SBORNÍK

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

Abstrakty

Vědecký seminář se uskutečnil dne 6. února 2020
pod záštitou děkana FIS
prof. Ing. Jakuba Fischera, Ph.D.

Sestavení sborníku
prof. Ing. Petr Doucek, CSc.
proděkan pro tvůrčí činnost a zahraniční vztahy

Nakladatelství Oeconomica – Praha 2020

ISBN 978-80-245-2352-1

OBSAH

Předmluva

„Den doktorandů“ patří mezi tradiční akce, které Fakulta informatiky a statistiky pořádá pro studenty doktorského studia – letos se jednalo již o dvacátý pátý ročník. Seminář se konal 6. února 2020 pod gescí děkana Fakulty informatiky a statistiky prof. Ing. Jakub Fischer, Ph.D. Pro mnohé z nich je to první vystoupení před odbornou veřejností, na němž získávají cenné zkušenosti a tříbí tak i formulace názorů a hypotéz. Kromě toho si vyzkouší presentaci závěrů výzkumné práce a argumentaci na jejich podporu. V letošním roce byly příspěvky, vzhledem k celkovému počtu devíti přihlášených účastníků ze všech studijních programů doktorského studia, sdruženy do jedné sekce.

Nedílnou součástí „Dne doktorandů“ je i práce hodnoticí komise, jejíž členové pečlivě sledují jednotlivá vystoupení a potom vybírají nejlepší práce k ocenění. Hlavními kritérii pro jejich rozhodování byly zejména kvalita a aktuálnost zpracovaného tématu, přístup k řešení vybraného problému, způsob použití metodiky, úroveň práce s reálnými daty a v neposlední řadě i schopnost prezentovat a argumentačně své výsledky obhájit v diskusi. Ti nejlepší z nich získávají prestižní „Cenu děkana FIS“, s níž je spojena i symbolická finanční odměna.

Za práci v hodnoticí komisi chci poděkovat všem jejím členům - prof. Ing. Haně Řezankové, CSc. (KSTP), prof. Ing. Josef Jablonský, CSc. (KEKO) a prof. Ing. Vojtěch Svátek, Dr. (KIZI). Komise se zhostila své práce na výbornou.

V letošním roce získali ceny za nejlepší příspěvky následující studentky a studenti:

1. místo: Ing. Michal Pieter: Application and comparison of Network DEA models in banking sector (školitel prof. Ing. Josef Jablonský, CSc.) – studijní program Ekonometrie a operační výzkum.

2. místo: Ing. Filip Hon: Fertility in the Czech Republic and its modeling (školitel doc. Ing. Jitka Langhamrová, CSc.) – studijní program Statistika.

3. místo: Ing. Jiří Zettel: Issues of data preprocessing for anomaly-based network intrusion detection systems (školitel prof. Ing. Petr Berka, CSc.) – studijní program Aplikovaná informatika.

Oceněným studentům doktorského studia upřímně blahopřeji a doufám, že získané zkušenosti uplatní při své další práci, ať už vědecké nebo v praxi. Uznání také patří všem vědeckým a pedagogickým pracovníkům FIS – školitelům doktorandů, kteří se „Dne doktorandů“ zúčastnili a svým vedením a radami byli nápomocni při zpracování 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 abstraktů.

prof. Ing. Petr Doucek, CSc.

proděkan pro tvůrčí činnost a zahraniční vztahy

STUDIJNÍ PROGRAM
APLIKOVANÁ INFORMATIKA

The evolution of Product Owner's activities

Daniel Remta

xremd03@vse.cz

Ph.D. student of Applied Informatics

Supervisor: doc. Ing. Alena Buchalcevová, Ph.D.,
(alena.buchalcevova@vse.cz)

Agile as an approach to software development is spreading in industry and being adopted also by large organizations. The most used Agile method is Scrum which defines three roles Product Owner, Scrum Master, and Development Team. The mentioned roles are undergoing an evolution due to the increased adoption rates. This paper focuses on the Product Owner's role and provides an overview to how the activities of the role have evolved over time. Using literature review the activities are identified, extracted, categorized and chronologically ordered.

Key words: Product Owner, Activities, PO, Agile

JEL Classification: L86

Analysis of Biomedical Ontologies

Jana Vataščinová

xvatj00@vse.cz

Ph.D. student of Applied Informatics

Supervisor: doc. Ing. Ondřej Zamazal, Ph.D.,
(ondrej.zamazal@vse.cz)

This paper introduces the analysis of selected biomedical ontologies with respect to biomedical ontology matching. The goal of the analysis was to discover characteristics of biomedical ontologies which influence the matching of these ontologies (finding corresponding entities in two different ontologies). The analysis was performed by querying the ontologies using the SPARQL language, by viewing the source code of the ontologies, and by browsing the ontologies in the ontology editor Protégé (e.g., for statistics such as number of classes in an ontology).

Eleven ontologies were analyzed which include all the biomedical ontologies used within the PubChem project. The paper demonstrates that biomedical ontologies have common characteristics which make them very specific. This can impact different processes which include biomedical ontologies, and these processes must take into consideration given nature of the ontologies. Characteristics of biomedical ontologies can have a substantial role in biomedical ontology matching. This must be reflected especially in the phase of choosing a suitable tool or approach for obtaining alignments between chosen ontologies. The characteristics that should be taken into consideration are especially large size of biomedical ontologies, codes as names, taxonomy character or upper ontology alignment.

Key Words: Biomedical ontology, biomedical ontology characteristics, ontology matching.

JEL Classification: O31

Issues of data preprocessing for anomaly-based network intrusion detection systems

Jiří Zettel

jiri.zettel@vse.cz

Ph.D. student of Applied Informatics

Supervisor: prof. Ing. Petr Berka, CSc., (berka@vse.cz)

In this work, I present several issues a researcher deals with when preparing the data to be suitable for the anomaly-based network intrusion detection systems. I introduce the techniques that such devices use, outline basic preprocessing steps, describe the research directions with the focus on data preprocessing and particularly anonymization techniques for the dataset. The dataset used with the experiments consists of the university information system log data. The goal is not to describe exhaustive list of preprocessing steps, but to illustrate one such case and give some practical information to selected steps, because the project for anomaly detection is in an early phase at the moment. Hence the experiments are mostly dedicated to the anonymization of the data. They bring deep insight into data preparation when implementing group-based anonymization techniques. Unlike other research in the field of anonymization, I don't focus on the design of new algorithms, but on the preprocessing steps and on exploring of applicability of existing algorithms. Each algorithm has specific requirements for the data, so preprocessing must be comprehensive. I present how such data can be transformed into relational data, introduce a novel approach for anonymization of IPv4 address in our dataset using several anonymization algorithms and discuss their principles, strengths, and weaknesses. Two ways of preprocessing of IPv4 for k-anonymity algorithms are presented: first, I split IPv4 into four parts and create generalization hierarchies and second I convert IPv4 to integer

values. Finally I propose an improvement in the Mondrian algorithm suitable for categorical attributes which gives better results than the original algorithm.

Key words: Anomaly Detection, Network Security, Data Preprocessing, Anonymization, K-Anonymity, IPv4, Privacy-Preserving

JEL Classification: Y40, C60

STUDIJNÍ PROGRAM EKONOMETRIE A OPERAČNÍ VÝZKUM

Inequalities in the Police Resources in the Regions of the Czech Republic: An Alternative Measurement Approaches

Jakub Hanousek

xhanj52@vse.cz

Ph.D. student Econometrics and operations research

Supervisor: prof. Ing. Mgr. Martin Dlouhý, Dr., MSc,
(dlouhy@vse.cz)

One of the major goals of the government policy is to minimize unwanted regional variations in the police resources in the country. The objective of this study is to present three alternative methods to a comparison of regional police resource capacities. Three alternative methods in this study are the separate evaluation, the common weights model, and the production frontier model based on data envelopment analysis (DEA). The police resources in this study are number of police employees, number of police vehicles and the amount of finance budget in regions. The data cover the year 2014 and come from the police statistics. The Czech Republic is divided into 14 regions. Inequalities in the police resources were measured against to the number of inhabitants in the region. The common weights and production frontier models that take into account the possibility of resource substitution give on average higher capacity scores and show lower differences between regional capacities. Product frontier model gives the lowest differences between regional police resources. I would like to compare inequalities measured against the number of inhabitants with inequalities measured against the number of crimes in the regions of the Czech Republic in the future research.

Keywords: Police resources, inequalities, Czech Republic

JEL Classification: C44

Application and comparison of Network DEA models in banking sector

Michal Pieter

michal.pieter@vse.cz

Ph.D. student Econometrics and operations research

Supervisor: prof. Ing. Josef Jablonský, CSc., (jablon@vse.cz)

Data envelopment analysis (DEA) is a popular mathematical tool used for evaluating the performance of an individual production unit (decision-making unit, DMU) among a set or grouping of like units. They all consume certain inputs and produce certain outputs (inputs and outputs are referred to as factors when grouped together), that are the same for all units. DEA calculates the efficiency of transforming these inputs into outputs, usually on the scale from 0 to 1. In traditional DEA, the internal structure of the unit is not considered. While such abstraction of the real system may often be justified, in some cases the internal processes are too significant to omit. Network data envelopment analysis (NDEA) was proposed as a way to model this structure and thus better capture in the model the intricacies of the real system and thus obtain more precise and meaningful results. Each unit is then composed of a series of interconnected processes (divisions, sub-processes), each of which acts as a DMU itself, with its own set of inputs and outputs. The processes can be connected to each other via links, factors that are output from one process and consumed as an input of another.

Often, the same real-world system can be modelled in multiple ways, depending on many factors, such as the availability of data or how realistic the model should be. In particular, adding further processes and more links, in an effort to bring the model closer to reality, translates to more variables and constraints in the model.

This added complexity requires more computational resources, longer time and is more difficult to construct and interpret. In a previous publication by the author, several methods were proposed to aid in comparing between these models and to measure both the complexity of the model and the quality of the results it gives. This paper serves as a follow-up and applies some of the methods to a practical example, based on data from the banking sector. Said data consists of 20 banks on the Czech market and contains values for a total of 11 factors, such as number of clients and branches, costs, deposits, loans and income.

Compared models consist of a single traditional (black-box) model using Tone's SBM non-oriented model, with both constant and variable returns to scale considered. The network models are based on Tone's and Tsutsui's NSBM model in its most comprehensive form – non-oriented, with link slacks in the objective and variable returns to scale. While all network models are based on this NSBM model, they all utilize it with their particular structure. Namely, a 2-stage and 3-stage models, both in series and a single 4-stage hybrid model that also has parallel processes and one shared input. Results for these 5 models are then calculated using LocalSolver solver.

After the results are obtained, the methods proposed in the previous paper are used to calculate various measures of both model complexity and their differentiating power, with regards to examined dataset. Measures of complexity include number of variables and constraints, number of factors, degree of serialization and parallelism, as well as computational intensity – total time and number of LocalSolver iterations. For differentiating power, number of inefficient units, standard deviation of efficiency scores and a measure based on cluster analysis is used.

Finally, to compute the compromising measure all pairs of complexity and differentiating power measures are considered. Multiple scenarios are also considered with regards to decision-maker preferences by adjusting an exogenous parameter. Finally

the tabulated compromising measures are interpreted, with the conclusion that while the black-box CRS model is preferred (in the context of given data) when complexity and differentiating power are equally important, the 4-stage model becomes more desirable as increased differentiating power is preferred over complexity. Interestingly, 3-stage model seems to be scored similarly to 4-stage model, suggesting negligible effect of introducing another stage.

Key words: Data envelopment analysis, DEA, Network DEA, NDEA, comparison, complexity, banking, banks.

JEL Classification: C38, C61, C65, C67, G21

STUDIJNÍ PROGRAM
STATISTIKA

Fertility in the Czech Republic and its modeling

Filip Hon

xhonf01@vse.cz

Ph.D. student of Statistics

Supervisor: doc. Ing. Jitka Langhamrová, CSc.,
(langhamj@vse.cz)

The article deals with the development of fertility according to the woman's age and in the Czech Republic insufficiently published possibilities of its modeling. The contribution shows the development of the curve of specific fertility rates during the second demographic transition. On the basis of statistical models the projection of the future development of the curve of specific fertility rates up to 2025 is performed. Statistical models methodologically described and for the purposes of calculation of projection used in the paper are Quadratic Spline model and Functional demographic model. The paper shows a big difference in projections when analysing the time series since 1989. The Quadratic Spline model would assume an ongoing shift in age when specific fertility is highest to older age. Overall, fertility rates should increase slightly, mainly due to higher specific fertility rates among women over the age of 30. On the other hand, younger women should achieve even lower age specific fertility rates. The Functional demographic model also anticipates a decline in fertility in women around the age of 20, but does not assume an increase in fertility in older age or overall fertility. However, it should be kept in mind that this is only a projection, which of course may differ from the actual future development.

Key words: Age specific fertility rates, fertility decline, postponing motherhood, statistical modeling.

JEL Classification: J11, J13

BTS: How do respondents in the industry understand the questionnaire?

Veronika Ptáčková

veronika.ptackova@vse.cz

Doktorand oboru statistika

Školitel: prof. Ing. Jakub Fischer, Ph.D., (jakub.fischer@vse.cz)

Having the data to help us outline the future of the economy is every analyst's dream. The Business Tendency Survey helps to describe the mood in the economy, both now and in the nearest horizon (the next three months). The questionnaire is designed as qualitative. But do the respondents understand it as expected? Do they know the definitions of basic terms such as economic situation, stocks or seasonality?

The present paper - with the help of a survey on the survey - opens up a discussion on how comprehensible the form is for respondents in the industry sector. We ask the respondents what economic situation, stock or capacity utilisation means for them. Thanks to the presented outputs, we will know better what we can predict with the help of partial questions and in case of major discrepancies, we can consider changing the methodology on the Czech and European level.

Key words: Business Tendency Survey, survey on survey, prediction ability, leading indicator

JEL Classification: C10, C22, C83

Equivalence of Fault Trees and Stochastic Petri Nets in Reliability Modelling

Ondřej Vozár

vozo01@vse.cz

Doktorand oboru statistika

Školitel: doc. RNDr. Luboš Marek, CSc., (marek@vse.cz)

There are two main methods of modelling reliability of complex industrial systems: i) fault trees, ii) Petri nets. Liu and Chiou (1997) described of equivalence of both methods for given system. Furthermore they found much simpler method to find critical cuts and minimal paths of the Petri net of the system in comparison with corresponding fault tree. The methods are illustrated on model example of three-masted vessel.

Key words: reliability, time to failure, fault tree, stochastic Petri net, exponential distribution.

JEL Classification: C10