(\mathrm{T}-1)$,

$\boldsymbol{X_1}$ je matice $(\mathrm{T}-1) \times 2$ obsahující jedničky a první diference a

$\boldsymbol{x_2}$ je (**T−1**) vektor samých jedniček.

Jako apriorní rozdělení předpokládáme normální rozdělení, u parametru β_1 modelu M_1 pak $\beta_1 \sim N(\boldsymbol{b_{11}}, \boldsymbol{\Sigma_{11}^{-1}})$ $s\ \boldsymbol{b_{11}} = [0, 0.5].^T$ $\boldsymbol{\Sigma_{11}}$ je kovarianční matice, jež má na diagonále 1 000 a 4, mimo diagonálu samé nuly.

Pro druhý model předpokládáme, že $\beta_1 \sim \mathrm{N}(\mathrm{b}_{22}, \sigma_{22}^2,)$, kde $\sigma_{22}^2 = 1\,000.$ Pro parametr τ (tzv. precision parametr, inverzní rozptyl) se předpokládá Gamma rozdělení $\tau_1 \sim \Gamma(a_{11}, s_{11})$a pro druhý model $\tau_2 \sim \Gamma(a_{22}, s_{22})$.

Přitom platí, že $a_{11} = s_{11} = a_{22} = s_{22} = 1.$ Dále je možné definovat hyper-parametry (parametry prioru, zde se jedná v případě prvního modelu o $\mathbf{b_{11}}$ a v případě druhého modelu σ_{22}^2).

Výsledné posteriorní rozdělení nemá formu žádného známého rozdělení, proto je nutné simulovat hodnoty z rozdělení plně podmíněného pomocí Gibbsova vzorkovače (Bijak, 2014), (Gelman,2007). Plně podmíněné rozdělení pro oba modely je pro model M_1:

$$p(\boldsymbol{\beta_1}|\tau_1; \boldsymbol{X_1}, \boldsymbol{y}, M_1) \sim N(\boldsymbol{\Omega_1^{-1}}(\boldsymbol{\Sigma_{11}^{-1}b_{11}} + \tau_1\boldsymbol{X_1^T y}), \boldsymbol{\Omega_1^{-1}}),$$

$$p(\tau_1|\boldsymbol{\beta_1}; \boldsymbol{X_1}, \boldsymbol{y_1}, M_1) \sim \Gamma\left(\frac{T-1}{2} + a_{11}, [0.5(\boldsymbol{y} - \boldsymbol{X_1\beta_1})^T(\boldsymbol{y} - \boldsymbol{X_1\beta_1}) + 1/s_{11}]^{-1}\right),$$

kde

$$\boldsymbol{\Omega_1} = (\boldsymbol{\Sigma_{11}^{-1}} + \tau_1\boldsymbol{X_1^T}\,\boldsymbol{X_1}).$$

Obdobně pak pro model M_2:

$$p(\boldsymbol{\beta_2}|\tau_2; \boldsymbol{x_2}, \boldsymbol{y}, M_2) \sim N(\omega_2^{-1}(\sigma_2^{-2}b_2 + \tau_2\boldsymbol{x_2^T}\Delta\boldsymbol{y}), \omega_2^{-1}),$$

$$p(\tau_2|\beta_2; \boldsymbol{x_2}, \boldsymbol{y_1}, M_2) \sim \Gamma(0.5(T-1) + a_{22}, [0.5(\Delta\boldsymbol{y} - \boldsymbol{x_2}\beta_2)^T(\Delta\boldsymbol{y} - \boldsymbol{x_2}\beta_2) + 1/s_{22}]^{-1}),$$

kde

$\omega_2 = [\sigma_2^{-2} + \tau_2(T-1)]$, druhý prvek v tomto součtu je jednotkový, protože $\boldsymbol{x_2}$ je vektor samých jedniček.

Předpovědi y_p (v bayesovském pojetí) jsou získávány pomocí výběru z posteriorního rozdělení.

$$p\left(y_p\left|\boldsymbol{\theta_1}, y_{p-1}, M_1\right.\right) = (2\pi)^{-0.5}\ \tau_1^{0.5} \exp[-\tau_1/2(y_p - \beta_{10} - y_{p-1}\beta_{11})^2],$$

kde

$$p = T + 1, \ldots . T + P.$$

$$p\left(y_p \left| \boldsymbol{\theta}_2, y_{p-1}, M_2 \right.\right) = (2\pi)^{-0.5}\ \tau_2^{0.5} \exp\left[-\tau_2/2(\Delta y_p - \beta_2)^2\right],$$

kde

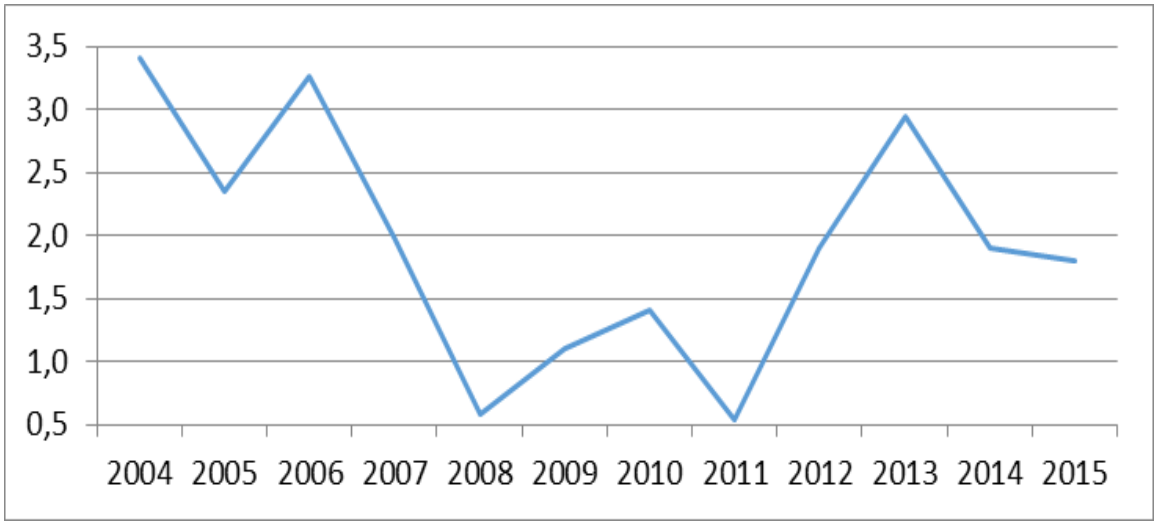
$$p = T + 1, \ldots . T + P.$$

Modelování intenzity emigrace v ČR

Cílem této práce je představení možnosti odhadu posteriorního rozdělení a předpovědí do budoucna u ukazatele, který měří intenzitu emigrace. Problematika modelování migrace je mnohem složitější a vyžaduje zavedení dalších metod, které budou v budoucnu představeny.

K analýze byla použita data vystěhovalých z České republiky v letech 2004 až 2015, jejíž hodnoty jsou vztaženy vždy ke střednímu stavu obyvatel daného roku a přepočítány na 1000 obyvatel. Emigrace bývá často ovlivňována stavem ekonomik v dané zemi a v zemích sousedních. Vývoj tohoto ukazatele v čase je možné shlédnout na Obrázku 1. Do výše definovaných modelů M_1 a M_2 pak vstupuje časová řada logaritmu tohoto ukazatele.

Obrázek 7: Intenzita emigrace v ČR v letech 2004 - 2015 na 1000 obyvatel středního stavu



Zdroj: ČSÚ, vlastní zpracování

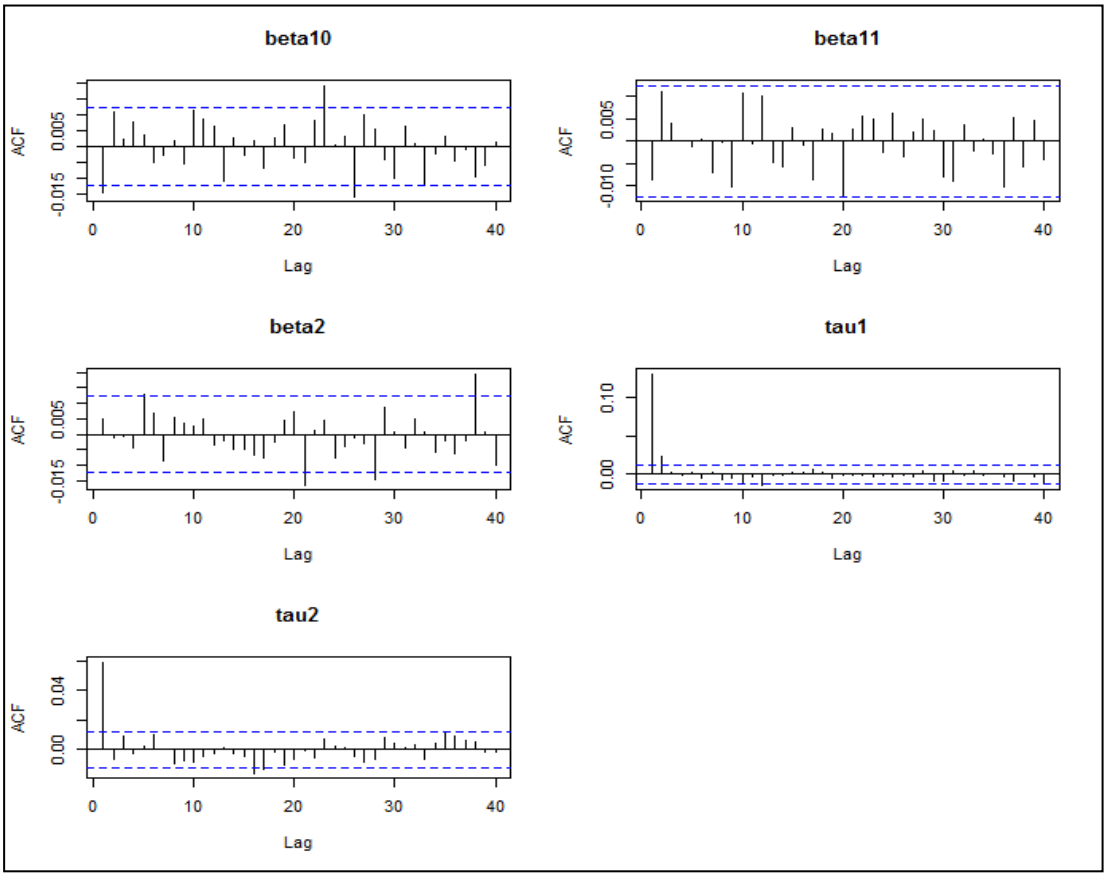
K tomu, aby bylo možné získat odhady hodnot jednotlivých parametrů β_{11}, β_{12} posteriorního rozdělení, jež nelze jednoduše spočítat, jsou jeho odhady získané na základě simulace hodnot z plně podmíněného pravděpodobnostního rozdělení. Tento postup je proveden nejprve pro model M_1 a následně pro model M_2 pomocí statistického programu R. Gibbsův vzorkovač (Bijak, 2011), (Gelman, 2007) si tedy zakládá na simulaci hodnot z předem daného rozdělení.

V první fázi simulace je nutné zvolit jakékoliv libovolné počáteční hodnoty (tzv. initial values), což mohou být například samé nuly jako v níže zmíněném příkladu. Počet simulací byl stanoven na 100 000 s burn-in vzorkem 75 000 (hodnoty, které se nezapočítávají kvůli ustálení odhadu). K vyhodnocení odhadů jednotlivých parametrů posteriorního rozdělení proto zůstává vzorek s počtem 25 000 hodnot. Statistiky posteriorního rozdělení (průměr, medián, kvantily) jsou vyobrazeny v následující tabulce (viz Tabulka 1). Pro kontrolu byla zavedena i autokorelací funkce jednotlivých parametrů viz Obrázek 2. Autokorelace by mohla být shledána u parametrů τ_1 a τ_2, ale vzhledem k tomu, že funkce křížových korelací nepotvrdila korelace ve vzorku, není důvod se tímto zaobírat. V Tabulce 2 jsou provedeny předpovědi (hodnoty byly zlogaritmované, jedná se o zpětný převod umocněním). Model M_2 se zdá být z hlediska jeho průměrné hodnoty reálnější odhadem, avšak s větším rozptylem. V případě konstrukce více modelů je možné kombinovat odhady jednotlivých modelů, což může někdy zlepšit předpovědi. Jednou z možností je i využít výsledky odhadů parametrů jako prior pro další analýzu. Nejrůznější možnosti využití jsou popsány v rozsáhlém množství literatury.

Tabulka 7: Statistiky posteriorního rozdělení jednotlivých parametrů

	průměr	rozptyl	q10%	q25%	medián	q75%	q90%
beta10	0,3085	0,0706	-0,0211	0,1409	0,3077	0,4757	0,6387
beta11	0,2817	0,1117	-0,0211	0,1409	0,3077	0,4757	0,6387
beta2	-0,0561	0,0680	-0,3763	-0,2208	-0,0568	0,1078	0,2648
tau1	2,6196	1,2360	1,3240	1,8162	2,4648	3,2532	4,1114
tau2	1,6368	0,4510	0,8551	1,1504	1,5442	2,0250	2,5302

Zdroj: R, vlastní zpracování



Obrázek 2: Autokorelační funkce parametrů posteriorního rozdělení

Zdroj: R, vlastní zpracování

Tabulka 2: Předpovědi intenzity emigrace pro model M1 a M2

M_1	průměr	rozptyl	q10%	q25%	medián	q75%	q90%
y_{2016}	0,6093	0,5381	-0,2922	0,1458	0,6056	1,0756	1,5077
y_{2017}	0,5445	0,6655	-0,4440	0,0307	0,5289	1,0288	1,5241
y_{2018}	0,4942	0,8008	-0,5172	-0,0291	0,4758	0,9807	1,5095
y_{2019}	0,4917	1,1004	-0,5454	-0,0479	0,4677	0,9867	1,5167
M_2	průměr	rozptyl	q10%	q25%	medián	q75%	q90%
y_{2016}	1,0136	0,8033	-0,0925	0,4467	1,0085	1,5765	2,1350
y_{2017}	0,9625	1,7165	-0,6587	0,1441	0,9632	1,7939	2,5871
y_{2018}	0,9037	2,8077	-1,1596	-0,1639	0,8968	1,9561	2,9834
y_{2019}	0,8434	3,9621	-1,6123	-0,4213	0,8422	2,0990	3,3293

Zdroj: R, vlastní zpracování

Závěr

Modelování migrace má svá úskalí, avšak ukazuje se, že se správným naložením vstupních dat a využitím vhodných metod by bylo možné získat přesnější odhady migračních toků. Metody bayesovské statistiky rozšiřují možnost, jak vnášet do odhadů konkrétní představy o vývoji (k tomu však potřebujeme často znát expertní názor). Jednou z ne příliš prozkoumaných oblastí se stává spolehlivost a zdroje migračních dat, a co by mělo být na vstupu konkrétních modelů. Kvalita migračních dat je tudíž velkým problémem a samotné modelování migrace se stává obtížným, protože je výsledkem kombinace mnoha špatně předvídatelných a obtížně kvantifikovatelných vlivů (ekonomických, politických, sociologických). Nicméně i přesto je možné se domnívat, že migrace je vázána na jiné dobře kvantifikovatelné veličiny, na základě nichž by bylo možné vnášet do modelů další informace a zpřesňovat kvalitu dat na vstupu.

Literatura

BATTAGLIA, Marianna. Migration, health knowledge and teenage fertility: evidence from Mexico. SERIEs [online]. 2015, 6(2), 179-206 [cit. 2016-12-26]. DOI: 10.1007/s13209-015-0124-3. ISSN 18694187. Dostupné z: http://link.springer.com/10.1007/s13209-015-0124-3

BIJAK, Jakub. *Forecasting international migration in Europe: a Bayesian view*. UK: Springer, 2011. ISBN 978-90-481-8897-0

BOLSTAD, William M. *Introduction to Bayesian statistics*. 2nd ed. Hoboken, N.J.: John Wiley, c2007. ISBN 0470141158.

Český statistický úřad [online].2015 [cit. 2017-01-01]. Dostupné z: https://czso.cz/csu/czso/obyvatelstvo_hu

GELMAN, Andrew. *Bayesian data analysis*. Third edition, 2014, ISBN 9781439840955.

HUBER, Peter, et al. Decomposing Welfare Wedges. An Analysis of Welfare Dependence of Immigrants and Natives in Europe. WIFO, 2014.

JOSIFIDIS, Kosta, et al. Eastern migrations vs western welfare states-(un) biased fears. Panoeconomicus, 2013, 60.3: 323.

ROUBÍČEK, Vladimír. Úvod do demografie. 1. vydání: CODEX Bohemia, s.r.o., Praha 1997, ISBN, 80-85963-43-4

JEL Classification: C11, F22

Summary

Migration models

The main aim of this article is to introduce basic problems of modeling migration, population data uncertainty and to show Bayesian approach to demographic modeling. In analytical part is shown how to calculate intensity of emigration and forecasts for it.

Keywords: migration, models, bayesian approach, emigration

Využití stavově-prostorových modelů v demografii

Martin Matějka

martin.matejka@vse.cz

Doktorand oboru statistika

Školitelka: doc. Ing. Ivana Malá, CSc., (malai@vse.cz)

Abstrakt: Stavově-prostorové modely (State-Space models) jsou velmi často využívané nejen v oblasti statistiky zejména pro svoji schopnost postihnout dynamiku vícerozměrných časových řad a složité závislosti mezi zkoumanými proměnnými. Modelování úmrtnosti vzhledem k věku i časové dimenzi je v demografii často asociováno s tradičně používaným Lee-Carterovým modelem. Je však možné ukázat, že Lee-Carterův model předpokládá, že v případě predikcí je dynamika změn úmrtnosti neměnná. Článek se zabývá problematikou klasických stavově-prostorových modelů s následným rozšířením na zobecněné stavově-prostorové modely. Následně je demonstrováno využití zobecněných stavově-prostorového modelu pro účely modelování a predikce českých specifických měr úmrtnosti. Model je aplikován na veřejně dostupná česká demografická data za roky od 1981 až 2011 a pro věky 10 až 95 let. Poslední dostupný rok 2014 je následně využit pro konfrontaci predikcí s reálnými měrami úmrtnosti. Následně je také provedena diagnostika modelu s ohledem na odhadnuté parametry a autokorelaci reziduí.

Klíčová slova: zobecněné stavově-prostorové modely, Lee-Carterův model, úmrtnost

Úvod

Stavově-prostorové modely (state-space models) získaly na popularitě koncem 90. let minulého století v oblasti systémové teorie, stěžejním rozmachem byla jejich aplikace v Apollo vesmírném programu (Hutchinson, 1984). Tyto modely se následně začaly využívat i v dalších oblastech nejen ekonomické teorie. Stavově-prostorové (dále SS) modely jsou založeny na předpokladu, že časová řada je realizací dynamického systému, který je narušován náhodnými chybami. SS modely představují obecný rámec, který zastřešuje významnou řadu statistických modelů. Mezi nejznámější aplikace patří modelování sezónnosti s proměnlivým charakterem, modelování výnosových křivek nejen tradičně přes oblast doby do splatnosti, ale také přes časovou dimenzi, čímž je možné predikovat výnosové křivky v čase do budoucna. Další významnou aplikací jsou dynamické (zobecněné) lineární modely (tj. zobecněné lineární modely s časově závislými parametry vysvětlujících proměnných). SS modely umožňují modelování jak jednorozměrných, tak vícerozměrných stacionárních či nestacionárních časových řad, které mohou obsahovat strukturální změny nebo časové nepravidelnosti. V neposlední řadě jsou jejich speciálním případem v současné době velmi populární ARMA modely.

Cílem předkládaného článku je využití SS modelů při modelování úmrtnosti. Tradičním přístupem k modelování úmrtnosti v čase je aplikace Lee-Carterova modelu (Lee a Carter, 1992). V textu budou diskutovány vlastnosti Lee-Carterova modelu, které slouží jako motivace pro využití SS modelu.

Současný stav poznání

Teoretickými východisky problematiky SS modelů v tradičním klasickém pojetí se zabývali Durbin a Koopman (2012). Hyndman (2008) zkoumal využití SS modelů pro exponenciální vyrovnávání a modelování sezónnosti s proměnlivým charakterem. Petris (2009) zkoumali bayesovský přístup k odhadu parametrů SS modelů.

V demografii jsou SS modely aplikovány v souvislosti s predikcí populačních stavů (Tavecchia, Besbeas, Coulson, Morgan a Clutton-Brock, 2009). Rozšíření Lee-Carterova modelu ve smyslu simultánního odhadu a predikce časově závislé sady parametrů se zabývalo několik studií (Reichmuth a Sarferaz, 2008, Husin, Zainol a Ramli, 2015, Fung, 2015).

Odhady parametrů SS modelů představují z praktického hlediska velmi náročnou výpočetní úlohu, proto je možné najít několik balíčků implementovaných v statistickém programovacím prostředí R (Team R, 2014), které se touto problematikou zabývají. Jsou to například KFAS (Helske, 2014) či dlm (Petris, 2010).

Metodologie stavově-prostorových modelů

SS modely je možné rozdělit na klasické (Gaussovské) a zobecněné SS modely. Zobecněné SS modely metodologicky vycházejí z klasických SS modelů, proto bude pozornost zaměřena na nejprve na metodologii

klasických SS modelů a následně na jejich rozšíření do podoby zobecněných SS modelů, které již bezprostředně souvisí s cílem tohoto článku, tedy modelováním úmrtnosti.

Gaussovské stavově-prostorové modely

SS modely se skládají z dvou základních soustav rovnic pozorování a stavů. První soustava *p* rovnic vysvětluje chování pozorovaných dat pomocí pozorovatelných nebo nepozorovatelných (latentních) proměnných

$$y_t = Z_t\alpha_t + \varepsilon_t, \tag{1}$$

kde časový index $t = 1,...,T$, y_t je vektor vícerozměrné časové řady pozorování typu (*p* x 1), $\varepsilon_t \sim N_p(0,H_t)$ je *p*-rozměrný vektor náhodných chyb. Vektor pozorovaných či nepozorovaných latentních proměnných α_t typu (*m* x 1) vysvětluje y_t pomocí stavové matice Z_t typu (*p* x *m*).

Druhá soustava *m* rovnic popisuje evoluci latentních proměnných v čase, na které opět působí náhodné chyby

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \tag{2}$$

kde $\eta_t \sim N_k(0,Q_t)$ je *k*-rozměrný vektor náhodných chyb působící na latentní proměnné α_{t+1}. Systémové matice T_t typu (*m* x *m*) a R_t typu (*m* x *k*) vymezující vztahy uvnitř stavových rovnic a rovnic pozorování. Kovarianční matice H_t a Q_t (obvykle předpokládáme, že jsou v čase invariantní a tedy uvažujeme pouze H typu (*p* x *p*) a Q typu (*k* x *k*)) určující kovarianční strukturu náhodných chyb pro jednotlivé rovnice (Durbin a Koopman, 2012).

SS modely umožňují modelovat pozorovanou časovou řadu y_t (případně vícerozměrnou časovou řadu) pomocí vektoru stavových proměnných. První rovnice tedy popisuje chování pozorovaných časových řad pomocí stavových proměnných, přičemž se předpokládá, že jsou pozorování y_t zatížena chybami ε_t. Druhá rovnice popisuje vývoj stavových proměnných v čase, která je řízena stochastickým procesem inovací η_t.

Gaussovské SS modely je možné rozšířit tak, aby bylo možné uvažovat pravděpodobnostní rozdělení exponenciálního typu. Rovnice pozorování je v takovém případě vyjádřena pomocí hustoty pravděpodobnosti náhodných veličin pozorované časové řady

$$\begin{aligned} p(y_t|\theta_t) &= p(y_t|Z_t\alpha_t) \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t \end{aligned}, \tag{3}$$

kde $p(y_t|\theta_t)$ je hustota pravděpodobnosti náhodných veličin pozorované časové řady a $\theta_t = Z_t\alpha_t$ se označuje jako signál. Signál je tedy lineárním prediktorem, který vysvětluje střední hodnotu $E(y_t) = \mu_t$ vysvětlované náhodné veličiny y_t s využitím spojovací funkce $l(\mu_t) = \theta_t$.

Pro práci s rovnicemi (2) můžeme využít standardní postupy, které jsou obecné známé z oblasti zobecněných lineárních modelů (Durbin a Koopman, 2012):

- Normální rozdělení se střední hodnotou μ_t a rozptylem u_t s identitou jako spojovací funkcí $\theta_t = \mu_t$.
- Poissonovo rozdělení se střední hodnotou λ_t a expozicí u_t s logaritmickou spojovací funkcí $\theta_t = \log(\lambda_t)$. Tedy $E(y_t|\theta_t) = D(y_t|\theta_t) = u_t e^{\theta t}$. Tento přístup je v dalších kapitolách uvažován.
- Binomické rozdělení s rozsahem výběru u_t, pravděpodobností úspěchu π_t a logitovou spojovací funkcí $\theta_t = \text{logit}(\pi_t)$. Tedy $E(y_t|\theta_t) = u_t\pi_t$ a $D(y_t|\theta_t) = u_t(\pi_t(1-\pi_t))$.
- Gamma rozdělení s parametry u_t (určující tvar rozdělení), μ_t (střední hodnota rozdělení) a logaritmickou spojovací funkcí $\theta_t = \log(\mu_t)$. Následně $E(y_t|\theta_t) = e^{\theta_t}$ a $D(y_t|\theta_t) = e^{2\theta_t}/u_t)$.

- Negativní binomické rozdělení s disperzním parametrem u_t, očekávanou hodnotou μ_t a logaritmickou spojovací funkcí $\theta_t = \log(\mu_t)$. Tedy $E(y_t|\theta_t) = e^{\theta_t}$ a $D(y_t|\theta_t) = e^{\theta_t} + e^{2\theta_t}/u_t$.

Odhad stavově-prostorových modelů

Základní procedury, využívané pro odhad parametrů klasických SS modelů, jsou Kalmanův filtr a Kalmanův vyhlazovač (Durbin a Koopman, 2012). Kalmanovým vyhlazovačem je označen proces konstrukce neznámých stavových proměnných α_t pro oblast pozorování (vícerozměrné) časové řady $y_1, \ldots, y_T$. Odhad stavových proměnných je proveden s ohledem na všechna (tj. i budoucí) pozorování. V případě zobecněných SS modelů je nutné aplikovat rozšířený Kalmanův filtr, který nejprve linearizuje zobecněný SS pomocí Laplaceovy aproximace a následně aplikuje Kalmanův filtr a vyhlazovač. Samotný odhad neznámých parametrů modelu (stavových proměnných, kovariančních matic či pouze jejich částí) je proveden v tradičním pojetí metodou maximální věrohodností s ohledem na uvažované pravděpodobnostní rozdělení exponenciálního typu. Numerická optimalizace je následně provedena nejčastěji metodou BFGS (zkratka vznikla z počátečních písmen jmen autorů Broyden, Fletcher, Goldfarb, Shanno), případně metodou Nelder-Mead (podle autorů Neldera a Meada, dále NM).

Lee-Carterův model

Velmi užitečným, dlouhodobě oblíbeným a stále používaným přístupem pro modelování měr úmrtnosti, je Lee-Carterův model (LC) log-bilineární model. Metoda navržená Lee a Carterem (1992) se stala základní statistickou metodou využívanou v demografických analýzách, LC model si klade za cíl popsat a predikovat věkově a časově specifické míry úmrtnosti (m_{xt}), které jsou vymezené následovně

$$\log(m_{xt}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{xt}, \tag{4}$$

kde t je čas, x věk, $m_{xt} = D_{xt} / E_{xt}$ reprezentuje časově-věkově specifické míry úmrtnosti a ε_{xt} jsou náhodné chyby, o kterých se předpokládá, že mají normální rozdělní. Počet zemřelých ve věku x a v čase t je označen D_{xt} a střední stav obyvatelstva ve věku x a čase t E_{xt}. První sada parametrů α_x vymezuje věkově specifický obecný profil úmrtnosti. Druhá sada parametrů β_x odchylky úmrtnosti od α_x dané působením obecného trendu úmrtnosti, který je určen třetí sadou parametrů κ_t. Vzhledem k faktu, že variabilita logaritmu specifických měr úmrtnosti narůstá s rostoucím věkem, je předpoklad normálně rozdělených chyb ε_{xt} nedostačující. Proto model předpokládá, že počet zemřelých je náhodná veličina, která se řídí Poissonovým rozdělením, tedy platí $D_{xt} \sim \mathrm{Po}(E_{xt} m_{xt} = E_{xt}(\alpha_x + \beta_x \kappa_t))$ (Delwarde a další, 2006).

Parametrizace modelu (3) není jednoznačná, neboť je model neidentifikovatelný především díky věkově specifické sadě parametrů β_x. Proto se zavádějí omezující podmínky

$$\sum_t \kappa_t = 0, \quad \sum_x \beta_x = 1, \tag{5}$$

které již umožňují získání jednoznačných odhadů parametrů LC modelu (Wilmoth, 1993). Odhady parametrů modelu je možné získat díky předpokladu Poissonova rozdělení s využitím metody maximální věrohodnosti a modelových specifických měr úmrtnosti.

Jednou z nevýhod LC modelu je fakt, že věkově specifická sada parametrů β_x je odhadnuta na základě historických dat a dále se v čase nevyvíjí. Z tohoto důvodu bude využit zobecněný SS model, který bude předpokládat, že parametry stavových rovnic se budou v čase vyvíjet a tento vývoj bude možné zohlednit v predikci do budoucna.

Aplikace zobecněného stavově-prostorového modelu

Data

Analýza v předkládaném článku využívá veřejně dostupná data z HMD (Human Mortality Database)[1], kde jsou k dispozici populační hodnoty počtu zemřelých a středních stavů. Zobecněný SS model bude založen na

[1] www.mortality.org

populačních úmrtnostních datech od roku 1981 do 2011 a pro věky od 10 do 90 let. HMD poskytuje úmrtnostní data až do roku 2014, která budou využita pro konfrontaci predikcí modelu se skutečností.

Aplikace zobecněného stavově-prostorového model úmrtnosti

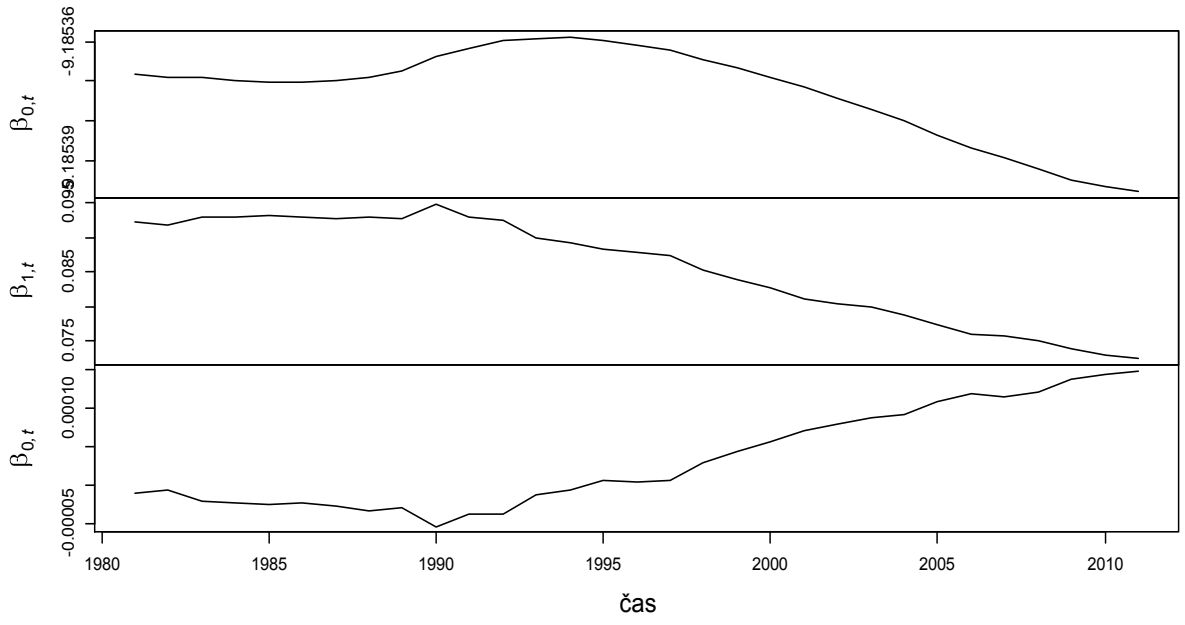
Zobecněný SS model byl použit na česká data (viz výše) od roku 1981 až do 2011 a pro věky 10 až 95 let. Specifikace zobecněného SS modelu je popsána následovně

$$\begin{aligned} p(y_t|\theta_t) &= p(y_t|Z\alpha_t) = MPoisson(E_t e^{\theta_t}) \\ \alpha_{t+1} &= \alpha_t + \eta_t \end{aligned}, \tag{6}$$

kde MPoisson je označení pro vícerozměrné Poissonovo rozdělení, y_t je vícerozměrná časová řada počtu zemřelých. V případě uvažování dynamického polynomu druhého stupně je možné rovnici signálu θ_t v (2) zapsat jako

$$\begin{bmatrix} \theta_{10,t} \\ \dots \\ \theta_{95,t} \end{bmatrix} = \beta_{0,t} + \beta_{1,t} \begin{bmatrix} 10 \\ \dots \\ 95 \end{bmatrix} + \beta_{2,t} \begin{bmatrix} 10^2 \\ \dots \\ 95^2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{x_{\min},t} \\ \dots \\ \varepsilon_{x_{\max},t} \end{bmatrix}. \tag{7}$$

Odhady dynamických parametrů modelu (6) $\beta_{k,t}$ pro k = 0, 1, 2 jsou znázorněny na obrázku 1. Vývoj odhadu parametrů $\beta_{0,t}$ demonstruje změnu obecné úrovně úmrtnosti v čase. Parametry $\beta_{1,t}$ a $\beta_{2,t}$ je možné interpretovat jako odchylky od tohoto obecného trendu vzhledem k věku.

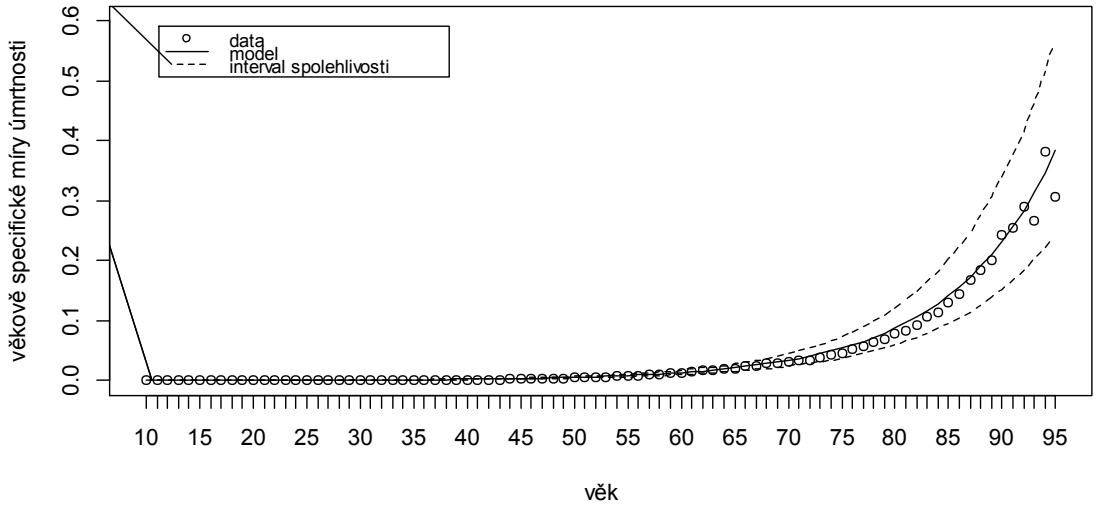


Obrázek 8: Vývoj parametrů zobecněného SS modelu Zdroj: vlastní výpočty

Nyní provedeme individuálními testy o parametrech. Bude nás zajímat, zda parametry v rovnici (6) jsou v modelu statisticky významné. Z celkového počtu 93 parametrů je na 5% hladině významnosti pouze 6 statisticky nevýznamných (konkrétně se jedná o parametry kvadratického členu pro roky 1981,1982 a 1994 až 1997.

Zobecněný SS model (6) byl odhadnut za období od 1981 až do 2011. Je tedy možné predikovat míry úmrtnosti například pro rok 2014 a porovnat tyto predikce se skutečnými pozorovanými měrami úmrtnosti, jak zachycuje následující obrázek. Predikce pro zobecněný SS model (5) jsou vypočítány pomocí MCMC (Markov chain

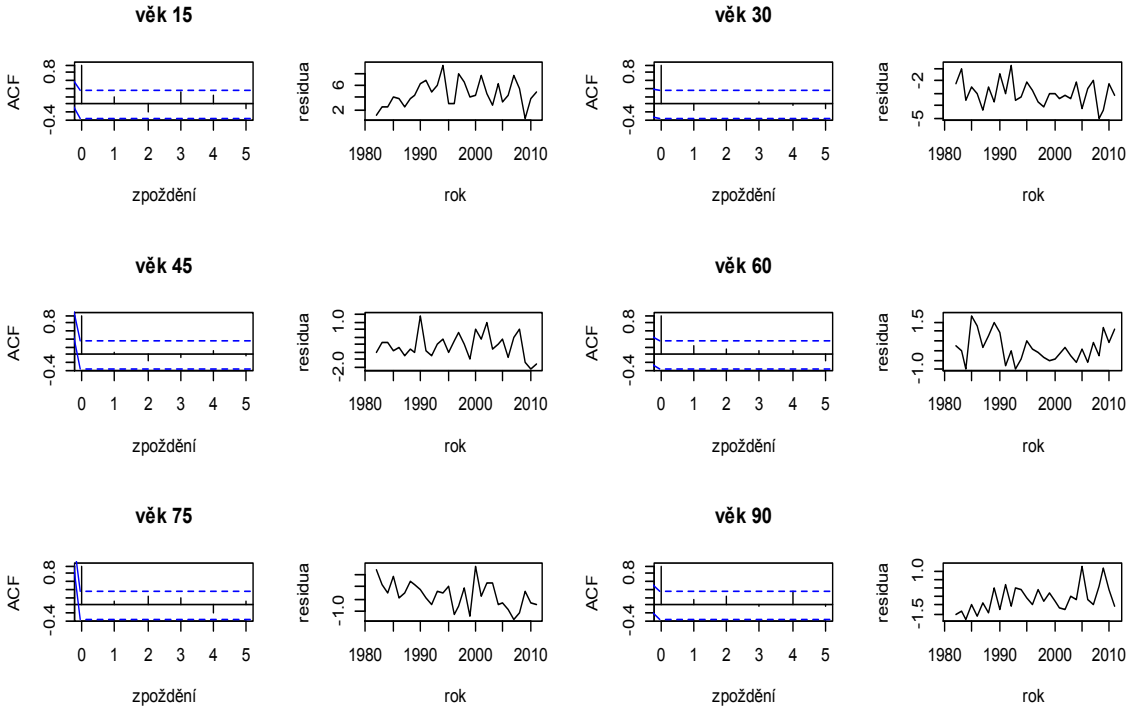
Monte Carlo) částicových filtrů, konkrétně metody importance sampling s počtem 1000 simulací. Intervaly spolehlivosti odpovídají empirickým kvantilům z provedených predikčních simulací (Helske, 2014).



Obrá

zek 9: Predikce zobecněného SS modelu pro rok 2014 Zdroj: vlastní výpočty

Z výše uvedeného porovnání predikcí (plná čára, respektive přerušované čáry) a pozorovaných dat (kolečka) je patrné, že odhad měr úmrtnosti pro věky od 10 do 95 let koresponduje s pozorovanými daty, nicméně je třeba upozornit, že v oblasti od 70 do 85 let je patrné mírné podhodnocení modelem predikovaných měr úmrtnosti při porovnání s reálnými měrami úmrtnosti v posledním dostupném roce 2014. Významnost případné autokorelační struktury reziduí je možné posoudit pomocí autokorelační funkce reziduí pro pevně zvolený věk přes zkoumané období. Pro každý věk od 10 do 95 let je zkoumáno, zda hodnota autokorelační funkce až do pátého zpoždění není statisticky významná. V 10 případech z 86 zkoumaných časových se projevila statisticky významná autokorelace prvního řádu Ilustrativní výběr autokorelačních funkcí a odpovídajících reziduí pro vybrané věky znázorňuje následující obrázek.



Obrázek 10: Autokorelační funkce a rezidua pro vybrané věky Zdroj: vlastní výpočty

Závěr

Předkládaný článek se zaměřuje na aplikace stavově-prostorových (SS) modelů pro modelování a predikování úmrtnosti. Nejprve je pozornost věnována klasickým Gaussovským SS modelům a jejich následnému rozšíření na zobecněné SS modely. Zobecněné SS modely svým charakterem korespondují s problematikou demografických analýz, neboť umožňují postihnout časovou dynamiku počtu zemřelých jako diskrétní náhodné veličiny. Specifikace modelu vychází z predikčních vlastností Lee-Carterova modelu a zvýšené potřeby postihnout změnu úmrtnosti v čase.

V praktické části článku je zobecněný stavově-prostorový model aplikován na česká úmrtnostní data od roku 1981 do roku 2011. Informace z posledního dostupného roku 2014 jsou následně využity k ověření predikčních schopností modelu, kdy 70 až 85 let model mírně podhodnocuje reálné specifické míry úmrtnosti.

Zobecněný stavově-prostorový model, tak jak byl specifikován v předkládaném článku, nebyl v oblasti demografických analýz nikdy aplikován, tudíž samotné představení modelu odkrývá široké možnosti využití stavově-prostorových modelů pro modelování úmrtnosti.

Literatura

DELWARDE, Antoine, et al. Application of the Poisson log-bilinear projection model to the G5 mortality experience. *Belgian Actuarial Bulletin*, 2006, 6.1: 54-68.

DURBIN, James; KOOPMAN, Siem Jan. *Time series analysis by state space methods*. Oxford University Press, 2012.

FUNG, Man Chung; PETERS, Gareth William; SHEVCHENKO, Pavel V. A State-Space Estimation of the Lee-Carter Mortality Model and Implications for Annuity Pricing. *Available at SSRN 2699624*, 2015.

HELSKE, Jouni. KFAS: Kalman filter and smoothers for exponential family state space models. *R package version*, 2014, 1: 4-1.

HYNDMAN, Rob, et al. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

HUSIN, Wan Zakiyatussariroh Wan; ZAINOL, Mohammad Said; RAMLI, Norazan Mohamed. Performance of the Lee-Carter State Space Model in Forecasting Mortality. In: *Proceedings of the World Congress on Engineering*. 2015.

HUTCHINSON, Charles E. The Kalman filter applied to aerospace and electronic systems. *IEEE transactions on aerospace and electronic systems*, 1984, 4: 500-504.

LEE, Ronald D.; CARTER, Lawrence R. Modeling and forecasting US mortality. *Journal of the American statistical association*, 1992, 87.419: 659-671.

PETRIS, Giovanni; PETRONE, Sonia; CAMPAGNOLI, Patrizia. Dynamic linear models. In: *Dynamic Linear Models with R*. Springer New York, 2009. p. 31-84.

PETRIS, Giovanni. dlm: Bayesian and likelihood analysis of dynamic linear models. *R package version*, 2010, 1.3.

REICHMUTH, Wolfgang, et al. Bayesian demographic modeling and forecasting: An application to US mortality. *Humboldt-Universitat zu Berlin, Germany*, 2008.

TAVECCHIA, Giacomo, et al. Estimating Population Size and Hidden Demographic Parameters with State-Space Modeling. *The American Naturalist*, 2009, 173.6: 722-733.

TEAM, R. Core. The R project for statistical computing. *Available at www. R-project. org/. Accessed October*, 2014, 31: 2014.

WILMOTH, John R. *Computational methods for fitting and extrapolating the Lee-Carter model of mortality change*. Technical report, Department of Demography, University of California, Berkeley, 1993.

JEL Classification: C32, J11

Summary

Application of State-Space models in Demography

In last decade, State-Space models became very frequently used modelling approach not only in statistical applications but also in other areas. State-Space models provide efficient framework for modeling sophisticated relationships in multidimensional time series as well as among random variables in general. Mortality modeling in respect to both age and time is very often associated with most commonly used Lee-Carter model. Nevertheless, predictions of mortality rates based on Lee-Carter model are based on assumption that age specific mortality change is preserved. This paper firstly very briefly describes theoretical background of Gaussian State-Space models with subsequent extension into Generalized State-Space models. The practical analysis aims to application of Generalized State-Space model for modeling and predicting Czech mortality rates. The model is based on publicly available demographic data from 1981 to 2011 years and ages from 10 to 95 years old. The last available year 2014 is then used for cross-validation of predictions with observed mortality rates. Finally, model parameters as well as autocorrelation structure of residuals are examined.

Keywords: Generalized State-Space models, Lee-Carter model, mortality.

Modelování sezónnosti s dlouhou délkou periody – aplikace na modelování počtu denních dopravních nehod nepojištěných vozidel

Jiří Procházka

xproj16@vse.cz

Doktorand oboru Statistika

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

Abstrakt: Denní počet dopravních nehod je pro pojišťovny důležitým indikátorem, s jehož pomocí jsou schopny například efektivně plánovat likvidaci pojistných událostí, ale i zlepšit proces rezervování. Pro výše zmíněné potřeby je mimo jiné relevantní vytvořit model predikující počet dopravních nehod. Nedílnou součástí tohoto modelu je i model denní sezónnosti. Denní sezónnost lze beze sporu označit jako fenomén sezónnosti s dlouhou délkou periody. Pro modelování tohoto fenoménu se používají především modely vycházející z rozkladu sezónního cyklu pomocí bázových funkcí. V tomto článku bude pozornost zaměřena na jednoduché modely, vycházející z rozkladu sezónního cyklu s využitím Fourierových bází. Dále bude zkoumán vliv externích proměnných, jako jsou průměrná denní teplota a denní úhrn srážek na počet dopravních nehod.

Klíčová slova: dopravní nehody, sezónnost s dlouhou délkou periody, Fourierovy báze

Úvod

Tento článek se zabývá predikcí denního počtu dopravních nehod způsobených vozidly bez pojištění odpovědnosti z provozu vozidel. Speciální pozornost je kladena na modelování denní sezónnosti počtu těchto událostí. Využitá data jsou data o počtu dopravních nehod způsobených vozidly bez pojištění odpovědnosti z provozu vozidel, poskytnutá Českou kanceláří pojistitelů, mezi léty 2007-2010. Z dat byly vyřazeny veškeré přestupné dny z důvodu zachování celočíselné sezónnosti, tedy sezónnosti o celočíselné délce. Délku sezónnosti budu v následujícím textu značit jako L. V tomto konkrétním případě $L = 365$. Délku časové řady budu v následujícím textu značit symbolem N. V tomto konkrétním případě $N = 1460$.

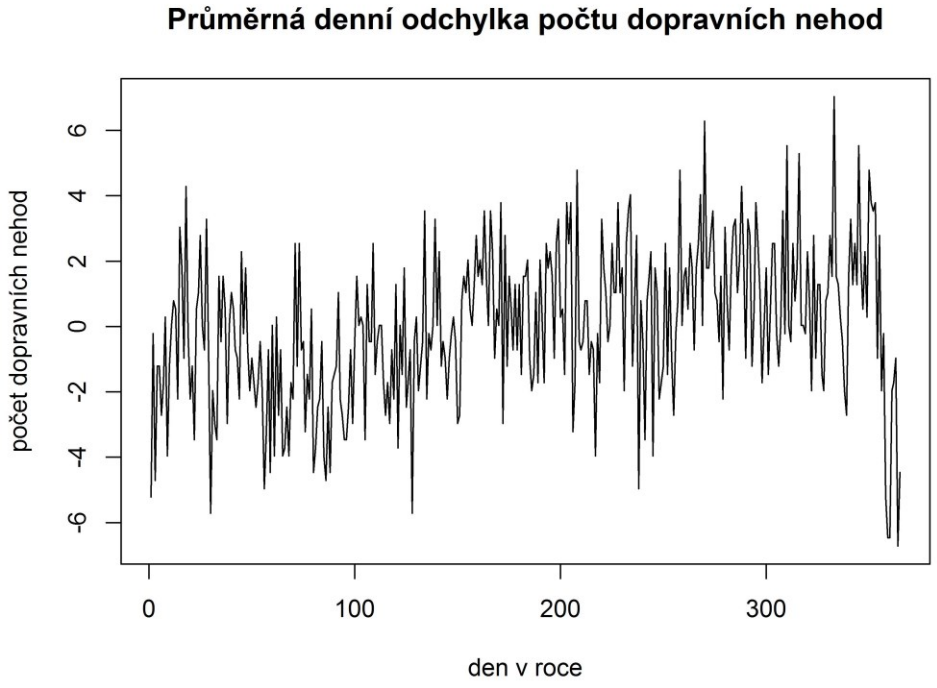
Z hlediska modelu predikce budu vycházet z velmi jednoduchého modelu sezónní časové řady délky N $\{X_t: t = 1, \ldots, N\}$,

$$X_t = S_t + B_t + E_t, \qquad t = 1, \ldots, N, \tag{1}$$

kde $\{S_t: t = 1, \ldots, N\}$, $\{B_t: t = 1, \ldots, N\}$ a $\{E_t: t = 1, \ldots, N\}$ reprezentují deterministickou sezónnost, ostatní deterministické komponenty (jako deterministický trend, externí vlivy ovlivňující vývoj časové řady atp.), a stacionární ARMA proces. Sezónnost jako taková bude v tomto konkrétním případě vyjadřovat jakousi expozici, kde je očekáváno, že jednotlivé dny v rámci roku budou vykazovat signifikantně nižší počet řidičů na silnicích a tedy i pravděpodobně nižší počet dopravních nehod. Ostatní deterministické komponenty budou reprezentovány indikátory extrémních výkyvů počasí. Budou sledovány především mrazy, zvýšené srážky a jejich kombinace. Klimatologická data použita v analýze byla poskytnuta Českým hydrometeorologickým ústavem.

Model deterministické sezónnosti

K efektivnímu modelování typu sezónnosti přítomného v denní časové řadě dopravních nehod je potřeba použít modely zabývající se modelováním sezónnosti s dlouhou délkou periody. Jak bylo diskutováno v (Procházka a spol., 2016), klasické modely v takovém případě selhávají. Důvodem tohoto selhání je například vysoký počet parametrů, výpočetní náročnost metod atd. Možností modelování sezónnosti s dlouhou délkou je například rozklad sezónního cyklu pomocí bázových funkcí.



Obrázek 1: Průměrná denní absolutní odchylka počtu dopravních nehod od průměrného počtu dopravních nehod.

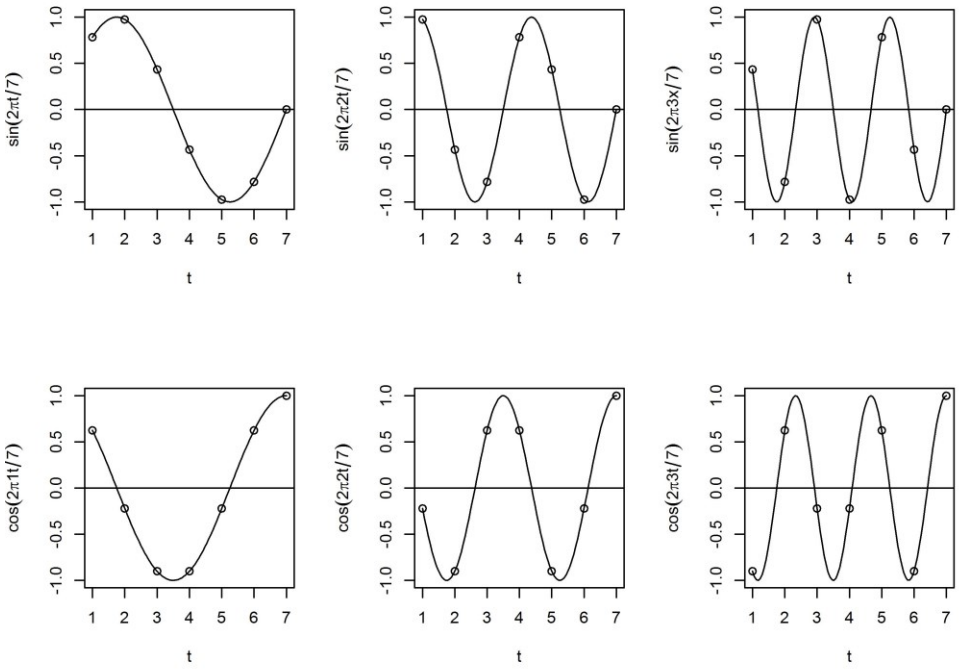
Na obrázku 1 je zobrazený hrubý tvar sezónního cyklu počtu dopravních nehod, vyjádřený jako odchylka průměrného denního počtu dopravních nehod od celkového průměru počtu dopravních nehod. Typické pro tento cyklus je pokles počtu dopravních nehod pozorovatelný na konci kalendářního roku. Tento pokles je i přes roční období a nepříznivé počasí způsoben nižším počtem řidičů na vozovkách. Modelování tohoto jevu je z hlediska rozkladu sezónního cyklu pomocí Fourierových bází velmi obtížné, a to speciálně kvůli tomu, že Fourierovy báze nejsou lokalizované v čase, zatímco tento jev je v čase lokalizovaný. Pro zachycení tohoto jevu s využitím rozkladu pomocí Fourierových bází je potřeba velké množství parametrů, což vzhledem k malému počtu pozorování může vést k značné variabilitě takto zkonstruovaného odhadu. Nicméně s výjimkou konce roku, může být rozklad sezónního cyklu pomocí Fourierovy báze použit jako vhodný model sezónnosti. V případě rozkladu sezónního cyklu pomocí Fourierových bází je možné sezónní komponentu časové řady $\{S_t\}$ rozložit v případě lichého L jako

$$S_t = \sum_{K=1}^{\lfloor L/2 \rfloor} \alpha_K \sin\left(2\pi \frac{K}{L} t\right) + \sum_{K=1}^{\lfloor L/2 \rfloor} \beta_K \cos\left(2\pi \frac{K}{L} t\right), \qquad t = 1,\ldots,N. \qquad \textbf{(2a)}$$

V případě sudého L jako

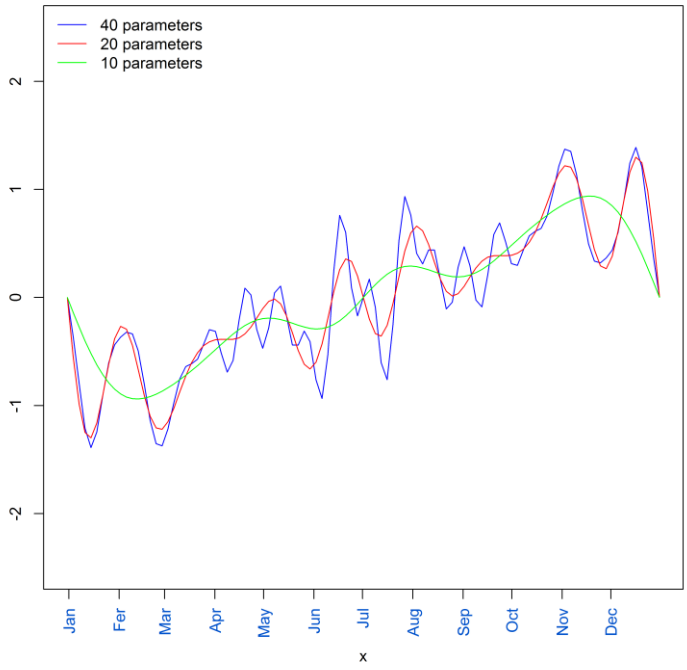
$$S_t = \sum_{K=1}^{L/2-1} \alpha_K \sin\left(2\pi \frac{K}{L} t\right) + \sum_{K=1}^{L/2} \beta_K \cos\left(2\pi \frac{K}{L} t\right), \qquad t = 1,\ldots,N, \qquad \textbf{(2b)}$$

kde α_K a β_K jsou v obou rovnicích parametry, které je potřeba odhadnout. Báze spojené s nízkými frekvencemi sinů a kosinů jsou reprezentovány nízkými hodnotami K, zatímco báze spojené s vysokými frekvencemi jsou reprezentovány vysokými hodnotami K viz obrázek 2.



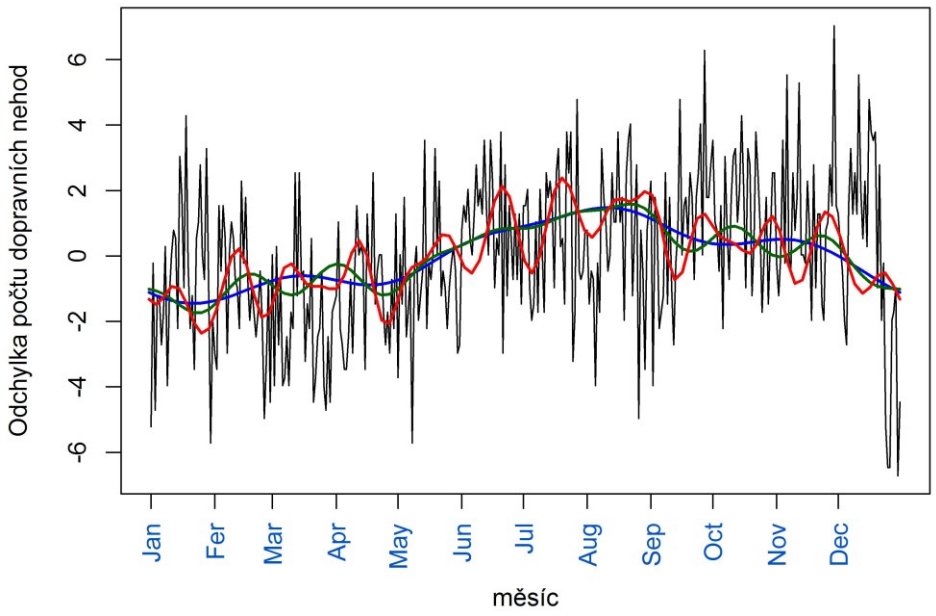
Obrázek 2: Šest bázových funkcí rovnice 2a pro $L = 7$. První řádek reprezentuje siny, druhý řádek kosiny. Sloupce pak reprezentují různou volbu K ($K = 1, 2$ a 3). Na obrázku je zobrazeno pouze prvních 7 hodnot funkce. Všechny zobrazené funkce jsou periodické s periodou rovnou 7.

V případě modelování sezónního cyklu jsou obvykle používány pouze nižší frekvence, což fakticky umožňuje do modelu zavést mírné vychýlení odhadu, ovšem za cenu velké redukce variability. Na obrázku 3 jsou zobrazeny rozklady sezónního cyklu s využitím Fourierových bází pro různou volbu K a tedy i různý počet parametrů.



Obrázek 3: Odhad sezónního cyklu s využitím rozkladu do Fourierových bází pro různou volbu *K*.

Na obrázku 3 je možné pozorovat, jak s rostoucím počtem parametrů klesá hladkost odhadnutého sezónního cyklu a roste variabilita odhadu. Jakési „optimální“ K je poté možné volit například na základě metody křížové validace, případně na základě různých informačních kritérií a podobně. Na následujícím obrázku je zachyceno, jak výše popsaný model vystihuje skutečný sezónní cyklus pro různou volbu K. Je vidět, že i pro $K = 5$ je model velice dobrou aproximací sezónního cyklu.



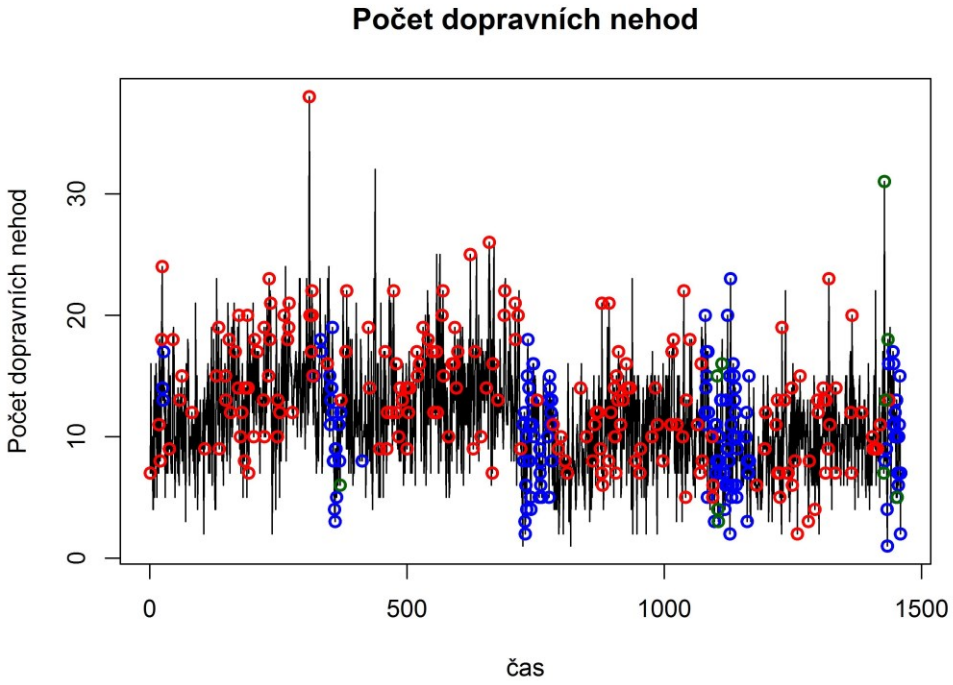
Obrázek 4: Odhad sezónního cyklu pomocí rozkladu do Fourierových bází dle vzorce 2a pro různou volbu *K*. Modrá křivka odpovídá volbě $K = 5$, zelená křivka volbě $K = 10$ a červená křivka volbě $K = 20$.

Model ostatních deterministických komponent

I když data vykazují mírný trend, byl pro účely sestavení predikčního modelu tento trend zanedbán. Do modelu byly přidány jako deterministické komponenty pouze indikátor mrazu $\{M_t: t = 1, \ldots, N\}$, indikátor zvýšených srážek $\{R_t: t = 1, \ldots, N\}$ a dále indikátor jejich kombinace $\{MR_t: t = 1, \ldots, N\}$. Indikátor mrazu je v čase t roven jedné tehdy, je-li teplota $\{T_t: t = 1, \ldots, N\}$, v čase t, $t - 1$ a $t - 2$ nižší než nula. Druhá indikátorová veličina, indikátor zvýšených srážek, je v čase t rovna jedné tehdy, přesáhne-li denní úhrn srážek v čase t hodnotu 2,6 mm/hod, což je obecná hranice zvýšených srážek, při které je již mírně snížená viditelnost. Třetí indikátorová veličina je v čase t rovna jedné pouze tehdy, jsou-li obě předchozí veličiny v čase t rovny 1. Model ostatních deterministických komponent $\{B_t: t = 1, \ldots, N\}$ bude mít tvar

$$B_t = \gamma_0 + \gamma_1 M_t + \gamma_2 R_t + \gamma_3 MR_t,$$

kde γ_0, γ_1, γ_2 a γ_3 jsou parametry. Na následujícím obrázku číslo 5 je zobrazen celkový počet dopravních nehod za jednotlivé dny mezi lety 2007-2010. Dále jsou zde vyznačeny body indikující mrazy a indikující zvýšené srážky.



Obrázek 5: Průběh časové řady s vyznačenými body indikace mrazu a indikace zvýšených srážek. Modrou barvou jsou vyznačeny body, kdy byl indikátor mrazu roven jedné, červenou barvou jsou vyznačeny body, kdy byl indikátor zvýšených srážek roven jedné a tmavě zelenou barvou jsou vyznačeny body, kdy byly napozorovány mrazy i zvýšené srážky.

Zatímco u zvýšeného počtu srážek lze pozorovat v průměru zvýšený počet dopravních nehod, pro indikátor mrazu tento fakt bez zahrnutí sezónnosti neplatí. Nicméně mráz obvykle nastává v obdobích, které jsou zároveň spojovány s poklesem počtu řidičů na vozovkách, a proto je potřeba neposuzovat vliv v mrazu nezávisle na sezónnosti.

Sestavení predikčního modelu

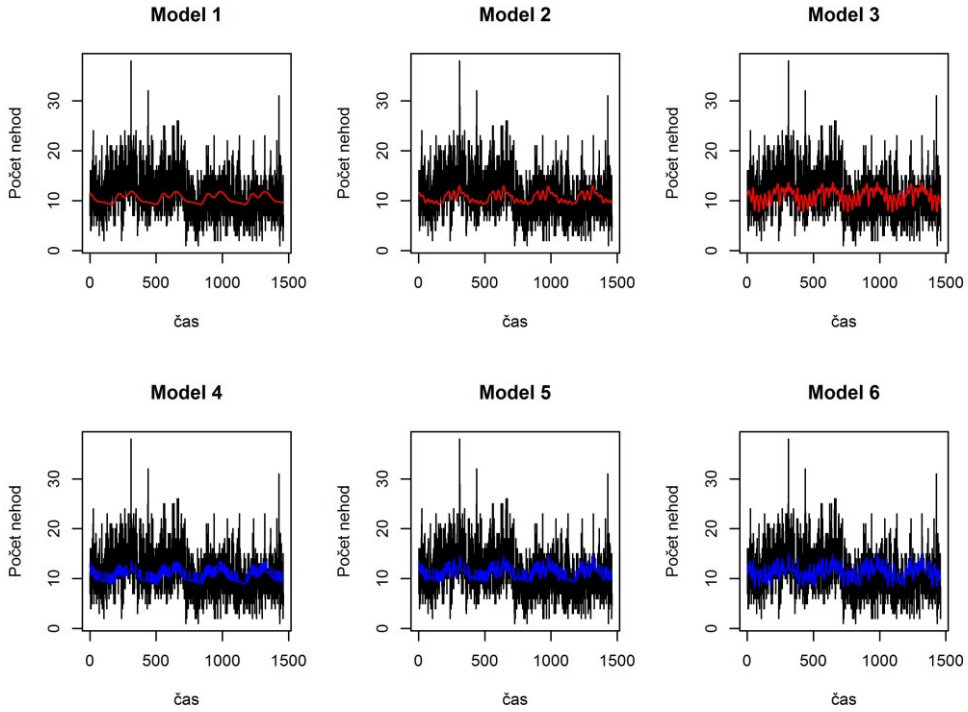
Celkem budu sestavovat šest různých modelů. První tři modely budou čistě jen modely sezónnosti časové řady rozložené pomocí Fourierových bází pro volbu $K = 5$, $K = 10$ a $K = 20$. Následující tři modely budou ty samé modely obsahující navíc externí informaci v podobě indikátoru mrazu, zvýšených srážek a jejich kombinace. Jednotlivé modely byly odhadnuty jednak na kompletních datech za roky 2007 – 2010, dále taktéž pouze za roky 2007 – 2009, kdy bude následně počítána na základě hodnot roku 2010 chyba na testovacích datech. Jako míra chyby bude využita odmocnina z průměrné čtvercové odchylky, tedy Root-mean-square error. Řád ARMA procesu chybové složky a odhady jeho parametrů budou určeny na základě automatizované procedury *auto.arima()* navržené Robem J. Hyndmanem viz (Hyndman, 2007).

Tabulka 8: Výsledky modelů. Jednotlivé sloupce tabulky reprezentují různé charakteristiky modelů. V každém sloupci je tučně vyznačena nejnižší hodnota napříč všemi modely Zdroj: Vlastní zpracování dat

Číslo modelu	Chyba na trénovacích datech	Chyba na testovacích datech	Počet parametrů modelu	AIC modelu
Model 1	4,59	**4,09**	**15**	8372,30
Model 2	4,54	4,13	23	8377,91
Model 3	4,48	4,23	45	8374,25
Model 4	4,54	4,28	17	**8344,04**
Model 5	4,43	4,37	25	8351,67
Model 6	**4,37**	4,38	44	8355,35

Přesto, že chyba na testovacích datech byla nejnižší v případě modelu bez externí informace, vyplývá z výsledků zachycených v tabulce 1, že externí informace o počasí má dodatečný vliv na počet dopravních nehod jak naznačuje snížené AIC. Speciálně silná závislost pak byla naměřena mezi indikátorem zvýšených srážek a počtem dopravních nehod.

Na obrázku 6 jsou do dat vykresleny vyrovnané hodnoty výše popsaných modelů. Z grafů i z modelů je jasné, že v tomto konkrétním případě hraje velkou roli náhoda. Na druhou stranu je možné z obrázku 6 pozorovat, že vyrovnané hodnoty vystihují v čase tvar časové řady.



Obrázek 6: Vizualizace vyrovnaných hodnot časové řady počtu dopravních nehod. První řádek matice grafů odpovídá modelům sestaveným bez externí informace. Druhý řádek matice grafů odpovídá modelům sestaveným na základě externí informace. První sloupec matice grafů odpovídá rozkladu sezónnosti dle vzorce 2a a $K = 5$, druhý sloupec volbě $K = 10$ a třetí sloupec volbě $K = 20$.

Závěr

Jak již bylo shrnuto v úvodu, modelování počtu dopravních nehod má bezesporný význam nejen v oblasti pojišťovnictví. Ke konstrukci predikčního modelu lze využít externí informace, jakými jsou například průměrná denní teplota, denní úhrn srážek a podobně. V poslední kapitole bylo naznačeno, že přidání externích informací může sloužit ke zlepšení prediktivních schopností modelu. I přesto, že chyba na testovacích datech vycházela nižší pro modely bez externích dat, je nutné si uvědomit, že tato chyba je taktéž pouze náhodnou veličinou a na základě získaného výsledku nelze vynášet definitivní závěry o predikční schopnosti modelu.

Jak bylo diskutováno v kapitole *Model deterministické sezónnosti*, rozklad sezónního cyklu do Fourierových bází nemusí být v tomto konkrétním případě nejlepším řešením. Jako alternativní řešení se nabízí rozklad pomocí kubických splinů s restrikcemi na periodicitu, případně pomocí kombinace Fourierových bází a lineárních splinů s restrikcemi na periodicitu. Tyto alternativní přístupy jsou detailněji rozpracovány a diskutovány v rozpracovaném článku *Insurance claims seasonality modelling*.

Reference

[1] DE LIVERA, Alysha M.; HYNDMAN, Rob J.; SNYDER, Ralph D. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 2011, 106.496: 1513-1527.

[2] HYNDMAN, Rob J., et al. *Automatic time series for forecasting: the forecast package for R*. Monash University, Department of Econometrics and Business Statistics, 2007.

[3] Procházka Jiří a kol., *Insurance claims seasonality modelling*. Working paper, 2017.

[4] Procházka, Jiří, Flimmel, Samuel, Jantoš, Milan, Bašta, Milan. Long seasonal periods modeling. In: *AMSE 2016 (Applications of Mathematics and Statistics in Economics)* [online]. Banská Štiavnica,

31.08.2016 – 04.09.2016. Banská Bystrica : Občianske združenie Financ, 2016, s. 292–301. ISBN 978-80-89438-04-4. ISSN 2453-9902.
Dostupné z: http://amse.umb.sk/proceedings/ProchazkaFlimmelJantosBasta.pdf.

[5] RAMSAY, James O.; SILVERMAN, Bernard W. *Applied functional data analysis: methods and case studies*. New York: Springer, 2002.

[6] RUPPERT, David; WAND, Matt P.; CARROLL, Raymond J. Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics 12. *Cambridge: Cambridge Univ. Press. Mathematical Reviews (MathSciNet): MR1998720*, 2003.

JEL Classification: C32, C53

Porovnání kalibrace modelů úmrtnosti na českých datech

Petr Sotona

petr.sotona@seznam.cz

Doktorand oboru statistika

Školitel: doc. RNDr. Luboš Marek, CSc., (marek@vse.cz)

Abstrakt: Článek porovnává stochastické modely úmrtnosti při aplikaci na česká historická data od roku 1920. První část uvádí přehled uvažovaných modelů a kalibračních metod. Druhá část poskytuje kvalitativní a kvantitativní metody použité pro porovnání uvedených modelů úmrtnosti. Porovnání je ukázáno z hlediska kvality modelu replikovat historický vývoj úmrtnosti (kvalita kalibrace), smysluplnosti výsledků, efektivity, úplnosti a robustnosti modelu. V závěru článku jsou shrnuté výsledky a uvedeny hlavní závěry porovnání modelů úmrtnosti na českých datech.

Klíčová slova: kalibrace modelů, stochastické modely úmrtnosti, úmrtnost.

Úvod

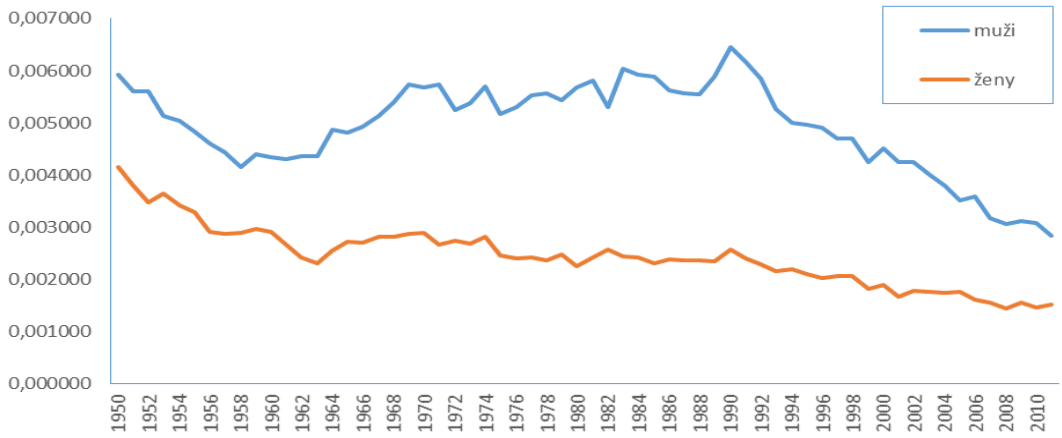
Cílem tohoto článku je použití teorie týkající se kalibrace modelů úmrtnosti, popsané např. v Pitacco et al. [5] a Wilmoth [7], na historická data o úmrtnosti v České republice. Nejprve odhadneme hodnoty parametrů porovnávaných modelů a následně modely na základě dosažených výsledků porovnáme pomocí kvantitativních i kvalitativních měr popsaných např. v Cairns et al. [2].

Pro výpočet použijeme data pro Českou republiku dostupná na internetu[1] (počty úmrtí a expozice pro muže a ženy, pro věky 0-95 a kalendářní roky 1920-2011), která ověříme s údaji Českého statistického úřadu.

Kalibraci modelů provedeme pro tři různé časové intervaly. V prvním případě uvažujeme všechna dostupná data, tj. období 1920 - 2011 s vyloučením let 1938 - 1944, pro které nejsou data k dispozici. Jako druhé období je zvoleno poválečné období 1950 - 2011. Tento časový horizont by měl být z velké části očištěn o vliv světových válek. Nicméně pro odstranění celkového vlivu válek by bylo potřeba vyloučit více let po konci druhé světové války. Přesto věříme, že vyloučení dat do roku 1949 včetně pomůže významně redukovat vliv válečného období. Poslední období představuje novou éru České republiky po Sametové revoluci, kdy uvažujeme kalendářní roky 1990 - 2011. Jak je vidět na obrázku 1 ukazujícím vývoj pravděpodobnosti úmrtí od roku 1950 pro muže a ženu ve věku 45 let, změna politické a sociální situace s pádem komunistického režimu se projevila i v poklesu úmrtnostních měr.

Pro každý ze zvolených časových intervalů lze najít důvod, proč může být kalibrace provedena právě na daném období. Nejkratší interval 1990 - 2011 se hodí pro modely, které jsou určené pro krátkodobé projekce, řekněme do deseti let. Delší časový interval 1950 - 2011 je vhodný pro dlouhodobé projekce, které by mohly být použity v pojišťovnách, penzijních společnostech, nebo jiných institucích. V těchto projekcích úmrtnosti nepotřebují (nechtějí) zohlednit války a další katastrofické události, jelikož mohou být tyto příčiny úmrtí vyčleněny z pojistného krytí. Nejdelší časové období 1920 - 2011 má využití pro dlouhodobé projekce, kde je vhodné zachytit vývoj úmrtnosti celé populace. Typické využití takových projekcí je ve státních studiích (vlády, ministerstva, statistické úřady), nemocnicích a mezinárodních projekcích. Projekce založené na nejdelším pozorovaném období jsou samozřejmě velmi důležité i pro pojišťovny a jiné soukromé instituce, pokud potřebují modelovat skutečný stav svého portfolia na dlouhém horizontu. V takovém případě je potřeba zohlednit úbytky portfolia i z důvodů, které nejsou pojištěny, ale mají vliv na budoucí přijaté pojistné.

[1] Human Mortality Database, www.mortality.org



Obrázek 1: Vývoj pravděpodobnosti úmrtí od roku 1950 do roku 2010 pro muže a ženu ve věku 45 let

Základní přehled

Následující přehled ukazuje modely úmrtnosti a metody kalibrace, které v našem srovnání uvažujeme (v závorce je uveden použitý předpoklad pro počty úmrtí popsané níže)

- **Lee-Carterův model (M1a)** - model definovaný v (1), kalibrovaný pomocí metody SVD[2] (předpoklad normálního rozdělení);
- **Lee-Carterův model (M1b)** - model definovaný v (1), kalibrovaný pomocí metody WLS[3] (předpoklad normálního rozdělení);
- **Lee-Carterův model (M1c)** - model definovaný v (1), kalibrovaný pomocí metody MLE[4] (předpoklad Poissonova rozdělení);
- **Renshaw-Habermanův model (M2)** - model definovaný v (2), kalibrovaný pomocí metody MLE (předpoklad Poissonova rozdělení);
- **Currieho "APC" model (M3)** - model definovaný v (3), kalibrovaný pomocí metody MLE (předpoklad Poissonova rozdělení);
- **Cairns-Blake-Dowdův model (M4a)** - model definovaný v (4), kalibrovaný pomocí metody WLS (předpoklad normálního rozdělení);
- **Cairns-Blake-Dowdův model (M4b)** - model definovaný v (5), kalibrovaný pomocí metody WLS (předpoklad normálního rozdělení);
- **Cairns-Blake-Dowdův model (M4c)** - model definovaný v (6), kalibrovaný pomocí metody WLS (předpoklad normálního rozdělení);
- **Platův model (M5)** - model definovaný v (7), kalibrovaný pomocí metody MLE (předpoklad Poissonova rozdělení).

Zavedeme následující značení. Funkce $m(t,x)$ označuje hrubou míru úmrtnosti osoby ve věku x let v kalendářním roce t, která je definována jako podíl $D(t,x,)$ počtu zemřelých osob ve věku x v kalendářním roce t a $E(t,x)$ počtu žijících osob ve věku x a kalendářním roce t. Funkce $q(t,x)$ označuje roční pravděpodobnost úmrtí osoby ve věku x v kalendářním roce t a $p(t,x)$ značí doplňkovou pravděpodobnost přežití. Níže jsou zjednodušeně uvedena teoretická vyjádření jednotlivých modelů úmrtnosti

M1a, M1b, M1c:
$$\ln m(t,x) = \alpha_x^{(1)} + \beta_x^{(1)}\kappa_t^{(1)}; \quad (1)$$

M2:
$$\ln m(t,x) = \alpha_x^{(1)} + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\gamma_{t-x}^{(2)}; \quad (2)$$

M3:
$$\ln m(t,x) = \alpha_x^{(1)} + \kappa_t^{(1)} + \gamma_{t-x}^{(2)}; \quad (3)$$

[2]Z anglického Singular value decomposition
[3]Z anglického Weighted least squares
[4]Z anglického Maximum likelihood estimation

M4a:
$$\ln \frac{q(t,x)}{p(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)} x; \quad (4)$$

M4b:
$$\ln \frac{q(t,x)}{p(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)} (\bar{x} - x); \quad (5)$$

M4c:
$$\ln \frac{q(t,x)}{p(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)} (\bar{x} - x) + \gamma_{t-x}^{(3)}; \quad (6)$$

M5:
$$\ln m(t,x) = \alpha_x^{(1)} + \kappa_t^{(1)} + \kappa_t^{(2)} (\bar{x} - x) + \kappa_t^{(3)} (\bar{x} - x)^+ + \gamma_{t-x}^{(4)}. \quad (7)$$

Podrobnější definici $m(t,x)$ i $q(t,x)$ lze nalést např v Pitacco et al. [5].

Přehled kalibrace

Existuje mnoho metod kalibrace modelů na pozorovaná data. V našem srovnání modelů úmrtnosti jsme využili především metodu maximální věrohodnosti (MLE) s využitím iteračního postupu Newton-Raphsonova algoritmu, popsaného např. v Pitacco et al. [5].

Kalibraci jsme prováděli zvlášť pro muže a pro ženy a pro výše zmiňované tři časové intervaly. Z pozorovaných historických dat o počtech žijících a zemřelých osob v jednotlivých kalendářních letech jsme spočetli hrubé míry úmrtnosti $m(t,x)$ a pro modely M4a, M4b a M4c jsme použili aproximaci $q(t,x) \approx 1 - \exp(-m(t,x))$ pro odvození pravděpodobnosti úmrtí $q(t,x)$.

Pro kalibraci modelů jsme využili Microsoft Excel a Visual Basic Application, kde jsme implementovali Newton-Raphsonův algoritmus. Pro metodu SVD jsme využili doplněk MATRIX and LINEAR ALGEBRA Package For EXCEL[5] pro MS Excel obsahující maticové operace.

Parametry odvozené z metody SVD modelu M1a jsme použili jako počáteční odhady pro Newton-Raphsonův algoritmus pro kalibraci modelů M1b a M1c. Počáteční hodnoty pro ostatní modely jsou uvedeny níže

M2: $\alpha_x^{(1)} = \widehat{\alpha_x^{(1)}}(M1a) \quad \forall x; \qquad \beta_x^{(1)} = \widehat{\beta_x^{(1)}}(M1a) \quad \forall x;$

$\kappa_t^{(1)} = \widehat{\kappa_t^{(1)}}(M1a) \quad \forall t; \qquad \beta_x^{(2)} = \frac{1}{m} \quad \forall x; \qquad \gamma_c^{(2)} = 0 \quad \forall c;$

M3: $\alpha_x^{(1)} = \widehat{\alpha_x^{(1)}}(M1a) \quad \forall x; \qquad \kappa_t^{(1)} = 0 \quad \forall t; \qquad \gamma_c^{(2)} = 0 \quad \forall c;$

M4a: $\kappa_t^{(1)} = \kappa_t^{(2)} = 0 \quad \forall t;$

M4b: $\kappa_t^{(1)} = \kappa_t^{(2)} = 0 \quad \forall t;$

M4c: $\kappa_t^{(1)} = \kappa_t^{(2)} = 0 \quad \forall t; \qquad \gamma_c^{(3)} = 0 \quad \forall c;$

M5: $\alpha_x^{(1)} = \widehat{\alpha_x^{(1)}}(M1a) \quad \forall x; \qquad \kappa_t^{(1)} = \kappa_t^{(2)} = \kappa_t^{(3)} = 0 \quad \forall t; \qquad \gamma_c^{(4)} = 0 \quad \forall c.$

Výše uvedené hodnoty byly použity pro všechny tři časové intervaly a neměly by mít vliv na konečné odhady parametrů. Nicméně tyto počáteční odhady ovlivňují počet iterací Newton-Raphsonova algoritmu.

Kalibraci metodou vážených nejmenších čtverců WLS jsme aplikovali s váhami $w(t,x)$ rovnými počtu zemřelých $D(t,x)$. Jelikož se nejdená o exogenní váhy, existuje riziko vychýlených odhadů. Nicméně používáme stejné váhy, jako byly navrženy v původní studii Wilmoth [7].

Jako kritérium pro ukončení iteračního procesu jsme definovali, že změna v hodnotě parametrů je menší než 10^{-12}, nebo relativní změna věrohodnostní funkce je menší než 10^{-20}. Upozorňujeme, že volba těchto

[5]Dostupný na http://digilander.libero.it/foxes.

technických kritérií je subjektivní a že tyto hodnoty ovlivňují i konečné hodnoty odhadovaných parametrů. Dané hodnoty jsou založeny na subjektivním pozorování a konvergenci iteračního procesu Newton-Raphsonova algoritmu.

Vzhledem k neexistenci jednoznačného řešení některých modelů jsme použili následující dodatečné podmínky pro kalibraci modelů

M1a: $\sum_{x=x_1}^{x_m} \beta_x^{(1)} = 1, \quad \sum_{t=t_1}^{t_n} \kappa_t^{(1)} = 0;$

M1b: $\sum_{x=x_1}^{x_m} \beta_x^{(1)} = 1, \quad \sum_{t=t_1}^{t_n} \kappa_t^{(1)} = 0;$

M1c: $\sum_{x=x_1}^{x_m} \beta_x^{(1)} = 1, \quad \sum_{t=t_1}^{t_n} \kappa_t^{(1)} = 0;$

M2: $\sum_{x=x_1}^{x_m} \beta_x^{(1)} = 1, \quad \sum_{t=t_1}^{t_n} \kappa_t^{(1)} = 0, \quad \sum_{x=x_1}^{x_m} \beta_x^{(2)} = 1, \quad \sum_{c=t_1-x_m}^{t_n-x_1} \gamma_c^{(2)} = 0;$

M3: $\sum_{t=t_1}^{t_n} \kappa_t^{(1)} = 0, \quad \sum_{t=t_1}^{t_n} \sum_{x=x_1}^{x_m} \gamma_{t-x}^{(2)} = 0, \quad \min_{\delta} S(\delta),$

$$\text{kde} S(\delta) = \sum_{x=x_1}^{x_m} (\alpha_x^{(1)} + \delta(x - \bar{x}) - \sum_{t_1}^{t_n} \frac{ln\ m(t,x)}{n})^2;$$

M4a: Není problém s neexistencí jednoznačného řešení;

M4b: Není problém s neexistencí jednoznačného řešení;

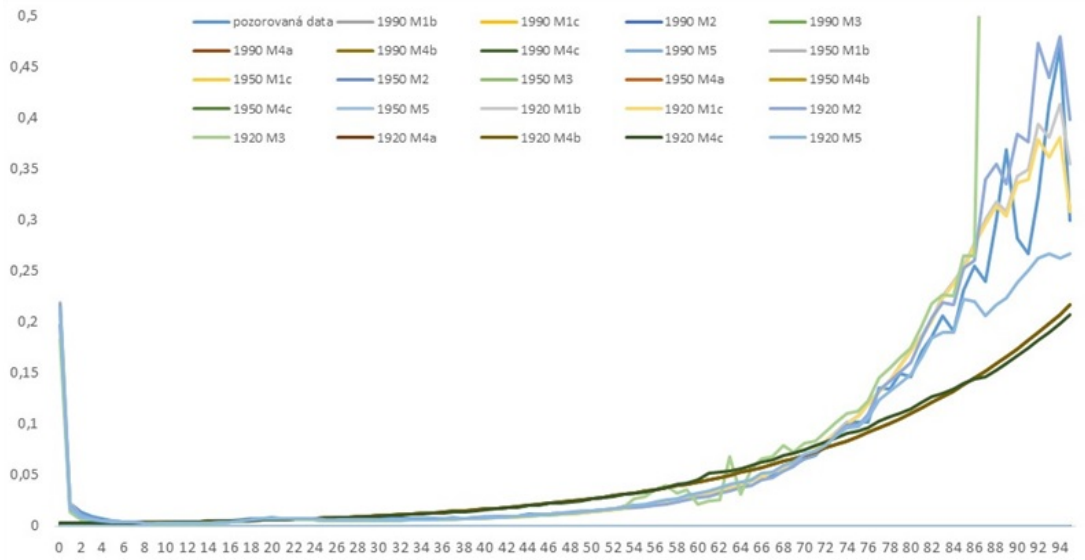
M4c: $\sum_{c=t_1-x_m}^{t_n-x_1} \gamma_c^{(3)} = 0, \quad \sum_{c=t_1-x_m}^{t_n-x_1} c\gamma_c^{(3)} = 0;$

M5: $\sum_{c=t_1-x_m}^{t_n-x_1} \gamma_c^{(4)} = 0, \quad \sum_{c=t_1-x_m}^{t_n-x_1} c\gamma_c^{(4)} = 0, \quad \sum_{t=t_1}^{t_n} \kappa_t^{(3)} = 0.$

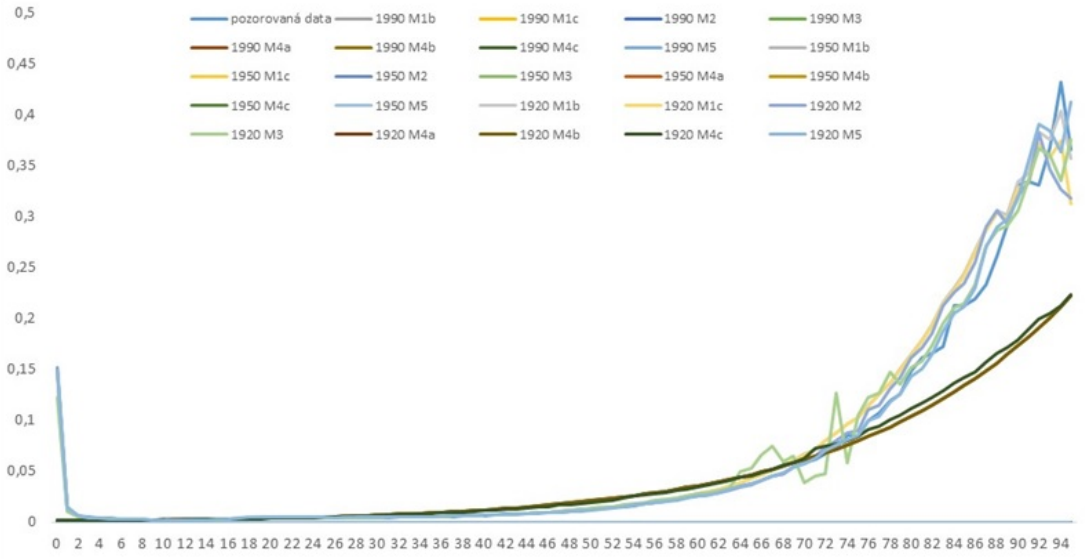
Před vlastním srovnáním modelů ještě poznamenáme podmínky pro zohlednění efektu kohort. Efekt kohort uvažujeme pouze v případě, kdy je alespoň deset dostupných pozorování a tato pozorování jsou minimálně pro věk padesát a výše. První podmínkou se vyhneme možnému zkreslení modelu díky volatilitě při malém počtu pozorování a druhou podmínku zavádíme z předpokladu, že tento efekt dle roku narození by se měl projevit spíše až ve vyšším věku, než v dětství a rané dospělosti. Při kalibraci na časový interval 1920 - 2011 jsme navíc vyřadili roky narození 1938 - 1944, jelikož pro dané kalendářní roky nebyla data k dispozici.

Výsledky kalibrace

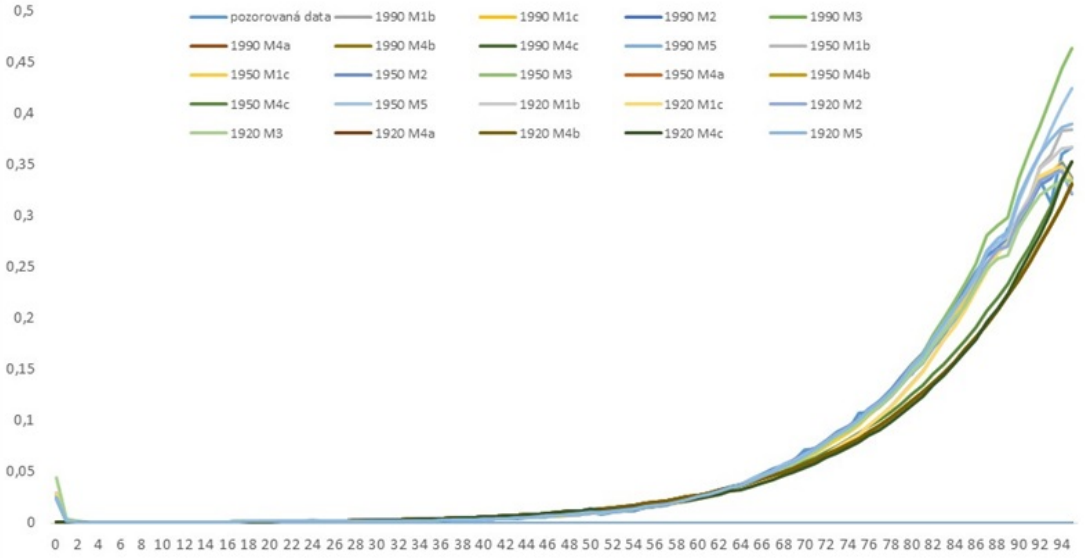
Předtím, než porovnáme modely podle jednotlivých měr, ukážeme zde výsledky kalibrace jednotlivých modelů. Níže uvedené obrázky 2 - 8 ukazují pozorované a odhadnuté míry úmrtnosti pro muže v rámci kalibrace na tři časové intervaly (konkrétně jsou uvedeny grafy celého věkového rozpětí pro kalendářní roky 1920, 1930, 1970 a 2000 a dále pro věky 40, 60 a 80 let pro jednotlivá období. Analogické grafy pro ženy ukazují podobné výsledky, proto je zde již neuvádíme.



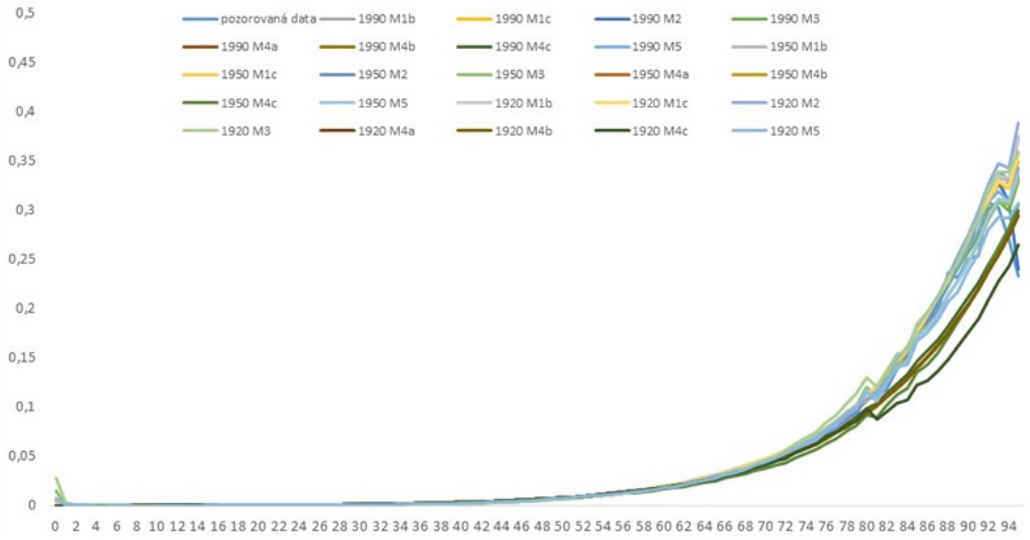
Obrázek 2: Pozorované a kalibrované míry úmrtnosti mužů pro kalendářní rok 1920



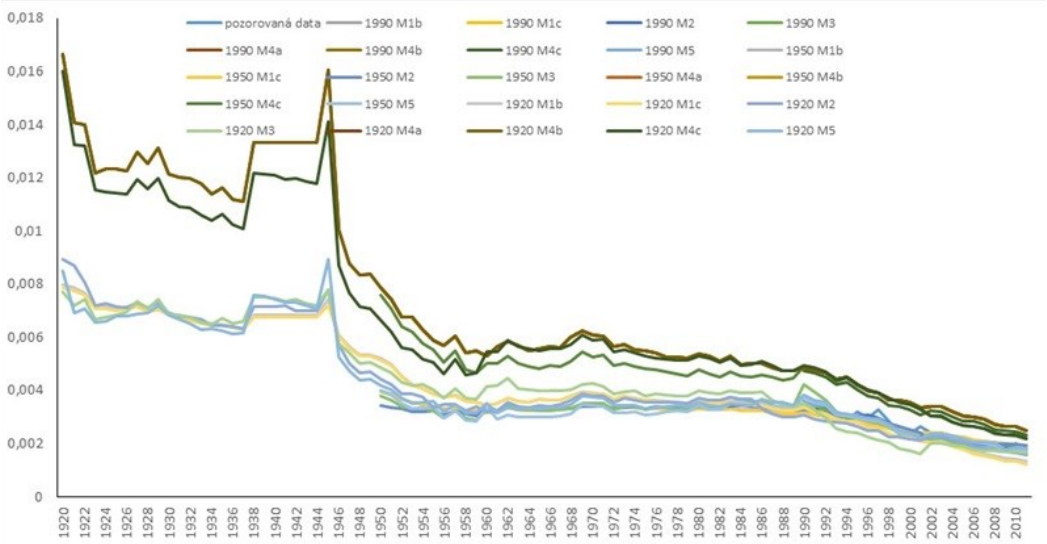
Obrázek 3: Pozorované a kalibrované míry úmrtnosti mužů pro kalendářní rok 1930



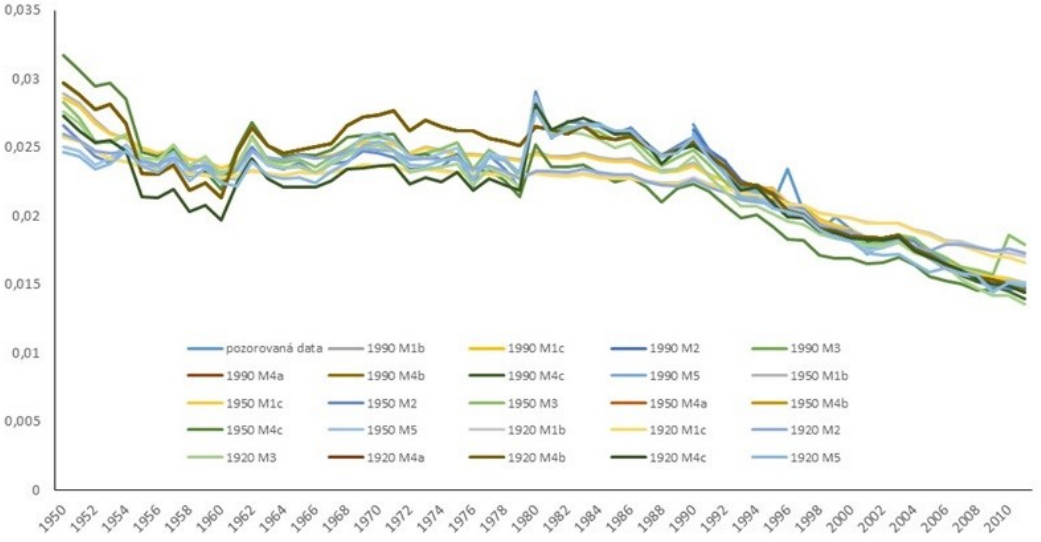
Obrázek 4: Pozorované a kalibrované míry úmrtnosti mužů pro kalendářní rok 1970



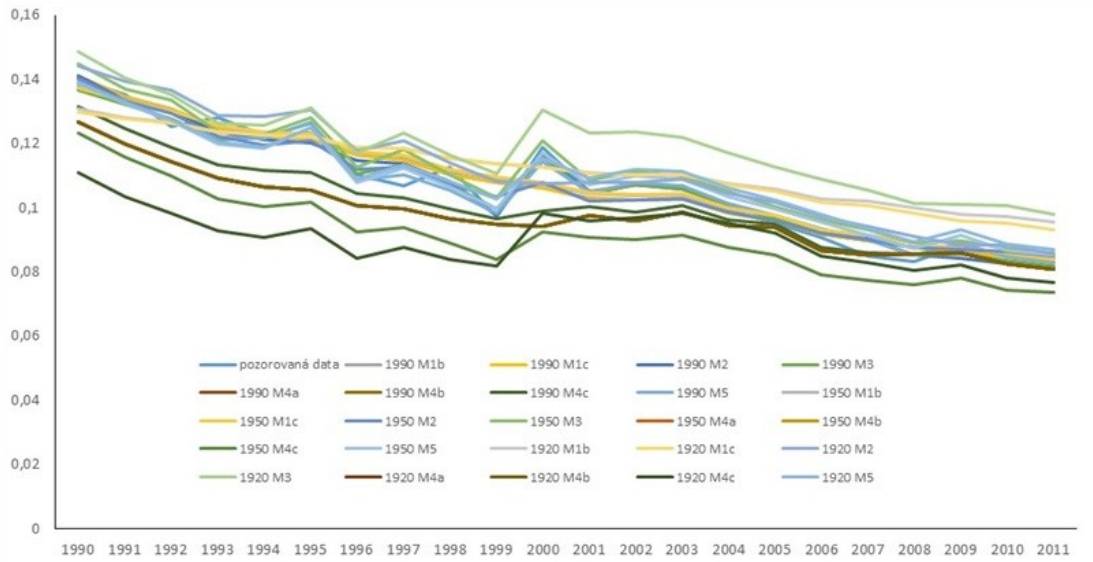
Obrázek 5: Pozorované a kalibrované míry úmrtnosti mužů pro kalendářní rok 2000



Obrázek 6: Pozorované a kalibrované míry úmrtnosti pro muže ve věku 40 let v kalendářních letech 1920 - 2011



Obrázek 7: Pozorované a kalibrované míry úmrtnosti pro muže ve věku 60 let v kalendářních letech 1950 - 2011



Obrázek 8: Pozorované a kalibrované míry úmrtnosti pro muže ve věku 80 let v kalendářních letech 1990 - 2011

Z grafů vidíme, že většina modelů replikuje pozorovaná data velmi přesně. To platí obzvláště v posledních letech, řekněme od roku 1990. Při kalibraci na delší časové období se ale více začínají projevovat rozdíly mezi modely a ukazovat jejich slabé stránky. Na první pohled nejhůře dopadly modely CBD (M4a, M4b, M4c). To je dáno strukturou těchto modelů, která je založena na modelování úmrtností pro užší věkové rozmezí (pro modelování důchodového pojištění). Při použití CBD modelů na celém věkovém rozpětí tak dostáváme horší výsledky. Vysokou nepřesnost dostáváme pro věk 0 a pro vysoké věky, kde je v pozorovaných datech vyšší míra volatility.

Kvalita kalibrace

Jako první porovnáme modely z hlediska kvality kalibrace jednotlivých modelů na historická data. První mírou je dosažená hodnota věrohodnostní funkce. Hodnoty jsou uvedeny bez konstantní části této funkce a pouze pro modely odhadnuté metodou maximální věrohodnosti MLE. Tabulka 1 ukazuje výsledky pro období 1990 - 2011.

Tabulka 1: Maximální věrohodnost modelů kalibrovaných na období 1990 - 2011

Model	Věrohodnost (muži)	Věrohodnost (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	-5 262 333	-4 857 121	3	3
M2	-5 261 843	-4 856 562	1	1
M3	-5 262 930	-4 857 847	4	4
M5	-5 262 217	-4 856 857	2	2

Nejvyšší hodnoty věrohodnostní funkce dosahuje model M2 a nejnižší hodnoty naopak model M3. To platí pro muže i ženy.

Následující tabulka 2 ukazuje pro srovnání výsledky při kalibraci modelů na období 1950 - 2011.

Tabulka 2: Maximální věrohodnost modelů kalibrovaných na období 1950 - 2011

Model	Věrohodnost (muži)	Věrohodnost (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	-15 219 292	-13 906 960	3	3
M2	-15 209 622	-13 904 360	1	1
M3	-15 224 184	-13 915 063	4	4
M5	-15 210 146	-13 904 888	2	2

Nejvyšší hodnoty dosáhl opět model M2 následovaný modelem M5 a M1c. Nejhůře opět dopadl model M3. Výsledky jsou tedy stejné jako v případě kratší historie a jsou primárně dané počtem odhadovaných parametrů v jednotlivých modelech úmrtnosti. V další části článku ukážeme i srovnání dle počtu parametrů a také kritéria BIC, které zohledňuje dosaženou maximální věrohodnost i počet parametrů modelu.

Pro období 1920 - 2011 již tabulku s dosaženou věrohodností neuvádíme. Nicméně dosažené výsledky jsou podobné těm, které jsou zde uvedeny.

Kvalitu kalibrace porovnáme ještě pomocí míry MAPE[6]. Tabulka 3 dále ukazuje srovnání pro období 1990 - 2011.

Tabulka 3: MAPE modelů úmrtnosti kalibrovaných na data 1990 - 2011

Model	MAPE (muži)	MAPE (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	9,1 %	11,8 %	2	3
M2	8,6 %	11,3 %	1	1
M3	10,7 %	14,1 %	4	4
M4a, M4b	19,9 %	27,0 %	6	6
M4c	17,8 %	26,0 %	5	5
M5	9,5 %	11,7 %	3	2

Podobně jako v předchozím případě vidíme, že se na předních místech pohybují modely s více parametry. CBD modely jsou na chvostu a dávají výrazně horší výsledky oproti ostatním modelům. Nejlepší výsledek má model M2 (pro muže i ženy), následuje M5 a M1c pro ženy, resp. M1c a M5 pro muže. Model M3 je opět na čtvrtém místě.

Pro srovnání ukazujeme v tabulce 4 stejné porovnání modelů na pozorovaném období 1950 - 2011 (výsledky na nejdelším období 1920 - 2011 opět neuvádíme s odvoláním na podobné závěry, jako ukazuje následující tabulka).

Tabulka 4: MAPE modelů úmrtnosti kalibrovaných na data 1950 - 2011

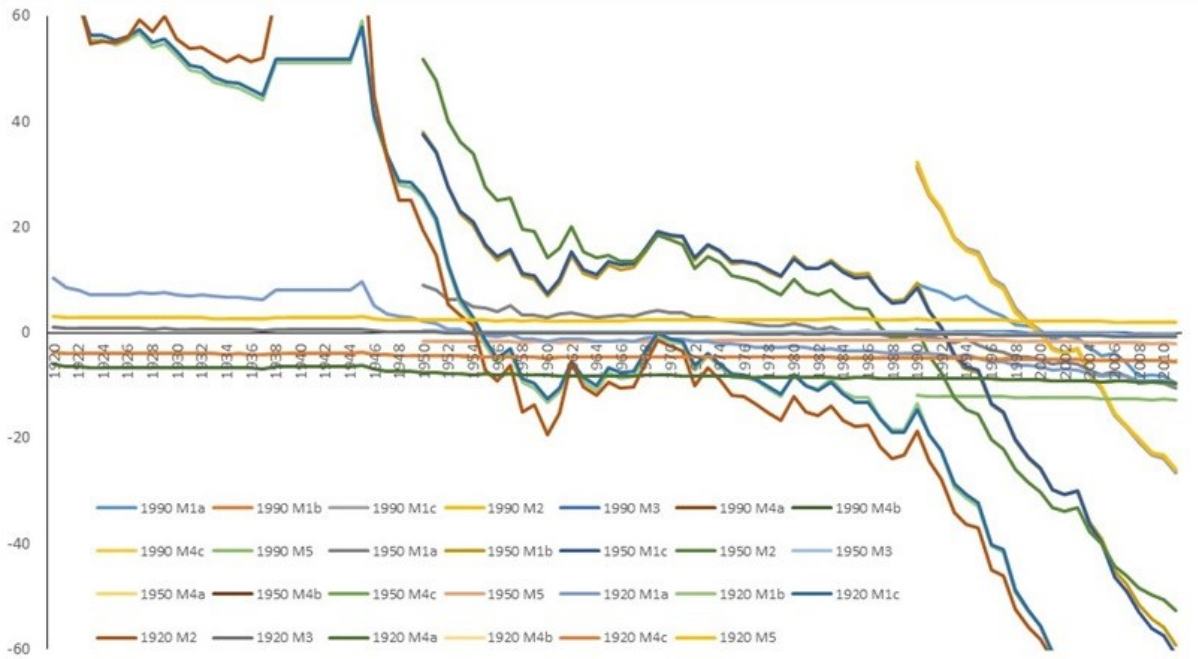
Model	MAPE (muži)	MAPE (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	12,2 %	11,8 %	3	3
M2	8,8 %	10,4 %	1	1
M3	15,4 %	17,0 %	4	4
M4a, M4b	29,5 %	32,5 %	5	5
M4c	35,0 %	43,6 %	6	6
M5	9,3 %	10,6 %	2	2

Výsledky se oproti kratšímu období liší jen minimálně. Nejlépe dopadl model M2 následovaný modely M5 a M1c s modelem M3 stále na posledním místě pokud nepočítáme CBD modely, které jsou snad ještě výrazněji za ostatními modely.

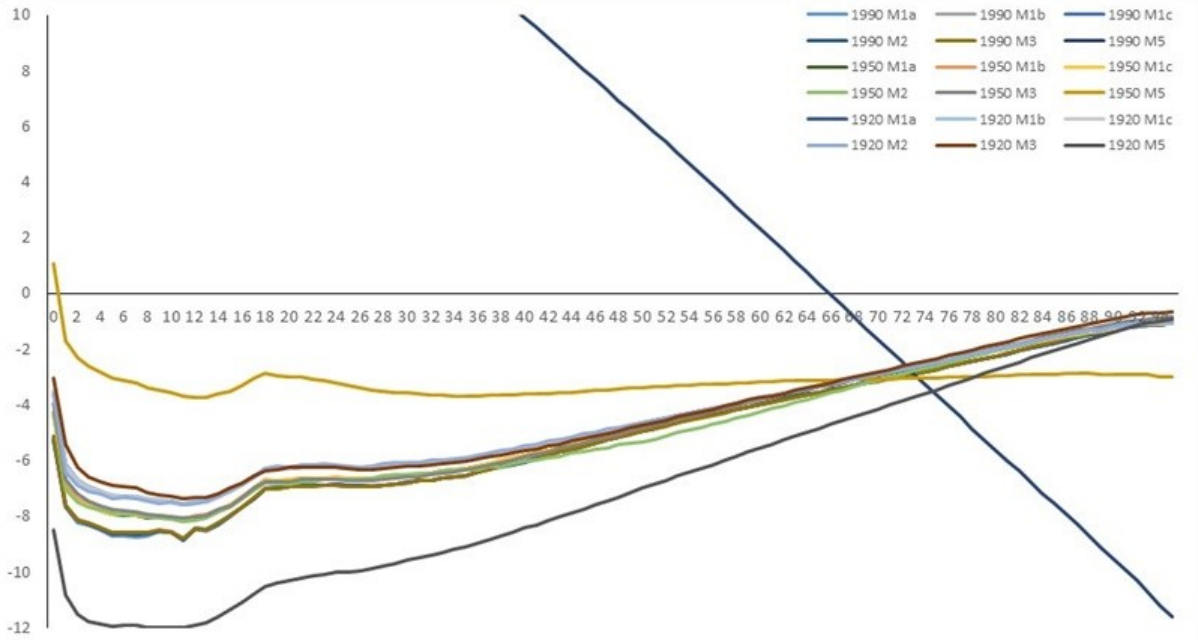
Smysluplnost

Smysluplnost modelů může být ohodnocena např. logickým zdůvodněním / interpretací hodnot parametrů a jejich pochopení vzhledem k daným biologickým a socio-ekonomickým skutečnostem. Na obrázcích 9 a 10 jsou dva hlavní typy parametrů v porovnávaných modelech úmrtnosti.

[6]Z angl. Mean Absolute Percentage Error, definice např. na www.wikipedia.org



Obrázek 9: Hodnoty parametrů $\kappa_t^{(1)}$ pro muže, jednotlivé modely a kalibrační období



Obrázek 10: Hodnoty parametrů $\alpha_x^{(1)}$ pro muže, jednotlivé modely a kalibrační období

Prvním pozorováním je, že u všech modelů můžeme pozorovat klesající trend u časově závislého parametru κ_t, což koresponduje s poklesem pravděpodobnosti úmrtí během posledních dvaceti, šedesáti, resp. devadesáti let.

Druhý závěr je velmi podobný tvar parametrů α_x, který odpovídá vývoji úmrtnosti vzhledem k věku jedince. Jedinou výjimkou je model M5, kde pozorujeme odlišný tvar, což může poukazovat na přeparametrizovanost tohoto modelu a kompenzaci základního efektu jinými parametry.

V dalším porovnání se zaměříme na předpoklad Poissonova rozdělení počtu úmrtí, který jsme použili pro modely odhadnuté metodou maximální věrohodnosti. K tomu využijeme standardizovaná rezidua $z(t,x)$, definovaná např. v Cairns et al. [2]. Za předpokladu, že počty úmrtí mají Poissonovo rozdělení, měly by tato rezidua být nezávislé a stejně rozdělené veličiny se standardním normálním rozdělením.

Následující tabulka 5 ukazuje střední hodnotu a rozptyl uvedených standardizovaných reziduí $z(t,x)$ při kalibraci modelů na data z období 1990 - 2011.

Tabulka 5: Standardizovaná rezidua $z(t,x)$ modelů úmrtnosti kalibrovaných na roky 1990 - 2011

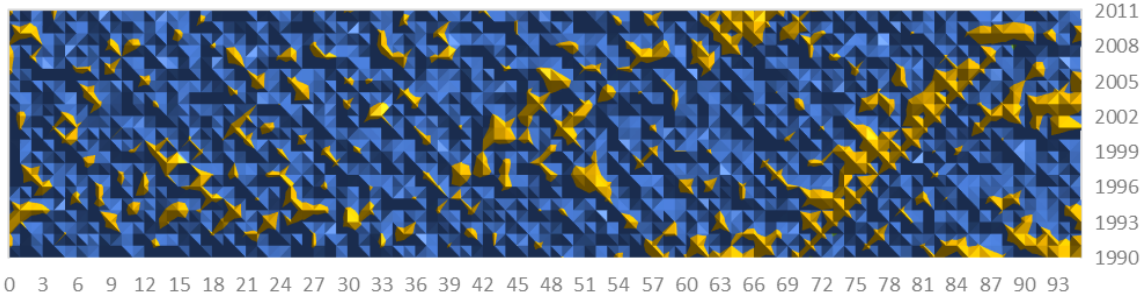
Model	Stř. hodnota (muži)	Rozptyl (muži)	Stř. hodnota (ženy)	Rozptyl (ženy)
M1c	-0,01	1,60	-0,01	1,54
M2	-0,01	1,13	0,00	1,01
M3	0,33	5,68	0,02	2,50
M5	0,04	1,50	0,07	1,31

Z uvedených hodnot vidíme, že standardizovaná rezidua $z(t,x)$ modelu M2 se svými hodnotami prvních dvou momentů nejvíce blíží momentům standardního normálního rozdělení. Ostatní modely vykazují výraznější rozdíly. Pro statistické zhodnocení jsme použili Kolmogovor-Smirnovův test na hladině spolehlivosti 95%. Tento test v případě všech modelů uvedených v tabulce 5 zamítl nulovou hypotézu, že tato standardizovaná rezidua $z(t,x)$ mají standardní normální rozdělení.

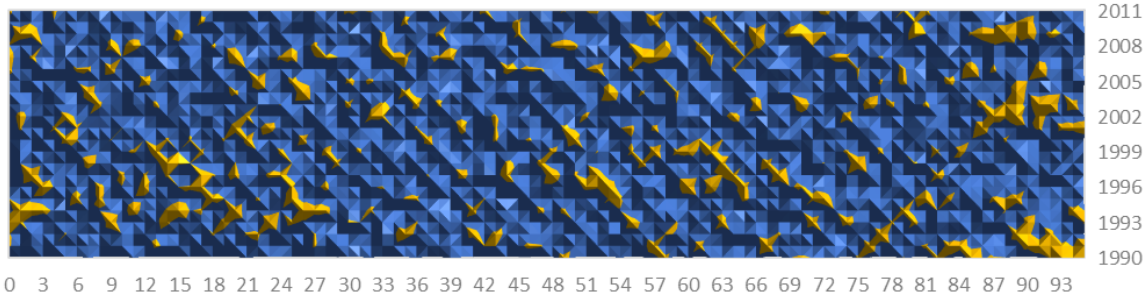
Pro delší kalibrované období jsme získali analogické výsledky na základě porovnání momentů i samotného otestování normality rozdělení.

Závěrem tohoto srovnání je, že se při použití na česká data nepotvrdil předpoklad Poissonova rozdělení počtu úmrtí. Nicméně podle článku Cairns et al. [2] je toto zjištění analogické pro mnoho zemí. V jejich práci, kde porovnávají vybrané modely na datech z Anglie, Walesu a Spojených států amerických, mají získaná standardizovaná rezidua ještě vyšší hodnoty volatility než v našem případě při aplikaci na česká data. Cairns et al. [2] ve své práci tvrdí, že to nemá vliv na odhady budoucího vývoje úmrtnosti, ale předpoklad Poissonova rozdělení může vést k podhodnocení budoucí variability měr úmrtnosti při použití skutečných základních pravděpodobností úmrtí.

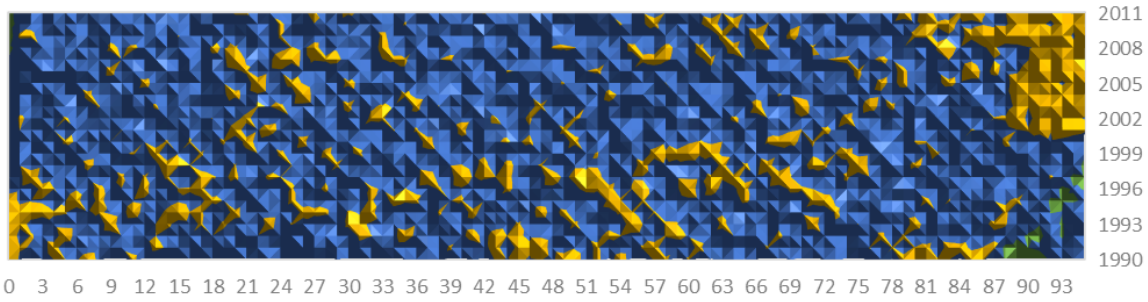
Dále se podíváme na předpoklad nezávislosti a stejného rozdělení standardizovaných reziduí $z(t,x)$ za pomocí povrchových grafů těchto reziduí. Za tohoto předpokladu bychom neměli pozorovat žádné shlukování, nebo systematické chování. Obrázky 11 - 14 ukazují standardizovaná rezidua modelů M1c, M2, M3a M5 pro muže a kalibrační období 1990 - 2011.



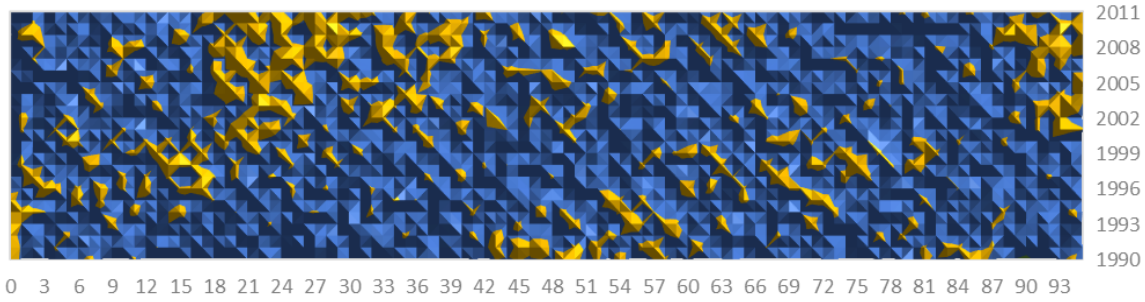
Obrázek 11: Hodnoty standardizovaných reziduí $z(t,x)$ modelu M1c pro muže a kalibrační období 1990 – 2011



Obrázek 12: Hodnoty standardizovaných reziduí $z(t,x)$ modelu M2 pro muže a kalibrační období 1990 - 2011



Obrázek 13: Hodnoty standardizovaných reziduí $z(t,x)$ modelu M3 pro muže a kalibrační období 1990 – 2011



Obrázek 14: Hodnoty standardizovaných reziduí $z(t,x)$ modelu M5 pro muže a kalibrační období 1990 - 2011

Pro ženy jsme odvodili velmi podobné výsledky. Celkově můžeme říct, že se rezidua chovají náhodně až na model M1c, kde je viditelná diagonální závislost, která ukazuje na chybějící zohlednění roku narození při kalibraci (tj. zohlednění kohortního efektu).

Efektivita

Efektivitou modelu myslíme především počet odhadovaných parametrů v souvislosti s vylepšením daného modelu, resp. s obtížností jeho kalibrace.

Nejprve se podíváme na prostý počet odhadovaných parametrů každého modelu. Následující tabulka 6 ukazuje počet parametrů pro všechny tři kalibrační období a pořadí jednotlivých modelů. Poznamenáváme, že v případě parametrů závislých na roku narození (kohortních parametrů) jsme uvažovali pouze roky, kde bylo alespoň deset pozorování a zároveň data alespoň pro věk padesát. To podporuje myšlenku, že by tento efekt měl ovlivňovat úmrtnost spíše ve vyšších věcích než v mládí a dospělosti. Z nejdelšího období 1920 - 2011 jsme vyloučili i roky narození 1938 - 1944, pro které nebyla dostupná podkladová data.

Tabulka 6: Počet odhadovaných parametrů jednotlivých modelů úmrtnosti

Období	1990 - 2011		1950 - 2011		1920 - 2011	
Model	Počet	Pořadí	Počet	Pořadí	Počet	Pořadí
M1a, M1b, M1c	214	4	254	3	284	2
M2	368	6	448	6	501	6
M3	176	3	256	4	309	3
M4a, M4b	44	1	124	1	184	1
M4c	102	2	222	2	312	4
M5	220	5	380	5	493	5

V tabulce snadno vidíme, že nejnižší počet parametrů má model M4a, resp. M4b. Rozšířená verze CBD modelu, model M4c, skončila na druhém místě až na nejdelší kalibrační období, kde se před tento model dostal LC model i APC model. Naopak nejhůře z hlediska počtu parametrů dopadl model M2 a jen o něco lépe skončil model M5. Zejména pro kratší časové intervaly se zdá být model M2 přeparametrizovaný vzhledem k ostatním modelům. S rostoucí dobou pozorování významně roste počet parametrů u modelu M5, což je dáno vysokým počtem parametrů závislých na čase.

Nyní porovnáme modely vzhledem ke kvalitě kalibrace se zohledněním počtu parametrů, konkrétně použijeme Bayesovo informační kritérium BIC, jehož definici je možné najít např. v Cairns et al. [2]. Tabulka 7 níže shrnuje BIC pro modely odhadnuté metodou maximální věrohodnosti v rámci kalendářních let 1990 - 2011.

Tabulka 7: BIC modelů kalibrovaných přes období 1990 - 2011

Model	BIC (muži)	BIC (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	-5 263 152	-4 857 941	2	2
M2	-5 263 251	-4 857 970	3	3
M3	-5 263 604	-4 858 520	4	4
M5	-5 263 059	-4 857 699	1	1

Při porovnání modelů z hlediska dosažené maximální věrohodnosti byl nejlepší model M2. Při zohlednění počtu parametrů klesl model M2 na třetí místo. Nejlépe se umístil model M5 následovaný modelem M1c. Modelu M3 nepomohl ani nejnižší počet parametrů (mezi modely odhadnutými metodou maximální věrohodnosti) a skončil nejhůře.

Analogické srovnání kalibrace modelů na datech z období 1950 - 2011 ukazuje následující tabulka 8.

Tabulka 8: BIC modelů kalibrovaných přes období 1950 - 2011

Model	BIC (muži)	BIC (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	-15 220 395	-13 908 063	3	3
M2	-15 211 569	-13 906 307	1	1
M3	-15 225 297	-13 916 176	4	4
M5	-15 211 797	-13 906 539	2	2

Při prodloužení pozorovaného období pozorujeme výměnu pořadí mezi modely M1c, M2 a M5. Model M3 skončil i v tomto případě na posledním místě. Změna pořadí je způsobena strukturou jednotlivých modelů v kombinaci se prodloužením časového intervalu a tím i zvyšováním počtu časově závislých parametrů.

Pro srovnání (v následující tabulce 9) uvádíme toto kritérium jednotlivých modelů i při kalibraci na nejdelší období 1920 - 2011.

Tabulka 9: BIC modelů kalibrovaných přes období 1920 - 2011

Model	BIC (muži)	BIC (ženy)	Pořadí (muži)	Pořadí (ženy)
M1c	-24 361 270	-22 382 106	3	3
M2	-24 341 992	-22 377 936	2	1
M3	-24 449 956	-22 483 332	4	4
M5	-24 338 328	-22 378 878	1	2

Oproti období 1950 - 2011 došlo jen k prohození pořadí mezi modelem M2 a M5 u mužů. U žen zůstalo pořadí zachováno. Celkově tedy můžeme shrnout toto srovnání tím, že nejhůře je na tom model M3 a nejlépe modely M2 a M5.

V rámci efektivity modelů se ještě podíváme na strukturu jednotlivých modelů. V našem seznamu modelů úmrtnosti jsou některé modely speciálním případem, resp. zobecněním jiného modelu na seznamu. K porovnání efektivity mezi těmito modely slouží test poměru věrohodností (popsán např v Cairns et al. [2]), který testuje nulovou hypotézu, že vnořený model (speciální případ jiného modelu) je pro kalibraci vhodnější, oproti alternativní hypotéze, že zobecněný model je ten správný pro modelování úmrtnosti při kalibraci na daná data.

Na našem seznamu nalezneme šest párů takovýchto modelů. Následující tabulka 10 tyto dvojce shrnuje.

Tabulka 10: Dvojce vnořených a zobecněných modelů

Vnořený model	Zobecněný model
M1c	M2
M3	M2
M3	M5
M4a	M4b
M4a	M4c
M4b	M4c

Test poměru věrohodností zamítl nulovou hypotézu pro muže i ženy a všechny tři kalibrační období až na dvě výjimky.

První výjimkou je dvojce modelů M4a vs. M4b kde je dosažená věrohodnost stejná pro oba modely a oba modely obsahují stejný počet parametrů (platí pro obě pohlaví i všechna tři kalibrační období). Proto můžeme říci, že jde o ekvivalentní modely, které se liší pouze koeficienty u těchto odhadovaných parametrů $\kappa_t^{(2)}$.

Druhou výjimkou je dvojce M4a vs. M4c, resp. M4b vs. M4c (viz. první výjimka) pro muže a kalibrační období 1990 - 2011, kde je dosažená věrohodnost vyšší pro vnořený model než pro zobecněný model. Na základě testu poměru věrohodností nezamítáme nulovou hypotézu a model M4a, resp. M4b se zdá být vhodnější pro modelování úmrtnosti mužů s daty za období 1990 - 2011.

Robustnost

V této části porovnáme modely vzhledem k jejich robustnosti, kterou posoudíme pomocí tří pozorovacích období (1920 - 2011, 1950 - 2011, 1990 - 2011). Pro ilustraci odkazujeme na obrázky 9 a 10.

V případě modelů M1a, M1b, M1c, M2 a M3 se zdají být hodnoty kalibrovaných parametrů robustní a neliší se výrazně pro jednotlivá kalibrační období. Pro CBD modely M4a a M4b jsou hodnoty časově závislých parametrů identické pro dané kalendářní roky v rámci všech tří období, což je založeno na faktu, že je kalibrace těchto parametrů provedena zvlášť pro každý kalendářní rok. Naopak u modelu M5 pozorujeme velké změny v hodnotách parametrů $\alpha_x^{(1)}$, $\kappa_t^{(1)}$ a $\kappa_t^{(2)}$ pro muže i pro ženy při změně kalibračního období. To ukazuje na nízkou robustnost tohoto modelu. Tyto změny v parametrech jsou pravděpodobně způsobeny vysokým počtem různých parametrů a jejich vzájemnou závislostí.

Úplnost

Všechny modely až na CBD modely (M4a, M4b, M4c) jsou vhodné pro použití modelování úmrtnosti v rámci celého věkového rozpětí. CBD modely byly vytvořeny pro modelování pravděpodobnosti úmrtí osob pobírajících starobní důchod. Nevhodnost použití CBD modelů na celé věkové rozpětí je podpořeno výsledky předchozího srovnávání, kde je vidět, že tyto modely jsou daleko za ostatními modely.

Zohlednění kohortního efektu může být rovněž použito k měření úplnosti modelu. Z definice každého modelu snadno vidíme, který model tento efekt zohledňuje a který ne. V našem srovnání mezi modely, které kohortní efekt berou v úvahu, patří M2, M3, M4c a M5. Lee-Carterovy modely (M1a, M1b, M1c) a základní CBD modely (M4a, M4b) naopak parametry podle roku narození neobsahují.

Diagonální závislosti v povrchových grafech standardizovaných reziduí $z(t,x)$ ukazují, že je v českých datech tento efekt pozorovatelný a je vhodné ho tedy zohlednit i v modelování úmrtnosti.

Shrnutí porovnání modelů úmrtnosti

Na základě provedeného srovnání můžeme určit model, který je nejlepší dle zvolených kritérií. Nicméně právě volba různých kritérií (měr) a především vah jednotlivých kritérií zásadně ovlivní celkový výsledek srovnání. Subjektivita při volbě konkrétních kritérií a jejich významnost při srovnání by měla vycházet z účelu, pro jaký potřebujeme úmrtnost modelovat. Jak již bylo uvedeno v úvodu tohoto článku, např. délka potřebné projekce může hrát důležitou roli při hodnocení modelů. Je zde potřeba poznamenat, že v článku jsme modely nehodnotili dle kvality projekcí daných modelů, což je další, možná nejdůležitější, kritérium hodnocení modelů. Další klíčový aspekt je, pro jaké věky potřebujeme úmrtnost modelovat. Jak jsme mohli vidět z našeho srovnání, CBD modely nelze využít při modelování celého věkového rozhraní. Nicméně, pokud bychom uvažovali pouze věkové rozmezí, pro než jsou tyto modely vytvořené, CBD modely by naopak mohly být nejvhodnější variantou. Pro tento případ odkazujeme např. na studii Cairns et al. [2], kde jsou i další rozšíření modelů CBD a srovnání těchto modelů s ostatními modely pro věky 60-89 let.

V našem srovnání na věkovém rozmezí 0-95 let při aplikaci na česká data a při zohlednění tří pozorovacích období bychom pro modelování úmrtnosti doporučili především model M2 navržený pány Renshawem a Hebermanem. Důvodem je jeho univerzálnost v rámci různých kalibračních období, použitelnost na celé věkové rozhraní a zachycení kohortního efektu, který lze v českých datech pozorovat. Komplexní model M5 se zdá být pro česká data již příliš podrobný a na použitých datech vykazoval známky přeparametrování. Proto bychom tento model zvolili spíše na populacích, kde je možné pozorovat veškeré efekty, které je schopný tento model zachytit. Opačný problém je u modelu M3, který má jednoduchou strukturu, ale zdá se být nedostatečný pro zachycení historického vývoje úmrtnosti v České republice. To je podloženo faktem, že tento model vykazoval ve většině případů nejhorší výsledky v porovnání s ostatními modely (až na modely CBD). Lee-Carterův model M1 je dobrou variantou v případě, že kohortní efekt nepovažujeme za významný, nebo chceme použít v praxi

lehce aplikovatelný model. Tento model bychom upřednostnili, pokud bychom potřebovali krátkodobé projekce založené na krátkodobé historii, kdy není kohortní efekt tak významný.

Literatura

[1] van Broekhoven, H.: *Market Value of Liabilities Mortality Risk: A Practical Model*, North American Journal, Volume 6, Number 3, str. 95-106, 2002.

[2] CAIRNS, A.J.G., BLAKE, D., DOWD, K. *A Quantitative Comparison of Stochastic Mortality Models Using Data from England and Wales and the United States*. North American Actuarial Journal, Vol.13, No.1, 2009.

[3] HABERMAN, S., RENSHAW, A. *Modelling Dynamics in Mortality Rates: Comparison of Forecasting Models*. KCL Financial Mathematics Seminar, Cass Business School, City University London, 22 March 2011.

[4] HUNT, A., BLAKE, D. *A General Procedure for Constructing Mortality Models*. Pension Institute, 2013.

[5] PITACCO, E., DENUIT, M., HABERMAN, S., OLIVIERI, A. *Modelling Longevity Dynamics for Pensions and Annuity Business*. Oxford University Press, 2009.

[6] PLAT, R. *On stochastic mortality modeling*. Insurance: Mathematics and Economics, Vol.45,p.123-132, 2009.

[7] WILMOTH, J.R. *Computational methods for fitting and extrapolating the Lee-Carter model of mortality change*. Technical report, 1993.

JEL Classification: C60. J10

Summary

Comparison of the calibration of mortality models on the Czech data

This article compares mortality models applied on the Czech historical data since 1920. The first part overviews all considered models and calibration methods used in the calibration. Second part provides the reader with qualitative and quantitative measures that are used for the comparison of mortality models. We compare them according to the quality of fit, reasonableness, parsimony, robustness and completeness. Finally we summarize the results and conclude the main results of our comparison on the Czech mortality data.

Keywords: model calibration, stochastic mortality models, mortality.

Kapitálové matice v regionální input-output analýze

Karel Šafr

karel.safr@vse.cz

Doktorand oboru statistika

Školitel: doc. Ing. Jaroslav Sixta, Ph.D. (jaroslav.sixta@vse.cz)

Abstrakt: Stěžejním tématem moderní strukturní analýzy je formulace rovnic, které reprezentují současný stav ekonomického poznání, a regionalizace těchto vztahů. Naopak méně diskutovanou součástí těchto modelů jsou dostupné datové zdroje, jež slouží jako podkladová data pro samotné odhady a kalibrace parametrů modelů. Za základní datový zdroj, který je potřebný pro tento typ analýz, lze považovat input-output tabulky a matice toků/stavů kapitálu. Tyto datové zdroje s nárůstem geografické podrobnosti členění dat se stávají nedostupné z oficiálních datových zdrojů. Zejména právě input-output tabulky a matice kapitálu v regionálním členění v případě České republiky neexistují. V rámci této práce jsem se zaměřil na druhý zmiňovaný datový zdroj – matice spotřeby fixního kapitálu. Používám dvě metody dopočtu na tato data, aplikuji a vyhodnocuji vliv na dva modelové přístupy. Tyto kapitálové matice zasadím do základního input-output modelu a modelu všeobecné rovnováhy. Ilustruji rozdíly výsledků v těchto modelech způsobené rozdílnými kapitálovými maticemi.

Klíčová slova: Kapitálové matice, Input-output analýza, Modely všeobecné rovnováhy, RAS metoda

Úvod

Nedílnou součástí ekonomické analýzy jsou strukturní ekonomické modely. Za základní rámec tohoto matematického přístupu lze pak považovat input-output (I-O) modely a přístupy z nich odvozené. Strukturní analýzy umožnují modelování široké škály dopadů v kontextu jednotlivých sektorů a produktů (popř. odvětví). I-O analýza se odlišuje od ostatních přístupů pro modelování dopadů právě svou podrobností výpočtu. Na input-output analýzu pak navázaly modely všeobecné rovnováhy (CGE). Výhodou CGE modelů je, že uvažují elasticity a jsou postavené na ekonomickém předpokladu maximalizace užitku spotřebitele/ů a zisků firem (popř. minimalizace nákladů). Oproti I-O modelů zde narážíme na problém agregace sektorů a produktů I-O tabulky (která tvoří podkladový zdroj dat), díky čemuž už nejsou výsledky tohoto přístupu detailní. Hlavním důvodem agregace je výpočetní náročnost strukturních CGE modelů a přehlednost výstupů.

Současný vývoj strukturních modelů spočívá také v regionalizaci modelů samotných. Modelové aplikace na národní úrovni narážejí na několik problémů. Je jím zejména rozdílnost regionálních produkčních funkcí, která pak způsobuje zkreslení v případě aplikace národní struktury na regionální problém (například analyzujeme-li čistě regionální dopady národohospodářských politik). Dále se jedná o nemožnost analýzy vlivu na ostatní regiony a regionální makroekonomické ukazatele. Zkreslení je pak o to významnější, čím je větší regionální diferenciace produkce. Významné zkreslení pak způsobí i regionálně jedinečné odvětví/produkty. Regionalizace dat pro strukturní analýzu v konečném důsledku umožní propojení regionálních makroekonomických ukazatelů a vyčíslí přesnější dopady do ekonomiky.

V rámci tohoto článku používám dva způsoby výpočtu regionální spotřeby fixního kapitálu. V první fázi jsem použil již odvozené metody pro dopočet struktury matic stavů kapitálu v třídění na produkty (Klasifikace CZ-CPA) pro národní úroveň. Dále jsem pak pomocí dopočtené struktury (za předpokladu homogenity opotřebení produktů podle typu produktu) vypočítal spotřebu fixního kapitálu, která je použita pro tvorbu produkce (představuje snížení hodnoty fixních aktiv jako důsledek fyzického a morálního opotřebení fixních aktiv – Český statistický úřad, 2017). Následně jsem takto dopočtené národní hodnoty regionálně alokoval pomocí upravené RAS metody.

Cílem tohoto článku je vyhodnotit vliv dvou metod dopočtu kapitálových matic na regionální úrovni pro modelové aplikace. Na dopočítaných maticích toků kapitálu za rok 2011 konstruuji I-O a CGE model a simuluji exogenní šoky. Článek poukazuje na citlivost těchto modelových aplikací na podkladová data pro kalibraci modelového aparátu. Pro výpočty využívám data Českého statistického úřadu, katedry ekonomické statistiky na Vysoké škole ekonomické v Praze a předcházejících vlastních dopočtů dat v oblasti toků produkce mezi regiony.

Současný stav poznání

Input-output modely a i CGE modely jsou v současné době běžně používané přístupy pro modelování dopadů. Aplikace je na regionální úrovni velmi rozpracovaná. První multiregionální model byl formulován Leontiefem

(Leontief a Strout, 1963) jako intra-regionální model. Od této chvíle se na tuto oblast začalo zaměřovat mnoho autorů. Za I-O aplikace stojí zejména zmínit práci Millera a Blaira (2009), ve které shrnuje základní principy jak regionálních I-O modelů, tak i konstrukce datových zdrojů. Z české a slovenské literatury je to pak Goga (2009), který ve své práci vysvětluje metody agregace a desagregaci dat pro regionální aplikace. Na tuto problematiku upozornil již Lenzen a kol. (2004), který shrnuje různé možnosti vztahů v regionálních I-O modelech. Tyto možnosti se dají shrnout do metod, kdy regiony mezi sebou vůbec neinteragují (bez regionálního vývozu a dovozu). Další situací je stav, kdy zkoumaný region má vazby se všemi ostatními regiony, ale tyto regiony již nejsou propojeny mezi sebou. Poslední možností je podle autora multiregionální přístup – kdy všechny regiony interagují pomocí regionálního vývozu mezi všema regiony. Rozdílnost těchto přístupů vyhodnocuji na českých datech již v předchozím článku (Šafr, Vltavská 2014).

Jak již bylo řečeno, CGE přístup navazuje svou úvahou na regionální input-output modely a tabulky. Na základě podobných předpokladů input-output modelu je CGE Model Sue Winga (2004), který v tomto zmíněném článku definuje CGE model, který má stejný rámec jako I-O model. Z tohoto důvodu bude tento model použit pro komparaci s I-O modelem a jeho výsledky. Navíc jednoduchost modelu, která by mohla být kritizovaná z ekonomické abstrakce od ostatních jevů, je vhodná pro evaluaci výsledků rozdílných datových zdrojů, jejichž efekty by se mohly v komplexním modelu se zahraničím skrýt. Za velmi podobný přístup lze označit model Roberta Ronsona (Ronson, 1991). Avšak oproti předcházejícímu přístupu neformuluje model v plné shodě s I-O rámcem. Jeho přístup slouží pro evaluaci efektů zdanění meziregionálních toků (které lze ale lehce simulovat i v modelu Sue Winga pomocí daní do mezispotřeby). Rutherford a Paltsev (1999) formulují velmi podrobný rámcový přístup jako má Sue Wing. Avšak tento model postihuje již komplexněji základní ekonomické vztahy a by mohlo již dojít k přenesení efektů rozdílných datových zdrojů. Za velmi komplexní ekonomické CGE modely na bázi I-O lze zmínit, například model pro analýzu multiregionálních CGE modelů na modelování dopadů dostavby dálnic (Kim, Hewings, Hong, 2004).

Hlavní metody dopočtu meziregionálních toků jsou Newtonův gravitační model, nebo metody na bázi maximalizace, či minimalizace entropie (Lahr, 1993; Sargento, Ramos, Hewings, 2012; Šafr, 2016a).

I-O rámec

Jádro strukturních analýz tvoří input-output tabulka. První input-output tabulku sestavil v moderní podobě ekonom W. Leontief (1941, 1986). Input-output tabulka znázorňuje rovnici zdrojů a užití v bilanční rovnováze po jednotlivých produktech tvořených v ekonomice. Tabulku lze zjednodušeně znázornit následovně (Miller a Blair, 2009):

Tabulka 1: Input-output tabulka Zdroj: Hronová a kol. (2009)

input-output tabulka	Mezispotřeba				Konečné užití		Celková produkce
Mezispotřeba	$x_{1,1}$	$\cdots$	$x_{1,n}$	$u_1^{[1]}$	$y_{1,1}$	$y_{1,K}$	x_1
	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_{n,1}$	$\cdots$	$x_{n,n}$	$u_n^{[1]}$	$y_{n,1}$	$y_{n,K}$	x_n
	$u_1^{[2]}$	$\cdots$	$u_n^{[2]}$				
Přidaná hodnota	$v_{1,1}$	$\cdots$	$v_{1,n}$				
	$v_{F,1}$	$\cdots$	$v_{F,n}$				
Celková produkce	x_1	$\cdots$	x_n				

Prvky x_{ij} představují tok mezispotřeby mezi odvětvím *i* a *j*. Proměnná y_{jk} znázorňuje konečné užití produktů odvětví *j*. Proměnná x_i představuje celkovou produkci/užití produktů odvětví i. Hrubou přidanou hodnotu znázorňuji pomocí proměnné v_{fj}. U hrubé přidané hodnoty (HPH) a konečného užití (KU) je nutno zdůraznit, že je v této zjednodušené I-O tabulce agreguji do vektoru hodnot, ale tradičně jsou znázorňovány jako matice

s náležitými řádky/sloupci struktury HPH či KU. Proměnná u_i znázorňuje sloupcové/řádkové součty matice mezispotřeby (**X** jejíž prvky jsou x_{ij}**)**. V případě této tabulky můžeme jednoduše znázornit základní vztahy modelu, které pak představují rámec I-O a i CGE modelů. První dvě rovnice jsou bilanční rovnováhou tabulky, první z pohledu užití jako:

$$\sum_{j}^{n} x_{ij} + \sum_{l=1}^{k} y_{ik} = x_i \,. \tag{1}$$

Dále lze zapsat pohled ze strany produkce:

$$\sum_{i}^{n} x_{ij} + \sum_{f=1}^{F} v_{fj} = x_j \tag{2}$$

Základním vztah v I-O a i CGE modelech pak znázorňuje vztah,

$$a_{ij} = \frac{x_{ij}}{x_j} \tag{3}$$

což je technický koeficient mezispotřeby. Tyto koeficienty se považují v I-O analýze za pevně dané vztahy, které jsou dlouhodobě platné. Vyjadřují svou definicí podíl vstupu produkce *i*-tého produktu na tvorbu *j*-tého produktu. Pomocí těchto vztahů lze znázornit základní I-O model pomocí Leontiefovy matice jako,

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{y}\,, \tag{4}$$

kde $\mathbf{x}$ je vektor celkové produkce, $\mathbf{y}$ vektor konečného užití, a $(\mathbf{I} - \mathbf{A})^{-1}$ je tzv. Leontiefova matice, jenž představuje inverzi z rozdílu jednotkové matice ($\mathbf{I}$) a matice technických koeficientů ($\mathbf{A}$). K těmto maticím přísluší výše zmíněné prvky I-O tabulky.

CGE v I-O rámci

Jak již bylo zmíněno v úvodu, za jedny z hlavních navazujících I-O přístupů lze označit CGE modely. CGE modely a I-O modely vycházejí z hlavního stejného rámce – obecné rovnováhy mezi zdroji a užitím. Tento vztah lze znázornit již zmíněnými rovnicemi (1) a (2). Tyto rovnice dále doplním o podmínky, které učiní CGE model konzistentní s I-O přístupem. Níže představený model vychází z rovnic podle Sue Wing (2003). Doplněním předcházejících rovnic o základní vztahy v CGE uzavřeném modelu:

$$\sum_{f}^{F} v_{fj} = \overline{V}_f \,, \tag{3}$$

tedy, že v ekonomice se vyskytuje předem daný počet statků $\overline{V}_f$, který je plně vyžit. Dále je nutné definovat rozpočtové omezení modelu. To lze vyjádřit jako:

$$\sum_{f}^{F} \overline{V}_f = \overline{m}\,, \tag{4}$$

což vyjadřuje, že domácnosti jsou jediným majitelem všech výrobních faktorů v ekonomice. Model je v tomto funkčním tvaru stále v konzistenci s I-O modelem. Jediným rozdílem je, že náš CGE model je definovaný na objemy výrobních faktorů a příjmu ($\overline{V_f}$, $\overline{m}$). Zafixováním těchto faktorů bychom v neelastické I-O analýze pak plně zastavili v I-O tabulce jako takové – z důvodu fixních /neelastických faktorů, jejichž struktura je pevně daná. Oproti tomu v CGE přístupu jsou technické koeficienty (3) chápány spíše ve smyslu elasticit ukazující rovnovážný vztah (za předpokladu kalibrace na samotnou I-O tabulku), což je dáno optimalizací, která se do CGE modelu dostává prostřednictvím agentů – firem a spotřebitele, popřípadě lze na straně KU rozšířit o vládu, zahraničí aj. Problém spotřebitele lze zapsat jako maximalizaci užitku. Vyjdeme-li ze základního modelu Sue Winga (2003) a rozšíříme-li ho jako Autor o daně, lze zapsat úlohu spotřebitele jako:

$$\max_{c_i} p_u U - \sum_{i}^{n} (1+\tau_i^y)(1+\tau_i^c) p_i c_i \,, \tag{5}$$

tedy maximalizujeme hodnotu užitku při nákladů užitku na omezení spotřeby, jež je samotnou užitkovou funkcí:

$$U = A_c \prod_{i=1}^{N} c_i^{\alpha_i} \,. \tag{6}$$

Parametr A_c je úrovňový parametr, který nám zabezpečuje konzistenci mezi úrovní užitku a produkcí statků. Parametr α_i představuje podíl hodnoty spotřeby *i*-tého statku na hodnotě celkové spotřeby (obojí i s daněmi). Spotřební funkci zapíšeme jako řešení předcházejících rovnic (5. a 6.):

$$c_i = \alpha_i \frac{m - \sum_{i=1}^{N} (1+\tau_i^Y) p_i s_i}{(1+\tau_i^C)(1+\tau_i^Y) p_i} \,, \tag{7}$$

Optimalizační úlohu jednotlivých firem lze pak zapsat jako:

$$\max_{x_{ij}, v_{fj}} p_j y_j - \sum_{i}^{n} (1+\tau_i^y)(1+\tau_{ij}^x) p_i x_{ij} - \sum_{f=1}^{F} (1+\tau_{fj}^v) w_f v_{fj} \,, \tag{8}$$

za předpokladu produkční funkce:

$$y_j = A_j \prod_{i=1}^{N} x_{ij}^{\beta_{ij}} \prod_{f=1}^{F} v_{fj}^{\gamma_{ij}} \,, \tag{9}$$

kde parametry β_{ij} a γ_{fj} představují technické koeficienty ve smyslu (3) rovnice input-output modelu (zase po započítání/odečtení vlivu daní aj). Pro tyto proměnné pak v případě uvažování daňových efektů již neplatí I-O předpoklad o součtu těchto parametrů na hodnotu jedné (tj.: $\sum_{f=1}^{F} \gamma_{fj} + \sum_{i=1}^{N} \beta_{ij} \neq 1$), což je způsobené právě uvažováním daní. Pro detailnější informace k tomuto problému kalibrace faktorů se lze podívat do originální práce autora Sue Wing (2003). Řešením rovnic (8) a (9) jsou pak poptávkové funkce po mezispotřebě a výrobních faktorech, které lze zapsat jako:

$$x_{ij} = \beta_{ij} \frac{p_j y_j}{\sum_{i}^{N} (1+\tau_{ij}^x)(1+\tau_i^y) p_i} \,, \tag{10}$$

$$v_{fj} = \gamma_{ij} \frac{p_j y_j}{\sum_{f}^{F} (1+\tau_f^v) w_f} \,. \tag{11}$$

Dosazením do původních rovnic modelu (1,2,3,4) lze získat řešení soustavy rovnic formou nelineární multikriteriální optimalizace. Důležité je zde zmínit, že je nutné rovnice (2) a (4) rozšířit o daňové efekty. Tedy aby platil součet na celkovou produkci v rovnici (2) a zároveň aby platilo, že spotřebitel má tyto daně jako příjem. Následně lze předělat a odvodit rovnice z původního modelu. Minimalizujeme tedy rozdíl vzniklý z rovnice (1) jako:

$$\Delta_i^c = \sum_{j=1}^{N} \beta_{ij} \frac{p_j y_j}{\sum_{i}^{N} (1+\tau_{ij}^x)(1+\tau_i^y) p_i} + \alpha_i \frac{\sum_{f=1}^{F} w_f V_f + T - \sum_{i=1}^{N} (1+\tau_i^Y) p_i s_i}{(1+\tau_i^C)(1+\tau_i^Y) p_i} + s_i - y_i \,, \tag{12}$$

kde proměnná T je příjem z vybraných daní a vychází z výše zmiňované úpravy rovnic. Tuto proměnou lze zapsat jako:

$$T = \sum_{j=1}^{N}\sum_{f=1}^{F}\tau_j^Y p_j y_j + \sum_{i=1}^{N}\tau_i^c p_i y_i + \sum_{i=1}^{N}\sum_{j=1}^{N}\tau_{ij}^x p_i x_{ij} + \sum_{f=1}^{F}\sum_{j=1}^{N}\tau_f^v w_f v_{fj} \,. \tag{13}$$

Rovnice (12) pak definuje minimalizační úlohu z pohledu užití produkce. Dále je nutné uvést v rovnováhu zdroje výrobních faktorů. Vyjdeme-li z rovnice (3), pak lze tuto úlohu formulovat jako:

$$\Delta_f^F = \sum_{j=1}^{F}\frac{\gamma_{fj} p_j y_j}{(1+\tau_f^v)w_f} - V_f \,. \tag{14}$$

Taktéž lze formulovat pohled ze strany produkce na I-O tabulku. Poté je možné vycházet z rovnice (2). V tomto případě je nutné podotknout, že součástí produkční strany rovnic jsou i daně, které je nutné zakomponovat do výchozí rovnice (2). Tedy můžeme zapsat:

$$\Delta_j^{\pi} = \sum_{i=1}^{N}\beta_{ij}\frac{p_j y_j}{\sum_{i}^{N}(1+\tau_{ij}^x)(1+\tau_i^y)} + \sum_{f=1}^{F}\gamma_{ij}\frac{p_j y_j}{\sum_{f}^{F}(1+\tau_f^v)} + t_j -$$

$$A_j\prod_{i=1}^{N}\left[\beta_{ij}\frac{p_j y_j}{\sum_{i}^{N}(1+\tau_{ij}^x)(1+\tau_i^y)p_i}\right]^{\beta_{ij}}\prod_{f=1}^{F}\left[\gamma_{ij}\frac{p_j y_j}{\sum_{f}^{F}(1+\tau_f^v)w_f}\right]^{\gamma_{ij}} p_j \,, \tag{15}$$

kde proměnná t_j představuje daňový výnos v odvětví j. Toto lze zapsat jako:

$$t_j = \sum_{f=1}^{F}\tau_j^Y p_j y_j + \tau_j^c p_j y_j + \sum_{j=1}^{N}\tau_{ij}^x p_i x_{ij} + \sum_{f=1}^{F}\tau_f^v w_f v_{fj} \,. \tag{16}$$

Finální rovnice minimalizace příjmů lze pak zapsat jako:

$$\Delta^m = \sum_{f=1}^{F} w_f V_f + \sum_{j=1}^{N}\sum_{f=1}^{F}\tau_j^Y p_j y_j + \sum_{i=1}^{N}\tau_i^c p_i y_i + \sum_{i=1}^{N}\sum_{j=1}^{N}\tau_{ij}^x p_i x_{ij} + \sum_{f=1}^{F}\sum_{j=1}^{N}\tau_f^v w_f v_{fj} - m \tag{17}$$

Tato mnou definovaná soustava rovnic (12,14,15,17) představuje minimalizační úlohu CGE modelu o 2N+F+1 rovnici. Oproti původnímu autorovi, který představil zjednodušené řešení bez daní, mám několik odlišností. Původní autor opomenul v rovnici minimalizace chyby ze strany produkce (má rovnice 15) mocninu u úrovňového parametru. V jeho případě zde mělo být A_j^{-1}. Tento model se snažili odvodit i jiní autoři, kteří ale přehlíželi zisky z daní v rovnici (12) a (15), což je příklad práce Kratochvíla (2012) na českých datech (v národním a ne regionálním třídění). S ohledem na to, že autor řeší model pomocí softwaru GAMS, předpokládám, že nastavení softwaru tomuto opomenutí zabránilo, protože soustava rovnic, jak ji definuje pro jeden typ daně, není kalibrovatelná přímo z datového zdroje. Správnost řešení lze lehce ověřit vytvořením simulované struktury I-O tabulky, kde již daň existuje. V případě, kdy tomu tak je, pak prostým dosazením původních dat do soustavy rovnic (12,14,15 a 17) musí vyjít hodnoty rovné nule pro všechny dílčí funkce.

Kapitálové matice

Jak již bylo zmíněno na začátku, kapitálové matice představují důležitý výrobní faktor v ekonomických modelech. Definovaný I-O a CGE model v předcházejících kapitolách vyžaduje rovnice pro spotřebu fixního kapitálu. Spotřeba fixního kapitálu (dále jen SFK) představuje snížení hodnoty fixních aktiv jako důsledek fyzického a morálního opotřebení fixních aktiv (ČSU, 2017). Matice vyjadřující strukturu tohoto opotřebení je nutné z národní úrovně přenést na regionální. Metoda RAS umožnuje aproximovat matici **H‘** (pro kterou jsou známé jen součty sloupců (i_j) a řádků (e_i)) z matice **H** za předpokladu minimalizace rozdílů struktury výsledné matice **H‘** od matice **H**. RAS metoda je nelineární metoda, která se počítá iteračně. Datově vycházím ze známých matic SFK za rok 2011 (Šafr, 2016b). Pokud vyjdeme ze stejného předpokladu, jako je v konstrukci

těchto matic, pak lze říct, že existuje dlouhodobý vztah mezi tvorbou kapitálu a spotřebou. Tento vztah musí být dlouhodobě v rovnovážný. Potom lze formulovat RAS metodu jako částečně optimalizační úlohu, kde v každé iteraci kromě kalibrace řádků a sloupců též kalibrujeme částečné součty sloupců a řádků. Podmínky této úlohy lze vyjádřit pro matici **H'** jako:

$$\sum_{i=1}^{q} h'_{ij} = i_j \text{ , a } \sum_{j=1}^{w} h'_{ij} = e_i \text{ ,} \tag{18}$$

kde pro matici **H'** (jejíž prvky jsou h'_{ij}), platí součty sloupců (i_j) a řádků (e_i). Pro tyto součty samozřejmě platí, že $\sum e_i = \sum i_i$. Za těchto podmínek se standardně aplikuje RAS metoda na podkladovou matici (**H**). S ohledem na regionalizaci matice je potřeba dále zabezpečit částečné součty submatic matice **H'**. Tedy rozložíme-li matici **H'** na submatice ($\mathbf{D}^{\varepsilon\phi}$ s prvky $d^{\varepsilon\phi}$):

$$\mathbf{H}' = \begin{bmatrix} \mathbf{D}^{11} & \cdots & \mathbf{D}^{1\delta} \\ \vdots & \ddots & \vdots \\ \mathbf{D}^{\chi 1} & \cdots & \mathbf{D}^{\chi\delta} \end{bmatrix} \text{ , } \sum\sum d^{\varepsilon\phi} = \varphi^{\varepsilon\phi} \text{ ,} \tag{19}$$

pro které platí, že celkový součet submatice je roven $\varphi^{\varepsilon\phi}$. Pak pro všechny součty submatic musí platit: $\sum_{\phi=1}^{\delta}\sum_{\varepsilon=1}^{\chi} \varphi^{\varepsilon\phi} = \sum e_i = \sum i_i$. Za této podmínky pak můžeme aplikovat upravenou RAS metodu třech kalibrací v iteraci na podkladovou matici **H**. První dvě kalibrace jsou standardními postupy – přes součet řádků a sloupců v matici **H** dané vzorcem (18), třetí iterace pak kontroluje součty submatic matice podle vzorce (19). Model v této formě je specifickou 3D RAS metodou (Holý, Šafr, 2017).

Zdroje dat

Pro regionální analýzu byly použity I-O tabulky konstruované na katedře ekonomické statistiky fakulty Informatiky a statistiky VŠE. Tyto tabulky (Sixta a Vltavská, 2016, Sixta a Fischer, 2015, katedra ekonomické statistiky, 2016) popisují strukturu produkce v jednotlivých regionech podle klasifikace (NUTS 3). Pro samotnou analýzu bylo potřeba data regionálně agregovat na NUTS 2 úroveň. Klasifikace samotného datového zdroje vychází podle CZ-CPA produktů a je shodná s publikovanými I-O tabulkami na Českém statistickém úřadě.

V těchto tabulkách v regionálním třídění není alokovaný regionální vývoz, dovoz podle regionu původu. Z tohoto důvodu jsem přebral výsledky z článků (Šafr, 2016a a Šafr, Vltavská 2016), kde jsem se zabýval dopočtem regionálních toků produkce.

Posledním chybějícím datovým zdrojem jsou stavy a nová produkce kapitálu. Výpočtem těchto dat jsem se zabýval v předcházející práci (Šafr, 2016b) a použil jsem tuto metodologii. Zde jsem navrhl dva přístupy jak aproximovat/dopočítat matici stavů kapitálu pro národní úroveň. Z obou přístupů lze vypočítat i matice opotřebení fixních aktiv v klasifikaci CZ-CPA. Tyto matice jsem následně použil do samotné analýzy, kterou popisuji v kapitole Kapitálové matice.

Výsledky

Dopočtené datové zdroje jsem následně aplikoval do výše diskutovaného I-O a CGE modelu. Od I-O modelu lze očekávat, že rozdíly zde budou výraznější než u CGE modelu. Toto bude způsobeno právě elasticitami CGE modelu. Jak v I-O modelu, tak v CGE modelu jsem simuloval šoky do cen SFK o výši 10% zdanění jednotlivých nákladů prostřednictví „ad-valorem" zdanění. Přesto, že přímo toto zdanění je těžko zdůvodnitelné z ekonomického hlediska, tak z pohledu testování vlastností SFK v modelech nabízí přímý nástroj na měření vlivu kapitálu v samotném modelu.

Zdanění způsobí zvýšení nákladů na primární vstupy do produkčních funkcí firem v jednotlivých odvětvích. V input-output modelu se daň promítne plně do ceny produktů, a tedy i do agregátní cenové hladiny v ekonomice. Naopak u CGE modelu budou mít firmy možnost toto zvýšení cen kompenzovat substitucí jednotlivých výrobních faktorů mezi odvětvími. Tato substituce bude však limitovaná celkovým fixním objemem jednotlivých druhů kapitálu, který musí být využit celý.

Následující tabulky porovnávají výsledky dopadů do agregátní cenové hladiny podle jednotlivých regionů a druhů kapitálu v I-O a CGE modelu (tab. 1 a tab. 4).

Tabulka 3: Agregované šoky do SFK po regionech (I-O model) v %. Zdroj: Autor

	K.1		K.2		K.3		K.4	
Regiony	Data	Model	Data	Model	Data	Model	Data	Model
CZ01	0,66	0,12	0,12	0,03	0,34	0,04	0,18	1,11
CZ02	0,33	0,09	0,22	0,02	0,25	0,03	0,21	0,86
CZ03	0,38	0,1	0,23	0,02	0,27	0,04	0,18	0,91
CZ04	0,40	0,07	0,09	0,02	0,20	0,03	0,11	0,69
CZ05	0,25	0,07	0,17	0,02	0,18	0,03	0,15	0,65
CZ06	0,37	0,09	0,21	0,02	0,27	0,03	0,18	0,87
CZ07	0,49	0,09	0,11	0,02	0,24	0,03	0,13	0,82
CZ08	0,26	0,07	0,17	0,02	0,20	0,03	0,17	0,68
Průměr	0,41	0,09	0,17	0,02	0,25	0,03	0,17	0,84

Vysvětlivky: Data – Matice kapitálu počítaná z dat pomocí opotřebení, Model – matice kapitálu počítaná podle dynamického I-O modelu (více Šafr, 2016b), CZ01: Praha, CZ02: Střední Čechy, CZ03: Jihozápad, CZ04: Severozápad, CZ05: Severovýchod, CZ06: Jihovýchod, CZ07: Střední Morava, CZ08: Moravskoslezsko; K.1: CZ-CPA 24-28+31-33, K.2: CZ-CPA 29-30, K.3: CZ-CPA 41-43, K.4: ostatní.

Z výsledků I-O modelu je vidět rozdílný vliv dopočtu matic spotřeby kapitálu na modelovou aplikaci. Naopak dopočet matic kapitálu podle dat ukazuje vyšší dopady do ekonomiky skoro ve všech typech kapitálu. Tato nerovnoměrnost je způsobená rozdílnou strukturou matic.

Tabulka 4: Agregované šoky do kapitálu po regionech (CGE model) v %. Zdroj: Autor

	K.1		K.2		K.3		K.4	
	Data	Model	Data	Model	Data	Model	Data	Model
CZ01	0,17	0,14	0,14	0,01	0,18	0,08	0,16	0,18
CZ02	0,21	0,15	0,24	0,02	0,11	0,08	0,21	0,28
CZ03	0,16	0,15	0,27	0,02	0,17	0,08	0,23	0,23
CZ04	0,03	0,11	0,12	0,01	0,07	0,07	0,10	-0,08
CZ05	0,09	0,12	0,18	0,01	0,06	0,07	0,12	0,01
CZ06	0,09	0,13	0,17	0,01	0,11	0,07	0,18	0,09
CZ07	0,14	0,13	0,19	0,02	0,14	0,08	0,12	0,15
CZ08	0,06	0,11	0,14	0,01	0,04	0,07	0,12	-0,06
Průměr	0,12	0,13	0,18	0,01	0,11	0,08	0,15	0,11

Vysvětlivky: Data – Matice kapitálu počítaná z dat pomocí opotřebení, Model – matice kapitálu počítaná podle dynamického I-O modelu (více Šafr, 2016b), CZ01: Praha, CZ02: Střední Čechy, CZ03: Jihozápad, CZ04: Severozápad, CZ05: Severovýchod, CZ06: Jihovýchod, CZ07: Střední Morava, CZ08: Moravskoslezsko; K.1: CZ-CPA 24-28+31-33, K.2: CZ-CPA 29-30, K.3: CZ-CPA 41-43, K.4: ostatní.

U CGE modelu se velké rozdíly mezi metody dopočtu SFK jako u I-O modelu neprojevily. Primárně je to dáno vlastnostmi modelu, jenž umožnuje substituovat jak vstupy, produkci, tak i konečnou spotřebu.

Závěr

Cílem příspěvku je evaluovat vliv dopočtených dat v regionálních modelech na samotné výsledky modelů. Tedy zda mohou ovlivnit závěry samotných studií data dopočtená jiným způsobem. Nikoliv evaluovat přesně vliv dopadu. Proto byl zvolen statický I-O model a CGE model, jenž svou specifikací neumožňuje tak jako I-O model ovlivňovat celkovou výši zapojených vstupů kapitálu, ale jen jejich strukturu alokace. Dva postupy výpočtu kapitálových matic jsem extrapoloval pomocí upravené metody RAS pro regionální úroveň. Ze samotné konstrukce modelu jsem očekával velké rozdíly mezi výsledky I-O modelu, ale malé rozdíly v CGE modelu.

Obě tyto úvahy se na výsledcích modelu potvrdily. U I-O modelu byl vysoký rozdíl způsoben nemožností substituovat zdaněný typu kapitálu jiným druhem kapitálu, model pak akceptoval celou výši zdanění do celkové cenové hladiny v ekonomice. Neelasticita substituce se dále projevila homogenními dopady napříč jednotlivými regiony. Naopak v případě CGE modelu lze pozorovat mírnější rozdíly na celkové cenové hladině, ale větší rozdíly mezi regiony, což je způsobeno možností substituce kapitálu jiným druhem kapitálu. Rozdílné regionální struktury umožnily jiné volby substitutů a model díky tomu dosáhl větší regionální diverzity, ale při nižší cenové hladině než v případě I-O modelu.

Tento článek jen naznačil velké téma regionálních strukturních dat. Kromě kapitálových matic, které jsou významnou součástí produkčních funkcí, článek vůbec neřešil regionální toky produkce. Dalším významným tématem v oblasti je vylepšení metod na konstrukci dat ve větší podrobnosti, než zde byla použita.

Literatura

Český statistický úřad 2017. Národní účty – metodika [Online]. Český statistický úřad. Praha: Český statistický úřad, 2017 [cit. 25. 1. 2017]. Dostupné z: https://www.czso.cz/csu/czso/10n1-05-_2005-narodni_ucty___metodika.

GOGA, M. 2009. *Input-Output analýza*. Vid. 1. Bratislava: Wolters Kluwer (Iura Edition), 202 s. ISBN 978-80-8078-293-1.

HOLÝ, Vladimír, ŠAFR, Karel, 2017. The Use of Multidimensional RAS Method in Input-Output Matrix Estimation, *Operations Research* (Odeslané od recenzního řízení).

HRONOVÁ, Stanislava, FISCHER, Jakub, HINDLS, Richard, SIXTA, Jaroslav, 2009. *Národní účetnictví: nástroj popisu globální ekonomiky*. V Praze: C.H. Beck. Beckova edice ekonomie. 359 s. ISBN 978-80-7400-153-6.

Katedra ekonomické statistiky, 2017. Regionalization of gross domestic product employing expenditure approach (data) [Online]. Fakulta informatiky a statistiky Vysoká škola ekonomická v Praze. [cit. 10. 1. 2017] Dostupné z: http://kest.vse.cz/veda-a-vyzkum/vysledky-vedecke-cinnosti/regionalizace-odhadu-hrubeho-domaciho-produktu-vydajovou-metodou/.

KIM, E., Hewings, D. J., HONG, Ch., 2004. An Application of an Intergrated Transport Network-Multiregional CGE Model: a Framework for the Economic Analysis of Highway Projects. *Economic Systems Research*, roč. 16, č. 3, s. 235-258. ISSN: 1469-5758.

KRATOCHVÍL, Petr, 2009. *Využití CGE modelů v environmentální oblasti* [Online]. Bakalářská práce, Fakulta sociálních věd: Institut ekonomických studií. [cit. 15. 1. 2017] Dostupné z: http://ies.fsv.cuni.cz/work/index/show/id/1091/lang/cs.

LAHR, Michael L. A Review of the Literature Supporting the Hybrid Approach to Constructing Regional Input-Output Models. *Economic Systems Research*. 1993, roč. 5, č. 3, s. 277-293. ISSN: 1469-5758.

LENZEN, M., PADE, L., MUNKSGAARD, J., 2004. CO2 Multipliers in Multi-region Input-Output Models. *Economic Systems Research*, roč. 16, č. 4, s. 391-412. ISSN: 1469-5758.

LEONTIEF, W., STROUT, A., 1963. *Multi-regional Input-Output Analysis, in Barna*. In: T. Barna (Ed.) Structural Interdependence and Economic Development, s. 119-150. London: Macmillan.

LEONTIEF, Wassily, 1941. *The Structure of American Economy, 1919-1929*: *An Empirical Application of Equilibrium Analysis.* Cambridge, Mass.: Harvard University Press, 2 vid., 1951, New York, Oxford University Press.

LEONTIEF, Wassily, 1986. *Input-Output Economics*. 2. vyd. New York, United States of America: Oxford university press, 448 s., ISBN 0195035275.

MILLER, Ronald E., BLAIR, Peter D., 2009. *Input-output analysis: foundations and extensions*. 2 vid., Cambridge: Cambridge University Press. ISBN 978-0-521-73902-3.

RONSON, Roberto, 1991. The adjustment of interregional input-output coefficients under heterogenous price sensitivity: A linearized model. *The annals of Regional Science*, roč. 2, č. 25, s. 101-114. ISSN 1432-0592.

RUTHERFORD T., PALTSEV, S., 1999. *From an Input-Output Table to a General Equilibrium Model: Assessing the Excess Burden of Indirect Taxes in Russia*. [Online]. Department of Economics, University of Colorado, mimeo. 57 s. Dostupné z: http://web.mit.edu/paltsev/www/docs/exburden.pdf

SARGENTO, A., L., M., RAMOS, P., N., HEWINGS, G., J., D., 2012. Inter-regional Trade Flow Estimation through Non-survey Models: An Empirical Assessment. *Economic Systems Research*, roč. 24, č. 2, s. 173-1973. ISSN: 1469-5758.

SIXTA, Jaroslav, FISCHER, Jan, 2015. Regional Input-Output Models: Assessment of the Impact of Investment in Infrastructure on the Regional Economy [Online]. *In: Mathematical Methods in Economy*, Czech Republic. s. 719-724. Dostupné z: http://mme2015.zcu.cz/downloads/MME_2015_proceedings.pdf.

SIXTA, Jaroslav, VLTAVSKÁ, Kristýna, 2016. Regional Input-output Tables: Practical Aspects of its Compilation for the Regions of the Czech Republic. *Ekonomický časopis*, roč. 64, č. 1, s. 56–69.

Sue Wing, I., 2004. CGE Models for Economy-Wide Policy Analysis [online]. *MIT Joint Program on the Science and Policy of Global Change technical Note,* roč. 6., 50 s. Dostupné z: http://web.mit.edu/globalchange/www/MITJPSPGC_TechNote6.pdf.

ŠAFR, Karel, 2016a. Allocation of commodity flows in the regional Input-Output tables for the Czech Republic. *In: Proceedings of 19th International Scientific Conference Application of Mathematics and Statistics in Economics*. 31. srpna – 4. září 2016, Banská Štiavnica.

ŠAFR, Karel, 2016b. Pilot Appliacation of the Dynamic Input-Output Model. Case Study of the Czech Republic 2005-2013. *Statistika*, 2016, roč. 96, č. 2., s. 15-31, ISSN 0322-788x.

ŠAFR, Karel, VLTAVSKÁ Kristýna, 2016. The evaluation of economic impact using regional Input-Output model: The Case study of Czech regions in context of national Input-Output tables. *In: Proceedings of 14th International Scientific Conference "Economic Policy in the European Union Member Countries"* 14-16. září 2016, Petrovice u Karviné.

JEL Classification: C67, C68, D57

Summary

Capital Matrices in Regional Input-output Analysis

Abstract: The key topic of modern structural analysis is formulation of equations that represent the current state of economic knowledge and also regionalization of such equations. However, available data sources represent a less frequently discussed part of models. These sources serve as supporting data for the actual estimation and calibration of model parameters. Next, input-output tables together with capital stock and flow matrices form a basic data source that is needed for this type of analysis. These data sources are increasingly more difficult to obtain from official data sources due to a more detailed level of regional data aggregation. Specifically, input-output tables and capital matrices do not exist on a regional level in the Czech Republic. In this article, I focus on the second data source – the matrix of consumption of fixed capital. I use two methods to process the data; that is, application and evaluation of the influence on two model approaches. Then I put the capital matrices in a basic input-output model and a general equilibrium model. The results of these models illustrate differences caused by various capital matrices.

Keywords: Capital matrices, Input-Output models, Computable general equilibrium models, RAS method.

Využití metod smíšených frekvencí při konstrukci čtvrtletního ukazatele výkonnosti zpracovatelského průmyslu ČR

Jan Zeman

jan.zeman@vse.cz

Doktorand oboru statistika

Školitel: doc. Ing. Jakub Fischer, Ph.D., (jakub.fischer@vse.cz)

Abstrakt: Odhady hrubého domácího produktu patří mezi velmi důležité charakteristiky vývoje ekonomické situace každého vyspělého státu. V mé analýze stojí ve středu zájmu akademický odhad čtvrtletního hrubého domácího produktu. Ten by měl poskytovat informaci dříve než bleskový odhad oficiálně publikovaný Českým statistickým úřadem. Bleskový odhad je stále považovaný za příliš pozdní. Akademický odhad, mající povahu signálního odhadu, by tak měl sloužit jako rychlejší alternativa bleskového odhadu. Signální odhad je sestavován s využitím zejména souběžných a předstihových ekonomických ukazatelů a podává informaci o směru vývoje hrubého domácího produktu.

Příspěvek se zabývá srovnáním metod smíšených frekvencí pozorování a jejich aplikací na data odvětví zpracovatelského průmyslu ČR s cílem konstruovat čtvrtletní ukazatel výkonnosti tohoto odvětví. Tento proces představuje část postupu, pomocí kterého by mělo být dosaženo signálního odhadu hrubého domácího produktu jako včasnější alternativy k výše uvedenému bleskovému odhadu.

V článku jsou představeny dvě metody specifickým způsobem kombinující měsíční a čtvrtletní data – MIDAS regrese a bridge modely. Jejich výsledky budou srovnány s výsledky Chow-Linovy interpolační metody. Proveden je i výběr vhodných ukazatelů, přispívající ke zlepšení předpovědí hrubé přidané hodnoty daného odvětví pomocí testování Grangerovy kauzality. Součástí je také srovnání vybraných metod a nástin dalších možných postupů.

Klíčová slova: zpracovatelský průmysl, hrubá přidaná hodnota, metody smíšených frekvencí pozorování, Grangerova kauzalita

Úvod

Odhady hrubého domácího produktu (dále HDP) patří mezi velmi důležité charakteristiky vývoje ekonomické situace každého vyspělého státu. Stále více však roste tlak na zkrácení doby do publikování bleskových odhadů tak, aby byla co nejdříve k dispozici informace nezbytná pro rozhodování na vrcholné úrovni jednotlivých států, zahrnující zároveň vhodný kompromis mezi včasností a kvalitou této informace. Ve srovnání s Českou republikou (45 dnů) některé státy publikují bleskové odhady o několik dnů dříve. Například ve Spojeném království 25 dnů po skončení referenčního čtvrtletí, ve Švédsku po 35 dnech. Ve srovnatelné době jako v ČR publikují své první odhady například Německo, Řecko či Itálie. Eurostat ve svých doporučeních apeluje na státy Evropské unie, aby své bleskové odhady publikovaly v rozmezí 45 až 60 dnů po skončení referenčního čtvrtletí, ale v důsledku sílící poptávky po co nejvčasnější informaci o vývoji ekonomiky se zaměřuje i na ještě dřívější odhad, který by byl k dispozici již 30 dnů po skončení referenčního čtvrtletí. Tento odhad Eurostat publikuje od dubna 2016 za eurozónu (Eurostat, 2016). Metodologicky se však z větší části drží postupů používaných pro bleskový odhad (t+45).

Vzhledem k tomu, že bleskové odhady jsou k dispozici dříve než pravidelné čtvrtletní odhady (t+70), musí dojít ke kompromisu mezi včasností a přesností a je tedy potřeba tyto dvě vlastnosti uvést do rovnováhy. Faktem však zůstává, že většina uživatelů, ekonomických analytiků a tvůrců hospodářské politiky výrazně více preferuje včasnost i za cenu nižší míry přesnosti a spolehlivosti příslušných odhadů.

Nejen v České republice, ale i v zahraničí se autoři často ve svých pracích věnovali odhadování čtvrtletního růstu HDP za pomoci nejrůznějších přístupů a modelů. Stěžejní je též vhodný výběr ukazatelů, na základě kterých je HDP odhadován. Volba těchto indikátorů bývá mnohdy poměrně subjektivní a každý autor k ní může přistupovat jinak. Na druhou stranu každý takový ukazatel by měl splňovat určité vlastnosti. Musí se jednat o ukazatel, který určitým způsobem věcně souvisí s odhadovaným makroagregátem, který nepodléhá příliš častým

revizím a který pochází z důvěryhodných a oficiálních zdrojů. Dalším důležitým prvkem je i včasné publikování takového ukazatele a jeho dostupnost.

Mezi používané ukazatele se řadí nejen ty, které jsou publikovány ve čtvrtletní periodicitě, ale i ukazatele v měsíční periodicitě. Ne všechny hodnoty měsíčních ukazatelů jsou však k dispozici v rámci referenčního čtvrtletí. Relativně bezproblémovými údaji jsou z hlediska časové dostupnosti data založená na subjektivních pocitech firem a spotřebitelů (kvalitativní data, například konjunkturální průzkumy). Tato data jsou publikována velmi krátce po skončení šetřeného období. Naopak, zpracování kvantitativních dat trvá delší dobu a následkem toho nejsou všechny hodnoty ukazatelů pro dané čtvrtletí k dispozici. Existují však metody, které se s tímto problémem dokážou vypořádat.

Cílem tohoto příspěvku je vytvořit modely vedoucí ke konstrukci čtvrtletního ukazatele výkonnosti zpracovatelského průmyslu, který bude reprezentován mezičtvrtletním reálným vývojem hrubé přidané hodnoty (dále HPH) tohoto odvětví. HPH představuje vhodnější makroagregát pro odhadování i s ohledem na to, že je sledována v odvětvovém členění klasifikace CZ-NACE a na rozdíl od HDP není zatížena údaji o daních a dotacích na produkty. Tato skutečnost následně umožňuje obdobným postupem získat odhady vývoje HPH i pro další odvětví. Ty by pak po vhodném způsobu agregování měly vytvořit signální odhad celkového vývoje HPH. Zpracovatelský průmysl je odvětvím, které se nejvíce podílí na tvorbě HDP (dlouhodobě z 25–27 %). Dílčím cílem příspěvku je i zhodnotit a porovnat zvolené metody, které jsou v praxi často k odhadům využívány.

Východiska metod smíšených frekvencí

Makroekonomické ukazatele jsou často sledovány a zjišťovány v rozdílných časových periodicitách. Některé jsou sledovány čtvrtletně (např. HDP), některé měsíčně (např. zaměstnanost, míra inflace, tržby v maloobchodě). V případě finančních ukazatelů jde často i o denní frekvenci. Vzhledem k dostupnosti dat sledovaných v měsíční periodicitě můžeme získat předpovědi o vývoji čtvrtletního HDP za již skončené čtvrtletí, případně i za čtvrtletí, ve kterém je tato předpověď konstruována (tzv. nowcasting[1]). Nevýhodou standardních regresních modelů je potřeba pracovat s daty se stejnou periodicitou sledování. Tedy závisle proměnná i nezávisle proměnné musí být sledovány buď čtvrtletně, nebo měsíčně (odhlédneme-li od jiných periodicit). Tradičním řešením tohoto problému je agregovat data s vyšší periodicitou na data s nižší periodicitou. K tomu se obvykle využívají techniky průměrování, součtu či použití hodnoty posledního měsíce v daném čtvrtletí jako reprezentanta tohoto čtvrtletí. To však záleží na charakteru dat, s kterými pracujeme. V případě, že přistoupíme k těmto technikám, musíme počítat s tím, že jejich vlivem dojde ke ztrátě určitého množství informace, kterou v sobě vysokofrekvenční data nesou a může se to projevit i v jiném charakteru dynamiky modelu. Tato skutečnost je obecně důsledkem časové agregace.

MIDAS regrese

Řešením, jak zkombinovat data s odlišnou periodicitou sledování a nedovolit tak ztrátu informace, je použití tzv. MIDAS[2] regrese (Franta a kol., 2014). MIDAS regrese představuje kompromis mezi dvěma přístupy. Tím prvním jsou jednoduché agregační techniky, jako součet či průměr, které přisuzují každému vysokofrekvenčnímu pozorování stejnou váhu (např. zprůměrování tří měsíců v rámci čtvrtletí). Druhým přístupem je umožnit, aby každé vysokofrekvenční pozorování mělo jiný (svůj vlastní) regresní koeficient. Zde je však problémem tendence získat velmi velké množství odhadnutých koeficientů. MIDAS regrese tak nabízí kompromis mezi těmito dvěma přístupy, který se projevuje tím, že jednotlivým pozorováním jsou přiřazeny nestejné váhy a je snížen počet odhadnutých koeficientů. Toho je dosaženo vyrovnáním modelu rozdělení zpoždění jako váhové funkce. Výhodou je zachování časové informace a flexibilita. Odhad je založen na nelineární metodě nejmenších čtverců (Kuzin a kol., 2009), která vychází z počátečních odhadů parametrů.

Základní rovnicí MIDAS regrese je:

$$Y_t^L = \sum_{i=1}^{q} \beta_i W_{t-i}^L + \lambda f\left(\gamma, X_{j,t}^H\right) + \varepsilon_t ,$$

kde

Y_t^L – závisle proměnná s nižší periodicitou sledování (čtvrtletí),

[1] Predikce pro současnost, velmi blízkou budoucnost nebo velmi blízkou minulost.
[2] Mixed Data Sampling

$W_{t\text{-}i}^{L}$ – soubor nezávisle proměnných se stejnou periodicitou jako závisle proměnná, případně závisle proměnná ve svých zpožděních,

$X_{j,t}^{H}$ – soubor nezávisle proměnných s vyšší periodicitou sledování (měsíce),

β_i, λ a γ – odhadované parametry,

$f(\ )$ – funkce, která transformuje data s vyšší periodicitou sledování na nižší periodicitu,

L (*low*) – nižší periodicita sledování, H (*high*) – vyšší periodicita sledování, t – časový okamžik, i – délka zpoždění.

Jako transformující funkce zde vystupuje vhodné váhové schéma. Literatura (Ghysels a kol., 2004, Kuzin a kol., 2009, Armesto a kol., 2010 nebo Schumacher, 2014) i software EViews nabízejí celou řadu schémat. Ve svém výzkumu jsem se zaměřil na následující dvě schémata:

1) Almonova metoda vážení

$$Y_t^L = \sum_{i=1}^{q} \beta_i W_{t-i}^L + \sum_{i=1}^{p} \gamma_i \sum_{j=0}^{k} j^{i\text{-}1} X_{t-j}^H + u_t ,$$

kde

k – počet zpoždění, p – stupeň polynomu.

2) U-MIDAS (Unrestricted Lag Polynomials)

$$Y_t^L = \sum_{i=1}^{q} \beta_i W_{t-i}^L + \sum_{j=0}^{m\text{-}1} \gamma_{t-j} X_{t-j}^H + u_t .$$

Tento typ váhového schématu je vhodný pro případy, kdy rozdíly v periodicitě sledování ukazatelů jsou malé (např. čtvrtletní a měsíční data). V této rovnici odhadujeme koeficienty pro každé zpoždění proměnné s vyšší periodicitou sledování.

Jako rozšíření základní MIDAS regrese navrhli Clements a Galvão (2008) tzv. autoregresní MIDAS model, který spočívá v přidání autoregresní složky do základní MIDAS rovnice. Tato složka může být spolu s dalšími odhadnuta nelineární metodou nejmenších čtverců. Autoři se domnívají, že rovnice v této formě může přispět ke zdokonalení odhadu.

Bridge modely

Alternativním způsobem využití dat s vyšší periodicitou sledování k odhadování či předpovídání časových řad s nižší periodicitou sledování jsou tzv. bridge[3] modely či rovnice. Tyto modely jsou hojně využívány centrálními bankami, které je aplikují ve svých modelech předpovědí (Arnoštová a kol., 2011). Spočívají v myšlence agregování dat s vyšší periodicitou sledování (měsíční) na data s nižší periodicitou sledování (čtvrtletí). Jedná se tedy o čtvrtletní regresní model rozšířený o čtvrtletní předpovědi měsíčních dat. Odhad je založen na klasické metodě nejmenších čtverců. Bridge model můžeme zapsat jako:

$$Y_t^L = \beta_0 + \sum_{i=1}^{j} \beta_i(L)\, X_t^L + u_t ,\ u_t \sim i.i.d.N\left(0;\sigma^2\right),$$

kde

X_t^L – vybrané měsíční ukazatele agregované na čtvrtletní periodicitu,

L – operátor zpoždění.

Stejně jako v případě MIDAS regrese, i do tohoto modelu je možné přidat autoregresní složku závisle proměnné.

Agregace dat s vyšší periodicitou sledování se odvíjí od toho, s jakými daty pracujeme (Schumacher, 2014). Agregační pravidla pro stacionární a nestacionární časové řady uvádí Stock a Watson (2002).

[3] V literatuře se nevyskytuje český ekvivalent. Výraz lze přeložit jako modely přemostění či můstkové modely.

Na základě práce Stocka a Watsona (2002) jsem pro agregaci měsíčních údajů zvolil vážený aritmetický průměr[4]:

$$x_t^q = \frac{1}{9}\left(x_t^m + 2\,x_{t-1/3}^m + 3\,x_{t-2/3}^m + 2\,x_{t-1}^m + x_{t-1-1/3}^m\right) \text{ pro } t = 3,6,9,12,\ldots,$$

kde

x_t^q – agregované mezičtvrtletní tempo růstu a x_t^m – meziměsíční tempo růstu.

Výhodou bridge modelů je jejich relativně snadná použitelnost, na druhou stranu se však potýkají s některými nevýhodami. Tou nejpodstatnější nevýhodou je fakt, že agregací měsíčních údajů na čtvrtletní údaje dochází ke ztrátě časové informace. Dále jsou tyto modely závislé na včasnosti použitých ukazatelů a často tak dochází k tomu, že v rámci odhadovaného čtvrtletí nejsou k dispozici všechny hodnoty měsíčního ukazatele.

Obě výše uvedené metody jsou založeny na podobném principu kombinace dat s vyšší a nižší periodicitou sledování. Přesto se v některých důležitých bodech odlišují.

MIDAS regrese je založena na přímém odhadu a v rámci předpovědí musí být model pro každý horizont znovu odhadnut. Bridge modely naopak odhadovány iterativně a je tedy nutné napřed odhadnout případná chybějící měsíční pozorování v rámci odhadovaného čtvrtletí a následně agregovat příslušné ukazatele na nižší periodicitu sledování. Odhady získané MIDAS regresí a bridge modelů se liší v důsledku použití odlišných polynomů zpoždění. V MIDAS regresi musí být použita k odhadu nelineární metoda nejmenších čtverců vzhledem k jejich nelineárnímu charakteru. U bridge modelů jsou koeficienty odhadovány s použitím klasické metody nejmenších čtverců.

MIDAS regrese na rozdíl od bridge modelu v sobě může zahrnovat také pozorování ukazatele s vyšší periodicitou sledování vztažená ke čtvrtletí, ve kterém je odhad prováděn. U bridge modelů jsou zahrnuty pouze pozorování těchto ukazatelů do konce referenčního čtvrtletí, za které odhadujeme.

Chow-Linova metoda

Chow-Linovu metodu řadíme do skupiny nepřímých metod sestavování čtvrtletních národních účtů, tedy skupiny metod, které využívají výsledky krátkodobých šetření. Jde o techniku retropolace, která v tomto rámci spočívá v propojení krátkodobě zjišťovaného ukazatele s ukazatelem ročních národních účtů. „*Model tohoto vztahu je založen na specifickým způsobem volené regresní funkci mezi krátkodobě zjišťovaným ukazatelem a ukazatelem ročních národních účtů. Takový model dává možnost rozvrhnout dosud známé roční hodnoty do jednotlivých čtvrtletí tak, aby jejich součet byl roven roční hodnotě, ale umožňuje též odhadnout hodnot příslušného agregátu pro běžné (popř. následující) čtvrtletí*.“ (Marek a kol., 2016) Dále Marek a kol. (2016) a pro konkrétní případ vztahu mezi čtvrtletním rozvrhovaným ukazatelem a referenčním měsíčním ukazatelem též Pečený (2006) uvádějí požadavky, které má vybraný referenční ukazatel splňovat. Mezi ně patří věcná souvislost rozvrhovaného a referenčního ukazatele, jeho dostupnost, rychlost a nízká finanční náročnost zjišťování, spolehlivost a další formální kritéria (např. délka časových řad).

Chow-Linův model můžeme zapsat jako jednoduchou regresní rovnici:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

kde

y – vektor rozvrhovaného ukazatele s nižší periodicitou sledování o rozměru T x 1,

X – matice p referenčních ukazatelů[5] s vyšší periodicitou sledování agregovaných na nižší periodicitu sledování o rozměru T x p,

$\boldsymbol{\beta}$ – vektor koeficientů o rozměru p x 1,

u – vektor náhodných veličin s nulovou střední hodnotou a kovarianční maticí $\mathbf{V} = E(\mathbf{u}\mathbf{u}')$.

[4] Další možností je použití prostého aritmetického průměru.

[5] V případě Chow-Linovy metody bez využití referenčního ukazatele by **X** byla jednotkovým vektorem T x 1

Výběr ukazatelů zpracovatelského průmyslu vhodných pro analýzu

Poměrně důležitou roli v celé analýze hrají i vhodně vybrané ukazatele. Ačkoliv se do značné míry jedná o subjektivní volbu, musí být dodrženy určité požadavky, které na data klademe. Mělo by se jednat o ukazatele, které věcně souvisí s odhadovaným makroagregátem. Dále by neměly podléhat příliš častým revizím a měly by pocházet z důvěryhodných a oficiálních zdrojů. Další důležitou vlastností je pak i jejich včasná publikace a dostupnost. Do analýzy jsem zahrnul kromě ukazatelů přímo souvisejících s odvětvím zpracovatelského průmyslu i ukazatele obecnější (např. podíl nezaměstnaných osob či index spotřebitelských cen), které taktéž mohou ovlivňovat vývoj odvětvové HPH.

V tabulce 1 jsou uvedeny vybrané měsíční ukazatele zpracovatelského průmyslu.

Tabulka 9: Ukazatele zpracovatelského průmyslu Zdroj: ČSÚ, ČNB, MPSV

Název ukazatele	
Index průmyslové produkce	Nové zakázky ze zahraničí
Tržby z průmyslové činnosti	Index cen průmyslových výrobců
Tržby z přímého vývozu	Podíl nezaměstnaných osob
Domácí tržby	Index cen vývozu
Podnikatelský indikátor důvěry	Index cen dovozu
Spotřebitelský indikátor důvěry	Vklady klientů u bank
Souhrnný indikátor důvěry	Bankami přijaté úvěry od klientů
Indikátor důvěry v průmyslu	Míra zaměstnanosti
Index spotřebitelských cen	Obecná míra nezaměstnanosti
Nové zakázky z tuzemska	

Tyto výše uvedené ukazatele tvoří výchozí soubor, ze kterého je potřeba vybrat ukazatele vhodné pro odhady HPH zpracovatelského průmyslu. K tomuto jsem využil test Grangerovy kauzality.

Grangerovu kauzalitu můžeme definovat podle Arlta a Arltové (2009) či podobně podle Lütkepohla a Krätziga (2004) následovně: jestliže řada *X* působí v Grangerově smyslu na řadu *Y*, pak řada *X* by měla pomoci zlepšit předpovědi řady *Y*. Jinými slovy, pro střední čtvercové chyby platí:

$MSE\left[Y_t\left(h|\Omega_t\right)\right] < MSE\left[Y_t\left(h|\Omega_t \setminus \left\{X_{t-i}, i \geq 0\right\}\right)\right]$ pro alespoň jeden z horizontů $h = 1, 2, ...,$

kde

Ω_t – soubor všech informací existujících v čase *t*,

$\Omega_t \setminus \left\{X_{t-i}, i \geq 0\right\}$ – soubor všech informací s výjimkou informací obsažených v minulosti a přítomnosti procesu $\left\{X_t\right\}$.

Pro testování byla uvažována ve všech případech 5% hladina významnosti, všechny řady byly pomocí rozšířeného Dickey-Fullerova testu (Arltová a Fedorová, 2016) testovány, zda jsou stacionární či nikoliv.

Všechny zahrnuté ukazatele byly transformovány na mezičtvrtletní tempa růstu a u všech byla testem prokázána stacionarita[6].

V tabulce 2 je uveden pro názornost výstup výsledku testu Grangerovy kauzality mezi HPH zpracovatelského průmyslu a indexem průmyslové produkce ze softwaru EViews. Obdobným způsobem byly testovány vztahy mezi HPH a dalšími ukazateli.

Tabulka 2: Test Grangerovy kauzality pro ukazatel indexu průmyslové produkce

Pairwise Granger Causality Tests
Sample: 2000Q2 2014Q3
Lags: 4

Null Hypothesis:	Obs	F-Statistic	Prob.
IPP does not Granger Cause HPH	53	3.74100	0.0105
HPH does not Granger Cause IPP		0.77243	0.5490

[6] Důležitá podmínka pro možnost odhadu parametrů stochastického procesu. Je-li proces stacionární, platí, že všechny střední hodnoty a rozptyly jsou v čase stejné a kovarianční a korelační funkce závisí pouze na časové vzdálenosti náhodných veličin.

Z výsledku je prokázáno, že mezičtvrtletní vývoj indexu průmyslové produkce (IPP) působí v Grangerově smyslu na mezičtvrtletní vývoj HPH zpracovatelského průmyslu (p-hodnota 0,0105 < 0,05). Jinými slovy zamítáme nulovou hypotézu o tom, že jedna řada (IPP) nepůsobí v Grangerově smyslu na druhou řadu (HPH).

Po otestování všech uvažovaných ukazatelů jsem došel k závěru, že Grangerova kauzalita byla prokázána u 9 z nich. Konkrétně se jedná o tyto ukazatele: index průmyslové produkce, domácí tržby, index cen průmyslových výrobců, index spotřebitelských cen, index důvěry v průmyslu, nové zakázky ze zahraničí, souhrnný indikátor důvěry, souhrnný podnikatelský indikátor důvěry a tržby z průmyslové činnosti. V případě ukazatele nových zakázek ze zahraničí však test vyšel poměrně těsně (p-hodnota 0,0497), a proto jej při konstrukci modelů uvažovat nebudu.

Odhady hrubé přidané hodnoty zpracovatelského průmyslu

Po výběru ukazatelů, které by zařazením do modelů měly zlepšit předpovědi, resp. odhady HPH zpracovatelského průmyslu, mohu přistoupit k samotnému jádru této analýzy. Odhady budou získány za období 3. čtvrtletí 2014 až 3. čtvrtletí 2016 s využitím dat zahrnujících období od 1. čtvrtletí 2000 (ve formě temp růstu pak od 2. čtvrtletí) do skončení referenčního čtvrtletí, ve kterém jsou s ohledem na odhadované období k dispozici potřebné údaje. Všechny t-testy a F-testy budou vyhodnocovány na hladině významnosti 5 %. U modelů bude taktéž provedena diagnostická kontrola reziduí[7]. Výsledné odhady byly hodnoceny podle několika kritérií, zejména pak podle kritéria Theilovo U_2:

$$U_2 = \sqrt{\frac{\sum_{i=1}^{T-1} \left(\frac{\hat{y}_{i+1} - y_{i+1}}{y_i}\right)^2}{\sum_{i=1}^{T-1} \left(\frac{y_{i+1} - y_i}{y_i}\right)^2}},$$

kde

y_i – pozorované hodnoty, $\hat{y}_i$ – odpovídající předpovědi.

Hodnoty tohoto kritéria menší než 1 představují zlepšení předpovědi oproti případu, kdy by nedošlo k žádné změně (předpověď beze změny).

MIDAS regrese stejně jako klasický regresní model umožňuje uvažovat více než jednu nezávisle proměnnou. Zde se však výrazně projevuje nevýhoda MIDAS regrese. Tou je skutečnost, že uvažování většího množství nezávisle proměnných, kromě toho, že se zde může projevit multikolinearita, vede k dramatickému zvýšení složitosti odhadu. Z tohoto důvodu se obvykle postupuje tak, že jsou odhadnuty modely pro každou nezávisle proměnnou zvlášť a následně z nich plynoucí předpovědi jsou zprůměrovány. Ve svém výzkumu jsem tedy konstruoval 8 modelů s využitím Almonovy metody vážení a 8 modelů U-MIDAS. Předpovědi jsem poté zprůměroval. U obou váhových schémat dosahovaly nejlepších výsledků modely s využitím indexu průmyslové produkce a indikátoru důvěry v průmyslu. Proto jsem následně konstruoval modely pro oba tyto ukazatele jako nezávisle proměnné najednou, tentokrát však pouze s využitím Almonovy metody vážení. U-MIDAS v tomto případě nebyla použita, vzhledem k tomu, že toto schéma vyžaduje testování stejného počtu zpoždění obou ukazatelů, což je poměrně přísný předpoklad.

V tabulce 3 je uveden výstup programu EViews pro model s použitím indexu průmyslové produkce.

V nabídce byla vybrána možnost automatického zvolení optimálního počtu zpoždění proměnné, zde konkrétně 6 zpoždění (včetně zpoždění 0), a to pro polynom 3. řádu. V dolní části výstupu jsou uvedeny již samotné odhadnuté koeficienty a graficky i jejich rozdělení s dopadem jednotlivých zpoždění na růst HPH.

Tabulka 3: Výstup MIDAS regrese (Almonova metoda vážení) mezi HPH a IPP

Dependent Variable: HPH
Method: MIDAS
Sample: 2001Q2 2014Q2
Included observations: 53
Method: PDL/Almon (polynomial degree: 3)
Automatic lag selection, max lags: 12

[7] Testování přítomnosti autokorelace, heteroskedasticity a normálního rozdělení.

Chosen selection: 6

Variable	Coefficient	Std. Error	t-Statistic	Prob.
HPH(-4)	-0.237331	0.113835	-2.084875	0.0423
Page: MONTH Series: IPP(-1) Lags: 6				
PDL01	-0.632499	0.240507	-2.629857	0.0114
PDL02	0.634852	0.156040	4.068521	0.0002
PDL03	-0.091060	0.021433	-4.248631	0.0001

R-squared	0.393739	Mean dependent var	1.013762
Adjusted R-squared	0.393739	S.D. dependent var	0.032682
S.E. of regression	0.025447	Akaike info criterion	-4.372507
Sum squared resid	0.033674	Schwarz criterion	-4.223806
Log likelihood	119.8714	Hannan-Quinn criter.	-4.315324
Durbin-Watson stat	1.784781		

MONTH(-1)	Lag	Coefficient	Distribution
	0	-0.088707	*
	1	0.272966	*
	2	0.452518	*
	3	0.449949	*
	4	0.265261	*
	5	-0.101548	*

V případě Chow-Linovy metody jsem jako referenční ukazatel použil index průmyslové produkce, který kromě toho, že splňuje výše uvedená kritéria, dosahoval mezi jednorozměrnými modely MIDAS regrese jednoznačně nejlepších výsledků a je obecně považován za dobrou proxy proměnnou při odhadech HPH či HDP. Po získání měsíčních temp růstu HPH zpracovatelského průmyslu jsem tato tempa odhadoval pro příslušná období na základě regresního vztahu právě s indexem průmyslové produkce.

Vyhodnocení odhadů

Jak jsem naznačil v předchozí kapitole, vyhodnocení odhadů získaných jednotlivými metodami bylo posuzováno podle kritéria U_2. Vzhledem k tomu, že pro každé časové období je proveden pouze bodový odhad, což neumožňuje provést relevantní vyhodnocení, vypočítal jsem výsledné charakteristiky na odhadech za celé časové období, za která jsou k dispozici data. V tabulce 4 jsou uvedeny dvě nejúspěšnější metody pro každé časové období spolu s kritériem U_2.

Z výsledků nelze jednoznačně určit, která z vybraných metod podává obecně kvalitnější odhady. Nicméně, na předních místech se vždy umístila metoda bridge modelů a metoda MIDAS regrese s Almonovým schématem vah s indexem průmyslové produkce a indikátorem důvěry v průmyslu jako nezávisle proměnnými (v tabulce označeno jako MIDAS (Almon) – mix). U obou přístupů se hodnota kritéria U_2 pohybuje v rozmezí 0,4 až 0,5, což potvrzuje, že odhady jsou poměrně úspěšné, ačkoliv prostor pro zlepšení zde samozřejmě stále je.

Tabulka 4: Vyhodnocení dvou nejúspěšnějších metod

Období	**I. nejúspěšnější metoda**	**U_2**	**II. nejúspěšnější metoda**	**U_2**
2014Q3	MIDAS (Almon) – mix	0,442	Bridge model	0,453
2014Q4	Bridge model	0,426	MIDAS (Almon) – mix	0,441
2015Q1	MIDAS (Almon) – mix	0,446	Bridge model	0,466
2015Q2	Bridge model	0,420	MIDAS (Almon) – mix	0,445
2015Q3	MIDAS (Almon) – mix	0,471	Bridge model	0,486
2015Q4	MIDAS (Almon) – mix	0,447	Bridge model	0,485
2016Q1	Bridge model	0,417	MIDAS (Almon) – mix	0,446
2016Q2	Bridge model	0,415	MIDAS (Almon) – mix	0,447
2016Q3	Bridge model	0,393	MIDAS (Almon) – mix	0,472

Další použité přístupy již s lehkým odstupem zaostávají. Konkrétně se jedná o regresi U-MIDAS (se zprůměrovanými odhady), kde kritérium U_2 osciluje mezi 0,52 a 0,55 (v případě modelu s indexem průmyslové produkce jako nezávisle proměnnou pak s mírně lepším výsledkem 0,50-0,51). Srovnatelně si vedla i MIDAS regrese s Almonovým schématem vah (v případě zprůměrování odhadů se jejich kvalita pohybuje v rozmezí 0,51-0,53) a analogicky v případě modelu s indexem průmyslové produkce pak v rozmezí 0,51-0,55. Naopak Chow-Linova interpolační metoda s referenčním ukazatelem indexu průmyslové produkce výrazně za ostatními zaostává, což je v souladu s očekáváními. Zde se hodnoty U_2 pohybují mezi 0,66 a 0,69.

Srovnáme-li výsledné odhady se skutečnými tempy růstu HPH zpracovatelského průmyslu[8] za období, pro která pořizujeme odhad, zjistíme, že v případě bridge modelů jsou odhady poměrně variabilní na rozdíl od odhadů pořízených ostatními přístupy. Z tohoto hlediska naopak nižší variabilitu vykazují zprůměrované odhady získané MIDAS regresí s Almonovým schématem vah, příp. U-MIDAS regresí.

Závěr

Provedená analýza potvrdila, že využití měsíčních časových řad ukazatelů v kombinaci s uvedenými metodami může do značné míry podat uspokojivé výsledky při odhadování HPH zpracovatelského průmyslu. Nelze jednoznačně určit, která z uvedených metod je tou nejlepší. Přihlédneme-li však k přímé konfrontaci odhadnutých hodnot s hodnotami skutečnými, mírně kvalitnější odhady jsou dosaženy MIDAS regresí, která využívá přímo měsíční informace, aniž by docházelo k jejich ztrátě případnou agregací. Je tedy podstatné posuzovat kvalitu odhadů nejen podle statistických kritérií, ale taktéž mírou souladu se skutečnými údaji, byť je nutné vypořádat se se skutečností, že oficiální odhady vývoje HPH jakéhokoliv odvětví jsou k dispozici až se značným zpožděním. Stejně tak obecně nelze předpokládat stejnou míru úspěšnosti jednotlivých přístupů při získávání odhadů vývoje v dalších odvětvích, protože každé odvětví je svým způsobem specifické a nemusí postihovat celkový vývoj ekonomické situace.

Dalším faktorem, který může ovlivnit výsledné odhady, je výběr statistických ukazatelů. V tomto článku byly použity ty, které jednak věcně souvisejí s odhadovaným makroagregátem a které jsou snadno dostupné a pocházející z oficiálních a věrohodných zdrojů.

Uvedené metody samozřejmě nejsou jedinými použitelnými alternativami. Poměrně využívaným prostředkem jsou i tzv. dynamické faktorové modely, které jsou obvykle konstruovány prostřednictvím tzv. state-space modelů s využitím rekurzivní techniky odhadu, tzv. Kalmanova filtru. V dalším výzkumu na uvedenou problematiku aplikuji taktéž metodu nejbližších sousedů, která se vyznačuje tím, že nepředpokládá obecný model dynamického procesu, který je typickým znakem časových řad.

Literatura

ARLT, Josef; ARLTOVÁ, Markéta. *Ekonomické časové řady*. 1. vydání. Praha: Professional Publishing, 2009. ISBN 978-80-86946-85-6.

ARLTOVÁ, Markéta; FEDOROVÁ, Darina. Selection of Unit Root Test on the Basis of Length of Time Series and Value of AR(1) Parameter. *Statistika: Statistics and Economy Journal*. Praha: Český statistický úřad, 2016, roč. 96, č. 3, s. 47–64. ISSN 0322-788x.

ARMESTO, Michelle; ENGEMANN, Kristie; OWYANG, Michael. Forecasting with Mixed Frequencies. *Federal Reserve Bank of St. Louis Review*. St. Louis: Federal Reserve Bank of St. Louis, 2010, roč. 92, č. 6, s. 521–536. ISSN 00149187.

ARNOŠTOVÁ, Kateřina; HAVRLANT, David; RŮŽIČKA, Luboš; TÓTH, Peter. *Short-Term Forecasting of Czech Quarterly GDP Using Monthly Indicators* [online]. Praha: Česká národní banka, 2011 [cit. 2016-12-19]. CNB Working Paper Series, č. 12, 32 s. ISSN 1803-7070. Dostupné z: https://www.cnb.cz/miranda2/export/sites/www.cnb.cz/en/research/research_publications/cnb_wp/download/cnbwp_2010_12.pdf

CLEMENTS, Michael; GALVÃO, Ana Beatriz. Macroeconomic Forecasting With Mixed-Frequency Data: Forecasting Output Growth in the United States. *Journal of Business & Economic Statistics*. Washington: American Statistical Association, 2008, roč. 26, č. 4, s. 546–554. DOI 10.1198/073500108000000015.

[8] Výsledky je potřeba brát s rezervou, jelikož získané výsledné odhady mají povahu bleskového odhadu, zatímco skutečné odhady jsou zveřejňovány až po přibližně 65 dnech po skončení referenčního čtvrtletí. Bleskový odhad odvětvových HPH však publikován není.

EUROSTAT. *Preliminary GDP Flash Estimate in 30 Days for Europe* [online]. Lucemburk: Eurostat, 2016 [cit. 2016-11-19]. Eurostat: Statistics Explained. Dostupné z: http://ec.europa.eu/eurostat/statistics-explained/index.php/Preliminary_GDP_flash_estimate_in_30_days_for_Europe

FRANTA, Michal; HAVRLANT, David; RUSNÁK, Marek. *Forecasting Czech GDP Using Mixed-Frequency Data Models* [online]. Praha: Česká národní banka, 2014 [cit. 2016-12-15]. CNB Working Paper Series, č. 8, 28 s. ISSN 1803-7070. Dostupné z: http://www.cnb.cz/en/research/research_publications/cnb_wp/download/cnbwp_2014_08.pdf

GHYSELS, Eric; SANTA-CLARA, Pedro; VALKANOV, Rossen. *The MIDAS Touch: Mixed Data Sampling Regression Models* [online]. Montreal: Centre interuniversitaire de recherche en analyse des organisations, 2004 [cit. 2016-12-15]. Scientific Series, č. 20, 36 s. ISSN 1198-8177. Dostupné z: http://www.cirano.qc.ca/pdf/publication/2004s-20.pdf

KUZIN, Vladimir; MARCELLINO, Massimiliano; SCHUMACHER, Christian. *MIDAS versus Mixed-Frequency VAR: Nowcasting GDP in the Euro Area* [online]. Frankfurt nad Mohanem: Deutsche Bundesbank, 2009 [cit. 2016-12-18]. Discussion Paper, č. 7, 27 s. ISBN 978-86558-509-7. Dostupné z: https://www.econstor.eu/dspace/bitstream/10419/27661/1/200907dkp.pdf

LÜTKEPOHL, Helmut; KRÄTZIG, Markus. *Applied Time Series Econometrics*. 1. vydání. New York: Cambridge University Press, 2004. ISBN 978-0-521-83919-8.

MAREK, Luboš, HRONOVÁ, Stanislava, HINDLS, Richard. Příspěvek k časnějším odhadům hodnot čtvrtletních národních účtů. *Politická ekonomie*. Praha: Vysoká škola ekonomická v Praze, 2016, roč. 64, č. 6, s. 633–650. ISSN 0032-3233.

PEČENÝ, Luboš. Bleskové odhady v systému národních účtů. Praha, 2006 [cit. 2016-12-29]. Diplomová práce. Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze, Katedra ekonomické statistiky. Vedoucí diplomové práce Jakub Fischer.

SCHUMACHER, Christian. *MIDAS and Bridge Equations* [online]. Frankfurt nad Mohanem: Deutsche Bundesbank, 2014 [cit. 2016-12-07]. Discussion Paper, č. 26, 28 s. ISBN 978-86558-509-7. Dostupné z: https://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Discussion_Paper_1/2014/2014_10_21_dkp_26.pdf?__blob=publicationFile

STOCK, James; WATSON, Mark. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*. Washington: American Statistical Association, 2002, roč. 20, č. 2, s. 147–162. DOI 10.1198/073500102317351921.

JEL Classification: C10, C32, C52, C53

Summary

The Use of Mixed-Data Sample Methods in the Construction of Quarterly Indicator of the Manufacturing Industry's Performance in the Czech Republic

The article deals with the introduction of methods combining low- and high-frequency data and its application to the problem of the construction of the quarterly indicator of the manufacturing industry's performance in the Czech Republic. For the calculations software EViews was used.

The article is motivated by the growing demand for timely data on the growth of the economy where it is desirable to obtain information on development approximately 30 days after the end of the reference quarter.

After the necessary pre-selection of indicators of the manufacturing industry the Granger causality test was performed to select those that can help improve forecasts of gross value added. Then, bridge models, MIDAS regression method (Almon and U-MIDAS weighting schemes) and Chow-Lin interpolation method were used and compared.

The quality of the estimates vary in different periods, however, bridge models and MIDAS Almon method perform the best and for all periods under consideration they are ranked at the forefront.

Keywords: manufacturing industry, gross value added, mixed-data sample methods, Granger causality