SBORNÍK

prací účastníků vědeckého semináře doktorského studia
Fakulta informatiky a statistiky
Vysoká škola ekonomická

Abstrakty

Vědecký seminář se uskutečnil dne 10. února 2022
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 2022

ISBN 978-80-245-2445-0

OBSAH

Předmluva

„Den doktorandů“ patří mezi tradiční akce, které Fakulta informatiky a statistiky pořádá pro studenty doktorského studia. V tomto roce se jednalo již o dvacátý sedmý ročník. Seminář se konal 10. února 2022 pod gescí děkana Fakulty informatiky a statistiky prof. Ing. Jakuba Fischera, Ph.D. Jednání proběhlo v hybridní formě, kdy se část vystoupení konala v on-line formě. Tento fakt ovšem nebyl na úkor kvality vystoupení doktorandů, byť pro mnohé z nich to bylo první vystoupení před odbornou veřejností, na němž získávají cenné zkušenosti a tříbí si tak i formulace názorů a způsob jejich obhajoby. Kromě toho si doktorandi vyzkoušeli presentaci závěrů výzkumné práce a argumentaci na jejich podporu.

V letošním roce byly příspěvky, vzhledem k počtu třinácti přihlášených účastníků ze všech studijních programů doktorského studia, rozděleny do dvou sekcí – Aplikovaná informatika a Kvantitativní metody – studijní programy Ekonometrie a operační výzkum a Statistika.

Nedílnou součástí „Dne doktorandů“ je i práce hodnoticích komisí, jejichž členové pečlivě sledují jednotlivá vystoupení a 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 i schopnost prezentovat a argumentačně své výsledky obhájit v diskusi. Ti nejlepší z účastníků získávají prestižní „Cenu děkana FIS“, s níž je spojena i symbolická finanční odměna.

Za práci v hodnoticí komisi studijního programu Aplikovaná informatika chci poděkovat všem jejím členům – prof. Ing. Vojtěchu Svátkovi, Dr. (KIZI), Mgr. et Mgr. Ing. Františku Sudzinovi, Ph.D. (KSA) a Ing. Filipu Vencovskému, Ph.D. (KIT), za práci v hodnoticí komisi pro Kvantitativní metody – studijní programy Ekonometrie a operační výzkum a Statistika pak prof. Ing. Josefovi Jablonskému, CSc. (KEKO), prof. Ing. Haně Řezankové, CSc. (KSTP) a doc. Ing. Jitce Langhamrové, CSc. (KDEM). Obě komise se zhostily své práce na výbornou.

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

Studijní program – Aplikovaná informatika

1. místo: Ing. Marcel Valový: Effects of Solo, Navigator, and Pilot Roles on Motivation: An Experimental Study

2.místo: **Ing. Markéta Smolníková:** The development of data literacy measurement

3.místo: Mgr. Jana Syrovátková: Sharing and forwarding messages via social networks

Studijní programy – Ekonometrie a operační výzkum a Statistika

1. místo: Ing. Petr Pokorný: Coordination contracts in multi echelon supply chains

2. místo: Ing. MUDr. Lubomír Štěpánek: Selected alternatives to statistical inference in survival analysis Coordination contracts in multi echelon supply chains: principles and some properties of methods comparing survival curves

3. místo: Ing. David Morávek: Life expectancy and it's past future development in Czechia

Oceněným studentům doktorského studia upřímně blahopřeji a pevně věřím, 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 Ing. Tereze 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

Risk of cloud computing

Jan Andraščík

xandj19@vse.cz

Ph.D. student of Applied informatics

Supervisor: prof. Ing. Petr Doucek, CSc. (doucek@vse.cz)

Risk management is a big area and managing risks in cloud environments has its own challenges. To make risk management and specifically risk assessment in the cloud easier, there are several methodologies which can be used. Most of the methodologies contain not only the methodology on how to perform the risk assessment, but also contain a bank of vulnerabilities, threats and sometimes risks. These banks can then help the assessors during the identification phase of the risk assessment.

While the banks help the assessors, they can also cause confusion as they can be contradicting among methodologies as the terms “vulnerability”, “threat” and “risk” are used interchangeably. This paper therefore focuses on unification of the terms and using literature review provides a list of vulnerability and threat groups which can then be used during the identification phase of the risk assessment and guide the assessor in the areas to evaluate.

Keywords: vulnerability, threats, risks, risk assessment, information security, cloud computing, cloud service providers

JEL Classification: L86

Literature review on Social Identity in the online environment

Jiří Korčák

jiri.korcak@vse.cz

Ph.D. student of Applied informatics

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

Social identity is a well-known concept in sociology. With the advent of the Internet, also this field of science is moving into the online environment and the relationship of people to other people and to technology is evolving. In order to understand recent research on social networks and the role of humans on these networks, it is first necessary to conduct a detailed systematic analysis of the existing knowledge and literature on this topic. This analysis will be conducted using the Systematic Literature Review method. The subsequent findings will be used as a basis for follow-up work regarding the relationship between humans and the community on social networks. This is the first step in understanding a complex two-way relationship that can provide many insights in the corporate, scientific and political worlds.

Keywords: Network, Social Innovation, Information and Internet Services, Belief

JEL Classification: D85, O35, D83, L86

Analysis of suitable frameworks for AI adoption in the public sector – a literature review

Václav Pechtor

pecv06@vse.cz

Ph.D. student of Applied informatics

Supervisor: prof. Ing. Josef Basl, CSc., josef.basl@vse.cz

Artificial intelligence is gaining momentum in the public sector. Governments and municipalities are trying to catch up to the private sector, where the adoption of AI is further advanced. Research about AI in the public field is still in the early stages, but publications have increased in the last few years. This paper analyzes the current literature regarding the adoption of AI in the public sector. The goal is to evaluate if there are suitable frameworks that help public institutions introduce, build, and run AI applications. To this goal, articles are evaluated how much the existing frameworks support the adoption AI process.

Keywords: Artificial intelligence; machine learning; literature review; public sector

JEL Classification: O33

Multi-Class Classification of biomedical research documents

Gollam Rabby

rabg00@vse.cz

Ph.D. student of Applied informatics

Supervisor: doc. Ing. Tomáš Kliegr, Ph.D. (tomas.kliegr@vse.cz)

In most research document repository platforms, biomedical document classification is a crucial task. It can assist researchers or readers in listing a research paper in an appropriate category. At first glance, biomedical document classification is merely an instance of text classification problems. However, biomedical documents possess some properties very different from general text classification because of the dependency on the title, abstract, body, and bibliometric data. A title and abstract is usually very short description, and sometimes the title is an incomplete sentence. A biomedical document classifier may need to be designed differently from a text classifier, although this issue has not been thoroughly studied. In this paper, using a large-scale real-world COVID-19 research paper data set, called CORD-19 data set, we investigate text classification procedures on title, abstract and bibliometric data. These procedures include data prepossessing such as word stemming, stop-word removal, then feature representation and multi-class classification. Our major findings include that TF-IDF and Binary document works better than entity-based representation. But it may be changed to a small fragment if possible to use the full-text entities. But because of the low computational power, it is not possible for us to use the full text or a full document. Also, there is not a colossal difference between using only the title and only the abstract for the biomedical documents. So, we can say that we can get a minimum summary in the title for the biomedical document. Further, multi-layer perception and random forest classifier work well for all over the document representation methods.

Keywords: COVID-19, Machine learning, Multi-class classification, Document classification

JEL Classification:

The Development of Data Literacy Measurement

Markéta Smolníková

xsmom00@vse.cz

Ph.D. student of Applied informatics

Supervisor: doc. Ing. Ota Novotný, Ph. D., novotny@vse.cz

The article summarizes the progress of the research project on Data Literacy Measurement. Despite the existence of several commercial or academic initiatives, there hasn’t been a comprehensive and objective tool for data literacy measurement available. The development of such measurement has to be preceded by the specification of what the data literacy encompasses. As our research focuses on business workforce, we executed a preliminary survey regarding the importance of work objectives that require working with data of the most proliferated job roles. The survey results confirmed assumed trends that different roles require different data literacy skills for accomplishing their work objectives. Linking these objectives with appropriate data competencies, knowledge and skills to handle data, validated our data literacy competency model that comprises of all competencies necessary for working with data at business positions like marketing specialists or mid-level managers. What is more, mapping the priority objectives of examined job roles and the adequate competencies offer the simplest manual to enhance the most importance competencies for the selected roles. The validated competency model subsequently served as a steppingstone for developing an automated and hands-on data literacy assessment tool which currently provides two versions of a questionnaire (based on different datasets) for tracking a progress of its respondents (e.g. before and after a data course). Both test versions were tested with about 300 first year business or IT students and wait for a statistical evaluation of its validity and reliability.

Keywords: IT Management; Training

JEL Classification: M15; M53

Sharing and forwarding messages via social media

Jana Syrovátková
syrj00@vse.cz

Ph.D. student of Applied informatics

Supervisor: prof. Ing. Petr Doucek, CSc. (petr.doucek@vse.cz)

Sharing and forwarding messages are one of the basic functions of social media. However, spreading fake news through social networks becomes increasingly pressing issue. In 2018, an interesting survey (N=362) focused on privacy and social networks have been conducted at the University of KwaZulu-Natal, Pietermaritzburg campus in Republic of South Africa (RSA), which inspired us to replicate in 2020 at the Prague University of Economics and Business. 450 respondents took part in the survey, of which 353 respondents were students of the Prague University of Economics and Business, 31 were students of other schools and 66 were non-students.

This article answers research questions: How much do students in the Czech Republic share news on social networks? How is it different from South African students and older non-students?

It turned out that students from University of KwaZulu-Natal take social media as a source of news more often than students of the Prague University of Economics and Business and recommend them more often. Students in the Czech Republic similarly take social media as a source of news more often than older non-students. Non-students tend to share specific individual messages. Overall, however, we can say that all groups are careful about sharing messages.

The limitation of the research was that the respondents themselves said how much they shared or did not share. However, it is still clear that respondents think about the topic of sharing and are careful about what they share or do not share.

Keywords: Economics, Social networking (online), Fake news, Safety

JEL Classification: D83, O35

Effects of Solo, Navigator, and Pilot Roles on Motivation: An Experimental Study

Marcel Valový

xvalm00@vse.cz

Ph.D. student of Applied informatics

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

[Context] We face a period in time that requires exploring alternative ways of increasing creativity and productivity in software teams. The main factor for their implementation is the effect on the motivation of developers. Motivated and cooperating team members are vital for any software project's success.

[Objective] This study aimed for an in-depth, socio-contextual, and detailed description and interpretation of the topic of pair programming. This agile practice in which programmers change from solo to roles of pilot and navigator was investigated concerning the possibility of increasing intrinsic motivation.

[Method] Using a mixed-methods approach, the present study examined a proposed nomological network of personality traits, pair programming, and motivation. Three experimental sessions produced data in two software engineering university classrooms and were quantitatively investigated by the non-parametric Kruskal-Wallis test and hierarchical cluster analysis. Consequently, the authors conducted semi-structured interviews with twelve experiment participants and utilized the thematic analysis method in an essentialist's way to produce themes overarching participants' attitudes towards pair programming.

[Results] The systematic coding of interview transcripts elucidated the research by producing seven themes ascertaining that pair programming fosters both positive and negative attitudes moderated by personality variables. Quantitative analysis of participants' psychometrics (N = 39) established the existence of three main clusters of software engineers. The data collected from 654 intrinsic motivation inventories confirmed that personality traits significantly affect the motivational effects of pair programming. The data revealed that the suitability of programmers for a given role could be determined by their predominant personality dimensions: (a) pilot – openness, (b) navigator - extraversion, agreeableness, (c) solo – neuroticism, low-extraversion.

[Conclusion] The executed experimental design has proven viable for providing the rich data corpus needed for inspecting the associations in our proposed nomological network. The aims of this study were reached; the results carry the potential to aid managers in deciding whether to try and introduce pair programming in their teams or not and deliver guidelines for assigning roles based on psychometrics.

Keywords: pair programming, agile development, intrinsic motivation, big five, thematic analysis, cluster analysis, software engineering

JEL Classification: L86, M15

Proposal of a methodology for the creation of a tool designed to support the work of the internal IT audit department

Ladislav Vaněk

xvanl24@vse.cz

Ph.D. student of Applied informatics

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

This paper is devoted to the design of a methodology for the creation of a tool designed to support the solution of thematic areas (audits) within the IT part of the internal audit department. The methodology is developed using a design type of research, the so-called Design Science Research (DSR). The result of this type of research should be a viable artifact, in this case the methodology and the tool created according to it. The usability of this tool for effective knowledge sharing within internal (IT) audit departments will then be tested in practice, including portability across a range of commercial insurance companies. The tool is created in addition to the standard audit trail tool that most internal audit departments already use. It should serve as a central place for knowledge management in specific areas addressed. The paper describes the method of building the methodology, shows the most important procedures in the development of the tool, and visually introduces its environment. The proposed future surveys will be key to assessing the effectiveness of the use of the proposed tool - this paper is one of the parts of the concept that the author focuses on throughout the dissertation.

Keywords: Internal audit, IT audit, knowledge sharing, methodology development, DSR

JEL Classification: O31

STUDIJNÍ PROGRAM EKONOMETRIE A OPERAČNÍ VÝZKUM

Multi-Echelon Closed Loop Supply Chain Coordination via Revenue Sharing Contract

Petr Pokorný
pokornyp@vse.cz

Ph.D. student of econometrics

Supervisor: prof. RNDr. Ing. Petr Fiala, CSc., MBA, (pfiala@vse.cz)

This paper deals with a revenue sharing based contract applied to a three-echelon closed loop supply chain, where the manufacturer produces new and remanufactured products, whose demands are interdependent. We attempt to improve the performance of a decentralized supply by applying a revenue sharing coordinating contract. We then test the effectiveness and the desirability of the solution. By effectiveness we strive to increase the total profit of the coordinated chain. Testing the desirability will show us, if the coordinated solution would be acceptable for all parties involved. We present that the solution reached via the contracts is indeed effective but hard to accept by the distributor, whereas both the retailer and the manufacturer more than double their profits. Reasons behind these findings lie in the demand functions and the way they affect the distributor's profit.

Keywords: Centralized, Closed Loop, Coordination, Decentralized, Game Theory, Multi-Echelon, Nash Equilibrium, Revenue Sharing, Stackelberg, Supply Chain

JEL Classification: C72

STUDIJNÍ PROGRAM STATISTIKA

Missing Data Imputation for Categorical Variables

Jaroslav Horníček
horj31@vse.cz
Ph.D. student of Statistics
Supervisor: prof. Ing. Hana Řezanková, CSc., (hana.rezankova@vse.cz)

Dealing with missing data is a crucial part of everyday data analysis. The IMIC algorithm is a missing data imputation method that can handle mixed numerical and categorical datasets. However, the categorical data are crucial for this work. This paper proposes the new improvements of the IMIC algorithm. The two proposed modifications consider the number of categories in each categorical variable. Based on this information, the factor, which modifies the original measure, is computed. The factor equation is inspired by the Eskin similarity measure known in the hierarchical clustering of categorical data. For the simulation, the real data with 17 categorical variables from the survey in the group of 395 students was used. The five different missing value ratios were created randomly (MCAR) in each of these categorical variables. For each of the missing value ratios and the modification of the IMIC algorithm, 20 iterations were run. It means 300 simulations in summary. The results show that as the missing value ratio in the dataset grows, better accuracy results are achieved using the second modification. From 35% missing value ratios, the algorithm utilized the second modification became significantly better based one-sided paired t-test. The variability in the results measured in all of the 20 iterations computed is lower using the second modification than the original IMIC algorithm. The paper also shortly analyzes the advantages and disadvantages of using the IMIC algorithm.

Keywords: IMIC algorithm, missing value imputation, categorical variables

JEL Classification: C38, C40, C80

Life expectancy and its past-future development in Czechia

David Morávek

david.moravek@vse.cz

Ph.D. student of Statistics

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

The article deals with the most used indicator of mortality in a given population – life expectancy. In evaluating its past development, using a decomposition algorithm of stepwise replacement, its differences are evaluated by age and sex. Furthermore, its differences at a final time point was split into additive components corresponding to the initial differences in the death rates and differences in trends in these underlying rates using a contour decomposition method that extends the stepwise replacement algorithm along an age-period demographic contour. When evaluating life expectancy in its future development, the Double Gap model is considered, which uses an algorithm based on high correlation of life expectancy between males and females. These association can be used to improve forecasts. We compared our results with forecasts based on population projections produced by the Czech Statistical Office and Eurostat. We focus on forecasting life expectancy at the age of 0 for males and females in Czechia.

Keywords: mortality, life expectancy, decomposition method, mortality forecasting, double-gap model

JEL Classification: C53

Comparison of models for simulation of synthetic microdata from the population census

Jiří Novák

xnovj159@vse.cz

Ph.D. student of Statistics

Supervisor: doc. Ing. Jaroslav Sixta, Ph.D. (jaroslav.sixta@vse.cz)

This article deals with the simulation of synthetic microdata from the population census. There is a great demand from the scientific community to publish more detailed outputs of official statistics for research and analysis purposes. The most detailed data of individual respondents are called microdata. In the case of population census, it is an area that needs to be handled with special care and needs special protection. In this contribution, the author aims to compare auspicious simulation-based methods for securing the confidentiality of the microdata, which is creating new, "synthetic" microdata from the original dataset. Personal data protection laws hinder the publication of raw microdata from the census. However, the proposed approach creates another dataset that does not contain the initial values but preserves the relationships between the variables and the hierarchical structure contained in the data. The two selected methods are compared: multinomial log-linear models and classification trees. For evaluating the information loss across models, demographic indicators and differences from the original data are used. The deviation in demographic indicators must be monitored because of their importance to the Czech Statistical Office. Otherwise, the data would not be analytically valid, and any analyses on them would be devalued from the very beginning. The paper analyses this approach on the microdata from the population census from 2011. The great advantage of these methods is that they allow statistical offices and agencies to disseminate datasets, which would otherwise have to remain hidden and confidential.

Keywords: population census, microdata, statistical disclosure control, synthetic, confidentiality

JEL Classification: C13, Cl8, C80

Selected alternatives to statistical inference in survival analysis: principles and some properties of methods comparing survival curves

Lubomír Štěpánek

lubomir.stepanek@vse.cz

Ph.D. student of Statistics

Supervisor: Luboš Marek, prof. RNDr. CSc (marek@vse.cz)

The task of comparing two or more time-event survival curves is very common in applied biostatistics, and several well-established methods are available. Depending on how many groups are supposed to be compared, the log-rank test, the score-rank test, the Cox proportional hazards model, or the Wilcoxon rank-sum test might be used. In this work, we propose more robust alternatives for two of these methods: We refine the log-rank test using combinatorial geometry and introduce a new alternative for comparing two or more time-to-event survival curves using the random forest algorithm. Regarding the comparison of two survival curves, we propose modelling of individual time-to-event survival curves in a discrete combinatorial way as orthogonal paths in a grid of survival plot, which enables a direct estimation of the p-value using its original definition of getting data at least the same extreme as the observed one. For comparing more than two survival curves, we propose a method using a random forest algorithm. Intuitively, the random forest containing a large proportion of trees with sufficient complexity, adjusted by tree pruning, can classify data into classes depicted by their survival curves, which tends to reject the null hypothesis about no statistical difference between the curves. We further present simulations that support our expectation that both the proposed methods are more robust in lower first-type error rates than traditional approaches. In the case of the random forest algorithm, we demonstrate that with increasing the tree pruning level, the first-type error rate of the method decreases, and robustness increases. Based on the simulations and preliminary analytical derivations, the methods seem to be promising alternatives for comparing two or more time-to-event curves.

Keywords: survival curves comparing, log-rank test, Cox proportional hazard model, robust alternative, random forests, combinatorial geometry, orthogonal paths in a grid, tree complexity, tree pruning level, first-type error rate decreasing, numerical simulation

JEL Classification: C10, C12, C14, C15