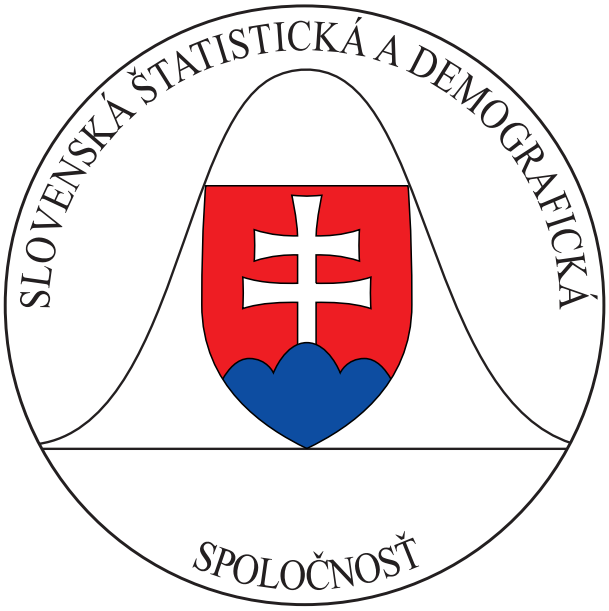
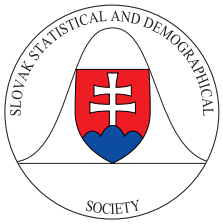
5/2013

FORUM STATISTICUM SLOVACUM



ISSN 1336-7420

Slovenská štatistická a demografická spoločnosť Miletičova 3, 824 67 Bratislava
www.ssds.sk



Naše najbližšie akcie:

(pozri tiež www.ssds.sk, blok Organizované akcie)

FERNSTAT 2013
12.9.-13.9., Banská Bystrica

Aplikácie metód na podporu rozhodovania 2013
jeseň, STU Bratislava

22. Medzinárodný seminár Výpočtová štatistika
5. – 6. 12. 2013, Bratislava

Prehliadka prác mladých štatistikov a demografov
5. – 6. 12. 2013, Bratislava

Regionálne akcie
priebežne

Slovenská štatistická konferencia
Jeseň 2014, 2 dni, Prešovský kraj

Slávnostná konferencia **50 rokov Slovenskej štatistickej a demografickej spoločnosti**
marec 2018, Slovenská republika

FOREWORD

Dear colleagues,

we propone the fifth issue of the nineth volume of the scientific peer-reviewed journal published by the Slovak Statistical and Demographical Society (SSDS). This issue comprises contributions that are content-compatible with the topics „Cluster analysis, Probability distributions, Hypotesis testing, Statistical models in economy, Applications of numerical analysis“.

Editors: Jana Kalická, Martin Kalina, Mária Minárová, Oľga Nánásiová, Jozef Chajdiak, Ján Luha.

Reviewed by: Jana Kalická, Martin Kalina, Mária Minárová, Oľga Nánásiová, and by other anonymous reviewers.

Assoc. Prof. Ing. Jozef Chajdiak, CSc.

Editor in chief

International Year of Statistics ("Statistics2013") is a global reminder of the importance of statistics. Slovak Statistical and Demographical Society joined the International Year of Statistics and will be mentioned at its professional events in the year 2013.

PREDHOVOR

Vážené kolegyne, vážení kolegovia,

predkladáme piate číslo deviateho ročníka vedeckého recenzovaného časopisu Slovenskej štatistickej a demografickej spoločnosti (SŠDS). Toto číslo je zostavené z príspevkov, ktoré sú obsahovo orientované v súlade s tematikou „Zhlukavá analýza, Rozdelenia pravdepodobnosti, Testovanie hypotéz, Štatistické modely v ekonómii, aplikácie numerickej analýzy“.

Editori: Jana Kalická, Martin Kalina, Mária Minárová, Oľga Nánásiová, Jozef Chajdiak, Ján Luha.

Recenzované: Jana Kalická, Martin Kalina, Mária Minárová, Oľga Nánásiová a ďalšími anonymnými recenzentami.

Doc. Ing. Jozef Chajdiak, CSc.
Šéfredaktor

STATISTICS2013.ORG
MEDZINÁRODNÝ ROK ŠTATISTIKY
ÚČASTNÍCKA ORGANIZÁCIA

Medzinárodný rok štatistiky ("Statistics2013") je celosvetové pripomenutie významu štatistiky. Slovenská štatistická a demografická spoločnosť sa pripojila k Medzinárodnému roku štatistiky a bude ho pripomínať pri svojich odborných akciách v roku 2013.

Dataminingová analýza na príklade jazykovej agentúry
Data mining analysis on the example of a language agency

Bohdalová Mária, Kurdyová Eva

Abstract: Clients have different needs and this is the reason why the aim of the companies is to settle down on the market and gain competitive advantage. It is necessary to find out and satisfy the customer`s needs with appropriate product or service. Data mining methods reveal not only the needs of the customers but also help to select the customers into larger units /segments with similar buying behaviour. The aim of this paper is the selected data mining methods, for example clustering apply on clients and purchase orders database of a no named language agency. By clustering the data, we have obtained the data distribution and observed the character of each cluster. In addition, cluster analysis usually acts as the pre-processing of other data mining operations. Therefore, cluster analysis has become a very active research topic in data mining.

Abstrakt: Každý klient má iné potreby, preto pre spoločnosti, ktorých cieľom je usadiť sa na trhu a tak získať konkurenčnú výhodu, je nevyhnutné poznať potreby zákazníkov a následne uspokojiť potrebu klientov vhodným produktom alebo službou. Metódy dataminingu odhaľujú nielen potreby zákazníkov, ale pomôžu aj zoskupovať klientov do väčších celkov/segmentov s podobným nákupným správaním. Cieľom príspevku je vybranú metódu dolovania dát, zhlukovú analýzu aplikovať na príklade zákazníckej databázy a na databáze objednávok nemenovanej jazykovej agentúry. Zhluková analýza často slúži aj ako predpríprava dát pred aplikovaním inej metódy dataminingu. Z tohto dôvodu sa zhluková analýza stala veľmi obľúbenou výskumnou témou v oblasti dolovania dát.

Key words: datamining, cluster analysis, statistical method
Kľúčové slová: dolovanie dát, zhluková analýza, štatistické metódy
JEL classification: C12, C38, C81

Úvod

Súčasný rozvoj informačných technológií prináša pre spoločnosti veľa možností ako využiť informácie o svojich zákazníkoch a akým spôsobom vykonávať marketing. Spoločnosti majú dostupné veľké množstvo údajov o zákazníkoch a toto im vytvára príležitosti a výzvy ako využiť svoje dáta a získať konkurenčnú výhodu. Na jednej strane, si treba uvedomiť, že znalosti skryté v týchto obrovských databázach sú kľúčovou podporou pre rôzne rozhodnutia. Napríklad znalosti o zákazníkoch z týchto databáz sú rozhodujúce pre marketingové funkcie. Aj napriek tomu zostáva veľa informácií skrytých a nevyužitých. Na druhej strane, intenzívna konkurencia a zvýšenie možností, ktoré sú pre zákazníkov k dispozícii, vytvára nové tlaky na marketingové rozhodovanie, pričom vzniká potreba spravovať a dlhodobo sa starať o zákazníkov. Vzniká nový fenomén riadenia vzťahov so zákazníkmi, ktorý vyžaduje, aby organizácia prispôsobila svoje produkty a služby zákazníkovi a najmä komunikovala so svojimi zákazníkmi na základe ich skutočných preferencií. Tu si treba uvedomiť, že efektívne riadenie vzťahov so zákazníkmi môžeme vykonať len na základe skutočného porozumenia potrebám a preferenciám zákazníkov. Za týchto podmienok, dataminingové nástroje môžu pomôcť odhaliť skryté znalosti a tak lepšie pochopiť zákazníka, zatiaľ čo systematické riadenie znalostí môže smerovať do efektívnej marketingovej stratégie. Vzhľadom na dôležitú

úlohu, ktorú zohrávajú marketingové rozhodnutia v aktuálnej orientácii na zákazníka, je potrebné prepojiť extrahované vedomosti o zákazníkoch s aplikáciou týchto poznatkov v kontexte s marketingovými rozhodnutiami. Riadenie vzťahov so zákazníkmi je možné iba integrovaním vedomostí získaných v procese riadenia a využívaním vedomostí pre marketingové stratégie. Toto pomôže spoločnosti vyjsť v ústrety potrebám zákazníkov na základe toho, čo vedia o svojich zákazníkoch. Veľké spoločnosti, ktoré majú niekoľko tisíc až milión zákazníkov a preto si nemôžu dovoliť osobný vzťah s každým z nich. Musia sa naučiť pracovať s údajmi, ktoré vznikajú pri styku so zákazníkmi. Datamining alebo dolovanie dát, zahŕňajú analytické techniky, ktoré pomáhajú firmám zmeniť zákaznícke dáta na „customer knowledge“ (Berry, Linoff, 2004).

Cieľom tohto príspevku je pomocou niekoľkých dataminingových techník odhaliť skryté informácie v dátach a cielene vybrať skupinu zákazníkov spoločnosti. V nasledujúcej časti uvádzame úvod do používania dataminingových techník, v kapitole 2 prezentujeme analýzu dát nemenovanej spoločnosti. V závere zhodnotíme naše zistenia.

1 Dataminingové techniky

Dolovanie dát je proces, ktorý používa paletu analytických nástrojov použitých na zistenie vzťahov a vzorov v údajoch, ktoré je možné použiť na odvodenie platných záverov alebo predpovedí. V dolovaní údajov používame štatistické metódy a metódy hraničiace s oblasťou umelej inteligencie. Tieto techniky sa aplikujú vo vzájomne veľmi odlišných oblastiach, ako je napríklad riadenie procesu výroby, marketing, riadenie ľudských zdrojov, atď (Kadora, 2006), (Shmueli et all., 2010).

Typické dataminingové šetrenie pozostáva z nasledujúcich krokov (Shmueli et all., 2010), (Berry, Linoff, 2004)

1. Preformulovanie (transformovanie) reálneho problému spoločnosti na dataminingový problém.
2. Výber vhodných údajov.
3. Predpripravenie údajov.
4. Redukovanie množstva údajov, ak je potrebné. Pre rozsiahle dátové množiny je vhodné vytvoriť tréningovú, validačnú a testovaciu sadu údajov.
5. Skonkretizovať cieľ dataminingovej analýzy prostredníctvom štatistických techník ako napríklad klasifikačných, predpovedných, zhlukovacích metód a pod.
6. Zvoliť konkrétnu dataminingovú techniku ako napr. regresnú analýzu, neurónové siete, hierarchickú zhlukovaciu analýzu a pod.
7. Vytvorenie modelu pomocou zvolenej štatistickej analýzy.
8. Vyhodnotenie modelu.
9. Overenie modelu.
10. Interpretovanie zistených výsledkov.
11. Použitie vytvoreného modelu, alebo úprava modelu, výber inej vhodnej metódy.

Treba podotknúť, že dataminingové šetrenie nie je priamočiary proces, ale každý krok je potrebné prehodnotiť a prípadne sa vrátiť na niektorý predchádzajúci krok.

Pri preformulovaní reálneho problému na dataminingový problém ide v podstate o riešenie jednej z nasledujúcich šiestich úloh (Berry, Linoff, 2004):

1. Klasifikácia
2. Odhady

3. Predpovede
4. Zoskupovanie na základe podobnosti
5. Zhlukovanie
6. Deskripcia a profilácia

Prvé tri úlohy patria k priamym data miningovým metódam, štvrtá a piata skupina úloh patrí k nepriamym data miningovým metódam, poslednú úlohu môžeme zaradiť k obom typom metód. Úlohou priamych dataminingových metód je nájsť pravidlá, ktoré vysvetľujú známu hodnotu cieľovej premennej, ktorú sme získali na základe historických údajov. Úlohou nepriamych dataminingových metód je nájsť všeobecné vzory, ktoré nie sú viazané na žiadne premenné. Najbežnejšiu formu nepriameho dolovania v dátach predstavujú zhlukovacie metódy (Řezanková et all., 2009), (Meloun, Militký, 2004), ktoré nájdu podobné skupiny, pričom základnou podmienkou je, aby bol čo najväčší rozdiel medzi nimi. V tomto príspevku využijeme zhlukovaciu metódu za účelom určenia vhodnej cieľovej skupiny klientov pre nemenovanú spoločnosť.

Dolovanie v dátach je nemysliteľné bez softvérovej podpory. Na trhu existuje viac softvérových produktov, pomocou ktorých vieme uskutočniť dataminingovú analýzu. K najpopulárnejším štatistickým softvérom na trhu, ktoré podporujú dataminingové techniky sú SAS a IBM SPSS. V tomto príspevku sme použili softvér IBM SPSS v. 20.

2 Zhluková analýza na príklade jazykovej agentúry

Využitie zhlukovej analýzy ukážeme na príklade údajov nemenovanej medzinárodnej jazykovej agentúry. Našou úlohou bolo určiť profil objednávok a zákazníkov a identifikovať zákazníkov s podobným nákupným a demografickým správaním. Na analýzu nám poslúžila dátová vzorka objednávok a zákazníkov získaná od danej spoločnosti. V programe IBM SPSS sme dáta vyčistili (nahradili sme české názvy dát slovenskými názvami, odstránili sme duplicity a pod.) a spracovali.

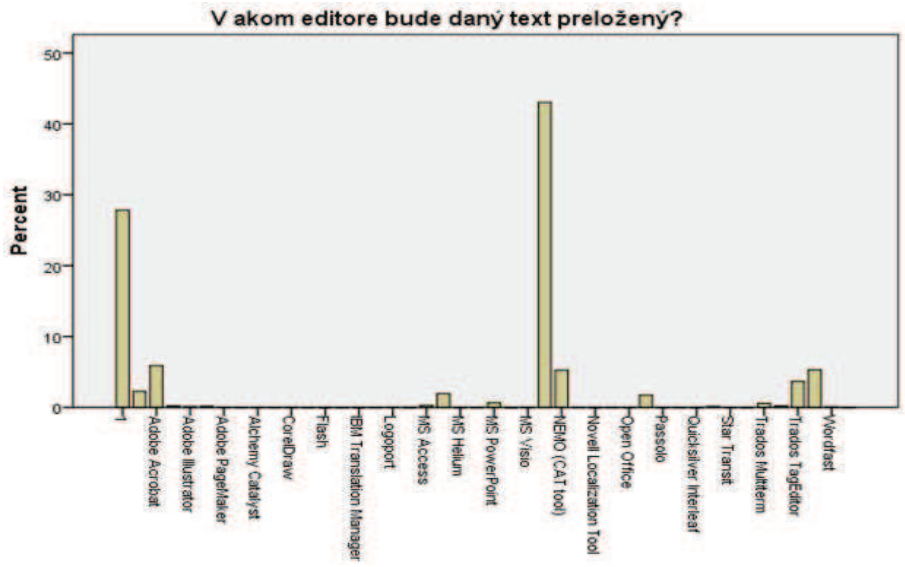
Profil objednávok sme skúmali pomocou frekvencie vo vzorkách a vypracovali sme χ^2 test za účelom zistenia súvislostí medzi odborom prekladu a použitým prekladateľským editorom.

Zo *vzorky objednávok* vyplýva:

81% objednávok sa týka prekladu a 15,6% korektúr. Najčastejšie boli klientmi objednané nasledujúce jazykové kombinácie: preklad z češtiny do angličtiny (9%), preklad z angličtiny do češtiny (5,5%), preklad z češtiny do angličtiny (4,4%), jazyková korektúra anglického jazyka (3,2%), preklad z nemčiny do češtiny (3,1%) a súdny preklad z angličtiny do češtiny (2,3%). Preklady sú rozdelené podľa prekladateľských odborov a môžeme ich charakterizovať ako náročnejšie všeobecné texty rozličných funkčných štýlov v danom jazyku. V našom prípade najvyhľadávanejšími odbormi boli: technika (19,2%), právo (16,2%), ekonomika a financie (12,7%), štylistické a literárne texty (11,5%) a lekárstvo (5,3%).

Väčšina prekladov bola objednaná za účelom publikovania (46,2%) alebo mali informatívny charakter (42,1%). Nadpolovičná väčšina (83,5%) prekladov bola doručená e–mailom, len 12,5% prekladov bolo odovzdaných osobne do rúk zákazníka. S 50%–nou pravdepodobnosťou môžeme tvrdiť, že texty boli určené pre všeobecnú verejnosť. Najpopulárnejšie používané editory pri prekladaní sú MS Word (43%), Adobe Acrobat (5,9%) a MS Excel (1,9%). Z CAT[1] nástrojov boli najpoužívanejšie nástroje Trados WorkBench (5,3%), Trados TagEditor (3,7%) a Across (2,3%).

[1] CAT je skratka slovného spojenia computer-aided translation, v slovenskom jazyku počítačom podporovaný preklad. CAT program uľahčuje prácu prekladateľa, lebo program spolupracuje s prekladovou pamäťou a terminologickou databázou. Pomocou tých nástrojov sa znižuje chybovosť prekladu ale aj čas, potrebný na prekladanie.



Obrázok 1 Výstupný graf z SPSS znázorňujúci najčastejšie používané editory pri preklade
Zdroj: spracovanie vlastné

Analýzou vzorky zákazníkov sme pomocou frekvenčných tabuliek zistili, že pre viac ako polovicu prípadov nebol určený sektor, v ktorom zákazník podniká. Ale vieme, že 8,2% odberateľov prekladov sa zaoberá obchodom, 7% zákazníkov sa zaoberá nehnuteľnosťami, prenájmom a službami, 6,7% podniká v oblasti spracovateľského priemyslu. 65% zákazníkov podniká samostatne alebo nevykazuje podnikateľskú aktivitu, 15% zákazníkov tvoria firmy s počtom zamestnancov 1 až 19. 12,6% klientov zamestnáva 20 až 499 zamestnancov. Faktúry boli najčastejšie vystavené v mene EUR alebo CZK, pričom klienti preferovali pobočky v Prahe 4 (14,7%), v Ostrave (5,3%), v Liberci (5,3%) a v Bratislave (5,1%).

Ďalej sme sa rozhodli zistiť či existuje závislosť medzi odborom prekladu a prekladateľským editorom. Použili sme χ^2test. Testovali sme zvlášť zákazníkov pochádzajúcich z Českej a zo Slovenskej republiky. Naše hypotézy, ktoré sme testovali na hladine významnosti $\alpha = 0,05$, boli nasledovné:

H_0: Zákazník sa náhodne rozhodne pre výber editora (medzi premennými „Odbor“ a „Editor“ nie je závislosť)

H_1: Zákazník cielene vyberá prekladateľský editor (medzi premennými „Odbor“ a „Editor“ je závislosť)

Tabuľka 1 χ^2 kvadrát test pre premenné „Odbor“ a „Editor“ podľa krajiny pôvodu

Chi-Square Tests				
Zákazník sa pochádza z CZ alebo SK?		***Value***	***df***	***Asymp. Sig. (2-sided)***
Zákazník CZ	Pearson Chi-Square	608225,636[b]	420	0,000
Zákazník SK	Pearson Chi-Square	66358,551[c]	380	0,000
Všetci zákazníci	Pearson Chi-Square	616065,182[a]	430	0,000

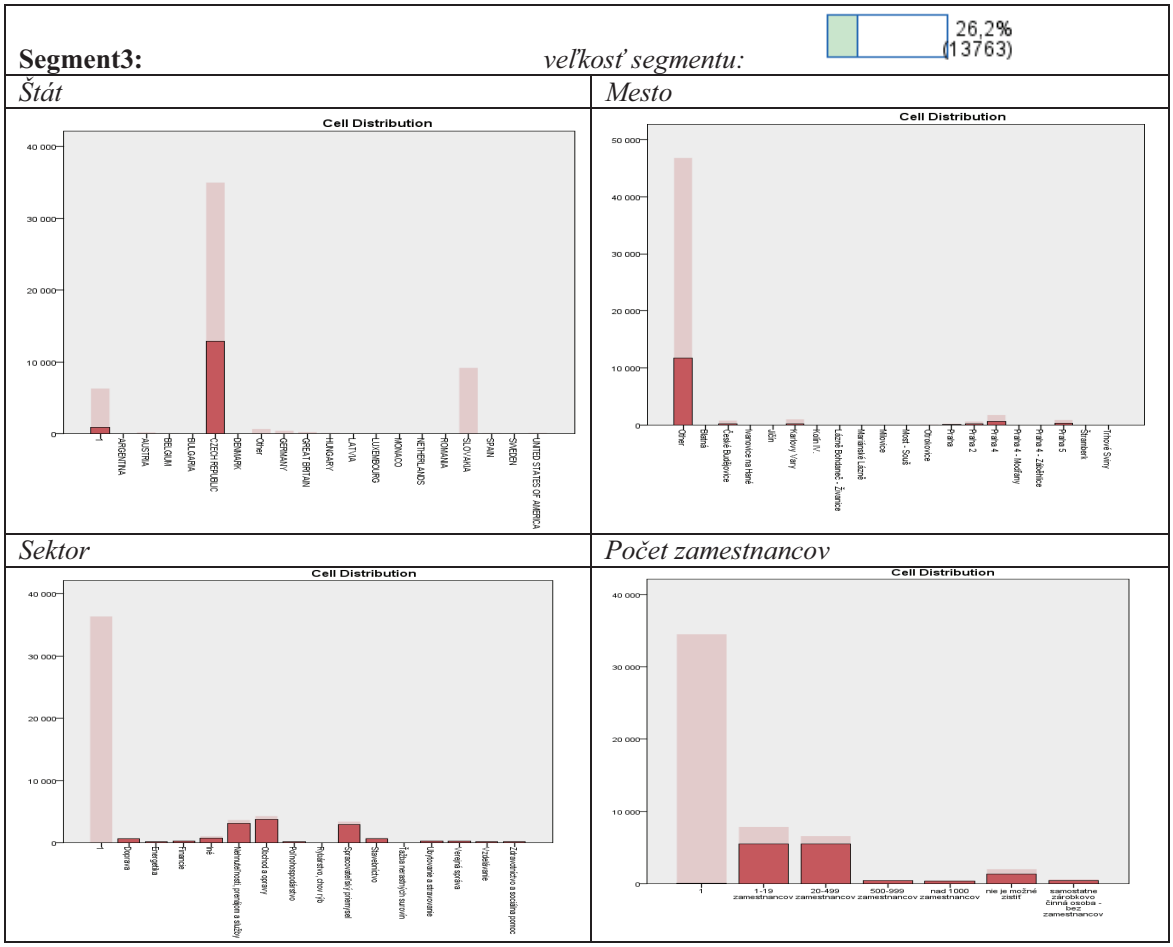
Zdroj: spracovanie vlastné pomocou IBM SPSS

Z tabuľky 1 vidíme, že vypočítaná *p*-hodnota je menšia v porovnaní s nami uvažovanou hodnotou štatistickej významnosti, ktorá je 5%. H_0, podľa ktorej sa zákazník náhodne rozhodne pri výbere prekladateľského editoru, sa zamieta (*p*–hodnota = 0,000). Potvrdila sa

alternatívna hypotéza, ktorá svedčí o závislosti medzi premennými „Odbor“ a „Editor“ v jednotlivých krajinách a aj v oboch krajinách.
χ^2 test nám potvrdil, že zákazníci (slovenskí aj českí) cielene vyberajú editor na základe odboru prekladu, napríklad slovenskí zákazníci medicínske preklady objednajú v editore Trados Studio (30,1%) alebo v Trados TagEditor–e (15,5%).

Tiež sme porovnali podiely prekladov, ktoré boli vyhotovené v jednotlivých editoroch a zistili sme, že českí zákazníci sú ochotnejší objednať CAT (počítačom podporované) preklady, pričom slovenskí zákazníci stále trvajú na tradičných technikách a preklady objednávajú bez nástroju CAT. Zo štatistického hľadiska sú významné aj objednávky, pri ktorých zákazník nedefinoval typ editoru. V týchto prípadoch rozhodnutie o výbere prekladateľskej techniky bolo na prekladateľskej agentúre (prípady sme evidovali ako „nezadané“).

Ďalej sme chceli vytvoriť segmenty zákazníkov s podobným nákupným správaním. Využili sme zhlukovú analýzu, pomocou ktorej sme identifikovali rôzne skupiny zákazníkov a ktoré majú podobné demografické alebo nákupné charakteristiky. Pri segmentácií trhu sme pracovali s dátami o zákazníkoch. Zhluky sme vytvorili na základe nasledujúcich podmienok: *štát, mesto, sektor* a *počet zamestnancov firmy*. Do segmentu č.1 patrí 46,3% zákazníkov, do segmentu č.2 27,5% a do tretieho segmentu 26,2%, pričom segmenty majú nasledujúce charakteristiky:



Obrázok 2: Výsledok zhlukovej analýzy zákazníkov: Segment3

Zdroj: spracovanie vlastné pomocou IBM SPSS

Segment1: Prvý, najpočetnejší segment zahŕňa zákazníkov z Českej republiky, z mesta Praha, České Budějovice alebo Karlove Vary. Pre klientov prvého segmentu nebolo možné definovať sektor, v ktorom podnikajú z dôvodu nezadaných hodnôt v databáze. Ak neberieme do úvahy nezadané hodnoty pri počte zákazníkov firiem, môžeme vyhlásiť, že väčšina firiem zamestnáva 1–19 zamestnancov. V porovnaní s ostatnými segmentmi, zhluk č.1 obsahuje najviac nezadaných hodnôt.

Segment2: Do zhluku patrí 27,5% zákazníkov firmy, ktorí cca s 60%–nou pravdepodobnosťou pochádzajú zo Slovenskej republiky. Segment v malom počte zahŕňa aj zákazníkov z Veľkej Británie a z Nemecka. 20,5% zákazníkov má svoje korene v Bratislave. Firmy patria do kategórie mikro a malých podnikov, ale ich oblasť podnikateľskej činnosti nie je možné zistiť.

Segment3: Ak sa nemenovaná jazyková agentúra rozhodne orientovať na tretí segment, jej cieľovou skupinou môžu byť zákazníci, ktorí pochádzajú z Českej republiky, z mesta Brno. Firmy v tom segmente s 40,4% pravdepodobnosťou zamestnávajú 1–19 zamestnancov a približne s 39% pravdepodobnosťou 20–499 zamestnancov. Najväčší počet klientov pracuje v oblasti obchodu a opravy (27,3%), alebo sa zaoberá nehnuteľnosťami a prenájmom. Ak nemenovaná jazyková agentúra chce získať konkurenčnú výhodu na trhu prekladateľských služieb musí vzhľadom na rastúci spracovateľský priemysel zabezpečiť spoluprácu s prekladateľmi z oblasti techniky a chémie, aby vedel uspokojiť potrebu firiem z tejto oblasti.

3 Záver

Výsledky analýzy poskytnutej databázy priniesli rozdelenie klientov do troch rôznych segmentov, z ktorých segment č.3 je možnou cieľovou skupinou spoločnosti. Z tohto zistenia vyplýva, že pre nemenovanú prekladateľskú agentúru by bolo vhodné orientovať sa na zákazníkov, ktorí pochádzajú z Českej republiky, z mesta Brno a pracujú v oblasti obchodu a opravy, alebo sa zaoberajú nehnuteľnosťami a prenájmom. Vzhľadom na rastúci spracovateľský priemysel by mala zabezpečiť spoluprácu s prekladateľmi z oblasti techniky a chémie, aby vedeli uspokojiť potrebu firiem z tejto oblasti. Nakoľko zákazníci cielene volia prekladateľské editory na základe odboru prekladu; firma by mala hľadať takých dodávateľov prekladov, ktorí majú nielen jazykové zručnosti, ale aj prekladajú špecifické odborné texty a poznajú vhodné odborné editory a majú ich k dispozícii.

Techniky dolovania v dátach môžu byť nápomocné pre spoločnosti pri riešení ich obchodných problémov tým, že nájdu vzory, zoskupenia a vzájomné vzťahy, ktoré sú skryté v obchodných informáciách uložených v dátových skladoch. Organizácie môžu používať tieto techniky pre získavanie nových zákazníkov, odhaľovať možné podvody v reálnom čase. Na základe výsledkov segmentácie sa môžu lepšie zacieliť na zákazníkov, na analýzu nákupného správania sa zákazníkov v čase a pod.. Lepšia detekcia nových trendov im umožní prijať proaktívny prístup vo vysoko konkurenčnom trhu a tak pridať oveľa väčšiu hodnotu existujúcim produktom a službám, prípadne zaviesť nové žiadané produkty a balíčky služieb. Bohužiaľ, spoločnosti sa zvyčajne stretávajú s rôznymi problémami pri realizácii výsledkov dataminingových analýz. Treba poznamenať, že žiadne techniky dolovania dát nemôžu vyriešiť všetky problémy spojené so vzťahom k zákazníkom. Ale správnym výberom techniky dolovania dát a jej správnym vykonaním môžu byť prínosom pre spoločnosť, ktorá bude schopná ponúknuť správny produkt pre správnu skupinu zákazníkov, prostredníctvom správnej ponuky a cez správne distribučné kanály, čo nakoniec vedie k lepšiemu riadeniu vzťahov so zákazníkmi.

4 Literatúra

[1] BERRY, J.A.M. – LINOFF, S. G. 2004. Data mining techniques : for marketing, sales, and customer relationship management. Indianapolis, Wiley Publishing, Inc.

[2] HAN, J. – KAMBER, M. 2006. Data Mining: Concepts and Techniques. Burlington, Elsevier Science & Technology.

[3] KADORA, A. 2006. Techniky dolovania údajov. [online]. Dostupné na internete: http://www2.fiit.stuba.sk/~kapustik/ZS/Clanky0607/kadora/index.html

[4] KURDYOVÁ, E., 2013. Marketingové kampane a analýza dát. Bakalárska práca FM UK, Bratislava (školiteľ: Bohdalová, M.)

[5] MELOUN, M. – MILITKÝ, J. 2004. Statistická analýza experimentálních dat. Praha, Academia

[6] PACÁKOVÁ, V. a kol. 2003. Štatistika pre ekonómov. Bratislava, Iura Edition.

[7] ŘEZANKOVÁ, H. – HÚSEK, D. – SNÁŠEL, V. 2009. Shluková analýza dat. Praha, Professional Publishing

[8] TEREK,M. – HORNÍKOVÁ, A. – LABUDOVÁ, V. 2010. Hĺbková analýza údajov. Bratislava, Iura Edition.

[9] SHAW, M. J. – SUBRAMANIAM, CH. – WOO TAN, G. 2001. Knowledge management and data mining for marketing. In: Decision Support Systems 31, s. 127–137, Elsevier

[10] SHMUELI, G. – PATEL, N. R. –BRUCE, P. C. 2010. Data Mining for Business Intelligence. New Yersey, John Wiley & Sons.

[11] ZHENG, Y. 2012. Clustering Methods in Data Mining with its Applications in High Education. IPCSIT vol.43 (2012), IACSIT Press, Singapore

Adresa autorov:

doc. RNDr. Mária Bohdalová, PhD.
Fakulta managementu UK v BA
Odbojárov 10
820 05 Bratislava
Maria.bohdalova@fm.uniba.sk

Bc. Eva Kurdyová, 3. ročník
Fakulta managementu UK v BA
Odbojárov 10
820 05 Bratislava
Eva.kurdyova@st.fm.uniba.sk

Skórová funkce rozdělení a korelace náhodných veličin
Score function of distribution and association of random variables

Zdeněk Fabián

Abstract. Score function of distribution, introduced recently by Fabián ([2], [4]), is a function describing relative influence of a realization of random variable with continuous distribution on its certain central characteristic. The score correlation coefficient of two random variables is the Pearson's coefficient of score functions of their marginal distributions. In this paper we study its properties by means of simulation experiments, in comparison with other correlation coefficients in current use.

Abstrakt. Skórová funkce rozdělení, nedávno zavedená autorem, charakterizuje vliv daného pozorování na určitou centrální charakteristiku spojitého pravděpodobnostního rozdělení. Skórový korelační koeficient dvou náhodných veličin je Pearsonův korelační koeficient hodnot skórových funkcí marginálních rozdělení v pozorovaných bodech. V tomto článku studuji jeho vlastnosti pomocí simulačních experimentů a porovnávám je s vlastnostmi korelačních koeficientů běžně užívaných.

Keywords: Correlation, simulation.

Klíčová slova: Korelace, simulace.

JEL classification: C13, C15, C63.

Introduction

Statistical estimation is typically based on averaging functions of the variables generally called *score functions* that measure sensitivity of the corresponding likelihood function of the assumed model with respect to its parameters. More in detail, a value $\psi(x_0;\theta)$ of a score function ψ at an observed point x_0 and for a parameter θ describes the relative influence of x_0 on the estimated characteristic under the chosen model.

Models of classical statistics are parametric families of distributions $\{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^m\}$ with parent F and density $f(x) > 0$ on an open interval $\mathscr{X} \subseteq \mathbb{R}$ and $f(x) = 0$ on $\mathbb{R} - \mathscr{X}$. $\mathscr{X}$ is called the support of F. Classical score functions with components

$$U_{\theta_j}(x;\theta) = \frac{\partial}{\partial \theta_j} \log f(x;\theta), \quad j = 1,...,m$$

are called *Fisher scores*. These scores naturally arise in the maximum likelihood method, that provides estimates of θ of minimum variance. However, they are not employed in correlation analysis since they have generally a vector nature. Instead, the classical Pearson's correlation coefficient of random variables X and Y is based on observed data only, without distinguishing a part of a real association of X and Y and a part which stems from properties of the marginals.

Generalized Fisher score functions are called in [4] *score functions of distribution*. We denote them by S and abbreviate by sfd. Sfd S of distribution F is a unique (scalar) function, a value $S_F(x_0)$ of which at $x_0 \in \mathscr{X}$ characterizes relative influence of $x_0 \in \mathscr{X}$ on a central characteristic of distribution F described below. If a parametric distribution has a parameter expressing this central characteristic, the sfd is identical with Fisher score for this parameter. It was shown that the sfd-based approach provides in some cases (as the heavy-tailed distributions) score functions that are bounded.

In the present paper we try to use sfd's for estimation of correlation of random variables with various marginal distributions and compare properties of the sample score correlation coefficient with some others currently used.

1. Score function of distribution

Let us briefly describe the main steps leading to a generalization of the Fisher score function.

The starting point is the identity valid for location distributions with support $\mathbb{R}$ and density $f(x-\mu)$ with location parameter $\mu \in \mathbb{R}$,

$$\frac{\partial}{\partial \mu} \log f(x-\mu) = -\frac{1}{f(x-\mu)} \frac{d}{dx} f(x-\mu).$$

The identity says that the Fisher score for location can be obtained by differentiating of $-\log f(x-\mu)$ with respect to the variable. By setting $\mu = 0$, Hampel et al. [5] concluded that the relative rate of the change of f

$$S_F(x) = -f'(x)/f(x) \tag{1}$$

describes, analogically to the Fisher score, the relative influence of $x \in \mathbb{R}$ with respect to the "center" (mode) of the distribution. This is the solution of equation $S_F(x) = 0$.

The score function (1) fails to describe distributions with support $\mathscr{X} \neq \mathbb{R}$ (cf. the uniform distribution with support $\mathscr{X} = (0,1)$ and $-f'(x)/f(x) = 0$ or the exponential one with support $\mathscr{X} = (0,\infty)$ and $-f'(x)/f(x) = 1$). We noticed in [1] that if a distribution with support $\mathscr{X} = (0,\infty)$ has a density in the form $\frac{1}{\tau} f(x/\tau)$, it holds that

$$\frac{\partial}{\partial \tau} \log \left(\frac{1}{\tau} f(x/\tau) \right) = \frac{1}{\tau} T_F(x;\tau) \tag{2}$$

where

$$T_F(x;\tau) = -\frac{\tau}{f(x/\tau)} \frac{d}{dx} \left[x \frac{1}{\tau} f(x/\tau) \right]. \tag{3}$$

By setting $\tau = 1$ in (3), an analogue of (1) for distributions with support $(0,\infty)$ was obtained in the form

$$T_F(x) = -\frac{1}{f(x)} \frac{d}{dx} [x f(x)]. \tag{4}$$

Formula (4) is interpreted in this way: if F is taken as transformed distribution $F = G \circ \eta$ with 'prototype' $G \in \mathscr{P}_{\mathbb{R}}$ and density

$$f(x) = g(\eta(x)) \eta'(x), \tag{5}$$

where g is the density of G, and if $\eta : (0,\infty) \to \mathbb{R}$ is $\eta(x) = \log x$, the term in square brackets of (4) is the density multiplied by the reciprocal Jacobian of the transformation.

This idea was generalized in [2] for parametric distributions with arbitrary interval support. Let F_θ be a distribution with support $\mathscr{X} \subseteq \mathbb{R}$ and density $f(x;\theta)$, and let $\eta : \mathscr{X} \to \mathbb{R}$ be continuous increasing. Set

$$T_F(x;\theta) = -\frac{1}{f(x;\theta)} \frac{d}{dx} \left[\frac{1}{\eta'(x)} f(x;\theta) \right]. \tag{6}$$

Let the solution $x^* = x^*(\theta)$ to the equation

$$T_F(x;\theta) = 0 \tag{7}$$

be unique. Function

$$S_F(x;\theta) = \eta'(x^*)T_F(x;\theta) \tag{8}$$

is called a *score function of distribution* (sfd of F_θ).

The choice of η is discussed in details in [4]. In this paper we study distributions with support $\mathbb{R}$ with $\eta(x) = x$ and sfd's in the form $S_F(x;\theta) = -f'(x;\theta)/f(x;\theta)$ and distributions with support $(0,\infty)$ with $\eta(x) = \log x$ and sfd's

$$S_F(x;\theta) = \frac{1}{x^*(\theta)}T_F(x;\theta).$$

Sfd appears to be a generalization of the Fisher score function for a central characteristic (typical value) x^* of the distribution. Sfd reflects main features of the distribution similarly as the Fisher score function does, but in contrast with it, it is a simple scalar function even in multiparameter cases. Its moments

$$ES_F^k = \int_{\mathscr{X}} S_F^k(x)dF(x), \tag{9}$$

in cases of regular distributions exist and are usually given by simple expressions (since S_F 'fits' the distribution F).

A typical value $y^* : S_G(y) = 0$ of 'prototype' distribution $G(y)$ with support $\mathbb{R}$ is the mode, a typical value $x^* : S_F(x) = 0$ of distribution $F = G \circ \eta$ is the 'image' of the mode of the prototype, $x^* = \exp(y^*)$.

By [2], the value ES_F^2 of any continuous distribution, where E stands for the expectation, is interpreted as Fisher information for x^*. Based on analogy with the Cramér-Rao theorem for variance of efficient estimators, the reciprocal value of Fisher information, called here *score variance*,

$$\omega^2 = \frac{1}{ES_F^2}, \tag{10}$$

was suggested as a measure of variability of distribution F. Its square root $\omega = \sqrt{\omega^2}$, a *score deviation*, represents a characteristic radius of the distribution. Values x^* and ω^2 can replace the mean value and variance, the values considered as representatives of the center and variability of distributions, which may not exist in cases of heavy-tailed distributions.

The *score moment estimator* $\hat{\theta}_{SM}$ is a solution to implicit estimating equations, the finite parametric versions of (9)

$$\frac{1}{n}\sum_{i=1}^{n} S_F^k(x_i;\theta) = ES_F^k(\theta), \qquad k = 1,...,m. \tag{11}$$

Having any estimate $\hat{\theta}$ of θ, results of the estimation can be described by means of $\hat{x}^* = x^*(\hat{\theta})$ and $\hat{\omega}^2 = \omega^2(\hat{\theta})$, which makes possible to compare results for differently parametrized models.

2. Score correlation coefficient

Let X, Y be random variables with supports $\mathscr{X}_X$ and $\mathscr{X}_Y$, respectively, with joint distribution F_{XY} and density $f_{XY}(x,y)$. Denote by f_X, f_Y the densities of marginal distributions and by $S_X(x), S_Y(y)$ their sfd's. By $\rho_P(X,Y)$ is denoted the Pearson's correlation coefficient.

In [3], the joint score moment of X and Y was defined by

$$ES_X S_Y = \int_{\mathscr{X}_X} \int_{\mathscr{X}_Y} S_X(x) S_Y(y) f_{XY}(x,y)\, dxdy, \tag{12}$$

and the *score correlation coefficient* by

$$\rho_F(X,Y) = \rho_P(S_X, S_Y). \tag{13}$$

The score correlation coefficient is a distribution-dependent measure of association of random variables. If marginal distributions are given in parametric forms, score correlation coefficient is a function of parameters, which are to be estimated. Sfd's of the normal, gamma and beta distribution $N(\mu,\sigma)$ are linear and the score correlation coefficient equals in these cases to the Pearson one.

Given an observed sample $(x_1,y_1),...,(x_n,y_n)$ from from $F_{XY}(\theta_X,\theta_Y)$, the statistical counterpart of ρ_F is the *sample score correlation coefficient*

$$r_F = \frac{\sum S_X(x_i;\hat{\theta}_X) S_Y(y_i;\hat{\theta}_Y)}{\sqrt{\sum S_X^2(x_i;\hat{\theta}_X) \sum S_Y^2(y_i;\hat{\theta}_Y)}}, \tag{14}$$

where $\hat{\theta}_X$ and $\hat{\theta}_Y$ are estimates of the corresponding vectors of parameters of marginal distributions, respectively.

3. A simulation study

In simulation experiments were generated couples (X,Z) using independently generated random samples of X and Z from one or two-parameter distributions $F(x;\theta_0)$ and $F(z;\theta_0)$ using routines from the MATLAB Statistics toolbox. To model an association of random variables X and Y, we set

$$Y = \alpha X + (1-|\alpha|)Z. \tag{15}$$

The theoretical value of $\rho \equiv \rho(X,Y;\alpha)$ was thus $\rho = \alpha/\sqrt{2\alpha^2 - 2|\alpha|+1}$. Due to symmetry we refer only positive values of α.

Samples were generated from distributions with increasing variability of distributions described by ω with $\theta_0 = \theta_0(\omega)$. Let us describe in details two of distribution used:

Weibull distribution with density

$$f_W(x;\tau,c) = \frac{c}{x}\left(\frac{x}{\tau}\right)^c e^{-(x/\tau)^c} \qquad \tau, c > 0. \tag{16}$$

with particular cases exponential ($c=1$), Rayleigh ($c=2$) and Maxwell ($c=3$) distributions. The sfd

$$S(x;\tau,c) = \frac{c}{\tau}[(x/\tau)^c - 1]$$

equals the Fisher score function for τ. Typical value is $x^* = \tau$ and $ES^2 = (c/\tau)^2$ so that $\omega^2 = \tau^2/c^2$ and $c_0 = x^*/\omega$ with a fixed x^*.

Beta-prime distribution (beta distribution of the second kind) with density

$$f_B(x;p,q) = \frac{1}{B(p,q)} \frac{x^{p-1}}{(x+1)^{p+q}}, \qquad p,q > 0 \tag{17}$$

is a heavy-tailed distribution. By (6),

$$T(x;p,q) = \frac{qx-p}{x+1}$$

is different from both partial scores for p and q. Typical value is

$$x^* = p/q, \tag{18}$$

sfd is $S(x;p,q) = (q/p)T(x;p,q)$. Since $ET^2 = pq/(p+q+1)$, the variability of the distribution is described by score variance

$$\omega^2 = \frac{p(p+q+1)}{q^3}. \tag{19}$$

p_0, q_0 are computed from (18) and (19) for a given x^* and ω. The density $f_B(x;1,c)$ appears to be the density of the *Pareto distribution* shifted to have support $(0,\infty)$. Its typical value is $x^* = 1/c$ and score variance is, by (19), $\omega^2 = (c+2)/c^3$.

Given a sample $(X_1,Y_1),...,(X_n,Y_n)$ from Weibull distribution, it is to compute the maximum likelihood estimate $\hat{c}$ of c and then

$$\hat{\tau} = (n^{-1}\sum_{i=1}^{n} x_i^{\hat{c}})^{1/\hat{c}}.$$

A value of a simplified inference function at x_i in (14) is thus $S_X(x_i) = (x_i/\hat{\tau})^{\hat{c}} - 1$. Given a sample from the beta-prime distribution, a typical value x^* is estimated from the first score moment equation (11), that is from

$$\sum_{i=1}^{n} \frac{x_i - x^*}{x_i + 1} = 0$$

and the value of the simplified inference function at x_i in (14) is $S_X(x_i) = (x_i - \hat{x}^*)/(x_i+1)$. The same procedures are used for Y.

Fig. 1. shows a sample (X,Y) generated from $f_B(1,1)$ with $\rho = 0.3$ (left), the corresponding sample $(S_X(X), S_Y(Y))$ and the values of estimated correlation coefficients.

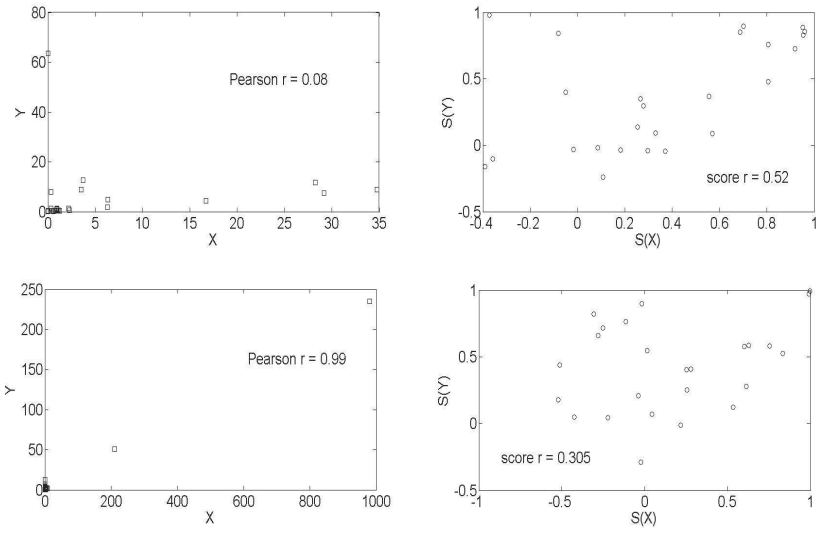


Fig. 1. Samples and correlation coefficients from $f_B(1,1)$ for $\rho = 0.3$.

Further, for each sample were computed the sample Pearson's correlation coefficient (r_P), the Spearman's rank correlation coefficient (r_S), Kendal's tau (the latter two were computed by means of code *corr* from the Matlab Statistical toolbox) and Huber's robust correlation coefficients (r_R). Denoting by $s_i = (x_i - m_X)/\omega_X$ and $t_i = (y_i - m_Y)/\omega_Y$, r_R is given by (see [6])

$$r_R = \frac{\sum \psi(s_i)\psi(t_i)}{[\sum \psi^2(s_i)\psi^2(t_i)]^{1/2}}, \tag{20}$$

where $\psi(z,k) = \max[-k, \min(z,k)]$ is the Huber score function. Since methods for robust estimates of location and scale of highly non-symmetric distributions are not available, we used in iterative solutions of estimation equations initial values $x_0^* = \text{median}(x_i)$ and $\omega_0 = 3 * MAD(x_i)$ (*MAD* stands for the median of absolute deviation from median).

Average values of correlation coefficients were estimated after 5.10^3 replications for each ω and plotted against ω for data generated from various distributions.

The results are as follows:

i) It appeared that the value of the sample correlation coefficient depend only weakly on the length n of samples if the length is about $n > 25$. In simulation experiments, the length of samples was $n = 75$.

ii) If F has support $\mathscr{X} = \mathbb{R}$, all the correlation coefficients are independent on ω. For any distribution, average r_P, r_S and r_R are roughly equal to the theoretical value ρ. We conclude that for distributions with 'mild' non-symmetry, the correlation properties overcome structures of distributions. The same conclusion follows from simulation experiments with 'mildly' non-symmetric distributions with support $\mathscr{X} = (0,\infty)$ (e.g. Maxwell, Rayleigh).

iii) Correlation coefficients in cases of highly non-symmetric distributions with support $\mathscr{X} = (0,\infty)$ are strongly dependent on their variability described by ω. Fig. 2 shows average sample correlation coefficients and their standard deviations for samples generated from Weibull and Pareto distributions with $\rho = 0.3$. The Pearson's r_P gives the best estimates of ρ for about $\omega < 2$ in the Weibull(1) case and for $\omega \ll 1$ in the Pareto case. The mean square errors (MSE) in right parts of the figure show that Pearson's r_P is of no use if samples are generated from heavy-tailed distributions. In what follows we do not present further results concerning r_P. Similarly, we do not present Kendall's r_K: in cases of heavy-tailed distributions it behaves quite unpredictably and usually underestimates the correct value.

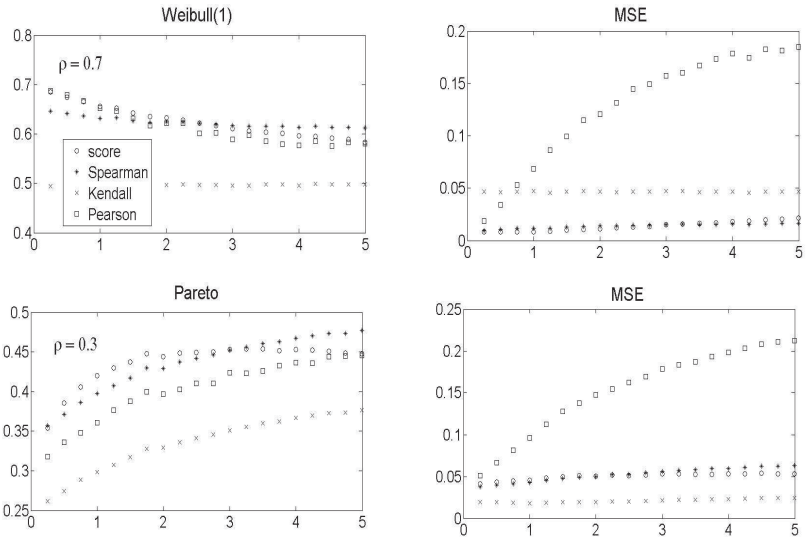


Fig. 2. Sample correlation coefficients for distributions with support $(0,\infty)$ as functions of ω.

iv) Fig 3. shows results when ρ was estimated by means of robust correlation coefficient r_R (20). Samples were generated from the Weibull(1) and

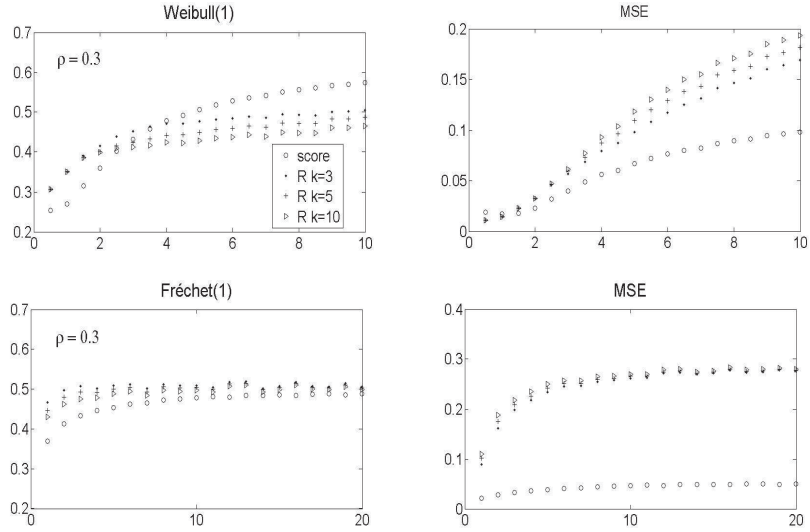


Fig. 3. Robust correlation coefficients as functions of ω.

Fréchet(1) distributions with $k = 3,5$ and 10 for $\rho = 0.3$. For low values of ω, r_F is less biased, for larger ω have robust estimates large variances. We conclude, that robust estimates of correlation coefficients are not appropriate for revealing an association of heavy-tailed random variables.

v) There remain only two correlation coefficients, the score r_F and the Spearman r_S, capable detecting an association of random variables with heavy-tailed distributions. A detailed study of their behavior with increasing variability of strongly non-symmetrical, heavy-tailed distributions is given in Fig. 4. Samples were generated from Weibull(1), beta-prime(1) and Pareto distributions

for four different values of ρ. The clearly apparent overall tendency with increasing variability of distributions is that both r_F and r_S posses a strong positive bias in cases of small values of ρ (with less biased r_F), and strong negative bias if ρ approaches to 1 (with less biased r_S). The splitting value is about $\rho = 0.5$.

This behavior is, unfortunately, unfavorable to attempts to find the true correlation of random variables with heavy-tailed distributions. However, after estimating ω, one can find an approximate 'true' r_F according the course of theoretical curves.

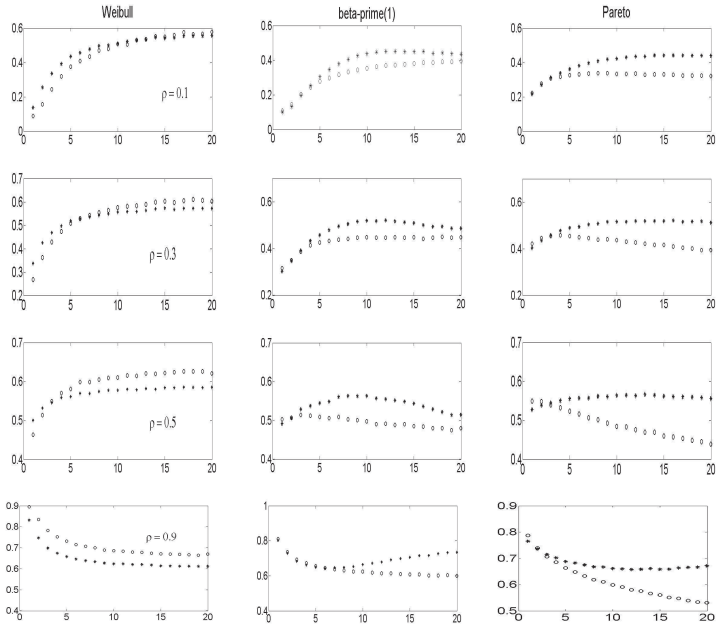


Fig. 4. Score and Spearman sample correlation coefficients as functions of ω. $o...r_F, *...r_S$

4. Conclusions

An introduction of the new measure of variability of continuous probability distributions make possible to study correlation coefficients as functions of their increasing variability. In the paper we have shown that the newly introduced score correlation coefficient as well as the Spearman's rho are able to detect association of random variables with heavy-tailed distributions supported by $(0, \infty)$.

5. Acknowledgements

The work has been done with institutional support RVO:67985807.

6. References

[1] Fabián, Z. 2001. Induced cores and their use in robust parametric estimation. Comm. Statist. Theory Methods, 30, pp. 537-556.

[2] Fabián, Z. 2007. Estimation of simple characteristics of samples from skewed and heavy-tailed distribution. In Recent Advances in Stochastic Modelling and Data Analysis, Singapore, World Scientific, pp. 43-50.

[3] Fabián, Z. 2010. Score correlation. Neural Network World, 20, pp. 793-798.

[4] Fabián, Z. 2013. Score function of distribution and revival of the moment method. Research report 1176, ICS ASCR (accepted to Comm. Statist. Theory-Methods).

[5] Hampel, F. R.-Rousseeuw, P. J.-Ronchetti, E. M.-Stahel, W. A. 1986. Robust statistic. The approach based on influence functions, Wiley, N.York.

[6] Shevlyakov, G.-Smirnov, P. 2011. Robust estimation of the correlation coefficient: An attempt of survey. Austrian J. of Statistics, 40, pp. 147-156.

Address of the author

Institute of Computer Science

Academy of Sciences of the Czech Republic

Pod vodárenskou věží 2, 18200 Praha

e-mail zdenek@cs.cas.cz

Kalibrácia váh štatistických zisťovaní v jazyku R
Calibration of weights of statistical surveys in R language

Boris Frankovič

Abstrakt
Dostupný softvér kalibrácie váh v štatistických zisťovaniach je v mnohých prípadoch príliš drahý, hľadanie lacnejších riešení je preto nevyhnutné. Francúzskym štatistickým úradom (INSEE) vytvorené SAS-makro Calmar2 efektívne rieši kalibráciu váh jednotiek vo výberových štatistických zisťovaniach. Rovnako tak sa touto témou zaoberá aj R balík *sampling*, konkrétne funkcia calib. Vychádzajúc z metodiky makra Calmar bol na Štatistickom úrade SR v jazyku R naprogramovaný alternatívny spôsob kalibrácie. Tento článok porovnáva všetky tieto tri kalibračné postupy.

Abstract
Available software for calibration of weights in statistical surveys is in many cases too expensive, looking for some cheaper solutions is therefore necessary. SAS-macro Calmar2, which was designed by French national statistical institute (INSEE), effectively solves the problem of calibration of weights. R package *sampling* (function calib) deals with this topic as well. Based on the methodology of macro Calmar Statistical Office of the Slovak Republic made an alternative technique for calibration (programmed in R). This article compares all these three calibration procedures.

Kľúčové slová: kalibrácia, zisťovanie, váhy, vzdialenostná funkcia, Lagrangeov multiplikátor, Calmar, sampling, calib
Key words: calibration, survey, weights, distance function, Lagrange multiplier, Calmar, sampling, calib
JEL classification: C61, C83

1. Úvod

Kalibrácia váh jednotiek vo výberových štatistických zisťovaniach zohráva čoraz dôležitejšiu úlohu pri tvorbe oficiálnych štatistík. Dostupnosť údajov z administratívnych zdrojov ako i údajov z cenzov umožňuje spresniť údaje v zisťovaniach s rovnakou populáciou, ale menšou výberovou vzorkou. Techniky kalibrácie váh sú do veľkej miery podporované i na európskej pôde, kde Eurostat pravidelne odporúča a apeluje na využívanie kalibrácie vo viacerých výberových zisťovaniach. Francúzsky štatistický úrad, INSEE, zhotovil v roku 1993 v programe SAS makro zvané Calmar (calibration on margins), ktoré následne v roku 2000 zdokonalil na makro Calmar2 a toto dokáže efektívne a účinne túto procedúru zvládnuť. Problémom je v tomto prípade však jednak cena samotného programu SAS, ako i jeho modulov potrebných na účinnosť tohto makra. Preto je pochopiteľné, že sa vytvorili rôzne snahy o nájdenie alternatívneho riešenia. R balík *sampling* [5] ponúka prostredníctvom svojej funkcie calib spôsob kalibrácie váh dátovej tabuľky. Rovnako tak sa aj Štatistický úrad SR pokúsil lacnejším a obdobne kvalitným spôsobom zalgoritmizovať celú procedúru kalibrácie váh v štatistických zisťovaniach (R kód s pracovným názvom Calif). Vychádzajúc z metodiky, ktorú vypracoval INSEE, bola v programe R [3] naprogramovaná prvá, a určite nie posledná, verzia nástroja určeného na kalibráciu. V tomto článku sa prostredníctvom dôkladne rozpracovanej teórie zhodnotia a porovnajú všetky tieto tri spôsoby kalibrácie váh vo výberových štatistických zisťovaniach.

2. Teoretická časť

Základnou ideou kalibrácie váh je upraviť pôvodné výberové váhy jednotiek štatistického zisťovania tak, aby spĺňali určité podmienky. Pod výberovou váhou sa rozumie prevrátená hodnota pravdepodobnosti zahrnutia jednotky do výberu, každá jednotka tak má iba jednu váhu a táto platí pre všetky jej hodnoty premenných. V praxi sa pod pojmom kalibrácia rozumie situácia, kedy máme k dispozícii úhrny určitých premenných z väčšieho, ideálne vyčerpávajúceho zisťovania, prípadne z údajov z administratívnych zdrojov a chceme, aby sme k týmto úhrnom dospeli aj po prevážení v zisťovaní s menšou výberovou vzorkou s tými istými kalibračnými premennými a tou istou populáciou za predpokladu, že sa minimalizuje vzdialenosť nových váh od tých pôvodných [1]. Zaveďme označenie ako v [1]:

S výberový súbor
n veľkosť výberového súboru
d_k pôvodná (prípadne modifikovaná o mieru neaktivity alebo mieru neodpovede) váha jednotky k
w_k váha jednotky k po kalibrácii
X_j vopred známy úhrn j-tej premennej v celej populácii
x_{jk} hodnota k-tej jednotky j-tej premennej
J počet kalibrovaných stĺpcov
$G(z)$ vzdialenostná funkcia

Pre výpočet novej váhy je potrebné vyriešiť rovnicu, v ktorej budeme minimalizovať vzdialenosť medzi váhami d_k a w_k za podmienky, že vážený úhrn premennej bude vopred želanou hodnotou. Hľadáme teda minimum Lagrangeovej funkcie (tak ako je navrhnuté v [2])

$$L = d^T G(w/d) - \lambda^T (x^T w - X)$$

kde $d^T = (d_1, \dots, d_n)$, $w = (w_1, \dots, w_n)^T$, $G(w/d)$ je vektorová funkcia, ktorá určuje vzdialenosť medzi pôvodnou a nakalibrovanou váhou, $\lambda^T = (\lambda_1, \dots, \lambda_J)$ je riadkový vektor Lagrangeovych multiplikátorov, $X = (X_1, \dots, X_J)^T$ je vektor vopred známych úhrnov a x je matica rozmeru n x J taká, že

$$x = \begin{pmatrix} x_{11} & \dots & x_{1J} \\ \dots & \ddots & \dots \\ x_{n1} & \dots & x_{nJ} \end{pmatrix}$$

[1]. Ak rozpíšeme horeuvedený zápis po zložkách, tak uvážiac, že $x_k = (x_{k1}, \dots, x_{kJ})^T$ je vektor riadku matice x, dostávame podrobnejší zápis [1]

$$L = \sum_{k \epsilon S} d_k G(w_k/d_k) - \lambda^T \left(\sum_{k \epsilon S} w_k x_k - X \right)$$

Pre vzdialenostnú funkciu $G(w_k/d_k)$ určujúcu vzdialenosť medzi oboma váhami, musí platiť, že ak $w_k = d_k$, teda ak $w_k/d_k = 1$, tak hodnota tejto funkcie je 0 (nulová vzdialenosť). Vzdialenostná funkcia musí byť kladná a konvexná, pričom v bode 1 nadobúda minimum (najmenšia možná vzdialenosť), t.z. derivácia v bode 1 je rovná nule a druhá derivácia v bode 1 kladná [2]. Ak použijeme substitúciu $w_k/d_k = r_k$, dostávame

$$L = \sum_{k \epsilon S} d_k G(r_k) - \lambda^T \left(\sum_{k \epsilon S} r_k d_k x_k - X \right)$$

Hľadáme minimum tejto funkcie pomocou parciálnych derivácií

$$\frac{\partial L}{\partial r_k} = d_k \frac{\partial G}{\partial r_k} - \lambda^T d_k x_k = 0$$

$$\frac{\partial G}{\partial r_k} = \lambda^T x_k$$

$$r_k = F(\lambda^T x_k)$$

kde $F(z)$ je inverzná funkcia k derivácii funkcie $G(y)$. Keďže funkcia G má zdola ohraničený a zhora neohraničený obor hodnôt a je konvexná, nájdený stacionárny bod musí byť minimom [1]. Nové váhy teda dostaneme ako

$$w_k = d_k F(\lambda^T x_k)$$

ako je uvedené v [2], pričom platí podmienka

$$\sum_{k \epsilon S} w_k x_k = X$$

Dostávame systém $n + J$ rovníc [1]. Druhým spôsobom je riešenie systému J rovníc, ako je to uvedené v [2]

$$\sum_{k \epsilon S} d_k F(\lambda^T x_k) x_{kj} = X_j$$

Táto optimalizačná úloha je riešiteľná rôznymi aproximačnými metódami, jednou z nich je napríklad Newtonova iteračná metóda, v ktorej je vhodné počiatočné hodnoty stanoviť na $\left(w_1^0, \dots, w_n^0, \lambda_1^0, \dots, \lambda_J^0\right) = (d_1, \dots, d_n, 0, \dots, 0)$ [2].

V praxi sa používajú nasledovné vzdialenostné funkcie (všetky ako sú definované v [2]):

- *lineárna*

$$G(r) = \frac{1}{2}(r-1)^2$$

Derivovaním dostávame $\frac{\partial G}{\partial r} = r - 1$, $\frac{\partial^2 G}{\partial r^2} = 1$, a teda v bode $r = 1$ má táto funkcia minimum. Inverzná funkcia k jej prvej derivácii bude

$$u = r - 1 \Rightarrow F(u) = 1 + u$$

Problém tejto funkcie je, že umožňuje výpočet aj záporných váh, čo v praxi nie je možné. Ak však existuje optimálne riešenie (so všetkými váhami väčšími alebo rovnými ako 1), tak je toto rýchlo nájdené. Je dobré túto funkciu použiť na začiatku, aby sme zistili algebraickú riešiteľnosť sústavy. Ak nie je nájdené riešenie použitím lineárnej vzdialenostnej funkcie, sústava nemá korene [1].

- *raking ratio*

$$G(r) = r \ln r - r + 1$$

Táto metóda je sprevádzaná podobným problémom ako lineárna. Záporné váhy už síce nedáva, avšak umožňuje výpočet váh menších ako 1, čo je obdobne problémom. Navyše nie v každom prípade umožní nájsť presné riešenie, nakoľko množina hodnôt jej derivácie je obmedzená. Máme $\frac{\partial G}{\partial r} = \ln r + 1 - 1 = \ln r$, $\frac{\partial^2 G}{\partial r^2} = \frac{1}{r}$, teda v bode $r = 1$ má globálne minimum. Inverzná funkcia F potom bude tvaru

$$u = \ln r \Rightarrow F(u) = e^u$$

Nakoľko $w_k = d_k F(\lambda^T x_k)$, je možné sa presvedčiť, že takto definovaná vzdialenostná funkcia umožní pre nájdenie presného riešenia dosiahnuť pri výpočte nereálne hodnoty. Ak toto presné riešenie existuje, tak je rýchlo nájdené a v prípade, že zahrňuje iba reálne hodnoty váh, je táto vzdialenostná funkcia pre výpočet optimálna (podobne ako aj lineárna).

Ak by sa nám však hodnoty nakalibrovaných váh zdali príliš vzdialené od tých pôvodných, môžeme pristúpiť k vzdialenostným funkciám, ktorých špecifikum je možnosť určenia maximálnych prípustných hraníc. Jedná sa o ohraničené verzie predošlých dvoch.

- *logit*

$$G(r) = \frac{1}{A}\left[(r-L)\ln\frac{r-L}{1-L} + (U-r)\ln\frac{U-r}{U-1}\right]$$

kde $A = \frac{U-L}{(1-L)(U-1)}$, L a U sú hranice pre r (teda platí, že spodná hranica pre novú váhu je $w = L.d$). Táto metóda je zovšeobecnením metódy raking ratio, čo sa dá dokázať výpočtom limít pre extrémne prípady, a to $L \to 0^+, U \to \infty$. Presný postup výpočtu je uvedený v [1], dostávame

$$\lim_{L\to 0^+, U\to\infty} G_{\text{logit}}(r) = \frac{1}{1}(r\ln r + 1 - r - 0) = r\ln r - r + 1 = G_{\text{rr}}(r)$$

Podobne ako v prípade raking ratio zistíme, či táto funkcia spĺňa podmienky vzdialenostnej funkcie. Počítame ako v [1]

$$\frac{\partial G}{\partial r} = \frac{1}{A}\left(\ln\frac{r-L}{1-L} - \ln\frac{U-r}{U-1}\right) = 0$$
$$(r-L)(U-1) = (1-L)(U-r)$$
$$r(U-L) = (U-L) \Rightarrow r = 1$$
$$\frac{\partial^2 G}{\partial r^2} = \frac{1}{A}\left(\frac{1}{r-L} + \frac{1}{U-r}\right)$$

čo pre stacionárny bod 1 dáva hodnotu 1, a teda sa jedná o minimum. Inverzná funkcia k prvej derivácii potom bude (odvodenie v [1])

$$F(u) = \frac{L(U-1) + U(1-L)e^{Au}}{(U-1) + (1-L)e^{Au}}$$

Všimnime si definičný obor funkcie $\frac{\partial G}{\partial r}$. Pri základnom predpoklade, a to, že L bude menšia ako 1 a U väčšia ako 1 máme $L < r < U$. Obor hodnôt $\frac{\partial G}{\partial r}$ je

$$\lim_{r\to L^+}\frac{1}{A}\left(\ln\frac{r-L}{1-L} - \ln\frac{U-r}{U-1}\right) = -\infty$$
$$\lim_{r\to U^-}\frac{1}{A}\left(\ln\frac{r-L}{1-L} - \ln\frac{U-r}{U-1}\right) = \infty$$

tak ako v [1]. Tento interval potom tvorí definičný obor funkcie $F(u)$, oborom hodnôt bude $(L;U)$ a výsledné nakalibrované váhy tak budú spadať do želaného rozpätia. Malou nevýhodou je, že v súčasnom stave nie je možné stanoviť si rozpätie pre každú váhu zvlášť. Teoreticky by sa možnosťou takéhoto nastavenia výrazne zvýšila teoretická hodnota kalibračného procesu (rovnako aj užívateľský komfort a voľnosť), prakticky by však takmer vždy viedla k neriešiteľnému systému s možným iba približným riešením. Ako sa uvádza v [1], „pri prísnom stanovení hraníc sa zavše môže stať, že program vypočíta iba približné

riešenie namiesto presného, nakoľko takéto obmedzenie mu výrazne znižuje množinu výsledkov. Preto treba s touto metódou zaobchádzať citlivo a vyskúšať viacero alternatívnych hodnôt pre nastavenie hornej a dolnej hranice, aby sa dospelo k schodnému výsledku.“ Rovnako je dôležité si uvedomiť, že čím viac váh ideme kalibrovať, tým lepšie riešenie sme schopní dosiahnuť. Prvé možné riešenie dostaneme, ak sa počet nezávislých stĺpcov matice x rovná počtu váh, pri stúpajúcom počte kalibrovaných jednotiek oproti nezávislým stĺpcom stúpa aj voľnosť algoritmu pri hľadaní takých váh, ktoré budú čo najbližšie k pôvodným.

- *lineárna ohraničená*

$$G(r) = \begin{cases} \frac{1}{2}(r-1)^2 & L \le r \le U \\ +\infty & \text{inak} \end{cases}$$

je ohraničená verzia klasickej lineárnej metódy. Podobne ako metóda logit je efektívnym riešením kalibrácie, kde sa nové váhy príliš odkláňajú od pôvodných a je potreba ich uzavrieť do prípustného intervalu. Aj tu však treba počítať s možnosťou nájdenia iba približného riešenia, jeho presnosť závisí na vôli a veľkosti nami zvoleného intervalu.

Tak ako stojí v [1], „kód Calif má v sebe zahrnuté všetky 4 typy vzdialenostných funkcií, obe metódy optimalizácie (ako prípad $n + J$ rovníc, tak i prípad J rovníc), pracuje ako s numerickými premennými, tak i s kategorickými, je možné mu zadať maximálny počet iteračných krokov i želanú presnosť, výsledky prezentuje v prehľadnej forme priamo v konzole R softvéru, rovnako ich zapisuje do .txt súboru a čo je najpodstatnejšie, umožňuje kalibrovať vzhľadom na zvolenú stratifikáciu a dokáže nájsť aj približné riešenie (narozdiel od makra Calmar2, ktoré stratifikáciu neposkytuje a je schopné nájsť iba presné riešenie).“ Podobne balík *sampling* so svojou funkciou calib nám ponúka plejádu nástrojov, od hore uvedených 4 vzdialenostných funkcií, cez možnosť zadať maximálny počet iterácií, schopnosť nájsť približné riešenie vo veľmi rýchlom čase až po zaujímavé grafické výstupy. Čo je však trochu jeho slabinou je práca s kategorickými premennými, kedy si musí užívateľ vopred pripraviť špeciálnu tabuľku vstupov (samotný výpočet už problémom nie je) a kalibrácia v stratifikovanom prípade (je nutné kalibrovať stratu po strate). Schopnosť nájsť približné riešenie je oproti Calmaru najsilnejšou zbraňou kódu Calif a funkcie calib. Naproti tomu Calmar2 sa môže pýšiť prostredím v ktorom pracuje (SAS), čo z neho robí veľmi atraktívny nástroj. Oboje Calmar2 i calib dokážu narábať so zovšeobecnenými kalibráciami v prípade neodpovedí v zisťovaní. Prehľad výhod a nevýhod jednotlivých programov:

Výhody:

Calmar2 – práca pod silným komerčným softvérom, dlhodobý vývoj (od začiatku 90-tych rokov), zovšeobecnené kalibrácie, špecializácia na sociálne štatistiky, 5 vzdialenostných funkcií

Calif – približné riešenia, stratifikácia, jednoduché vstupy, najprehľadnejšie výstupy zo všetkých troch, chybové hlášky v slovenčine, otvorenosť pre ďalší vývoj, bezplatný

calib – najrýchlejší a najpresnejší riešiteľ spomedzi všetkých troch, grafické výstupy, bezplatný

Nevýhody:

Calmar2 – nemožnosť nájsť často nutné približné riešenie, bez stratifikácie, cena

Calif – občasné problémy s metódou logit, dlhší čas riešenia prípadov s ohraničenými metódami, približné riešenia sú niekedy až príliš približné (nie najlepšie možné)

calib – komplikovaný systém vstupov pre kategorické premenné, neprehľadné, hoci pekné výstupy

Čo sa približných riešení týka, tieto sú často potrebné, najmä v prípade použitia ohraničených metód. Povedzme napríklad, že chceme kalibrovať štatistické jednotky tak, aby sme v sume dostali povedzme 10 000 000 € a presné algebraické riešenie neexistuje. V takomto prípade Calmar2 riešenie nenájde, avšak Calif i calib sú schopní nájsť riešenie napríklad 10 000 002 €, čo je v reáli plne postačujúce.

Kód Calif má zabudovanú procedúru pre dvojitú stratifikáciu, ktorá sa dá aplikovať na kalibráciu váh v zisťovaní o inováciách realizované s dvojročnou periodicitou. V praktickej časti ukážeme aj príklad kalibrácie z tohto zisťovania.

3. Praktická časť

Pre funkčnosť kódu Calif je potrebné načítať balík *BB* [4], pre funkciu calib načítame balík *sampling* [5], oba v programe R [3].

1. príklad

Majme jednoduchú tabuľku s dvoma numerickými premennými

Tab. 1. Dáta k 1. príkladu

X	Y	VAHA
143	15	10
412	29	7,5
312	19	8,2
711	54	2,3
1050	102	1
178	12	11
255	19	8
201	22	7,5

a chceme ich navážiť na hodnoty 18 000 a 1 100 použitím metódy logit. Vyskúšame a porovnáme všetky tri programy, hranice určíme $L = 0, U = 6{,}81$.

a) Calmar2 (výpočet bol realizovaný v programe SAS Enterprise Guide 4.2 [6])

Calmar2 si s touto tabuľkou pre tieto hranice poradil, pričom nakalibrované váhy sú

Tab. 2. Nakalibrované váhy Calmarom

X	Y	VAHA
143	15	4,90E-13
412	29	0,000468984
312	19	55,80144169
711	54	3,35E-17
1050	102	2,87E-79
178	12	3,313150479
255	19	6,22098E-05
201	22	8,53E-21

z čoho ihneď vidíme ich nereálnosť (spôsobenú príliš obmedzujúcimi výslednými sumami a nie chybou výpočtu). Okrem tohto výstupu poskytuje Calmar aj ďalšie, ako napríklad:

```
COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE)
                      ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)

                              MARGE          MARGE      POURCENTAGE POURCENTAGE
         VARIABLE    MODALITÉ    ÉCHANTILLON    POPULATION  ÉCHANTILLON POPULATION

            X                       15269.2         18000           .          .
            Y                        1198.5          1100           .          .
MÉTHODE : LOGIT, INF=0.00, SUP=6.81
```

PREMIER TABLEAU RÉCAPITULATIF DE L'ALGORITHME :
LA VALEUR DU CRITČRE D'ARRĘT ET LE NOMBRE DE POIDS NÉGATIFS APRČS CHAQUE ITÉRATION

ITÉRATION	CRITČRE D'ARRĘT	POIDS NÉGATIFS
1	2.61927	0
2	1.93272	0
3	1.11729	0
4	0.63567	0
5	0.36783	0
6	0.34621	0
7	0.17127	0
8	0.03974	0
9	0.00888	0
10	0.00290	0
11	0.00061	0
12	0.00004	0

MÉTHODE : LOGIT, INF=0.00, SUP=6.81
DEUXIČME TABLEAU RÉCAPITULATIF DE L'ALGORITHME :
LES COEFFICIENTS DU VECTEUR LAMBDA DE MULTIPLICATEURS DE LAGRANGE APRČS CHAQUE ITÉRATION

VARIABLE	MODALITÉ	LAMBDA1	LAMBDA2	LAMBDA3	LAMBDA4	LAMBDA5	LAMBDA6
X		0.02102	0.04045	0.07141	0.10875	0.14980	0.20002
Y		-0.26046	-0.55436	-1.02267	-1.59568	-2.23438	-3.01986

VARIABLE	LAMBDA7	LAMBDA8	LAMBDA9	LAMBDA10	LAMBDA11	LAMBDA12
X	0.24615	0.27700	0.29769	0.30950	0.31277	0.31297
Y	-3.73476	-4.20162	-4.50907	-4.68436	-4.73275	-4.73576

The SAS System 15:29 Wednesday, May 15, 2013

MÉTHODE : LOGIT, INF=0.00, SUP=6.81

The UNIVARIATE Procedure
Variable: __WFIN (PONDÉRATION FINALE)

Basic Statistical Measures

Location		Variability	
Mean	7.389390	Std Deviation	19.59575
Median	0.000031	Variance	383.99330
Mode	.	Range	55.80144
		Interquartile Range	1.65681

Quantiles (Definition 5)

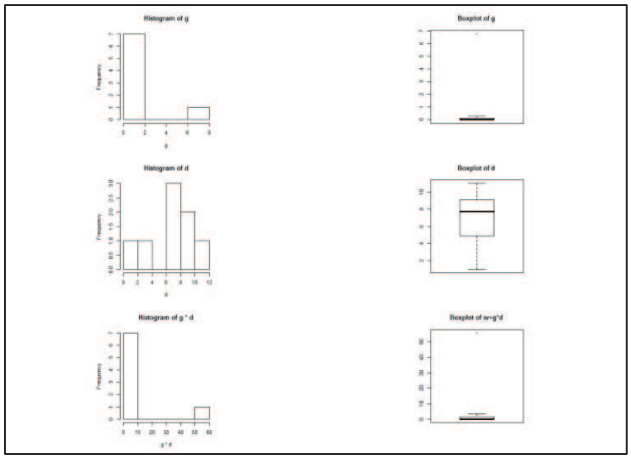
Quantile	Estimate
100% Max	5.58014E+01
99%	5.58014E+01
95%	5.58014E+01
90%	5.58014E+01
75% Q3	1.65681E+00
50% Median	3.11049E-05
25% Q1	1.67611E-17
10%	2.86530E-79
5%	2.86530E-79
1%	2.86530E-79
0% Min	2.86530E-79

```
Extreme Observations

        ---------------Lowest---------------        ---------------Highest--------------

              Value          ID_HD       Obs               Value          ID_HD       Obs

         2.86530E-79              5         5         4.90173E-13              1         1
         8.52519E-21              8         8         6.22098E-05              7         7
         3.35137E-17              4         4         4.68984E-04              2         2
         4.90173E-13              1         1         3.31315E+00              6         6
         6.22098E-05              7         7         5.58014E+01              3         3


                    Stem Leaf                     #             Boxplot
                       5 6                        1                *
                       4                                           |
                       3                                           |
                       2                                           |
                       1                                           |
                       0 0000003                  7             +--+--+
                         ----+----+----+----+
                     Multiply Stem.Leaf by 10**+1


                                    Normal Probability Plot
                    55+                                                *       ++++
                      |                                              +++++
                      |                                          +++++
                      |                                      +++++
                      |                                  +++++
                     5+           *    *   * +*+++*  *   *
                      +----+----+----+----+----+----+----+----+----+----+----+
                                    -2        -1         0        +1        +2
```

b) calib

Výpočet pomocou funkcie calib priniesol rovnaké hodnoty váh, výstupy vyzerajú nasledovne:

```
calib(Xs=data1[,1:2],d=data1[,3],total=c(18000,1100),method="logit",bounds=
c(0,6.81),description=TRUE,max_iter=20000)
number of iterations  7663
summary - initial weigths d
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   6.200   7.750   6.938   8.650  11.000
summary - final weigths w=g*d
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00000  0.00000  0.00003  7.38900  0.82860 55.80000
[1]  4.919680e-14  6.258334e-05  6.805047e+00  1.463159e-17  2.926959e-79
3.011954e-01 7.785823e-06 1.143263e-21
```



Obr. 1. Grafický výstup funkcie calib

Vypočítané hodnoty však nereprezentujú skutočné hodnoty váh, ale hodnoty $\frac{w_k}{d_k} = F(\lambda^T x_k)$. Pre získanie váh teda potrebujeme vypočítané hodnoty prenásobiť vektorom pôvodných váh.

```
calib(Xs=data1[,1:2],d=data1[,3],total=c(18000,1100),method="logit",bounds=
c(0,6.81),description=FALSE,max_iter=20000)*data1[,3]
[1]   4.919680e-13   4.693751e-04   5.580139e+01   3.365266e-17   2.926959e-79
3.313149e+00 6.228658e-05 8.574472e-21
```

c) Calif

Spôsob výpočtu naprogramovaný ako R kód na Štatistickom úrade SR (s pracovným názvom Calif) na základe metodiky Calmaru nemal s touto úlohou problém

```
kalibracia(data=data1,margins=c(18000,1100),met=3,L=0,U=6.81)
  Successful convergence.
$vahy
[1]   0.000000   0.000469 55.801441   0.000000   0.000000   3.313151   0.000062
0.000000

$sumy
    X     Y
18000  1100

$min
[1] 0

$max
[1] 55.80144

$podiely_vah
[1] 0.00 0.00 6.81 0.00 0.00 0.30 0.00 0.00

$L
[1] 2.866373e-79

$U
[1] 6.805054

$L_w1
[1] 1

$suma_rozdielov
[1] 91.58776

$info
  upozornenie
1 Uspesna konvergencia
```

Legenda:

$vahy – nakalibrované váhy
$sumy – dosiahnuté sumy
$min – najmenšia hodnota spomedzi nakalibrovaných váh
$max – najväčšia hodnota spomedzi nakalibrovaných váh
$podiely_vah – podiely nakalibrovaných a pôvodných váh pre každú váhu zvlášť
$L – najnižší podiel
$U – najvyšší podiel
$L_w1 – spodná hranica, pre ktorú všetky váhy budú väčšie alebo rovné ako 1 (môže sa meniť v závislosti od nastavených súm)
$suma_rozdielov – celková suma rozdielov medzi pôvodnými a nakalibrovanými váhami (miera odchýlky)
$info – upozornenia procesu priebehu kalibrácie

Zobrazené váhy sú zaokrúhlené, skutočné hodnoty s presnosťou na cca 18 desatinných miest Calif automaticky vygeneruje ako externý súbor.

Vezmime si však teraz trochu prísnejšie ohraničenie, napríklad $L = 0{,}2\,, U = 2$. Calmar2 v tomto prípade nenašiel riešenie, rovnako ani calib pre metódu logit, avšak našiel riešenie pre lineárnu ohraničenú metódu s týmito hranicami.

```
a<-
calib(Xs=data1[,1:2],d=data1[,3],total=c(18000,1100),method="truncated",bou
nds=c(0.2,2),description=TRUE,max_iter=100000)
No convergence in 1e+05 iterations with the given bounds.
The bounds for the g-weights are: 0.2  and  2
 and the g-weights are given by g
number of iterations  1e+05
summary - initial weigths d
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   6.200   7.750   6.938   8.650  11.000
summary - final weigths w=g*d
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.200   1.240   4.239   8.005  15.350  22.000
[1] 0.2000000 2.0000000 2.0000000 0.2000000 0.2000000 2.0000000 0.8098355
0.2000000
a*data1[,3]
[1]  2.000000 15.000000 16.400000  0.460000  0.200000 22.000000  6.478684
1.500000
(a*data1[,3])%*%as.matrix(data1[1:2])
           X        Y
[1,] 17989.42 1241.935
```

Môžeme vidieť výsledné nakalibrované váhy, ako i sumy, ku ktorým sme dospeli. Calif našiel približné riešenie použitím metódy logit ako i lineárnej ohraničenej, konkrétne však lepšie pre prvú menovanú

```
kalibracia(data=data1,margins=c(18000,1100),met=3,L=0.2,U=2)
    Unsuccessful convergence.
$vahy
[1]  2.000000 15.000000 16.400000  0.460000  0.200000 22.000000  3.838588
1.500000

$sumy
        X         Y
17316.200  1191.773

$min
[1] 0.2

$max
[1] 22

$podiely_vah
[1] 0.20 2.00 2.00 0.20 0.20 2.00 0.48 0.20

$L
[1] 0.2

$U
[1] 2

$L_w1
[1] 1
```

```
$suma_rozdielov
[1] 47.50141

$info
  upozornenie
1 Neuspesna konvergencia
2 Skuste rozsirit pripustny interval (L,U)
```

2. príklad (ako je uvedený v [1])

Uvažujme tabuľku

Tab. 3. Dáta k 2. príkladu

X	Y	Z	VAHA
1	1	1	10
2	1	1	11
2	1	3	13
2	2	2	7
2	2	2	8
1	2	2	8
2	1	2	9
2	2	2	14
1	1	3	4,5
1	2	1	2,7
2	2	3	11,2
1	1	1	8
2	1	2	3
1	1	1	6
2	1	1	5
2	2	2	4
2	1	1	4
2	1	1	5
2	2	2	4
2	2	1	6
1	1	2	10
2	1	3	5
1	2	1	6
2	2	1	6
1	2	2	5
2	2	2	5

$$x = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 3 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 \\ 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 & 2 \\ 1 & 0 & 1 & 0 & 3 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 3 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 2 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 0 & 3 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 & 2 \end{pmatrix}$$

kde X a Y sú kategorické premenné a Z je numerická. VAHA je pôvodná váha pred kalibráciou. Chceme dostať výsledné sumy nasledovne

Tab. 4. Sumy k 2. príkladu

X1	X2	Y1	Y2	Z
95	150	125	140	375

Pre kategorickú premennú Q a jej k-tu kategóriu má prvok x_{ik} matice $x^{(Q)}$ hodnotu

$$x_{ik} = \begin{cases} 1 \text{ ak } Q_i = k \\ 0 \quad \text{inak} \end{cases}$$

Matica x bude mať v tomto prípade horeuvedený tvar. V prípade Calmaru a Califu stačí načítať dátovú tabuľku a určiť kategorické premenné, pre načítanie do funkcie calib potrebujeme skonštruovať maticu x. Použitím metódy raking ratio Calmar nedokázal nájsť

presné riešenie, iteračným nepohodlným postupom (po jednotlivých premenných) sme sa dopracovali k váham, ktoré nám dali výsledné sumy

Tab. 5. Sumy vypočítané Calmarom

X1	X2	Y1	Y2	Z
104,9609	160,0391	125	140	375

Je vidieť, že suma odchýlok od želaných hodnôt súm má hodnotu 20. Calif vyriešil systém s rovnakou odchýlkou, avšak s menším rozdielom oproti pôvodným váham. Kým v prípade Calmaru sa celkový rozdiel vyšplhal na 136,606, v prípade Califu iba na 117,34. Výsledné sumy boli

Tab. 6. Sumy vypočítané Califom

X1	X2	Y1	Y2	Z
100	155	120	135	375

```
kalibracia(data=data2,kat=c("X","Y"),margins=c(95,150,125,140,375),met=2,ri
es=TRUE)
  Unsuccessful convergence.
$vahy
 [1] 19.386485 20.082394  5.031815  8.824595 10.085252 10.709347  7.565621
17.649190   1.849567   7.849781   6.501207 15.509188   2.521874 11.631891
9.128361   5.042626   7.302689   9.128361   5.042626 16.427397   8.926441
1.935313 17.443958
[24] 16.427397  6.693342  6.303282

$sumy
X.1 X.2 Y.1 Y.2   Z
100 155 120 135 375

$min
[1] 1.849567

$max
[1] 20.08239

$podiely_vah
 [1] 1.94 1.83 0.39 1.26 1.26 1.34 0.84 1.26 0.41 2.91 0.58 1.94 0.84 1.94
1.83 1.26 1.83 1.83 1.26 2.74 0.89 0.39 2.91 2.74 1.34 1.26

$L
[1] 0.3870627

$U
[1] 2.907326

$L_w1
[1] 0.3703704

$suma_rozdielov
[1] 117.3363

$info
  upozornenie
1 Neuspesna konvergencia
```

Systém nebol plne riešiteľný preto, lebo neplatilo X1+X2 = Y1+Y2 (čo by malo byť principiálne vylúčené). Použili sme ho na otestovanie situácie, kedy nie je možné dostať presné riešenie. Funkcia calib metódou raking ratio riešenie nenašla, v prípade lineárnej dokázala nájsť identický výsledok ako kód Calif (pre lineárny prípad). Bolo však najskôr potrebné si vytvoriť maticu x

```
data2x<-matrix(0,26,5)
data2x[data2[1]==1,1]<-1
data2x[data2[1]==2,2]<-1
data2x[data2[2]==1,3]<-1
data2x[data2[2]==2,4]<-1
data2x[,5]<-data2[,3]
calib(Xs=data2x,d=data2[,4],total=c(95,150,125,140,375),method="raking")
NULL
calib(Xs=data2x,d=data2[,4],total=c(95,150,125,140,375),method="linear")*da
ta2[,4]
 [1]  19.7799308  20.7606501   1.2688125  10.0313004  11.4643433  12.1896333
8.9321975 20.0626008  0.8471800  6.5301379  6.0275878 15.8239446  2.9773992
11.8679585   9.4366591   5.7321717   7.5493273   9.4366591   5.7321717
13.9674501 10.8312765
[22]  0.4880048 14.5114176 13.9674501  7.6185208  7.1652146
(calib(Xs=data2x,d=data2[,4],total=c(95,150,125,140,375),method="linear")*d
ata2[,4])%*%data2calib
     [,1] [,2] [,3] [,4] [,5]
[1,]  100  155  120  135  375
```

Z výpočtu vidíme, že calib sa dostal na rovnaké sumy ako Calif (hoci inou metódou).

3. príklad

V tomto príklade porovnáme calib a Calif v prípade stratifikovaného výberu. Pri takomto výbere nemajú všetky jednotky výberového súboru pred kalibráciou rovnakú pôvodnú váhu. Tieto sa líšia v závislosti od príslušnosti do konkrétnej straty. Je preto potrebné zadať kalibračné sumy pre každú stratu zvlášť. Použijeme metódu raking ratio.

Tab. 7. Dáta k 3. príkladu

A	B	STRATY	VAHA
165477	34	100	1,5
205511	11	100	1,5
94539	52	100	1,5
40663	9	100	1,5
42185	9	100	1,5
10845	52	100	1,5
6357	47	100	1,5
45547	95	100	1,5
214456	21	100	1,5
83796	98	105	1,87
8749	20	105	1,87
1023532	50	105	1,87
3126	63	105	1,87
3784	22	110	6,325
30000	17	110	6,325
51494	71	110	6,325
19497	45	110	6,325
12600	20	110	6,325
18887	29	110	6,325
3032	92	110	6,325
31491	58	110	6,325
34215	33	110	6,325

Vidíme, že v tabuľke máme 3 straty, označené ľubovoľnými príslušnými hodnotami. Kalibrácia v Calife zohľadní túto stratifikáciu a vypočíta nám váhy tak, aby sedeli sumy za jednotlivé straty zvlášť.

Tab. 8. Sumy k 3. príkladu

STRATY	A	B
100	1100000	320
105	2200000	416
110	1600000	2000

```
kalibracia(data=data3,sumy=sumy3,stratifikacia=TRUE,met=2)
  Successful convergence.
  Successful convergence.
  Successful convergence.
$vahy
 [1]  1.237746  1.590983  0.985368  1.422043  1.423768  0.921816  0.963349
0.629067   1.456718   1.721987   1.837183   1.987350   1.764034   3.955234
14.008058   7.457960   4.002879   6.107045   6.172158   0.513876   4.603869
10.599052

$sumy
                A           B
100 1100000.0165    319.9999
105 2199999.9238    416.0001
110 1600000.0374   2000.0000

$min
[1] 0.513876

$max
[1] 14.00806

$podiely_vah
 [1] 0.83 1.06 0.66 0.95 0.95 0.61 0.64 0.42 0.97 0.92 0.98 1.06 0.94 0.63
2.21 1.18 0.63 0.97 0.98 0.08 0.73 1.68

$L
[1] 0.08124517

$U
[1] 2.214713

$L_w1
[1] 0.6666667

$suma_rozdielov
[1] 29.14026

$info
  upozornenie
1 Uspesna konvergencia
```

Táto metóda nám rýchlo vypočítala váhy, ktoré však, ako môžeme vidieť, nie sú všetky väčšie alebo rovné ako 1. Pristúpime preto k metóde logit a zadáme minimálnu a maximálnu hranicu pre pomer w_i/d_i . Parameter L_w1 nám napomáha zvoliť takú hranicu, pre ktorú budú určite všetky váhy zmysluplné.

```
kalibracia(data=data3,sumy=sumy3,stratifikacia=TRUE,met=3,L=2/3,U=4)
  Unsuccessful convergence.
  Successful convergence.
  Unsuccessful convergence.
$vahy
 [1]  1.000000  2.328894  1.000000  1.000000  1.000000  1.000000  1.000000
1.000000   1.000000   1.723043   1.835495   1.987281   1.762981   4.216667
25.300000   4.216667   4.216667   4.216667   4.216667   4.216667   4.216667
5.919663

$sumy
                A           B
100 1098682.3276    344.6178
```

```
105 2199999.9893     416.0000
110 1555184.6730    2046.4655

$min
[1] 1

$max
[1] 25.3

$podiely_vah
 [1] 0.67 1.55 0.67 0.67 0.67 0.67 0.67 0.67 0.67 0.92 0.98 1.06 0.94 0.67
4.00 0.67 0.67 0.67 0.67 0.67 0.67 0.94

$L
[1] 0.6666667

$U
[1] 4

$L_w1
[1] 0.6666667

$suma_rozdielov
[1] 39.37333

$info
  upozornenie
1 Neuspesna konvergencia
2 Skuste rozsirit pripustny interval (L,U)
```

Vypočítané váhy sú teraz všetky väčšie alebo rovné ako 1. To má však za následok, že niektoré výsledné sumy nemôžu vyhovovať nášmu výpočtu (príliš obmedzujúce hranice). Je preto na našom úsudku, a to prípad od prípadu, zvoliť vhodný kompromis medzi presnosťou výsledných súm a minimalizáciou rozdielu pôvodných a nakalibrovaných váh.

Tu je dôležité zdôrazniť, že pre najpresnejší možný výpočet v prípade použitia ohraničenej metódy je vhodné kalibrovať stratu po strate, kedy môžeme pre každú jednotlivo zadať iné hranice. V našom príklade sme pre všetky tri straty museli zadať rovnaké hranice, čo mohlo spôsobiť prílišné obmedzenie v niektorých z nich. Kalibrácia cez všetky straty súčasne je tak vhodná najmä v prípade použitia lineárnej alebo raking ratio metódy.

Ten istý prípad sme vyskúšali kalibrovať aj pomocou funkcie calib. Použili sme metódu truncated (nakoľko táto v calibe funguje lepšie ako logit), pre všetky straty sme stanovili rovnaké hranice $L = \frac{2}{3}, U = 4$. Keďže calib neponúka možnosť stratifikácie, počítali sme pre každú stratu zvlášť.

```
a1<-
calib(Xs=data3[data3$STRA==100,1:2],d=data3[data3$STRA==100,4],total=c(1100
000,320),method="truncated",bounds=c(2/3,4),max_iter=100000)*data3[data3$ST
RA==100,4]
No convergence in 1e+05 iterations with the given bounds.
The bounds for the g-weights are: 0.6666667  and  1.55687
 and the g-weights are given by g
a2<-
calib(Xs=data3[data3$STRA==105,1:2],d=data3[data3$STRA==105,4],total=c(2200
000,416),method="truncated",bounds=c(2/3,4),max_iter=100000)*data3[data3$ST
RA==105,4]
a3<-
calib(Xs=data3[data3$STRA==110,1:2],d=data3[data3$STRA==110,4],total=c(1600
000,2000),method="truncated",bounds=c(2/3,4),max_iter=100000)*data3[data3$S
TRA==110,4]
No convergence in 1e+05 iterations with the given bounds.
The bounds for the g-weights are: 0.1012234  and  5.188147
```

```
 and the g-weights are given by g
a1
[1] 1.000000 2.335306 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
1.000000
a2
[1] 1.721477 1.838044 1.987383 1.764527
a3
[1]   4.2166667  32.8150277   4.2166667   4.2166667   4.2166667   4.2166667
4.2166667  4.2166667  0.6402382
a1%*%as.matrix(data3[data3$STRA==100,1:2])
           A        B
[1,] 1100000 344.6884
a2%*%as.matrix(data3[data3$STRA==105,1:2])
           A   B
[1,] 2200000 416
a3%*%as.matrix(data3[data3$STRA==110,1:2])
           A    B
[1,] 1600000 2000
```

Výsledné váhy sú veľmi podobné tým, ktoré boli vypočítané v Calife, avšak zásadný rozdiel je v poslednej z nich. Tú totižto calib vypočítal menšiu ako 1, a tým aj docielil presné výsledné sumy. Takéto riešenie by však nebolo možné.

4. príklad

Ukážeme si funkčnosť kódu Calif v zisťovaní o inováciách. Pre toto zisťovanie je špecifický netradičný postup kalibrácie, ktorá sa vykonáva na triedení strata x región, t.z. sumy by mali byť docielené v každom regióne v každej strate. Ak sú však v strate výberového súboru údaje za iné regióny ako v požadovaných sumách (stačí keď sa líšia v jednom regióne), kalibruje sa na vyššej úrovni, teda na úrovni straty namiesto regiónu. Rovnako sa postupuje aj v prípade, že pre niektorý región neexistuje algebraické riešenie.

Majme dve kategorické premenné X a Y, numerickú premennú Z a región SUJ.

Tab. 9. Dáta k 4. príkladu

X	Y	Z	SUJ	STRA	VAHA
1	1	1	1	25	10
2	1	1	1	25	11
2	1	3	2	25	13
2	2	2	2	25	7
2	2	2	2	25	8
1	2	2	3	25	8
2	1	2	3	25	9
2	2	2	3	25	14
1	1	3	1	30	4,5
1	2	1	1	30	2,7
2	2	3	2	30	11,2
1	1	1	1	35	8
2	1	2	1	35	3
1	1	1	1	35	6
2	1	1	1	35	5
1	1	1	1	35	3
2	2	2	2	35	4
2	1	1	2	35	4
2	1	1	2	35	5
2	2	2	2	35	4

2	2	1	2	35	6
1	1	2	3	35	10
2	1	3	3	35	5
1	2	1	3	35	6
2	2	1	3	35	6
1	2	2	3	35	5
2	2	2	3	35	5
2	1	1	3	35	3

Výsledné sumy majú byť

Tab. 10. Sumy k 4. príkladu

STRATY	SUJ	X1	X2	Y1	Y2	Z
25	1	10	15	15	0	20
25	2	0	20	5	30	80
25	3	10	25	10	20	40
30	1	5	5	3	10	30
30	2	5	5	3	4	20
30	3	2	4	2	4	21
35	1	15	12	27	0	30
35	2	0	30	10	20	40
35	3	25	18	20	23	72

Vidíme, že v strate 30 sú vo výberovom súbore údaje za 2 regióny, avšak v požadovaných sumách za 3 regióny, budeme teda kalibrovať na úrovni celej straty 30 (napr. X2 sa bude rovnať 14). Problémom však je, že údajov za túto stratu je primálo, algebraické riešenie neexistuje. Program túto stratu obíde.

```
kalibracia(data=data4,sumy=sumy4,kat=c("X","Y"),stratifikacia=TRUE,inovacie
=TRUE,met=1)
  Successful convergence.
  Successful convergence.
  Successful convergence.
  Successful convergence.
$vahy
 [1]  9.925053 12.208485  2.133546  9.637080 11.013806 10.074951  5.732915
19.274161  4.500000  2.700000 11.200000  7.058817  3.000009  5.294113
9.000001  2.647057  5.000000  4.444444  5.555556  5.000000 10.000000
12.338628  3.812863
[24]  7.615653  6.348702  5.045719  3.989926  3.848508

$sumy
          X1        X2        Y1        Y2         Z
25  20.00000  59.99999  30.00000  50.00000 140.00000
35  14.99999  12.00001  27.00000            30.00001
35            30.00000  10.00000  20.00000  40.00000
35  25.00000  18.00000  20.00000  23.00000  72.00000

$min
[1] 2.133546

$max
[1] 19.27416

$podiely_vah
 [1] 0.99 1.11 0.16 1.38 1.38 1.26 0.64 1.38 1.00 1.00 1.00 0.88 1.00 0.88
1.80 0.88 1.25 1.11 1.11 1.25 1.67 1.23 0.76 1.27 1.06 1.01 0.80 1.28

$L
[1] 0.1641189

$U
[1] 1.8
$L_w1
```

```
[1] 0.3333333

$suma_rozdielov
[1] 48.81141

$info
  upozornenie
1 Kvoli malym rozsahom, alebo neriesitelnosti systemu neboli nakalibrovane
tieto jednotky : c(9, 10, 11)
2 Uspesna konvergencia
```

Nastavením parametra ries=TRUE program nájde aj približné riešenie pre stratu 30.

```
kalibracia(data=data4,sumy=sumy4,kat=c("X","Y"),stratifikacia=TRUE,inovacie
=TRUE,met=1,ries=TRUE)
  Successful convergence.
  Unsuccessful convergence.
  Successful convergence.
  Successful convergence.
  Successful convergence.
$vahy
 [1]  9.925053 12.208485  2.133546  9.637080 11.013806 10.074951  5.732915
19.274161  8.178571  3.857143 14.178571  7.058817  3.000009  5.294113
9.000001  2.647057  5.000000  4.444444  5.555556  5.000000 10.000000
12.338628  3.812863
[24]  7.615653  6.348702  5.045719  3.989926  3.848508

$sumy
          X1         X2         Y1         Y2          Z
25  20.000004  59.999993  29.999999  49.999999 140.000003
30  12.035714  14.178571   8.178571  18.035714  70.928571
35  14.999987  12.000010  26.999997             30.000007
35             30.000000  10.000000  20.000000  40.000000
35  25.000000  18.000000  20.000000  23.000000  72.000000

$min
[1] 2.133546

$max
[1] 19.27416

$podiely_vah
 [1] 0.99 1.11 0.16 1.38 1.38 1.26 0.64 1.38 1.82 1.43 1.27 0.88 1.00 0.88
1.80 0.88 1.25 1.11 1.11 1.25 1.67 1.23 0.76 1.27 1.06 1.01 0.80 1.28

$L
[1] 0.1641189

$U
[1] 1.81746

$L_w1
[1] 0.3703704

$suma_rozdielov
[1] 56.6257

$info
  upozornenie
1 Neuspesna konvergencia
```

4. Záver

V tomto článku sme predstavili program vytvorený Štatistickým úradom SR v bezplatnom R softvéri určený na kalibráciu váh jednotiek štatistických zisťovaní a porovnali sme ho s makrom Calmar2 a funkciou calib, ktorá je súčasťou balíka „sampling“. Rozpracovali sme matematické pozadie kalibrácie váh, ktorej stavebným prvkom je nájdenie

viazaných extrémov. Z uvedených príkladov môžeme vidieť výsledky výpočtov jednotlivých programov.

Je nutné vysloviť, že kód Calif je stále vo vývoji a je priestor na jeho zlepšenie. Potenciál momentálne vidíme v zapracovaní novej metódy, ktorá bude automaticky počítať všetky váhy väčšie alebo rovné ako 1 bez zadávania ohraničení, GUI rozhrania a najmä v implementácii funkcie calib priamo do kódu Calif ako tretiu možnosť optimalizačného výpočtu. Pre potreby sociálnych štatistík by bolo užitočné, ak by užívateľ mohol vopred zadefinovať, akú presnosť by si želal na vážené sumy hodnôt jednotlivých premenných (napríklad z predošlého príkladu – presnosť na A 100%, presnosť na B 80%).

5. Literatúra

[1] FRANKOVIČ, B. 2013. *Kalibrácia váh v štatistických zisťovaniach*. V *Slovenská štatistika a demografia 1/2013.* Bratislava: ŠÚ SR, 2013. ISSN 1210-1095.

[2] SAUTORY, O. 1993. *La macro CALMAR*. Paríž: INSEE, 1993. Dostupné na: http://www.insee.fr/fr/methodes/outils/calmar/doccalmar.pdf

[3] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

[4] VARADHAN, R., GILBERT, P. 2009. BB: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function. Journal of Statistical Software, 32(4), 1-26. Dostupné na: http://www.jstatsoft.org/v32/i04/

[5] TILLÉ, Y., MATEI, A. 2012. sampling: Survey Sampling. R package version 2.5. Dostupné na: http://CRAN.R-project.org/package=sampling

[6] SAS Enterprise Guide, Version 4.2. Copyright © 2006 – 2008 by SAS Institute, Inc., Cary, NC, USA. All rights reserved.

Kontakt

Boris Frankovič
Štatistický úrad SR
Odbor metód štatistických zisťovaní
Miletičova 3
824 67 Bratislava 26
boris.frankovic@statistics.sk

Exact Test for Dispersion

Presný test rozptylu

Michal Illovský

Abstract: The presented article proposes a non-parametric alternative for testing for equal variance between two samples. The test is applicable under the assumptions, that both samples are random samples from their respective populations, the measurement scale is continuous and there is independence assumed between and within the samples.

Firstly, we conducted three sets of experiments (for normal, heavy-tailed and skewed data) to compare the power of our test to that of the F-test and Brown-Forsythe test as proportion of correctly rejected null hypotheses while alternating both level of significance and the effect size. Secondly, we varied the sample sizes to see, how the test would behave with increasing number of data. Thirdly, we simulated non-equal sample sizes.

We conclude that under certain conditions our test can be viewed as a preferable alternative to the F-test (non-normality) and the Brown-Forsythe test (both lower and unequal sample sizes).

Abstrakt: V príspevku som navrhol neparametrickú alternatívu testu rovnosti rozptylu dvoch štatistických súborov. Test možno použiť za podmienok, že vzorky sú náhodnými výbermi zo svojich základných súborov, že veličina je spojitá, a za predpokladu nezávislosti pozorovaní medzi vzorkami aj v rámci vzorky.

Vykonal som sériu troch experimentov (pre normálne dáta, rozdelenie s ťažkými chvostmi a pravostranne zošikmené rozdelenie) za účelom porovnania sily testu s F-testom a Brown-Forsythovým testom. Sledoval som podiel správne zamietnutých nulových hypotéz pri meniacej sa hladine významnosti a veľkosti rozdielu (effect size). V ďalšom som posudzoval správanie sa testu s meniacou sa veľkosťou vzorky. Napokon som simuloval nerovnaký rozsah oboch vzoriek čo do počtu pozorovaní.

Výsledky simulácií viedli k záveru, že za určitých okolností je navrhovaný test vhodnejšou alternatívou v porovnaní s F-testom (nenormálne dáta) a s Brownovým-Forsythovým testom (malá veľkosť vzoriek a nerovnosť ich rozsahov).

Key words: statistical test, dispersion, variance, F-test, Brown-Forsythe test, binomial distribution

Kľúčové slová: štatistický test, disperzia, rozptyl, F-test, Brownov-Forsythov test, binomické rozdelenie

JEL classification: C14

Introduction

The presented article proposes a non-parametric alternative for testing for equal variance between two samples. The test is applicable under the assumptions, that both samples are random samples from their respective populations, the measurement scale is continuous and there is independence assumed between and within the samples.

Firstly, we conducted three sets of experiments (for normal, heavy-tailed and skewed data) to compare the power of our test to that of the F-test and Brown-Forsythe test as proportion of correctly rejected null hypotheses while alternating both level of significance and the effect size. Secondly, we

varied the sample sizes to see, how the test would behave with increasing number of data. Thirdly, we simulated non-equal sample sizes.

We conclude that under certain conditions our test can be viewed as a preferable alternative to the F-test (non-normality) and the Brown-Forsythe test (both lower and unequal sample sizes).

1. Data

The data consist of two random samples. Let $x_1, x_2, \ldots x_n$ denote the of size n from population 1 and let $y_1, y_2, \ldots y_m$ denote the random sample of size m from population 2.

2. Assumptions

- Both the samples are random samples from their respective populations
- The measurement scale is continuous
- There is independence assumed between and within the samples

3. Test Statistic and Null Distribution

Let med_X and med_Y stand for medians of the respective populations. If they are not known, sample medians are used instead and the test still is approximately valid. Convert each x and y to its absolute deviation from the median using:

$$k[i] = |x - medX|;\ i = 1, \ldots, n$$
$$l[i] = |y - medY|; i = 1, \ldots, m \tag{1}$$

Join the absolute median deviations together in one set and sort them by magnitude. This can either be done in ascending or descending order. We assumed the continuous scale, as we do not want any value k_i and l_i to be exactly equal. Now, let R be a merged set of the sets K and L and let every value thereof be represented by a rank r_i with i going from 1 to $N=n + m$ and there are no ties.

$$R = K \cup L$$
$$r[i]; i = 1, \ldots, n + m = N \tag{2}$$

The test statistics is obtained by halving the set R in two and counting the occurrences of K and L in the first half. Under the null hypothesis the variable K has the binomial distribution with the parameter p that stand for probability of success equal to n/N, where N stands for total sample size, and the parameter number of trials equal to $N/2$ (or alternatively, L has binomial distribution with parameters p equal to m/N and $N/2$, respectively).

$$K \sim Binomial\left(\frac{N}{2}, \frac{n}{N}\right)$$
$$L \sim Binomial\left(\frac{N}{2}, \frac{m}{N}\right) \tag{3}$$

Should the half of the total sample size $N/2$ be large enough, then the skew of the distribution shall not be too great. In such case a reasonable approximation to the Binomial($N/2$, n/N) is given by the normal distribution:

$$K \sim Normal\left(\frac{N}{2} * \frac{n}{N}, \sqrt{\frac{N}{2} * \frac{n}{N} * \left(1 - \frac{n}{N}\right)}\right) \tag{4}$$

This basic approximation can be improved in a simple way by using a suitable continuity correction. The basic approximation generally improves as *N/2* increases (to at least 20) and is better when *p* is not near to 0 or 1.

The normal approximation can be applied, where the following criteria (5 and 6) are met:[1]

$$\frac{N}{2} * \frac{n}{N} > 5$$
$$\frac{N}{2} * \left(1 - \frac{n}{N}\right) > 5 \tag{5}$$

$$\left|\left(\frac{1}{\sqrt{\frac{N}{2}}}\right) * \left(\sqrt{\frac{1 - \frac{n}{N}}{\frac{n}{N}}} - \sqrt{\frac{\frac{n}{N}}{1 - \frac{n}{N}}}\right)\right| < 0{,}3 \tag{6}$$

4. Hypotheses

A) Two-Tailed Test
H0: *X* and *Y* are identically distributed except for possibly different means
H1: $Var_X \neq Var_Y$
Reject the null hypothesis H0 at the level of significance of *α*, if *K* is less than *α/2* quantile or greater than *1-α/2* quantile of the binomial distribution (or its normal approximation) with the parameters as specified above.

B) Lower-Tailed Test
H0: *X* and *Y* are identically distributed except for possibly different means
H1: $Var_X < Var_Y$
Reject the null hypothesis H0 at the level of significance of *α*, if *K* is less than *α* quantile of the binomial distribution with the parameters as specified above.

C) Upper-Tailed Test
H0: *X* and *Y* are identically distributed except for possibly different means
H1: $Var_X > Var_Y$
Reject the null hypothesis H0 at the level of significance of *α*, if *K* is more than *1-α* quantile of the binomial distribution with the parameters as specified above.

5. Theory

Whenever *X* and *Y* are independent and identically distributed except for having different means, every assignment of ranks to *R* that emerged as union from the sets *K* and *L* is equally likely and the same count of *x* and *y* in either half of it can be excepted depending solely upon the parameters *n/N* and *m/N*.

6. Simulation

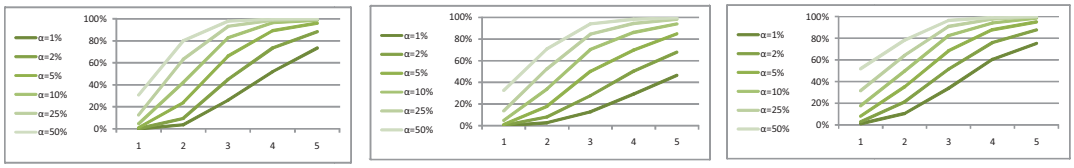
We compared our test with the notorious F-Test of equal variances between two samples and the more robust Brown-Forsythe test.

[1] [2] p. 130

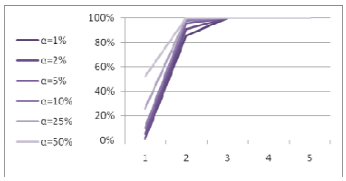
We conducted three sets of experiments to compare the power of our test (listed as "My test" in the tables), the F-test and Brown-Forsythe test as proportions of correctly rejected null hypotheses while alternating both level of significance and the effect size.

In these experiments we worked with equal sample sizes of $n = m = 30$. The effect size on the horizontal axes was expressed as the proportion of standard deviations of the two populations having been compared (1 was for equal variance). The level of significance was from 1% up to 50%. In the first experiment we simulated standard normal distribution. In the second experiment we simulated heavy tails situation using t-distribution with one degree of freedom. In the third experiment we simulated skew by using log-normal distribution corresponding to standard normal distribution from our first experiment. In all three experiments we compared results for our test (green) with that of the F-test (purple) and Brown-Forsythe test (red).

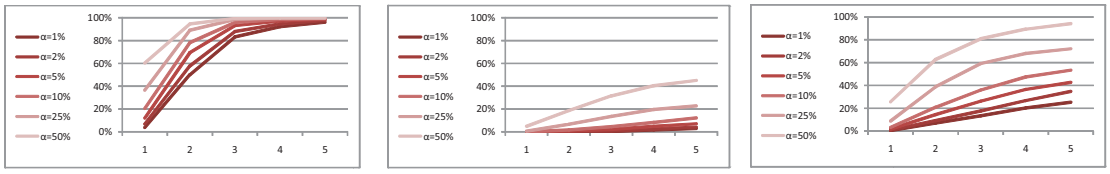
Tab.1 – Our test: Normal, Heavy-Tails and Skewed Distributions



Tab. 2 - F-test: Normal Distribution



Tab. 3 - B-F test: Normal, Heavy-Tails and Skewed Distributions



As opposed to the previous set of experiments, here we examined only effect sizes 1 and 2 (where 1 = equal variances, the null hypothesis is true) and varied the sample sizes (from 30 to 1000) to see, how the test would behave with increasing number of data. In this case we only worked with normal and skewed distributions. The horizontal axes represent the level of significance.

Tab.4 – Varying Sample Size: Normal Distribution

NORMAL DISTRIBUTION

EQUAL SAMPLE SIZES

EQUAL DISPERSION

	MY TEST							F-TEST							BROWN-FORSYTHE TEST						
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
	30	0,1%	0,3%	1,6%	5,4%	15,0%	32,8%	30	1,0%	2,2%	6,1%	10,8%	25,1%	50,1%	30	4,3%	6,5%	13,6%	20,4%	36,8%	58,9%
	50	0,1%	0,1%	0,8%	2,0%	11,2%	42,3%	50	0,8%	1,7%	5,4%	10,7%	26,4%	49,4%	50	4,7%	7,0%	13,1%	20,5%	39,7%	61,8%
n1=n2	100	0,1%	0,3%	0,8%	2,7%	9,6%	40,3%	100	1,3%	2,4%	5,4%	10,9%	25,5%	50,2%	100	4,9%	7,0%	13,5%	22,2%	39,2%	61,8%
	300	0,0%	0,1%	0,5%	2,2%	10,2%	34,3%	300	1,2%	2,0%	5,1%	9,8%	24,4%	50,8%	300	5,0%	8,4%	14,0%	21,8%	39,7%	62,7%
	1000	0,1%	0,1%	0,5%	2,0%	11,6%	35,7%	1000	1,2%	2,5%	5,4%	10,4%	24,2%	49,7%	1000	4,9%	7,2%	13,2%	22,0%	28,5%	61,4%

UNEQUAL DISPERSION

	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
	30	0,4%	2,6%	8,1%	19,1%	38,1%	60,4%	30	86,6%	90,8%	95,7%	97,8%	99,4%	99,9%	30	51,0%	57,9%	69,7%	79,3%	88,3%	95,2%
	50	1,9%	5,1%	12,8%	23,5%	54,2%	82,5%	50	98,7%	99,2%	99,8%	100,0%	100,0%	100,0%	50	75,8%	81,5%	88,7%	92,4%	96,9%	99,0%
n1=n2	100	12,1%	18,7%	38,8%	61,4%	80,4%	96,7%	100	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100	96,9%	98,3%	99,2%	99,7%	99,7%	100,0%
	300	83,5%	90,2%	96,3%	99,0%	99,8%	100,0%	300	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	300	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%
	1000	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	1000	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	1000	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Tab. 5 - Varying Sample Size: Skewed Distribution

SKEWED DISTRIBUTION (Log-Normal)
EQUAL SAMPLE SIZES
EQUAL DISPERSION

	MY TEST							F-TEST							BROWN-FORSYTHE TEST						
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
n1=n2	30	1,2%	4,1%	8,8%	17,5%	31,8%	49,5%	30	41,6%	46,8%	54,1%	61,1%	72,9%	84,7%	30	0,4%	0,8%	1,4%	2,8%	8,2%	26,1%
	50	1,9%	3,2%	7,0%	12,2%	30,9%	61,3%	50	43,8%	49,1%	55,9%	62,4%	73,6%	83,9%	50	0,7%	1,4%	3,4%	6,8%	18,3%	41,4%
	100	2,8%	4,1%	8,6%	16,6%	29,6%	61,4%	100	51,4%	54,8%	61,9%	68,3%	77,2%	86,2%	100	0,5%	0,8%	2,2%	4,2%	12,7%	35,0%
	300	3,8%	6,2%	10,6%	20,0%	32,4%	57,8%	300	52,0%	57,2%	61,6%	67,8%	77,0%	86,6%	300	0,2%	0,2%	0,4%	1,3%	5,8%	22,8%
	1000	2,4%	4,0%	9,3%	17,8%	29,9%	55,4%	1000	65,0%	68,5%	73,5%	78,5%	83,0%	90,5%	1000	0,0%	0,0%	0,0%	0,0%	3,0%	15,0%
UNEQUAL DISPERSION																					
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
n1=n2	30	10,5%	21,0%	34,5%	53,1%	68,2%	81,3%	30							30	5,1%	6,6%	11,4%	19,1%	35,2%	61,4%
	50	25,3%	37,2%	49,7%	62,3%	82,1%	94,5%	50							50	5,6%	9,0%	15,7%	25,1%	49,1%	76,2%
	100	58,4%	66,7%	80,4%	90,0%	95,4%	99,1%	100			N/A				100	9,5%	15,0%	26,9%	41,4%	69,6%	92,4%
	300	99,6%	99,6%	99,8%	100,0%	100,0%	100,0%	300							300	39,2%	54,0%	73,4%	87,6%	98,0%	99,6%
	1000	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	1000							1000	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

We also simulated non-equal sample sizes in this set of experiments (the difference between samples was expressed as % of combined sample size, that is *Abs(n-m)/(n+m)%*, where *n+m* was constant 100 (so 10% difference stood for sample sizes 45 and 55, and so on). In these circumstances we worked with normal and skewed distribution (the skewed again represented by log-normal distribution) and compared results of our test with that of the F-test and Brown-Forsythe test.

Tab.6 – Varying Inequality between the Sample Sizes: Normal Distribution

NORMAL DISTRIBUTION
UNEQUAL SAMPLE SIZES
EQUAL DISPERSION

	MY TEST							F-TEST							BROWN-FORSYTHE TEST						
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
Abs(n1-n2) / (n1+n2=100) %	5%	0,0%	0,0%	0,6%	2,3%	6,9%	30,4%	5%	1,0%	2,1%	5,3%	10,7%	25,7%	52,9%	5%	4,4%	6,4%	12,9%	21,2%	39,4%	63,0%
	10%	0,0%	0,1%	0,5%	1,7%	10,6%	43,0%	10%	1,2%	2,1%	4,7%	9,9%	24,1%	50,3%	10%	5,2%	7,7%	13,8%	20,7%	38,2%	61,1%
	25%	0,0%	0,0%	0,5%	2,1%	11,0%	44,8%	25%	0,9%	1,6%	4,9%	10,1%	24,4%	48,4%	25%	3,7%	5,9%	11,6%	18,9%	36,9%	59,4%
	50%	0,0%	0,1%	0,3%	1,1%	9,9%	41,2%	50%	1,2%	2,2%	5,1%	10,7%	25,3%	50,4%	50%	3,9%	6,6%	12,1%	19,8%	38,1%	60,6%
	100%	0,1%	0,2%	0,7%	2,4%	9,8%	25,1%	100%	1,0%	2,1%	5,1%	9,6%	26,7%	51,8%	100%	4,0%	6,1%	12,0%	21,2%	37,5%	60,5%
UNEQUAL DISPERSION																					
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
Abs(n1-n2) / (n1+n2=100) %	5%	21,3%	28,4%	60,7%	73,7%	84,5%	96,7%	5%	98,8%	99,4%	99,9%	99,9%	100,0%	100,0%	5%	74,6%	80,0%	87,6%	91,7%	97,3%	98,8%
	10%	18,7%	32,6%	49,2%	66,8%	89,3%	98,2%	10%	99,0%	99,6%	99,8%	99,9%	100,0%	100,0%	10%	74,3%	80,9%	87,6%	92,5%	96,3%	99,0%
	25%	17,8%	32,5%	47,6%	64,6%	87,5%	97,3%	25%	98,3%	99,0%	99,5%	99,8%	99,9%	100,0%	25%	75,0%	80,9%	89,2%	93,3%	96,8%	99,3%
	50%	13,0%	25,9%	42,6%	60,3%	86,9%	97,7%	50%	98,2%	99,2%	99,9%	99,9%	100,0%	100,0%	50%	71,0%	77,6%	85,8%	91,8%	96,3%	98,8%
	100%	9,6%	19,4%	35,2%	53,6%	79,8%	90,1%	100%	93,0%	95,5%	98,0%	99,3%	99,8%	99,9%	100%	50,4%	52,4%	56,2%	60,3%	69,4%	81,8%

Tab.7 – Varying Inequality between the Sample Sizes: Skewed Distribution

SKEWED DISTRIBUTION (Log-Normal)
UNEQUAL SAMPLE SIZES
EQUAL DISPERSION

	MY TEST							F-TEST							BROWN-FORSYTHE TEST						
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
Abs(n1-n2) / (n1+n2=100) %	5%	2,4%	4,4%	12,3%	20,0%	29,3%	55,5%	5%	45,3%	49,8%	57,0%	63,0%	72,4%	82,6%	5%	0,0%	0,3%	0,4%	1,4%	6,4%	23,6%
	10%	1,8%	3,5%	8,9%	16,0%	30,5%	51,4%	10%	42,8%	47,7%	55,0%	61,3%	73,4%	84,5%	10%	0,1%	0,3%	0,4%	1,1%	6,0%	22,1%
	25%	2,0%	4,2%	7,6%	14,3%	29,0%	58,7%	25%	44,3%	48,3%	55,3%	61,1%	72,7%	84,3%	25%	0,0%	0,2%	0,6%	2,0%	6,3%	23,2%
	50%	1,5%	2,9%	6,6%	11,8%	31,1%	61,1%	50%	43,6%	48,2%	55,3%	61,9%	73,0%	84,9%	50%	0,3%	0,4%	0,7%	1,5%	6,8%	23,5%
	100%	1,3%	3,1%	6,9%	13,9%	29,5%	47,1%	100%	41,3%	46,6%	53,3%	60,2%	72,0%	84,0%	100%	0,2%	0,4%	1,3%	2,4%	6,7%	23,7%
UNEQUAL DISPERSION																					
	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%	α=	1%	2%	5%	10%	25%	50%
Abs(n1-n2) / (n1+n2=100) %	5%	25,7%	29,6%	54,9%	66,2%	75,8%	90,2%	5%							5%	6,6%	9,7%	16,3%	25,1%	47,8%	75,9%
	10%	23,5%	34,8%	46,7%	59,4%	80,0%	92,7%	10%							10%	6,3%	9,6%	17,1%	25,7%	46,9%	74,1%
	25%	23,3%	35,1%	47,2%	60,0%	81,7%	93,1%	25%			N/A				25%	5,2%	8,1%	13,9%	22,4%	45,7%	74,8%
	50%	19,1%	28,9%	42,1%	53,9%	76,0%	91,3%	50%							50%	6,6%	9,4%	16,2%	24,4%	45,3%	73,7%
	100%	14,9%	24,6%	36,2%	49,5%	68,1%	80,1%	100%							100%	7,5%	10,0%	15,1%	21,9%	39,7%	67,5%

7. Conclusions and Recommendations

The series of experiments have shown that our test is lacking statistical power in most of the situations where the normality holds compared to the widely used parametric alternative. However, this is not the whole story .The findings can be summarized in the following:

1. **<u>When normality holds</u>**
   a) The Exact test has the least power
   b) The Exact test is the most conservative
   c) F-test is always preferable to B-F test
   d) The power of all tests increases with greater sample sizes. The power of the Exact test at size 300 roughly equals to that of F-test at size 30 and B-F at size less than 100

2. **<u>With non-normal data</u>** the Exact test has more power than B-F at lower sample sizes and with more unequal sample sizes

3. Because of the exact nature of our test it is always **preferable choice with small samples**, unless it is too small
4. The Exact test is **computationally less intensive** as it:
   a) Does work with medians and not means
   b) Does not calculate variances but absolute deviations
   c) Does calculate binomial expansion, unless replaced by Normal approximation

With regard to what's been concluded we propose the following decision tree to facilitate dealing with the three considered test in practical applications:

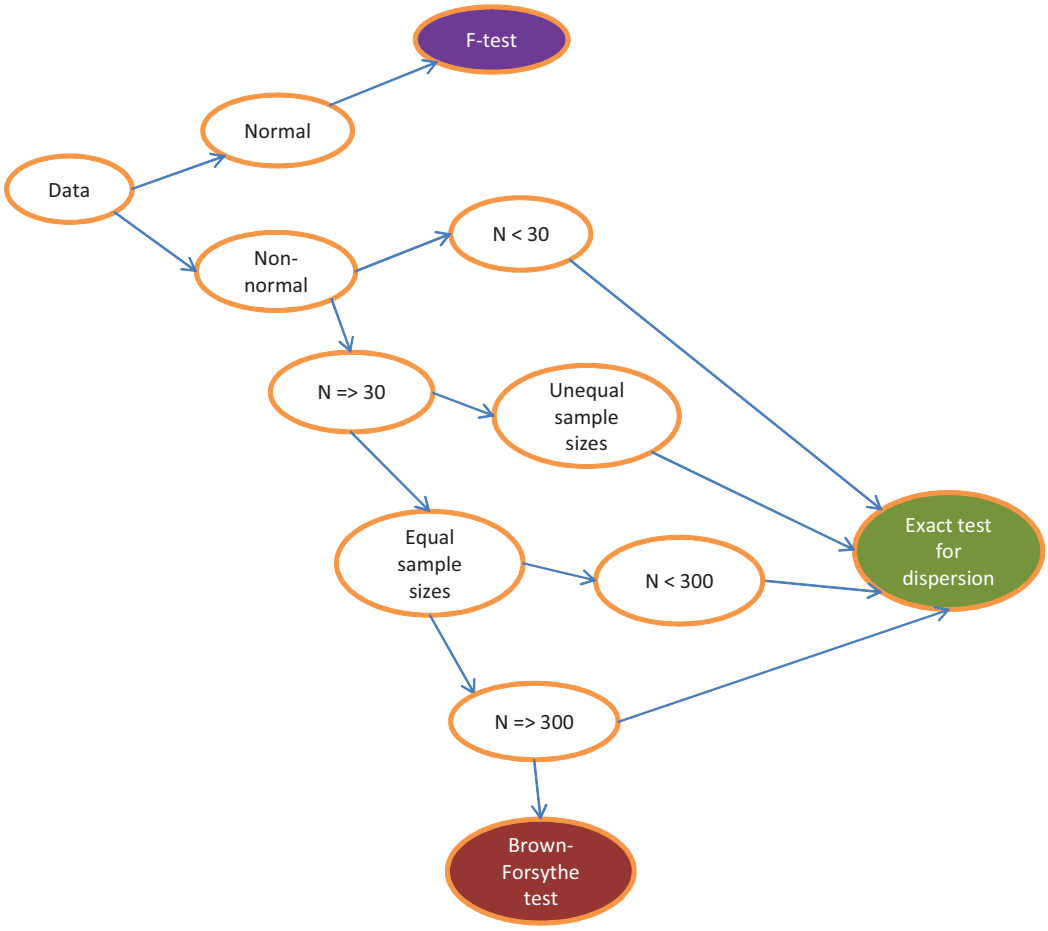


Fig. 1 - Decision Tree

8. Possible Issues

1. More than two samples

To test whether counts K_1, K_2, K_3, ... within the critical part of the joined set R be proportionate to their relative sample sizes. If not, than the homogeny of the dispersion will be put in question. This whole issue calls for a generalization.

2. Ties in data

The test assumption states that the data is continuous and no real ties occur. However, would the test be inapplicable in a discrete case with ties? That has not been answered, yet.

References

1. BOX, G., E., P., HUNTER, W., G., HUNTER, J., S.: Statistics for Experiments, John Wiley & Sons, 1978, ISBN-10: 0471093157
2. RÉNYI, A.: Probability Theory, Dover Publications, Incorporated, 2007, ISBN 0486458679

Address of the Author

Illovský Michal, Mgr.

Slovenská Technická Univerzita v Bratislave

Radlinského 11, 813 68 Bratislava

michal.illovsky@stuba.sk

michal.illovsky@gmail.com

Measurement of portfolio credit risk according to CreditMetrics model (Meranie portfóliových kreditných rizík podľa modelu CreditMetrics)

Jozef Jackuliak

Abstract: This article deals with an issue of portfolio credit risk measurement. The main goal is to highlight the importance of measurement credit risk according to empirical researches as well as to practical example. Core of the work consists of depiction of credit risk measurement methodology based on CreditMetrics model and application of this model on illustrative portfolio. CreditMetrics methodology is a tool to measure credit risk in a portfolio due to changes in the value of debt and analyses the migration of assets within a defined rating system. The article analyses credit risk measurement of model example and compares it to the value at risk for various confidence levels. Based on obtained distribution of portfolio values for one year risk horizon the paper refers to a "non-normality" of distribution of the investigated portfolio which is caused by a credit event.

Abstrakt: Predkladaný príspevok sa zaoberá problematikou portfóliového kreditného rizika. Jeho cieľom je poukázať na dôležitosť merania daného rizika na základe empirických štúdií ako aj na základe praktického výpočtu. Jadro príspevku pozostáva z popisu metodiky výpočtu kreditného rizika podľa modelu CreditMetrics a jeho aplikácie na praktickom príklade. Metodológia CreditMetrics je nástroj na meranie kreditného rizika portfólia v dôsledku zmeny hodnoty dlhu, pričom analyzuje migráciu aktív v rámci definovaného ratingového systému. Príspevok analyzuje kreditné riziko modelového portfólia a porovnáva ho s hodnotou v riziku pre rôzne hladiny spoľahlivosti. Na základe dosiahnutej distribúcie hodnôt portfólia v rizikovom horizonte jeden rok poukazuje na „nenormalitu“ rozdelenia výnosov skúmaného portfólia, spôsobenú vznikom kreditných udalostí/zlyhaní.

Key words: CreditRisk model, Monte Carlo, VaR

Kľúčové slová: CreditRisk model, Monte Carlo , VaR

JEL classification: C63, G10, G17

Úvod

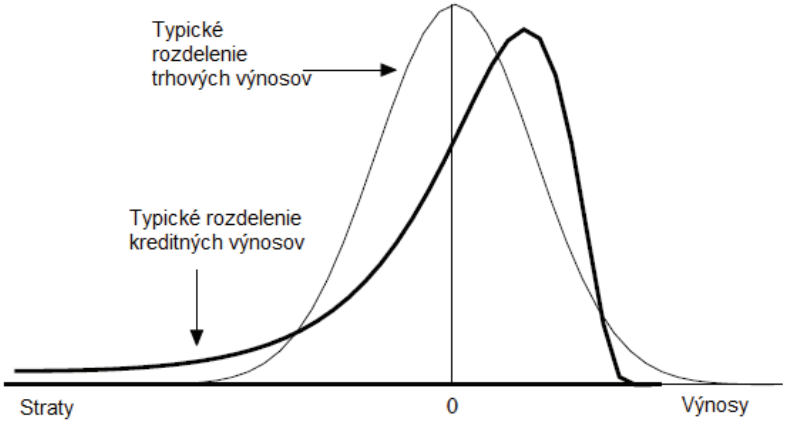
Portfóliové kreditné riziká sú objektom skúmania prednostne vo finančných inštitúciách (bankách), pre ktoré je charakteristické, že vo svojich účtovných knihách spravujú úverové portfóliá. Kreditné riziká predstavujú viac, ako len riziko straty z dôvodu zlyhania protistrany. Do kreditného rizika sa zaraďuje aj strata vzniknutá v dôsledku zmeny ratingu dlžníka. Prvé inštitúcie, zaoberajúce sa problematikou kreditných rizík, boli finančné inštitúcie, ktoré financovali spotrebiteľov a domácnosti. Neskôr sa použiteľnosť týchto modelov rozšírila aj na korporácie a malé podniky. Pri meraní kreditných rizík vychádzame z hlavných atribútov. Sú to definícia udalosti zlyhania, expozícia v prípade zlyhania, pravdepodobnosť zlyhania v čase t a miera návratnosti alebo ozdravenia.

V predkladanom príspevku sa zameriame na meranie kreditného rizika vyplývajúceho z držania aktív počas horizontu jedného roka. Na základe metodológie CreditMetrics si vysvetlíme spôsob merania portfóliového kreditného rizika a následne ukážeme jej použiteľnosť na prípadovej štúdii.

1. Portfóliové meranie kreditných rizík podľa modelu CreditMetrics.

Model CreditMetrics bol prvýkrát publikovaný v roku 1997 J.P Morganom ako nástroj na meranie kreditného rizika portfólia v dôsledku zmeny hodnoty dlhu. Zmena hodnoty je spôsobená zmenou kvality dlžníka. Zmena kvality dlžníka sa však nemení len po spozorovaní udalosti zlyhania. Podstatou modelu CreditMetrics je analýza migrácie aktív v rámci definovaného ratingového systému [1]. Zmeny hodnoty dlhu nie sú teda vyvolané len v prípade udalosti zlyhania, ale aj v prípade zmeny kredibility dlžníka. CreditMetrics meria riziko pomocou metódy VaR (Value at Risk – hodnota v riziku) v rizikovom horizonte 1 rok.

Problém v modelovaní kreditného rizika spočíva v nesymetrickosti kreditných výnosov. Na rozdiel od trhovo obchodovateľných nástrojov, ktorých výnosy môžu mať Gaussovo rozdelenie, pri kreditných výnosoch to neplatí. Rozdelenie úverových výnosov je zľava zošikmené s dlhým chvostom (viď obrázok č.1). Tento dlhý chvost na ľavej strane je tvorený stavmi zlyhania.



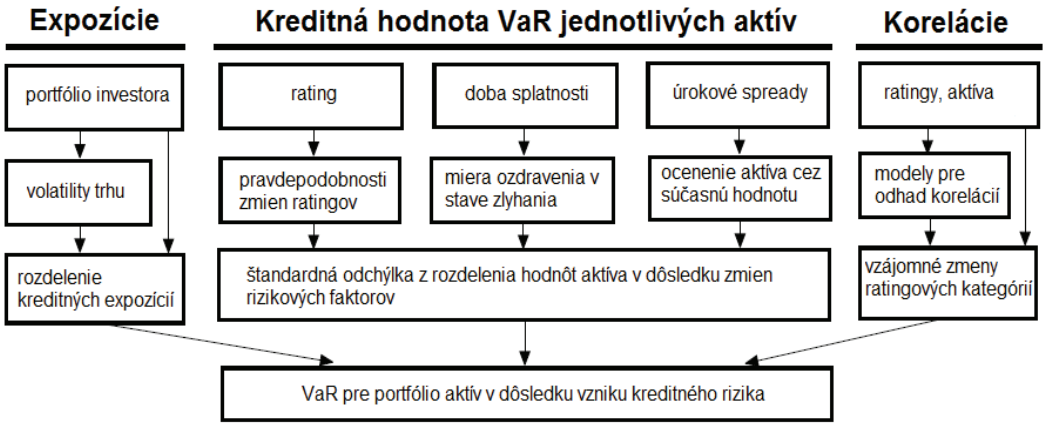
Obrázok 1: Porovnanie trhového rozdelenia výnosov s kreditným rozdelením *(Zdroj: [2])*

Hlavnou myšlienkou merania kreditných rizík, podľa tohto modelu, je zistiť hodnoty pravdepodobnosti zlyhania a určiť celý systém kvalitatívnej (ratingovej) migrácie aktív v súvislosti so zmenou kreditného rizika. Pri meraní kreditného rizika podľa CreditMetrics vychádzame z nasledujúcich predpokladov [2]:

1. *„Všetky aktíva klasifikované v rovnakej rizikovej kategórii majú rovnakú mieru pravdepodobnosti zlyhania a rovnakú výnosovú krivku úrokového rozpätia. Podobne to platí aj pre pravdepodobnosti migrácie aktíva do iných rizikových kategórií. Rating sa okamžite zmení v prípade, že sa zmení pravdepodobnosť zlyhania.*
2. *Súčasná miera pravdepodobnosti zlyhania sa rovná historicky vypočítanej priemernej hodnote rozdelenia pravdepodobností podľa jednotlivých rizikových kategórií.“*

Modelovanie kreditného rizika pomocou CreditMetrics pozostáva z troch častí, ktoré môžeme nazvať *expozícia, kreditná hodnota VaR jednotlivých aktív a korelácie* (viď obrázok č. 2). Postup pri modelovaní kreditného rizika je daný krokmi zľava doprava. ***Prvým krokom*** je dekompozícia portfólia podľa druhu aktív (dlhopisy, úvery, swapy...). Toto uskutočňujeme podľa expozície a volatility trhu na objemy expozícií. ***Druhým krokom*** je definovanie ratingového systému aktív a stanovenie ich pravdepodobností migrácie medzi jednotlivými rizikovými kategóriami za určité obdobie. Toto sa uskutočňuje pomocou tzv. matice

pravdepodobností prechodu. Jednotlivým aktívam sa podľa ich kvality priradí rating a na základe matice pravdepodobností prechodu sa analyzujú jednotlivé stavy rizikovosti ako aj stav zlyhania. ***Tretím*** a zároveň posledným krokom je odhad koeficientov korelácie pohybov cien aktív v portfóliu. Tieto zmeny sú odvodené od zmien rizikových faktorov jednotlivých aktív. Na výpočet korelácií aktív používame simulácie spoločných pohybov. Vychádzame z modelovania viacfaktorového rozdelenia za predpokladu, že rozdelenie výnosov je predmetom normálneho rozdelenia. Pri simuláciách CreditMetrics používa metódu Monte Carlo, ktorá generuje očakávané výnosy aktíva na základe viacfaktorového rozdelenia s odhadnutými hodnotami korelácií. Po simulácií sa každý scenár pridelí určitej rizikovej kategórii, a tým sa dosiahne rozdelenie strát podľa rizikovosti nastania určitého vzťahu [1].



Obrázok 2: Postup pri modelovaní kreditného rizika pomocou CreditMetrics *(Zdroj: [2])*

2. Prípadová štúdia

V tejto časti si metodológiu CreditMetrics uvedieme na praktickom príklade. Pre naše portfólio sme si definovali 10 aktív (dlhopisov). Pre každé aktívum sme si zistili nominálnu hodnotu, aktuálnu trhovú hodnotu (zo dňa 20.4.2012), prislúchajúci kvalitatívny rating podľa agentúry Standard and Poor's a kupón vyplácaný každoročne z nominálnej hodnoty. Definovali sme si nákupné množstvá, na základe ktorých po vynásobení trhovou hodnotou sme dostali celkovú súčasnú hodnotu portfólia 130 883 195 €

Tabuľka 1: Zloženie portfólia *(Zdroj: vlastné spracovanie)*

Číslo aktíva	Názov	Rating	Nominálna hodnota	Kupón	m*	Počet kusov	Súčasná hodnota	Celková súčasná hodnota
1	BUNDES0BL	AAA	100,00 €	0,75%	5	100000	100,55 €	10 054 500,00 €
2	REP OF AUSTRIA	AA+	100,00 €	4,30%	2	200000	107,92 €	21 584 000,00 €
3	CZECH GB	AA	100,00 €	4,00%	5	250000	107,96 €	26 990 000,00 €
4	SLOVGB	A	100,00 €	3,50%	4	350000	104,44 €	36 553 650,00 €
5	SLOREP GB	A	100,00 €	4,38%	2	10000	104,77 €	1 047 740,00 €
6	HUNGARY GVT	BB+	100,00 €	8,00%	3	80000	98,63 €	7 890 240,00 €
7	REP OF POLAND	A-	100,00 €	3,63%	4	15000	104,92 €	1 573 815,00 €
8	COMERZBANK AG	BBB	100,00 €	6,25%	2	100000	103,75 €	10 375 000,00 €
9	ALPHA CREDIT	CCC	100,00 €	6,00%	2	50000	83,63 €	4 181 250,00 €
10	BMW FINANCE	A	100,00 €	4,00%	2	100000	106,33 €	10 633 000,00 €
	*m= maturita						**SPOLU**	**130 883 195,00 €**

Ďalej, potrebujeme zistiť prechodové pravdepodobnosti zmeny ratingov daných aktív. Pri tomto kroku nám bola ústretová ratingová agentúra Standard and Poor's, ktorá nám poskytla aktualizovanú prechodovú maticu platnú pre rok 2011.

Okrem zistenia pravdepodobností prechodu do iného ratingu potrebujeme zistiť senioritu (nadradenosť voči ostatným dlhopisom) a zaistenie jednotlivých aktív. V prípade zlyhania dlžníka je dôležité správne odhadnúť mieru ozdravenia daného aktíva, ktorá vychádza práve z uvedenej nadradenosti a zaistenia aktíva. Pre tento krok sme použili štúdiu, ktorú vykonali autori Saunders a Allenc na základe zozbieraných dát za roky 1978 – 2009.

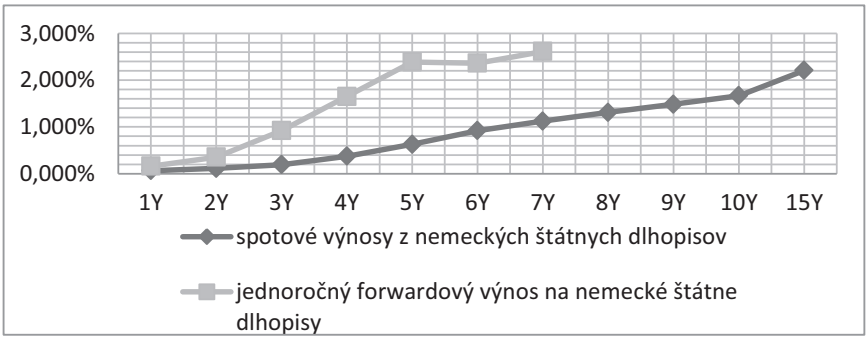
Tabuľka 2: Vážené priemery mier ozdravenia aktív *(Zdroj: [3])*

Aktívum	nadradené- zaistené	nadradené- nezaistené	podriadené - nezaistené
Vážený priemer	57,55 €	36,24 €	30,85 €

Po aplikovaní týchto hodnôt, ďalším dôležitým krokom pri meraní portfóliového kreditného rizika je výpočet hodnoty na konci rizikového horizontu (v našom prípade o jeden rok), pre každú ratingovú kategóriu. Pre tento výpočet použijeme rovnicu pre výpočet súčasnej hodnoty budúcich peňažných tokov do splatnosti, diskontovaných požadovanými forwardovými výnosmi pre konkrétny rating.

$$V = c + \frac{C1}{1+f1} + \frac{C2}{1+f2\ ^{2}} + .. + \frac{Cn}{1+fn\ ^{n}} \tag{1}$$

Tento krok je potrebné urobiť pre každé aktívum a pre každú možnú ratingovú kategóriu. Jednoročnú forwardovú výnosovú krivku sme si odvodili od spotovej krivky európskeho benchmarku (nemecké bezkupónové štátne dlhopisy). Ostatné forwardové krivky vypočítame pripočítaním kreditných spreadov medzi spotovým výnosom benchmarku a spotovým výnosom predstaviteľa danej ratingovej kategórie.



Graf 1: Spotové a jednoročné forwardové výnosy z nemeckých štátnych bezkupónových dlhopisov
(Zdroj: vlastné spracovanie)

Pri odhade korelácií aktív v našom portfóliu sme vychádzali z historických dát ročných výnosov dlhopisov aktualizovaných na dennej báze. Naše aktíva v portfóliu sú väčšinou pozitívne korelované, až na aktívum číslo 6 (maďarské štátne dlhopisy), ktoré zdieľajú so všetkými ostatnými aktívami negatívnu závislosť.

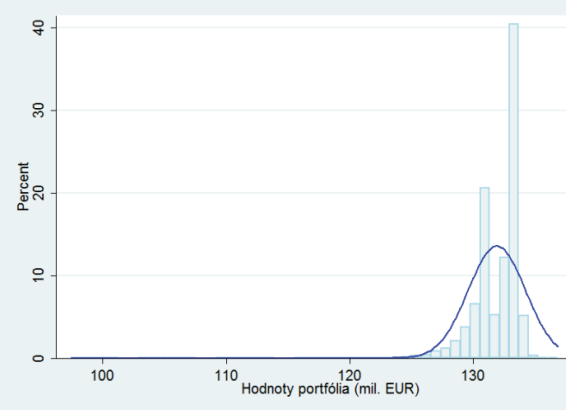
Na základe pravdepodobností aktuálnej prechodovej matice sme si vytvorili tzv. „thresholds" (hranice výnosov - Z) pre dlžníkov v našom portfóliu. Za predpokladu, že naše výnosy sú normálne rozdelené, potrebné hraničné výnosy, napr. pre rating CCC, sme vypočítali nasledovne:

$$P\ \ CCC\ = P\ \ Z_{Def} < V < Z_{CCC}\ = \phi\ \ \frac{Z_{CCC}}{\sigma}\ - \phi\ \ \frac{Z_{Def}}{\sigma}\ , \qquad (2)$$

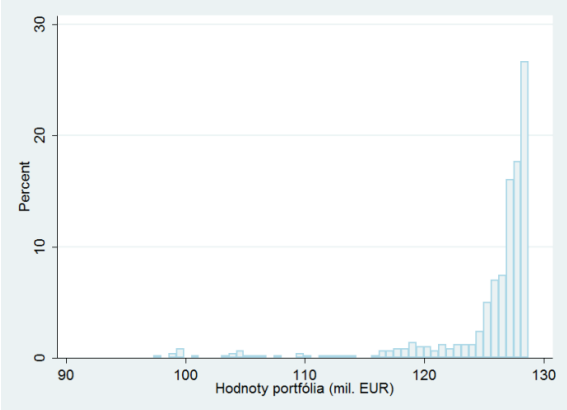
kde P predstavuje pravdepodobnosť, V výnosy aktív a ϕ kumulatívnu distribúciu pre štandardné normálne rozdelenie. To znamená, že v prípade ak výnos nášho desiateho aktíva s ratingom CCC bude väčší ako hraničná hodnota výnosu pre default Z_{Def}, ale menší ako hraničná hodnota výnosu pre Z_{CCC}, naše aktívum zostane v rovnakej ratingovej kategórií. Rovnako sme pristupovali aj pri výpočte hraníc pre ostatné ratingy. Pomocou simulácie sme vygenerovali 10000 nezávislých náhodných výnosov z normálneho rozdelenia N(0,1). Pre určenie korelovaných výnosov sme použili Choleského rozklad korelačnej matice [4]. Korelačnú maticu sme vynásobili s nezávislými výnosmi a získali sme závislé výnosy pre naše portfólio, ktoré spadali do prislúchajúcich ratingov. Vypočítali sme hodnoty portfólia vzhľadom na vygenerované scenáre a získali sme 10000 rôznych hodnôt skúmaného portfólia v horizonte jedného roku.

3. Výsledky

Po vypočítaní hodnôt portfólia pre 10000 scenárov sme získali rozdelenie výnosov v priebehu jedného roku (pozri graf 2). Na rozdelení nasimulovaných budúcich hodnôt portfólia môžeme vidieť typické zošikmenie, ktoré je charakteristické pre rozdelenie kreditných portfólií. Zošikmenie vytvárajú vygenerované náhodné zlyhania aktív, ako aj zníženia ratingových kategórií. Ak porovnáme naše rozdelenie hodnôt s normálnym (viď plnú čiaru v grafe) vidíme, že sa výrazne odlišuje. S najväčšou pravdepodobnosťou (40 %) sa naše portfólio na konci jednoročného horizontu pohybovalo okolo hodnoty 133 mil. EUR. S pravdepodobnosťou 20% u žiadneho dlžníka nedošlo k zníženiu ratingu a portfólio nadobudlo hodnotu približne vypočítanej strednej hodnoty. Pozitívnou charakteristikou opisovaného portfólia je, že vo viac ako v polovici prípadov sme nezaznamenali výraznejšiu kreditnú udalosť a portfólio zaznamenalo zhodnotenie oproti súčasnej trhovej hodnote.



Graf 2: Rozdelenie budúcich hodnôt portfólia
(Zdroj: Vlastné spracovanie)



Graf 3: Piaty percentil rozdelenia hodnôt portfólia
(Zdroj: Vlastné spracovanie)

Graf č. 3 zobrazuje prvých 500 najmenších hodnôt simulácie (piaty percentil). Napriek tomu, že naša najpravdepodobnejšia hodnota bola 133 mil. EUR, niektoré hodnoty portfólia boli pod hranicou 100 mil. EUR, a teda v prípade vzniku podobného scenára naše portfólio stratí viac ako 30 mil. EUR, čo predstavuje stratu 23%. Pravdepodobnosť vzniku tohto scenára je však veľmi nízka a pre lepší odhad kreditného rizika portfólia sme si vypočítali základné štatistické ukazovatele. Pri vygenerovaní 10000 budúcich hodnôt sa stredná hodnota rovnala 131 899 616 EUR. Štandardná odchýlka sa rovnala 2,3 mil. EUR. Rozptyl medzi

najvyššou a najnižšou dosahovanou hodnotou bol až 39,5 mil. EUR. Tento rozdiel bol oveľa väčšou mierou ťahaný do ľavej strany. To poukazuje na veľké extrémy v rozdelení strát portfólia.

Tabuľka 3: Popisná štatistika *(Zdroj: Vlastné spracovanie)*

Popisná štatistika	
Stredná hodnota	131 899 616,04 €
Štandardná odchýlka	2 320 623,26 €
Max. hodnota	136 873 920,00 €
Min. hodnota	97 382 496,00 €

Stredná hodnota ako aj štandardná odchýlka nie sú však najlepšie meradlá rizika pre podobné portfólia. Rozdelenie strát, resp. výnosov, nespadá do normálneho rozdelenia a hladina spoľahlivosti sa na základe týchto parametrov nedá správne definovať. Hodnota pre piaty percentil predstavuje 128 742 256 EUR. Čo je o 2 140 939 EUR menej ako súčasná hodnota portfólia Pre investora to znamená, že s pravdepodobnosťou 5% bude mať skúmané portfólio v horizonte jedného roku hodnotu nižšiu o viac ako 2,1 mil. EUR. Podobne sme si vypočítali hodnoty nášho portfólia pre ostatné percentily a porovnali sme ich s hodnotami percentilov za predpokladu normálneho rozdelenia výnosov. Hodnoty medzi aktuálnym a normálnym rozdelením sa výrazne odlišujú.

Tabuľka 4: Percentily budúcich hodnôt portfólia *(Zdroj: Vlastné spracovanie)*

Percentil	Aktuálne rozdelenie	Normálne rozdelenie	
	Hodnota portfólia	Rovnica	Hodnota portfólia
99%	134 105 664,00 €	μ + 2,33σ	137 306 668 €
95%	133 728 904,00 €	μ + 1,65σ	135 728 644 €
50%	132 772 360,00 €	μ	131 899 616 €
5%	128 742 256,00 €	μ - 1,65σ	128 070 588 €
2,50%	127 293 038,39 €	μ - 1,96σ	127 351 194 €
1%	124 854 648,00 €	μ - 2,33σ	126 492 564 €
0,50%	119 540 839,51 €	μ - 2,58σ	125 912 408 €

Použitím výsledkov zo scenárov odhadujeme, že s pravdepodobnosťou 1% hodnota portfólia počas budúceho roku klesne na hodnotu rovnú alebo menšiu ako 124,85 mil. EUR. V prípade predpokladu normálneho rozdelenia, hodnota portfólia by klesla len na hodnotu rovnú alebo menšiu 126,58 mil. EUR. Stredná hodnota pri normálnom rozdelení je však nižšia ako pri aktuálnom rozdelení. V prípade 99% hladiny spoľahlivosti je hodnota v riziku počas horizontu jedného roku približne 7 mil. EUR. Pri hladine spoľahlivosti 95% je to 3,2 mil. EUR.

4. Záver

Ak by finančná inštitúcia držala opisované portfólio počas jedného roku, tak na základe výpočtov VaR podľa aktuálneho rozdelenia, na hladine spoľahlivosti 95% by ekonomické kapitálové požiadavky pre neočakávané budúce straty predstavovali približne 3,1 mil. EUR. Na hladine spoľahlivosti 99%, by požiadavky tvorili približne 7 mil. EUR.

Literatúra

[1] SIVÁK, R., GERTLER, Ľ., KOVÁČ, U.: *Riziká a modely vo financiách a v bankovníctve.* Bratislava: Sprint dva, 2010, 2.vydanie. 346 s. ISBN 978-80-89393-44-2

[2] GUPTON, G. M., FINGER, C. C., BHATIA M.: *CreditMetrics – Technical Document.* New York: J.P. Morgan, 1997

[3] SAUNDERS, A., ALLEN*C, L.: Credit Risk Measurement In and Out of the FinancialCrisis.* New Jersey: JohnWiley&Sons, Inc, 2010. 380 s. ISBN 978-0-470-47834-9

[4] BOHDALOVÁ, M. Štatistické metódy vo finančných službách [Dizertačná práca] - Univerzita Komenského v Bratislave., Fakulta Managementu, Katedra informačných systémov. - Školitelia:doc. RNDr. Oľga Nánásiová, Csc., doc. RNDr. Michal Greguš, PhD. - Bratislava 2006

[5] BOHDALOVÁ, M., GREGUŚ, M.:Stochastické analýzy finančných trhov. Univerzita Komenského, Bratislava, 2012, 183 s., ISBN 978-80-223-3318-4

[6]LOFFlER, G., POSCH, P., N.: Credit risk modeling using Excel and VBA.Chichester: JohnWiley&Sons, Ltd, 2007. s. 278 ISBN 978-0-470-03157

[7] BOHDALOVÁ, M.: A comparison of value-at-risk methods for measurement of the financial risk. In: The Proceedings of E-Leader - New York : CASA, 2007. - nestr. [6 s.], E-Leader conference. Praha, 11.-13.6.2007

Adresa autora:

Jozef Jackuliak, Mgr.

Univerzita Komenského - Fakulta managementu

Odbojárov 10 P.O.BOX 95, 820 05 Bratislava

jozef.jackuliak@fm.uniba.sk

Zobrazenie kategoriálnych dát do R
Mapping of categorical data into real number interval

Dušan Janál

Abstract: Cluster analysis is one of the multidimensional statistical methods. Its goal is to partition objects into groups so that objects in one group (cluster) are as much similar as possible and any two objects from two different groups differ from each other as much as possible. The article tries to find a convenient mapping of categorical data into real number interval (i.e, zero-one interval). This mapping should enable the clustering categorical data using methods for numerical data.

Abstrakt: Zhluková analýza je jednou z viacrozmerných štatistických metód, ktorej cieľom je zoskupiť objekty do skupín (zhlukov) tak, aby objekty jednej skupiny boli na seba čo najviac podobné, a zároveň ľubovoľné dva objekty z rôznych skupín maximálne odlišné. Článok sa zaoberá hľadaním vhodného zobrazenia kategoriálnych dát do množiny reálnych čísel, napr. z intervalu (0,1). Toto zobrazenie má umožniť zhlukovanie kategoriálnych dát pomocou zhlukovacích metód pre numerické dáta.

Key words: categorical data, clustering, mapping, dissimilarity measures
Kľúčové slová: kategoriálne dáta, zhlukovanie, zobrazenia, miery nepodobnosti
JEL classification: C650

Úvod

Zhluková analýza je jednou z viacrozmerných štatistických metód, ktorej cieľom je zoskupiť objekty do skupín (zhlukov) na základe ich podobnosti. Každá zhlukovacia metóda používa algoritmus, ktorý ovplyvňuje výsledky zhlukovej analýzy. Uvedieme základné rozdelenie algoritmov.

Hierarchické algoritmy postupne delia jednu množinu obsahujúcu *n* objektov postupne na *n* jednoprvkových množín (divizívne), alebo tieto jednoprvkové množiny postupne spájajú až do vytvorenia jednej množiny s *n* objektmi (aglomeratívne), napr. LIMBO [1].

Nehierarchické algoritmy, medzi ktoré patrí napr. k-modusový algoritmus, delia tú istú množinu objektov na vopred stanovený počet zhlukov pri minimalizácii druhej mocniny vzdialenosti od centra zhluku. Centrom zhluku je napr. v prípade k-modusového algoritmu modus daného zhluku.

Výsledkom každej zhlukovej analýzy je delenie $\boldsymbol{C}$ zhlukovaných objektov do zhlukov $C_1, C_2, ..., C_k$. Zhlukovať môžeme, okrem objektov o, aj atribúty $\boldsymbol{A}$ alebo atribútové hodnoty $\boldsymbol{a}$. Atribúty predstavujú sledovanú vlastnosť. V prípade kategoriálnych dát môžu nadobudnúť konečné, pri numerických dátach nekonečné množstvo atribútových hodnôt. Ak by zobrazenie atribútových hodnôt do reálnych čísel bolo náhodné, nemá pre zhlukovanie žiadny význam. Rovnako to platí aj v prípade, ak zobrazí všetky kategoriálne atribútového hodnoty do jednej numerickej hodnoty.

Každý objekt môžeme zapísať ako *m*-rozmerný vektor $\boldsymbol{o} = (o_1, o_2, ..., o_m)$, kde o_j je prvok z množiny atribútových hodnôt $a_{j,i}$ pre *j*-tu sledovanú vlastnosť, atribút $\boldsymbol{A_j}$. Objekt zobrazený do R môžeme zapísať ako $\boldsymbol{o_R} = (x_1, x_2, ..., x_m)$, kde $o_j \rightarrow x_j$ a x_j je z R.

V článku sa snažíme nájsť jedno také zobrazenie kategoriálnych dát do množiny reálnych čísel z intervalu (0,1), ktoré by umožnilo zhlukovanie kategoriálnych dát pomocou zhlukovacích metód pre numerické dáta a zároveň zachovalo vzťahy medzi atribútovými hodnotami. Aby sme sa vyhli zacykleniu, musíme si vopred stanoviť zobrazenia dvoch

atribútových hodnôt. Použitie nenulovej korekcie λ nám umožní výpočet všetkých ostatných zobrazení, pretože odstráni vplyv nulovej početnosti $n_{j,i}$, ak sa niektorá atribútová hodnota v dátach nevyskytne. Zmena λ ovplyvňuje výsledné zobrazenia (Obr. 1), preto si ju tiež pevne stanovíme. Jej hodnota by mala byť blízka nule, aby bolo skreslenie čo najmenšie. Pre danú λ existuje množina riešení $\mathbf{X} = \lfloor \mathbf{x_{j,i}} \rfloor$. Z nich sa potom vyberie riešenie s $\max(Var(\boldsymbol{x}_{j,i}))$.

1. Miery nepodobnosti

Rovnaký algoritmus však môže rôznu dvojicu objektov považovať za podobné alebo diametrálne odlišné objekty. Rozhoduje o tom použitá miera nepodobnosti, resp. podobnosti. [6], pre kategoriálne dáta [4].

Mierou nepodobnosti nazývame takú funkciu, ktorá dvom totožným objektom z množiny priradí hodnotu 0 a rôznym objektom nenulovú kladnú hodnotu [3]. Pri normalizovanej miere je táto funkcia zhora ohraničená jednotkou. Ak normalizovaná miera dvoch objektov je rovná 1, vtedy takéto dva objekty nazývame maximálne nepodobné na danej miere nepodobnosti. Pre normalizovanú mieru podobnosti platia vlastnosti vychádzajúce z duality.

Nech Y je neprázdna množina, potom funkcia $dis : Y \times Y \to [0, \infty)$ s vlastnosťou, že $\forall y_1, y_2 \in Y, \quad dis(y_1, y_2) \geq dis(y_1, y_1)$ sa nazýva mierou nepodobnosti.

Duálnou k miere nepodobnosti je miera podobnosti a platí pre ňu vlastnosť vychádzajúca z duality, že $\forall y_1, y_2 \in Y, \quad sim(y_1, y_1) \geq sim(y_1, y_2)$.

Špeciálnym prípadom miery nepodobnosti je miera vzdialenosti, ktorá priradí dvojici objektov 0 práve vtedy, ak sú totožné. Miera vzdialenosti je nezávislá na poradí objektov. Ak navyše platí trojuholníková nerovnosť nazývame túto vzdialenosť metrikou na množine Y.

Pri kategoriálnych dátach najčastejšie zisťujeme mieru nepodobnosti

- medzi jednotlivými atribútmi (cieľom je zníženie počtu sledovaných atribútov)
- medzi atribútovými hodnotami (podrobnejšie v ďalších častiach)
- medzi objektmi

Na zisťovanie miery nepodobnosti medzi atribútovými hodnotami možno využiť aj poznatky z teórie informácie, najmä entropiu a z nej odvodenú vzájomnú informáciu (Mutual Information). Tá je mierou podobnosti dvoch delení, výsledkov zhlukovania. Pre dve nezávislé delenia má hodnotu 0, maximálnu kladnú hodnotu dosahuje pre dve totožné delenia. Metódy, ktoré ju využívajú, združujú hodnoty jedného atribútu tak, aby po zlúčení najviac kopírovali delenie objektov podľa druhého atribútu [5], teda dve atribútové hodnoty sú podobné vtedy, ak pri ich nerozlišovaní stratíme najmenej informácii o atribútovej hodnote druhého atribútu. Delenie v tomto prípade predstavuje rozdelenie objektov podľa príslušnosti k atribútovým hodnotám.

Druhý prístup využíva kosínusovú nepodobnosť (1), teda rozhodujúca je vnútorná štruktúra objektov majúcich vlastnosť $a_{j,i}$ (*i*-ta hodnota *j*-teho atribútu) a $a_{j,k}$ (*k*-ta hodnota *j*-teho atribútu).

$$dis_{\cos}(a_{j,i};a_{i,k}) = 1 - \cos(\varphi) = 1 - \frac{\sum_{h} \frac{\boldsymbol{g}_{j,i,h} \cdot \boldsymbol{g}_{j,k,h}}{\left|\boldsymbol{g}_{j,i,h}\right| \left|\boldsymbol{g}_{i,k,h}\right|}}{m-1} \quad , \tag{1}$$

kde

$\boldsymbol{g}_{j,i,h}$ - vektor súčasného výskytu atribútových hodnôt a atribútových hodnôt h-teho atribútu

$\boldsymbol{g}_{j,k,h}$ - vektor súčasného výskytu atribútových hodnôt a atribútových hodnôt h-teho atribútu

m – počet sledovaných atribútov

Vyberieme atribútové hodnoty toho istého atribútu, ktorých podobnosť chceme zistiť. Objekty, ktoré majú tieto hodnoty v danom atribúte, rozdelíme ďalej podľa hodnôt iného atribútu. Ak pomer veľkostí takto vzniknutých skupín je pre obe porovnávané atribútové hodnoty rovnaký, platí $p_{h,t,i} = p_{h,t,k}$, kde $p_{h,t.i}$ a $p_{h,t,k}$ sú podmienené pravdepodobnosti vzhľadom na porovnávané atribútové hodnoty, kde $p_{h,t,i} = p(a_{h,t} \left| a_{j,i} \right.)$. Potom atribútové hodnoty $a_{j,k}$ a $a_{j,i}$ majú normovanú mieru podobnosti 1, teda sú podľa tohto prístupu chápané ako totožné.

2 Zobrazenie do R

Pre praktické účely by bolo najvhodnejšie nahradiť atribútové hodnoty číselnou hodnotou z intervalu (0,1) tak, aby zachovávala vzťahy medzi jednotlivými atribútovými hodnotami a teda by neboli tieto hodnoty prideľované náhodne. Tento prístup tiež závisí od vnútornej štruktúry objektov. Zachovanie vzťahov nám zabezpečí použitie podmienených pravdepodobností. Pre $\lambda > 0$ môžeme množinu riešení zapísať pomocou sústavy lineárnych rovníc (2), ktoré priradia atribútovým hodnotám čísla z intervalu (0,1).

$$x_{j,i} = \frac{1}{m-1} \sum_{h=1,\, h \neq j}^{m} \sum_{t=1}^{|A_h|} \frac{\left(p_{\,h,t,i} \,.\, n_{j,i} + \lambda\right).x_{h,t}}{n_{j,i} + \left|A_h\right|.\lambda} \,, \tag{2}$$

kde

$x_{j,i}$, $x_{h,t}$ – hľadané zobrazenia atribútových hodnôt do $\mathbb{R}$

m – počet atribútov

$p_{\,h,t,i|}$ - pravdepodobnosť výskytu t-tej hodnoty v h-tom atribúte medzi objektmi majúcimi v j-tom atribúte i-tu hodnotu

$n_{j,i}$ – počet objektov majúcich v j-tom atribúte i-tu hodnotu

λ – korekcia (pre jednu množinu riešení je konštantná)

$\left|A_h\right|$ – počet atribútových hodnôt h-teho atribútu

Pre $\lambda = 0$ môžeme (2) zjednodušiť na tvar (3)

$$x_{j,i} = \frac{1}{m-1} \sum_{h=1,\, h \neq j}^{m} \sum_{t=1}^{|A_h|} p_{h,t,i} \,.\, x_{h,t} \,, \tag{3}$$

kde

$x_{j,i}$, $x_{h,t}$ – hľadané zobrazenia atribútových hodnôt do $\mathbb{R}$

m – počet atribútov

$p_{t,i}$ - pravdepodobnosť výskytu t-tej hodnoty v h-tom atribúte medzi objektmi majúcimi v j-tom atribúte i-tu hodnotu

Tento postup však spôsobí zacyklenie (vznik singulárnych matíc) a jeho použiteľnosť si vyžiada pevné stanovenie hodnôt (zobrazení do R) pre rôzne atribútové hodnoty. Ich počet je nevyhnutné minimalizovať. Použiteľné výsledky sa dosiahli až pri stanovení dvoch hodnôt. Aby sme mohli zobrazenie do $\mathbb{R}$ uskutočniť, bolo potrebné zodpovedať na tri základné otázky.

Aké hodnoty vybrať?

Ktorým atribútovým hodnotám ich priradiť?

Ako z množiny riešení vybrať to správne?

Ak chceme, aby sa hodnoty nachádzali vo vopred stanovenom intervale a používame podmienené pravdepodobnosti, teda hodnoty z intervalu (0,1), stanovené hodnoty by mali byť hranice stanoveného intervalu. Do úvahy prichádzali dve možnosti intervalov (0,1) a (-1,1). Overovali sme prvý z nich. Pri veľkom počte atribútov dochádzalo k „hromadeniu“ hodnôt v blízkosti jednotky. Je preto vhodný pre použitie pri dvoch, maximálne troch atribútoch, inak dostávame skreslené výsledky.

Príklad 1: Majme 10 objektov charakterizovaných atribútmi A_1 a A_2, kde v oboch atribútoch je zastúpenie atribútových hodnôt rovnomerné, a dvom z nich ($a_{2,1}$ a $a_{2,2}$) sú priradené atribútové hodnoty 1 a 0.

Pre náš konkrétny prípad môže zobrazenie $x_{1,1}$ atribútovej hodnoty $a_{1,1}$ zapísať ako (4).

$$x_{1,1} = \frac{1}{m-1}\left(\frac{\left(p_{2,1,1}.n_{1,1}+\lambda\right).x_{2,1}}{n_{1,1}+\left|A_2\right|.\lambda} + \frac{\left(p_{2,2,1}.n_{1,1}+\lambda\right).x_{2,2}}{n_{1,1}+\left|A_2\right|.\lambda} \right) \tag{4}$$

Z (2) môžeme každé zobrazenie vyjadriť pomocou sústavy lineárnych rovníc (5a)-(5d)

$$0 = \left(1-m\right).x_{1,1} + 0.x_{1,2} + \left(\frac{p_{2,1,1}.n_{1,1}+\lambda}{n_{1,1}+\left|A_2\right|.\lambda} \right).x_{2,1} + \left(\frac{p_{2,2,1}.n_{1,1}+\lambda}{n_{1,1}+\left|A_2\right|.\lambda} \right)x_{2,2} \tag{5a}$$

$$0 = 0.x_{1,1} + \left(1-m\right).x_{1,2} + \left(\frac{p_{2,1,2}.n_{1,2}+\lambda}{n_{1,2}+\left|A_2\right|.\lambda} \right).x_{2,1} + \left(\frac{p_{2,2,2}.n_{1,2}+\lambda}{n_{1,2}+\left|A_2\right|.\lambda} \right)x_{2,2} \tag{5b}$$

$$1 = 0.x_{1,1} + 0.x_{1,2} + 1.x_{2,1} + 0.x_{2,2} \tag{5c}$$

$$0 = 0.x_{1,1} + 0.x_{1,2} + 0.x_{2,1} + 1.x_{2,2} \tag{5d}$$

Ako toto zobrazenie závisí od hodnoty λ a od podmienených pravdepodobností si ukážeme na Obr. 1 resp. Obr. 2. Pre rôzne λ dostávame rôznu množinu riešení $\boldsymbol{X}$. Pre $\lambda \approx 0$

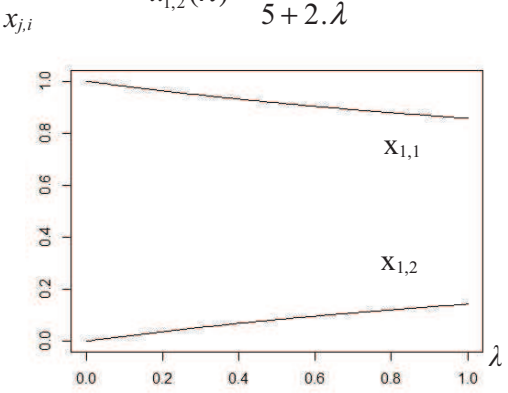
dostávame riešenie približujúce sa hľadanému. Optimálna λ je maximálna hodnota, pri ktorej ďalšom znižovaní sa už hodnoty riešení nebudú výrazne meniť.

Po dosadení všetkých vstupov a úprave dostávame v našom prípade výraz (6a).

$$x_{1,1}(\lambda) = \frac{5+\lambda}{5+2.\lambda} \tag{6a}$$

Podobne zobrazenie druhej atribútovej hodnoty ($x_{1,2}$) môžeme pomocou λ zapísať ako (6b).

$$x_{1,2}(\lambda) = \frac{\lambda}{5+2.\lambda} \tag{6b}$$

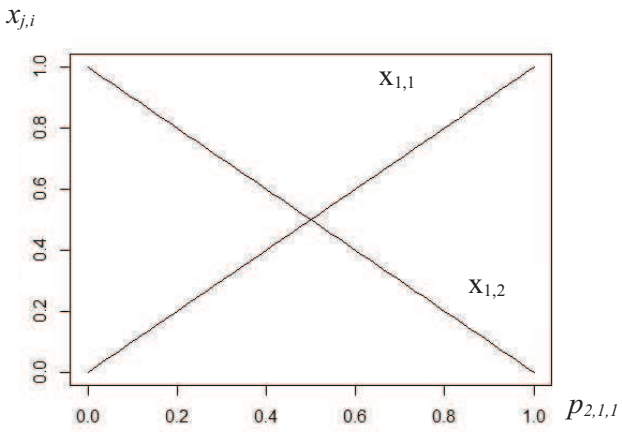


Obr. 1: Závislosť zobrazení od λ

Ak by sme chceli použiť ako premennú iba jednu podmienenú pravdepodobnosť, museli by sme ostatné vyjadriť pomocou Bayesovej vety. Potom môžeme vyjadriť zobrazenia $x_{1,1}$ (7a) a $x_{1,2}$ (7b) nasledovne

$$x_{1,1}\left(p_{2,1,1}\right) = \frac{5.\,p_{2,1,1}+0{,}01}{5{,}02} \tag{7a}$$

$$x_{1,2}\left(p_{2,1,1}\right) = \frac{5{,}01-5.\,p_{2,1,1}}{5{,}02} \tag{7b}$$



Obr. 2: Závislosť zobrazení od $p_{2,1,1}$

Odpoveď na druhú otázku nemala jednoznačné riešenie, preto bolo potrebné vyskúšať všetky možné kombinácie pre jednotlivé atribúty. Hornú hranicu sme pridelili atribútovej hodnote umiestnenej v dotazníku vyššie, dolnú nižšie umiestnenej z dvojice. Dostali sme $\sum_j \binom{|A_j|}{2}$ riešení.

Ako kritérium rozhodovania sme zvolili maximálny rozptyl, pretože sme chceli zamedziť zobrazeniu dvoch rôznych atribútových hodnôt do tej istej hodnoty z určeného intervalu.

Po zobrazení objektov ako *m*-rozmerných vektorov je výhodné zisťovať nepodobnosť medzi objektmi pomocou manhattanskej vzdialenosti (8), ktorá najlepšie zachováva vzťahy medzi atribútovými hodnotami, pretože sčítava rozdiely medzi zobrazeniami atribútových hodnôt toho istého atribútu.

$$d(o_1, o_2) = \sum_{j=1}^{m} |\mathbf{x}_j - \mathbf{z}_j| \ , \tag{8}$$

kde

x - *m*-rozmerný vektor charakterizujúci objekt o_1,
z - *m*-rozmerný vektor charakterizujúci objekt o_2.

3 Porovnanie mier nepodobnosti

Výber vhodnej miery nepodobnosti je pre výsledok zhlukovej analýzy veľmi dôležitý ako si ukážeme v Pr. 2, kde sú objekty charakterizované dvoma atribútmi (farbou a tvarom) a jeden z atribútov môže nadobudnúť 3 hodnoty. Výsledky si ilustrujeme na (Obr.3-Obr.5)

Príklad 2: Majme 13 objektov, kde jednotlivé farby a tvary sú zastúpené v súlade s kontingenčnou tabuľkou (Tab. 1). Z nej je vidieť, že v konečnom dôsledku zhlukujeme 6 objektov (2 tvary x 3 farby = červený kruh CK a štvorec CS, modrý kruh MK a štvorec MS a žltý kruh ZK a štvorec ZS) Ďalej viac objekty rozlíšiť nevieme.

Tab. 1: Kontingenčná tabuľka pre atribúty farba a tvar

	žltá	červená	modrá
■	**3**	**1**	**3**
●	**0**	**3**	**3**

Zostavme si maticu nepodobnosti pre počet rozdielnych znakov (Tab. 2)

Tab. 2**: **Počet rozdielnych znakov medzi objektmi

	CK	CS	MK	MS	ZK	ZS
CK	0	0-1 (1)	1-0 (1)	1-1 (2)	1-0 (1)	1-1 (2)
CS		0	1-1 (2)	1-0 (1)	1-1 (2)	1-0 (1)
MK			0	0-1 (1)	1-0 (1)	1-1 (2)

MS				0	1-1 (2)	1-0 (1)
ZK					0	0-1 (1)
ZS						0

Ak priraďujeme pri zhlukovaní rôznu váhu rôznym atribútom (vlastnostiam objektov), rozdielnosť medzi objektmi môžeme definovať ako váženú overlap dissimilarity (vážený priemer počtu rozdielnych znakov pre každý atribút). Tab. 3 je zostavená pre váhu prvého atribútu w_1=0,1.

Tab. 3: Vážená overlap dissimilarity measure (w_1=0,1)

	CK	CS	MK	MS	ZK	ZS
CK	0	0,9	0,1	1	0,1	1
CS		0	1	0,1	1	0,1
MK			0	0,9	0,1	1
MS				0	1	0,1
ZK					0	0,9
ZS						0

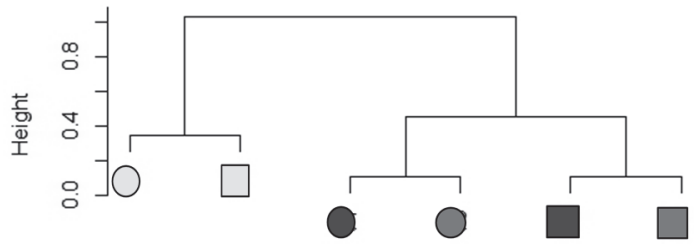
Výsledkom hierarchického zhlukovania pri použití váženej overlap dissimilarity s váhou prvého atribútu w_1 ako miery nepodobnosti bude rozdelenie objektov podľa tvaru na štvorce a kruhy, v ďalšom kroku dostaneme jeden zhluk obsahujúci všetky objekty.

Ak by sme za mieru nepodobnosti medzi atribútovými hodnotami zvolili kosínusovú nepodobnosť a medzi objektmi súčet týchto nepodobností pre jednotlivé atribúty, potom by nepodobnosť medzi jednotlivými objektmi vyzerala nasledovne (Tab. 4),

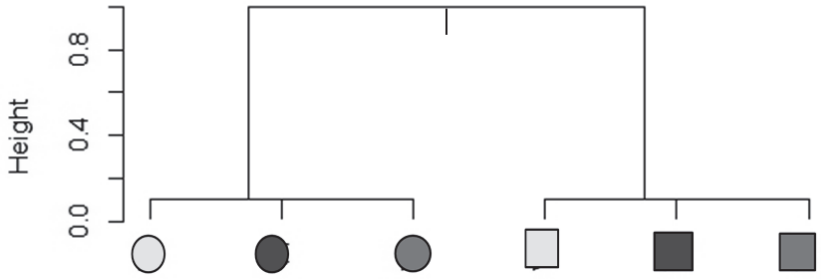
Nepodobnosť medzi žltým kruhom a modrým štvorcom je súčet kosínusov uhlov medzi vektormi (v našom prípade ($\boldsymbol{a}_{f,1} = (3;0)$ a $\boldsymbol{a}_{f,2} = (3;3)$, pre atribút „farba“) a vektormi ($\boldsymbol{a}_{t,1} = (3;1;3)$ a $\boldsymbol{a}_{t,2} = (0;3;3)$, pre atribút „tvar“) odpočítaný od počtu atribútov.

Tab. 4: Miera nepodobnosti pri kosínusovej nepodobnosti

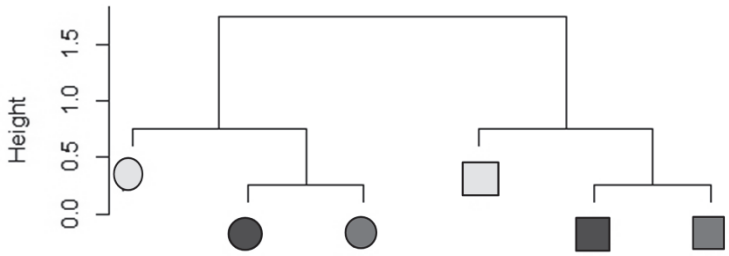
	CK	CS	MK	MS	ZK	ZS
CK	0	0,351	0,106	0,457	0,684	1,035
CS		0	0,457	0,106	1,035	0,684
MK			0	0,35 1	0,293	0,644
MS				0	0,644	0,293
ZK					0	0,351
ZS						0



Obr. 3: Hierarchické aglomeratívne zhlukovanie s využitím kosínusovej nepodobnosti



Obr. 4: Hierarchické aglomeratívne zhlukovanie s využitím Overlap dissimilarity measure



Obr. 5: Hierarchické aglomeratívne zhlukovanie s využitím manhattanskej vzdialenosti a zobrazení do R

Porovnávaním použitých mier nepodobnosti sa v našom príklade zistil nízky Pearsonov koeficient korelácie medzi kosínusovou a overlap nepodobnosťou. Ak sme ich porovnávali s navrhovanou mierou nepodobnosti, tento koeficient sa zvýšil. V našom prípade teda platí (9).

$$r(\mu_i,\mu) > r(\mu_i,\mu_j), \tag{9}$$

$r(\mu_i,\mu)$ - Pearsonov korelačný koeficient pôvodnej a novej miery nepodobnosti

$r(\mu_i,\mu_j)$ - Pearsonov korelačný koeficient pôvodných mier nepodobnosti

Ak porovnáme výsledky zhlukovania pri všetkých troch mierach nepodobnosti, zistíme, že pri navrhnutom zobrazení kategoriálnych premenných a navrhovanej miere nepodobnosti sme v našom konkrétnom prípade dostali akúsi kombináciu dvoch hierarchických aglomeratívnych zhlukovacích metód, kde prvá z nich využíva ako mieru nepodobnosti

kosínusovú nepodobnosť a druhá overlap nepodobnosť a priemerný korelačný koeficient α možno vyjadriť (10)

$$\alpha = \frac{\sum_{i=1}^{n_{DSM}} r(\mu_i, \mu_j)}{n_{DSM}}, \tag{10}$$

kde

$r(\mu_i, \mu_j)$ - Pearsonov korelačný koeficient pôvodných mier nepodobnosti

n_{DSM} - počet dvojíc mier nepodobnosti

Optimálnu mieru nepodobnosti teda získame maximalizáciou funkcie (11) pre všetky $i = 1,2,...,n$, alebo po úprave pre dve miery nepodobnosti (12)

$$\beta = \frac{\sum_{i=1}^{n_{SM}} r(\mu_i, \mu)}{n_{SM}}, \tag{11}$$

kde

$r(\mu_i, \mu)$ - Pearsonov korelačný koeficient pôvodnej a novej miery nepodobnosti

n_{SM} - počet mier nepodobnosti

$$\beta = \frac{s_{\mu_B}\left(\overline{\mu_A.\mu} - \overline{\mu_A}.\overline{\mu}\right) + s_{\mu_A}\left(\overline{\mu_B.\mu} - \overline{\mu_B}.\overline{\mu}\right)}{n_{SM}.s_{\mu_A}.s_{\mu_B}.s_{\mu}} \tag{12}$$

s_{μ_A}, s_{μ_B} - odhady rozptylov pôvodných mier nepodobnosti

$\overline{\mu}$ - odhad strednej hodnoty novej miery nepodobnosti

$\overline{\mu}_A, \overline{\mu}_B$ - odhady stredných hodnôt pôvodných mier nepodobnosti

n_{SM} – počet pôvodných mier nepodobnosti

Z takto získanej miery by sme mohli spätným postupom získať ešte zodpovedajúcejšie zobrazenie do $\mathbb{R}$. Jeho existencia však nie je zaručená.

Záver

Prostredníctvom podmienených pravdepodobností sa nám podarilo nájsť jedno z možných zobrazení, ktoré sme ilustrovali na príklade dvoch atribútov. Dve z použitých mier nepodobnosti neboli korelované. V porovnaní s treťou však obe výraznejšie korelovali. S využitím hierarchického aglomeratívneho algoritmu a troch rôznych mier nepodobnosti sme dosiahli rôzne výsledky. Ak sme použili tretiu mieru nepodobnosti, výsledok po prvom kroku predstavoval zjemnenie prvých dvoch delení.

Poďakovanie: Tento článok bol podporený z grantu VEGA 1/0143/11.

Literatúra

[1] ANDRITSOS, P. et al. LIMBO: Scalable Clustering for Categorical Data, In: Lecture Notes in Computer Science, 2004, s.123-146

[2] KHORSHIDPOUR, Z. – HASHEMI, S. – HAMZEH, A. CBDL: Context-Based Distance Learning for Categorical Attributes, In: International Journal of Intelligent System

[3] NEUBRUNN, T.- DRAVECKÝ, J. Vybrané kapitoly z matematickej analýzy:Základy teórie miery a integrálu Bratislava: Alfa, 1990, 206 s .

[4] ŘEZÁNKOVÁ, H. Cluster analysis and categorical data. Praha : Vysoká škola ekonomická v Praze, 2009.

[5] SHAMIR, O. - SABATO, S. – TISHBY, N. Learning and generalization with the information bottleneck. In: Theoretical Computer Science, 2010, s. 2696-2711

[6] TVERSKY, A. Features of similarity In: Psychological Review, 1977, s. 327-352

Adresa autora:

Dušan Janál, Ing.

Stavebná fakulta STU

Radlinského 11, 813 68 Bratislava

janal@math.sk

Základný algorimus simulovaného žíhania
Simulated annealing – the basic algorithm

Zuzana Krivá

Abstract: The paper deals with the basic version of the simulated annealing – the so called Boltzmann annealing – the stochastic algorithm to solve the global optimization problems of applied mathematics, namely locating a good approximation to the global minimum of a given function in a large search space. At the beginning we present the sample problem, then the paper introduces the particular steps of the solution and the problems we must face when we use the Boltzmann annealing. The paper represents the complemental text for the lecture for which several macros in Visual Basic have been created aiming to help the students understand the algorithm and the way how we interpret the results. The paper is accompanied with the images showing these program tools.

Abstrakt: Článok sa zaoberá základnou verziou simulovaného žíhania, tzv. boltzmannovským žíhaním, ktorý sa používa v aplikovanej matematike na hľadanie dobrých aproximácií globálnych miním alebo maxím v rozľahlom prehľadávacom priestore. Na začiatku článku sa predstaví vzorová úloha a článok nás postupne prevádza problémami, ktoré nastávajú pri jej riešení a podrobne sa zaoberá jednotlivými krokmi jej riešenia pomocou boltzmannovského žíhania. Článok predstavuje sprievodný text k prednáške, pre ktorú bolo vytvorených niekoľko makier vo Visual Basicu v Exceli, ktoré si kládli za cieľ pomôcť študentom k pochopeniu algoritmu a spôsobu, akým treba interpretovať jeho výsledky. Článok je doplnený obrázkami z vývojového prostredia.

Key words: Monte Carlo optimization, stochastic algorithm, combinatorial optimization, hill climbing algorithm, Metropolis algorithm, simulated annealing.

Kľúčové slová: metóda Monte Carlo, stochastický algoritmus, kombinatorický problém, horolezecký algoritmus, Metropolisov algoritmus, simulované žíhanie.

JEL classification: C02, C15, C63.

Úvod

Tento článok sa zaoberá algoritmom simulovaného žíhania, a to jeho základnou verziou, tzv. boltzmannovským žíhaním. Simulované žíhanie je stochastický algoritmus na riešenie problému minimalizácie. O simulovanom žíhaní existuje množstvo literatúry, ktorá je väčšinou alebo jednoducho čitateľná a nezodpovedá všetky základné otázky, alebo veľmi zložitá, ktorá zodpovedá všetky možné otázky, okrem tých ktoré nás zaujímajú, prípadne ak ich zodpovedá, tak komplikovaným spôsobom.

V ďalších odsekoch sa budeme snažiť o vypichnutie najdôležitejších vlastností algoritmu simulovaného žíhania, a tak poskytnúť čitateľovi základnú predstavu o princípoch jeho fungovania. Použijeme na to vzorový príklad, sa snaží popísať jednotlivé fázy riešenia.

1. Problémy riešiteľné simulovaným žíhaním

Pri riešení praktických úloh často potrebujeme nájsť dobré aproximácie globálneho minima účelovej funkcie v rozľahlom prehľadávacom priestore. Za predpokladu, že nám stačí prijateľne dobré riešenie v rámci nejakého časového intervalu a nepotrebujeme skutočne najlepšie možné riešenie:

- simulované žíhanie môže byť efektívnejšie ako náročné výpočty,
- vôbec umožňuje približne riešiť úlohu v prijateľnom čase,
- programátorsky nie je náročné.

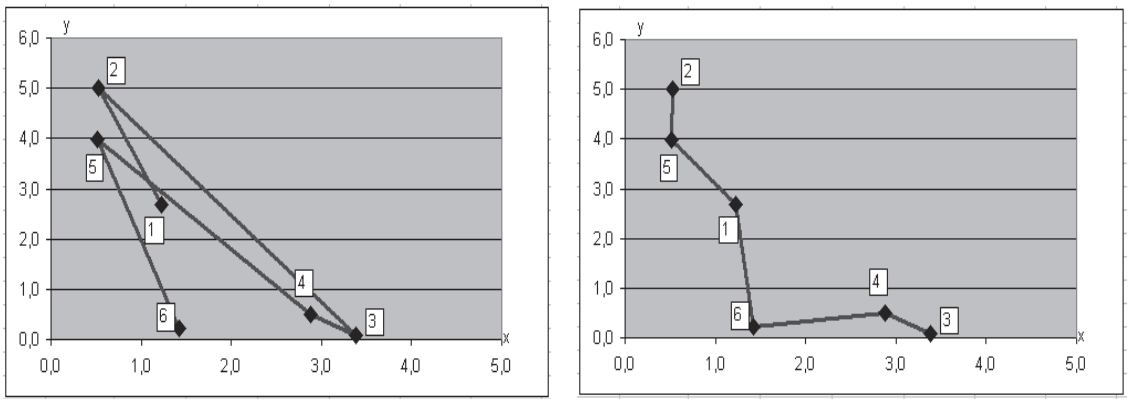
Jeho veľkým problémom je však nastavovanie parametrov. Jeden náš štatistik sa o ňom vyjadril, že je to algoritmus síce ľahko pochopiteľný a vysvetliteľný, avšak nastavenie jeho parametrov tak, aby dával dobré výsledky vyžaduje veľké množstvo kumštu.

Pre mnohé úlohy sú sformulované aj podmienky ako nájsť globálne minimum, nezaručujú však, že je to rýchlejšie, ako keby sme prešli celý prehľadávací priestor. Mnohí výskumníci však vedia ako zľaviť z prísnych podmienok tak, aby dostali uspokojivé výsledky. Predmetom intenzívneho výskumu je hľadanie menej prísnych teoretických podmienok [4].

2. Predstavenie vzorovej úlohy - príklad zložitého kombinatorického algoritmu

Je zadaných N bodov v rovine. Budeme zostrojovať lomenú čiaru, ktorá prechádza každým bodom práve raz a je otvorená. Budeme sa pýtať, ktorá lomená čiara má najmenšiu dĺžku.

V teórii grafov budeme hľadať minimálnu hamiltonovskú cestu v grafe, v tomto prípade špeciálnom, kde ohodnotenie hrán je dané vzdialenosťou vrcholov. Slabou modifikáciou – spojením prvého a posledného segmentu by sme dostali úlohu hľadania minimálnej hamiltonovskej kružnice. O týchto úlohách je známe, že nie sú ekvivalentné, ale majú rovnakú zložitosť riešenia. Všeobecná úloha hľadania minimálnej hamiltonovskej kružnice je špeciálnym prípadom problému obchodného cestujúceho, ktorý býva vzorovým príkladom tzv. NP-úplného problému, t.j. problému riešeného nedeterministickými Turingovými strojmi v polynomiálnom čase (t.j. prakticky exponenciálne), a teda nie v reálnom čase pre väčšie úlohy.



Obr. 1: Body v rovine sú očíslované, cestu môžeme reprezentovať permutáciou triedy 6. Na obr. vľavo je permutácia 123456, ktorú budeme často využívať ako počiatočnú, na obrázku vpravo je permutácia 251643 (resp. 346152) zodpovedajúca ceste s najkratšou dĺžkou.

Tab. 1: Súradnice vrcholov.

1,220	0,532	3,381	2,876	0,515	1,422
2,669	4,997	0,079	0,500	3,994	0,228
1	**2**	**3**	**4**	**5**	**6**

Niektoré charakteristiky: počet permutácií 720, najkratšia cesta 7.09, najdlhšia cesta 22.366.

3. Niektoré metódy riešenia vzorovej úlohy

Pripomeňme si, že účelovou funkciou, ktorú ideme minimalizovať, bude dĺžka čiary.

Metódy hrubej sily. V tomto prípade by metóda hrubej sily predstavovala vygenerovanie všetkých permutácií, prípadne prehľadávanie grafu – (branch and bound). Generovanie všetkých permutácií je problém zložitosti O(N!). Nasledujúca tabuľka poukazuje na prudký

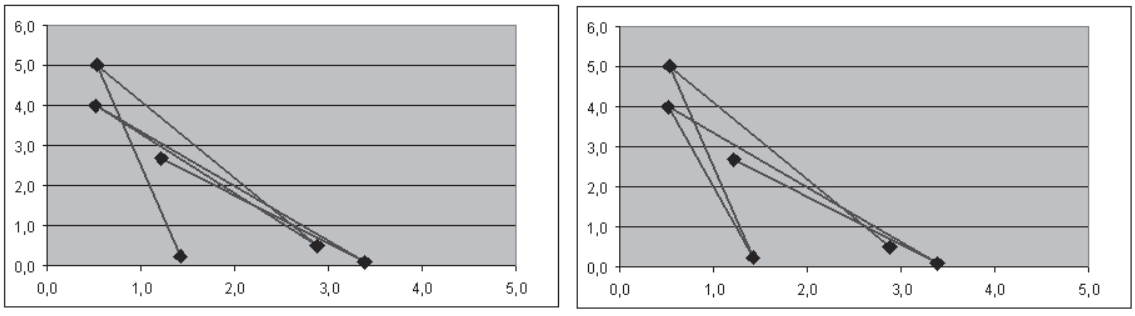
nárast permutácií. Generovanie 13! už predstavuje úlohu, ktorá sa, naprogramovaná v C na bežnom počítači, rieši niekoľko dní. Pritom v praktických úlohách potrebujeme často stovky vrcholov.

Tab.2 Nárast počtu permutácií.

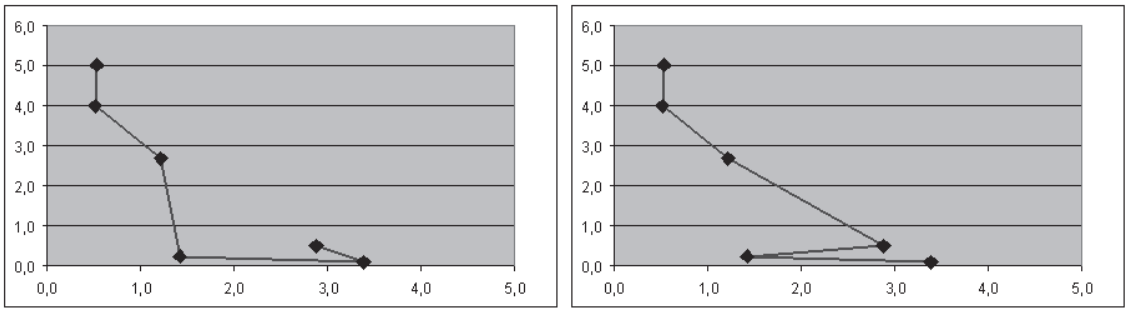
Počet permutácii:	
5	120
6	720
7	5 040
8	40 320
9	362 880
10	3 628 800
11	39 916 800
12	479 001 600
13	6 227 020 800
14	87 178 291 200
15	1 307 674 368 000

Predstavme si teraz, že vrcholy sú mestá a chceme ich pospájať cestou, pričom chceme minimalizovať náklady. Nemusíme nutne nájsť skutočné minimum, postačí nám, ak je dosť dobré, teda že od skutočného nie je príliš vzdialené. Ďalej si uvedieme niektoré metódy, ktoré vracajú tzv. suboptimálne riešenia.

Heuristiky – nerieši a problém exaktne, ale na základe nejakej skúsenosti, často krát spresňujú nejaký hrubý odhad. V našom príklade napr. môže heuristika pracovať len s takými čiarami, ktorých úseky sa nepretínajú (obr.2), alebo vždy si bude vyberať najbližší nenavštívený vrchol (obr.3).



Obr.2 Vľavo cesta s maximálnou dĺžkou, vpravo s dĺžkou blízkou maximálnej. Tieto cesty obsahujú veľa pretínajúcich sa úsekov a veľa dlhých segmentov



Obr.3 Cesty s dĺžkou blízkou minimálnej (cesta s minimálnou dĺžkou je na Obr.1 vpravo). Tieto cesty obsahujú krátke segmenty (t.j. spájajú blízke vrcholy), ktoré sa nepretínajú

Horolezecký algoritmus možno zaradiť medzi gradientné metódy, pokiaľ ho však spúšťame opakovane s iným náhodne vygenerovaným počiatočným riešením, možno ho považovať aj za stochastickú metódu.

Na začiatku sa vygeneruje počiatočné riešenie x_o. V ďalších krokoch sa generujú použitím konečného počtu operácií riešenia ležiace v určitom okolí východzieho riešenia (napr. vo vzorovom príklade ich môžeme získať výmenou 2 vrcholov v permutácii). Z týchto riešení sa vyberie to, ktoré má z hľadiska účelovej funkcie najlepšie ohodnotenie. Akceptujú sa len rovnako dobré alebo lepšie riešenia, preto sa metóda môže dostať k nevýraznému lokálnemu minimu blízko počiatočného náhodne vygenerovaného riešenia x_o a nikdy sa nedosiahne optimálne riešenie. Problém môžeme čiastočne odstrániť tak, že ho budeme opakovane spúšťať s náhodne voleným počiatočným riešením. Stochastičnosť metódy spočíva na náhodnom výbere počiatočného riešenia, lebo horolezecký algoritmus postupuje systematicky bez akejkoľvek náhodnosti.

Problémom tejto metódy je teda uviaznutie v lokálnom minime, a práve preto sa často spomína v súvislosti so simulovaným žíhaním. Algoritmus simulovaného žíhania sa práve tento problém snaží riešiť tým, že pripúšťa aj prechod do stavu s horšou hodnotou účelovej funkcie.

Simulované žíhanie patrí medzi stochastické algoritmy a je založený na analógii medzi žíhaním tuhých telies a optimalizačným problémom. Existuje mnoho verzií a mnoho modifikácií základného algoritmu. V tomto článku sa budeme zaoberať len jeho základnou verziou, tzv. boltzmannovským žíhaním.

4. Žíhanie tuhého telesa a simulované žíhanie vo fyzike

Žíhanie vo fyzike predstavuje proces, ktorého cieľom je odstraňovanie vnútorných defektov telies alebo dosiahnutie menej zrnitej štruktúry. Hnacím motorom procesu je dostatočne vysoká teplota žíhania. Materiál sa najprv zahreje na predpísanú teplotu tak, aby bol celý rovnomerne zahriaty. Teplota žíhania je často blízka teplote začiatku tavenia materiálu, nikdy ju však nesmie prekročiť. Vyžaduje sa malá rýchlosť zmien teploty. Žíhané predmety sa z teploty žíhania ochladzujú veľmi pomaly. Pri žíhaní sa využíva schopnosť difúzie atómov, ktoré z pôvodne nerovnovážneho stavu na začiatku difundujú do stavu rovnovážneho. Teplota sa musí znižovať tak pomaly, aby sa pri každej teplote T všetky častice telesa mali možnosť dostať do rovnovážnej polohy - ekvilibria, charakterizovanej určitou strednou hodnotou energie $\bar{E}$, závislou na teplote T. Energia telesa sa postupne znižuje. Pod ekvilibriom si nepredstavujeme statickú situáciu. Je to situácia, keď systém náhodne mení svoj stav z jednej konfigurácie do druhej, tak že pri každej teplote T, pravdepodobnosť w_T, že nájdeme systém v konkrétnej konfigurácii častíc i s energiou E_i , je daná vzorcami

$$w_T(i) = \frac{1}{\boxed{Q(T)}} exp\left(-\frac{E_i}{kT}\right), \qquad \boxed{Q(T)} = \sum_i exp\left(-\frac{E_i}{kT}\right) \quad (1)$$

kde, Q(T) je tzv. partičná funkcia alebo normalizačný faktor, teda v ekvilibriu je systém opísaný *tzv. boltzmannovským rozdelením.* Príklad tohto rozdelenia pre vzorovú úlohu, kde namiesto energie E bude vystupovať hodnota účelovej funkcie, bude uvedený neskôr.

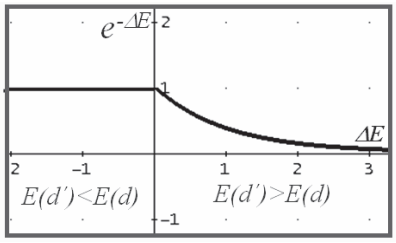
Hlavná myšlienka Metropolisovho algoritmu. Chceme napríklad vypočítať teplotný priemer systému. Keď má systém len malú množinu stavov, dá sa vypočítať presne, ale keď veľkú, hocijaký náhodný systém konfigurácií by neviedol k dobrému výsledku. Chceme vybrať reprezentačnú vzorku konfigurácií nie celkom náhodne, ale tak, aby vzorka bola zviazaná s konfiguráciami, ktoré sú významne zastúpené v ekvilibriu.

Pre simuláciu správania sa takéhoto systému na počítači, Metropolis a spol. (1953) navrhli metódu Monte Carlo, (t.j. metódu založenú na generovaní náhodných alebo pseudonáhodných čísel), ktorá simuluje evolúciu systému nasledovne: začne z nejakého počiatočného stavu a postupne, pomocou nejakej prechodovej funkcie generuje nové stavy (tzv. poruchy). Tieto stavy prijíma alebo zamieta na základe tzv. Metropolisovho kritéria, kde pravdepodobosť prijatia je vlastne daná podielom boltzmannonvých pravdepodobností vygenerovaného a aktuálneho stavu. Označme *d* ako aktuálny stav a *d´* ako vygenerovaný stav. Pravdepodobnosť *p* prijatia je daná vzťahom

$$p(d' \leftarrow d) = min\left(1, e^{\frac{-\Delta E}{kT}}\right)$$

$$\Delta E = E_{d'} - E_d \qquad (2)$$

(Metropolisovo kritérium)



Obr.4 Vľavo Metropolisovo kritérium, vpravo hodnoty pravdepodobností pre T=1

V podiele dvoch boltzmannovských pravdepodobností vypadol normalizačný faktor, takže v ňom vystupuje iba hodnota rozdielu energie (účelovej funkcie) v aktuálnom a novo vygenerovanom (porušenom stave). Ak je tento rozdiel záporný, t.j. nový stav má menšiu energiu, prijímame ho vždy. Nový stav však s určitou pravdepodobnosťou prijímame, aj keď má horšiu energiu, pričom pravdepodobnosť je daná jednak veľkosťou rozdielu a jednak teplotou. Pri vysokých teplotách môže byť prijatý aj dosť veľký rozdiel, naopak, pri nízkych sú prijímané len malé rozdiely energií.

Aplikovaním veľkého počtu porúch (pri fixnej teplote T) a ich akceptovaním do ďalšieho procesu s pravdepodobnosťou (3) dostaneme systém v tepelnej rovnováhe, pričom distribúcia pravdepodobnosti rozloženia stavov sa asymptoticky blíži k boltzmannovskej distribúcii. Tento tvar metódy Monte Carlo sa v štatistickej fyzike nazýva Metropolisov algoritmus [5].

Simulované žíhanie je vlastne zreťazenie Metropolisových algoritmov, kde výstup z jedného (pri danej teplote T po aplikovaní dostatočného množstva porúch *kmax*) je vstupom do druhého. Môžeme ho popísať nasledovným algoritmom.

```
vstup=počiatočný stav;
T = tmax; /* nastavenie T na počiatočnú – maximálnu teplotu */

while                              (T>tmin)
{vystup=Metropolis(vstup,kmax,T);                    (A1)
                                                vstup=vystup;
T=zniz_teplotu(T); /*zníženie teploty podla planu chladenia*/
}
```

Takto riadený systém sa bude postupne vyvíjať k stavom s nižšou energiou.

5. Simulované žíhanie na riešenie všeobecného minimalizačného problému

V 80. rokoch dostali nezávisle dve skupiny ľudí [1,2] nápad, že spôsobom podobným simulovanému žíhaniu tuhého telesa môžeme hľadať globálne minimum ľubovoľnej funkcie, pretože algoritmus pracuje iba s konceptom:

- konfigurácia x
- hodnota účelovej funkcie v konfigurácii x - $f(x)$.

Vo vzorovom príklade jednej konfigurácii zodpovedá jedna permutácia, $f(x)$ je dĺžka čiary zodpovedajúca energii. Čas sa stane formálnym parametrom, a teda konštanta k splymie s časom T. Boltzmannovské pravdepodobnosti potom sú dané vzorcami:

$$w_T(x) = \frac{1}{Q(T)} exp\left(-\frac{f(x)}{\boxed{T}}\right), \qquad Q(T) = \sum_x exp\left(-\frac{f(x)}{\boxed{T}}\right) \qquad (3)$$

Metropolisov algoritmus s akceptáciou stavov danou (2) na obr.4, kde f(x) nahradí $E(x)$ produkuje distribúciu pravdepodobnosti stavov, ktorá sa asymptoticky blíži k boltzmannovskej distribúcii.

6. Metropolisov algoritmus pri hľadaní hamiltonovskej cesty

Pozrime sa najprv na Boltzmannovo rozdelenie, keď účelová funkcia predstavuje dĺžku čiary. Pre jednoduchosť zoberieme zo šiestich bodov vzorového príkladu prvé štyri. Máme *24* konfigurácií predstavujúcich *12* rôznych ciest, najkratšia cesta má dĺžku *5,17,* najdlhšia má dĺžku *10,11*. Máme konečný a malý počet konfigurácií, ich Bolzmannovské pravdepodobnosti môžeme vyčísliť presne. Najprv vypočítame pre každú konfiguráciu výrazy *exp(-f(x)/T).* Ich spočítaním dostaneme normalizačný faktor *Q(T)* a môžeme vypočítať pravdepodobnosti pre všetky konfigurácie. Na obr.5 máme znázornené pravdepodobnosti konfigurácie s dĺžkou f(x) pri rôznych teplotách, zvislá modrá čiara znázorňuje presné stredné hodnoty. Pri *T=5* je aj pravdepodobnosť čiary s dĺžkou 10 dosť vysoká a v postupnosti generovanej Metropolisovým algoritmom sa po dosiahnutí ekvilibria budú často vyskytovať aj stavy s maximálnou dĺžkou čiary alebo tejto dĺžke blízke. Postupne, so znižovaním teploty sa bude zvyšovať pravdepodobnosť stavov s menšou dĺžkou čiary (obrázokv strede), až pre *T=0,1* by sa v postupnosti stavov ekvilibria malo vyskytovať už len samotné globálne minimum.

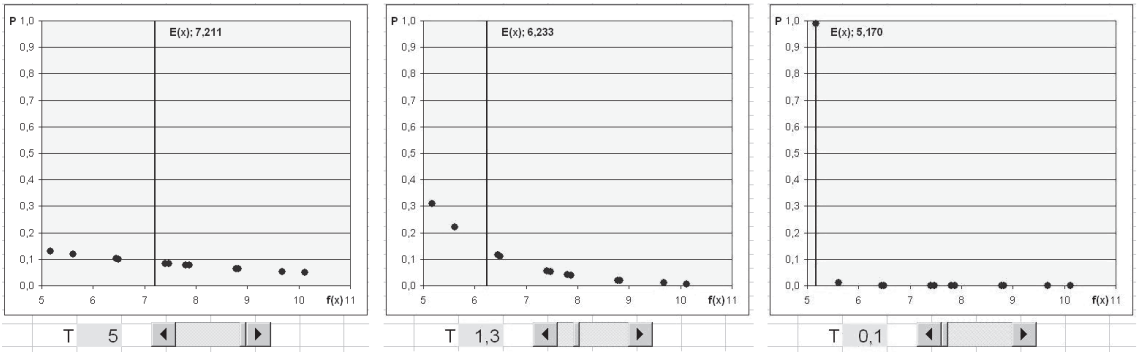
Pri simulácii Metropolisovým algoritmom musíme vyriešiť jeden problém, ktorý sme zatiaľ nespomínali, a to ako budeme generovať nový (porušený) stav, čiže musíme stanoviť stochastický operátor poruchy pre prechod do susedného stavu. Susedný stav by mal byť v istom zmysle blízky a využívať informáciu o probléme z predchádzajúcich pokusov – v tom je rozdiel oproti základnej metóde Monte Carlo, ktorá generuje konfigurácie nezávisle od predchádzajúceho pokusu.

Môžeme uvažovať napríklad takéto možnosti:

- výmena dvoch susedných vrcholov v permutácii,
- výmena dvoch ľubovoľných vrcholov v permutácii,

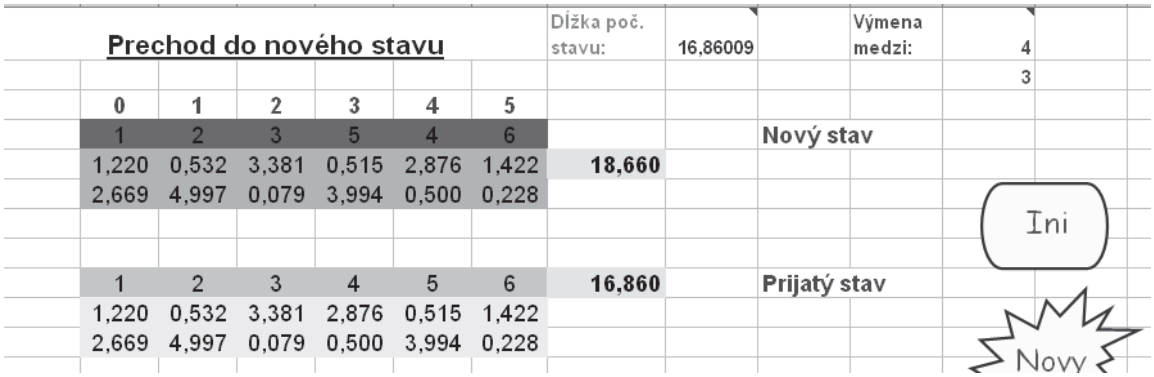
- výmena dvoch ľubovoľných vrcholov v permutácii s prevrátením cesty medzi nimi.

V rôznych fázach riešenia môžeme používať rôzne operátory poruchy. V. Černý (z UK Bratislava) vo svojej práci používal operátor podobný tretiemu, ale s tým, že keď sa algoritmus blížil ku koncu, vymieňali sa vrcholy, ktoré neboli od seba veľmi vzdialené, a teda operátor poruchy sa viac podobal prvému. Výber tohto operátora môže veľmi ovplyvniť efektívnosť algoritmu. Ďalej by mal tento operátor dodať dostatočne krátku cestu z počiatočného stavu do ľubovoľného stavu, v ktorom môže byť globálne optimum.



Obr.5 Boltzmannovské pravdepodobnosti pre rôzne teploty

7. Praktická realizácia Metropolisovho algoritmu



Obr.6 Akceptácia stavov. Identifikátory bodov sú čísla 1 až 6 (tieňované záhlavia), vrcholy lomenej čiary sú číslované od 0. Na obrázku sú vymenené body na pozíciach lomenej čiary 3 a 4, vymenené sú body s identifikátorom 4 a 5.

Popísať proces prijímania a zamietania stavov nám pomôžu obrázky 6 a 7, ktoré predstavujú jeden hárok programu Excel rozdelený na dve časti. Na obr. 6 vľavo hore vidíme predovšetkým pole súradníc vrcholov. Vrcholy sú očíslované od 1 po 6 v tieňovaných záhlaviach, tieto čísla sú vlastne identifikátory bodov. Horné čísla v bielom poli sú očíslované pozície (0 až 5) v čiare. V dolnej časti vidíme počiatočnú cestu zodpovedajúcu permutácii 123456 (na pozícii čiary 0 je bod s identifikátorom 1, na pozícii čiary 1 je bod s identifikátorom 2 atď.) Táto čiara bola vygenerovaná pomocou tlačidla *Ini,* za ktorým sa skrýva funkcia pre vygenerovanie nového stavu a na začiatku je súčasne aj prijatým stavom. Pomocou tlačidla *Nový* sa generuje jedno náhodné číslo predstavujúce číslo pozície v čiare, kde sa udeje výmena (vpravo hore Výmena medzi:), na obrázku je to 4. Vymenia sa vrcholy na vygenerovanej a predchádzajúcej pozícii, tu štvrtej a tretej a pozícii, v prípade nuly meníme s piatou pozíciou. Výmenou sa čiara zodpovedajúca *prijatému stavu* zmení na nový

stav s novou dĺžkou čiary, teda s novou hodnotou *f(x),* ktorý je zobrazovaný v časti *Nový stav*. Prijatie alebo zamietnutie nového stavu bude realizovať tlačidlo *Rozhodni* (obrázok 7). Ak stav bude zamietnutý, *nový stav* sa nahradí predchádzajúcim, teda tu *prijatým stavom*. Ak stav bude prijatý, *prijatý stav* sa nahradí *novým stavom*.

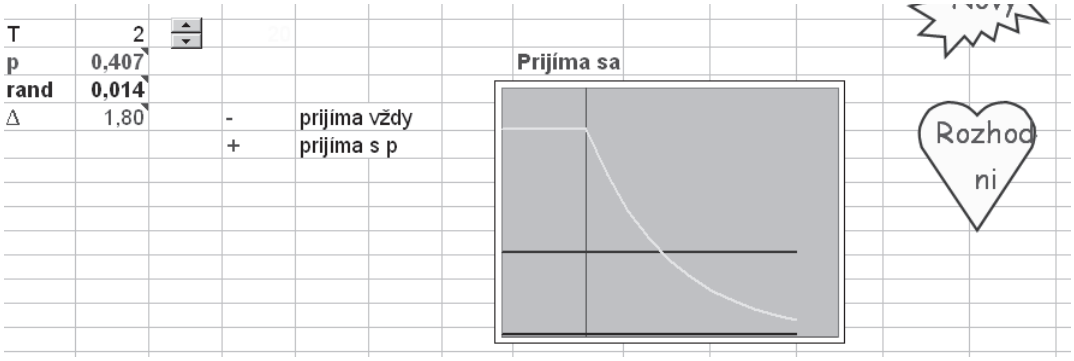
Opäť vygenerujeme nový stav a opäť rozhodneme o jeho prijatí alebo neprijatí. Počet opakovaní je jedným zo vstupných parametrov algoritmu.

Obrázok 7 sa podrobnejšie venuje procesu rozhodovania o prijatí alebo zamietnutí. Na grafe je bledou farbou znázornený tvar akceptačnej funkcie pre *T=2*, tvar tejto funkcie sa mení v závislosti od teploty. Číslo v políčku *p* predstavuje pravdepodobnosť prijatia pre danú dĺžku čiary, pričom do akceptačnej funkcie vstupuje rozdiel dĺžok čiar nového a predchádzajúceho stavu. Pripomeňme, že ak je záporný, *p* je rovné jednej a nový stav (s menšou dĺžkou) je prijímaný vždy. Tu je rozdiel dĺžok rovný 1,80 a zodpovedajúca pravdepodobnosť 0,407. Keď je teplota vyššia dostaneme pre tento rozdiel väčšiu pravdepodobnosť, pri nižšej teplote menšiu.

To, že nový stav prijímame s pravdepodobnosťou *p* realizujeme tak, že vygenerujeme náhodné číslo *rand* v rozmedzí 0 a 1. Na obrázku sú znázornené pravdepodobnosť prijatia a vygenerované náhodné číslo dvoma vodorovnými úsečkami. Podľa geometrickej pravdepodobnosti nový stav prijímame, ak *rand* je menšie ako *p*.

V jazyku C môžeme metropolisov algoritmus zapísať nasledovne:

```
while (kmax) // kmax - pocet opakovaní
   {kmax--;
    kopia_ciary(pciara,ciara);//ciara sa uchova v pciara
    dlzka_noveho=novy_stav();
    p = exp(-(dlzka_noveho-dlzka_prijateho)/T);
    rand = ((double) rand())/(RAND_MAX );//nah.cislo v <0,1>
    if (rand < p)
       {dlzka_prijateho= dlzka_noveho;}
    else
       kopia_ciary(ciara,pciara); //v prip.neprijatia navrat do
                                  //predchadzajuceho stavu
  }
```

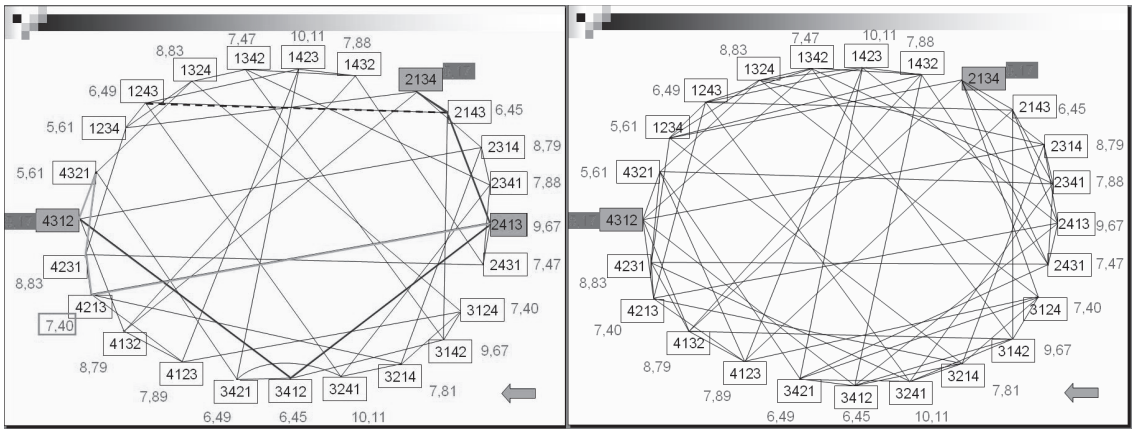


Obr.7 Akceptácia stavov

(A2)

Zastavme sa ešte pri stochastickom operátore poruchy. Tu je dôležité splniť, že sa z každého stavu vieme dostať do ľubovoľného iného stavu podľa možnosti na čo najmenší počet krokov. Porovnajme si prvý a druhý prípad, teda keď vymieňame dva susedné a dva ľubovoľné vrcholy. Opäť si vezmime iba štyri body a pozrime sa na Obr.8. Pri výmene dvoch susedných vrcholov (obrázok vľavo) je stupeň vrcholu prehľadávacieho grafu 4 a z jedného vrcholu do druhého sa dostaneme na najviac 6 preklopení. Všeobecne je stupeň vrcholu prehľadávacieho grafu *n* a z jedného stavu do druhého sa dostaneme maximálne na *n(n-1)/2* preklopení. Pri výmene dvoch ľubovoľných vrcholov (obrázok vpravo) je stupeň vrcholu prehľadávacieho grafu 6 a z jedného vrcholu do druhého sa dostaneme na najviac 3 preklopenia. Všeobecne je stupeň vrcholu prehľadávacieho grafu *n(n-1)/2* a z jedného stavu do druhého sa dostaneme maximálne na *(n-1)* preklopení. Druhý spôsob umožňuje väčšie zmeny účelovej funkcie.

Na záver tejto časti poznamenajme, že vybrať vhodný stochastický operátor poruchy vyžaduje náhľad do problému a nemusí byť samozrejmý. V. Czerny v [2] radil, aby sme ich mali niekoľko a pre daný problém všetky vyskúšali. Výber tohto operátora nemá vplyv na rovnovážnu kánonickú distribúciu, ale môže mať drastický vplyv na rýchlosť, akou je ekvilibrium dosiahnuté. Pri praktickej realizácii na počítači boli napr. lepšie výsledky dosahované ak sme menili ľubovoľné dva stavy ako keď sme menili dva susedné stavy.



Obr.8 Priestor stavov pre prvý a druhý stochastický operátor poruchy. Cieľové stavy, teda stavy s najmenšou dĺžkou čiary 5,17 sú zvýraznené farebne. Na obrázku vľavo sú naznačené niektoré možnosti pre počiatočný stav 2413. Všimnime si napríklad, že pokiaľ výmenou prejdeme do stavu 4213, všetky susedné stavy majú väčšiu dĺžku čiary, a teda pri nízkej teplote sa s vysokou pravdepodobnosťou z tohto stavu nedostaneme

8. Praktická realizácia simulovaného žíhania

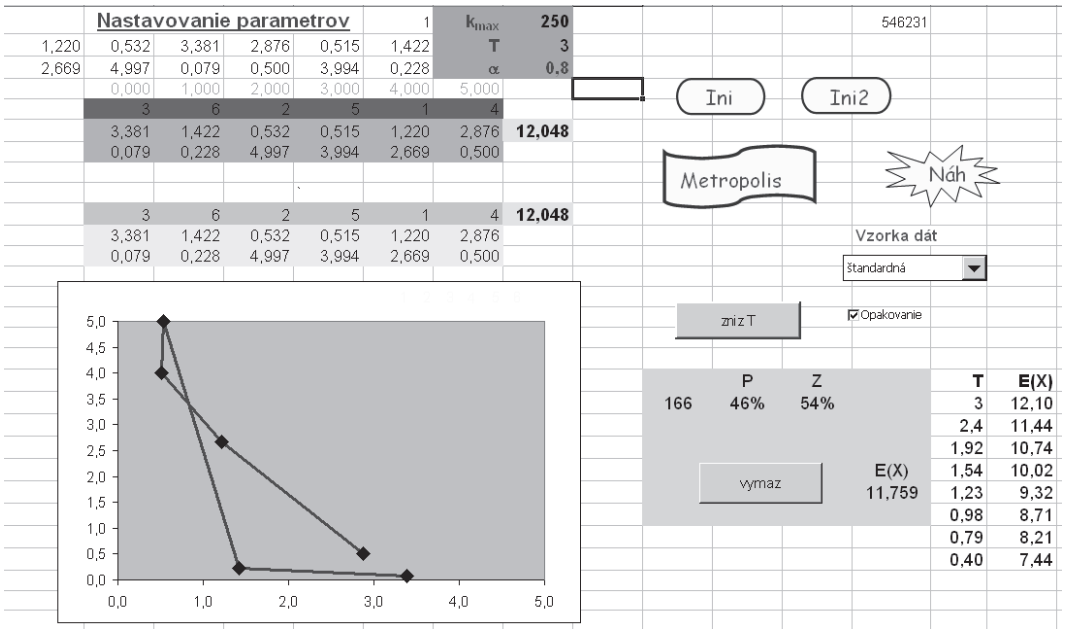
Ako sme už povedali, simulované žíhanie je vlastne zreťazenie metropolisových algoritmov (viď. algoritmus A1). Skôr, ako ho začneme realizovať, treba stanoviť:

- Stavový priestor a účelovú funkciu
- Operátor pre generovanie nových porúch (stavov)
- Počiatočná teplota T_{max} a plán chladenia:
  - Rýchlosť znižovania teploty a
  - Počet vygenerovaných porúch pre T *kmax* (čas relaxácie)

Prvými dvoma bodmi sme sa už zaoberali, venujme sa teraz plánu chladenia. S boltzmannovským simulovaným žíhaním by mal byť konzistentný logaritmický plán

chladenia, t.j. $T(k) = \frac{T_0}{\ln(k)}$, kde k je časový index chladenia. Väčšinou sa však používa exponenciálny plán chladenia $T_k = \alpha^k T_0$, alebo inak $T_{k+1} = \alpha T_k$. V angličtine sa niekedy používa výraz *simulated quenching* namiesto *simulated annealing* [4]. Tento plán chladenia budeme používať aj my.

Skôr ako si povieme niečo o najošemetnejšej časti – nastavovaní parametrov, pozrime sa na Obr.9. Tlačidlá so zvýrazneným orámovaním (*INI, INI2, Metropolis*) a *ZnizT* predstavujú tlačidlá pre hlavné kroky simulovaného žíhania: počiatočnú inicializáciu dát a Metropolisov algoritmus, ktorý sa spúšťa opakovane, pričom po jeho prebehnutí sa vždy α-krát zníži teplota tlačidlom *zniz T*. V zvýraznenom poli hore v strede vidno parametre výpočtu. Vpravo dole je presne vypočítaná stredná hodnota boltzmannovskej distribúcie pre dané T. Môžeme ju porovnať so strednou hodnotou vygenerovaných (prijatých) stavov pre danú teplotu, ktorá sa zobrazuje vľavo vo farebnom poli. Štatistika ďalej ukazuje počet prijatých a zamietnutých vygenerovaných porúch s horšou hodnotou účelovej funkcie.

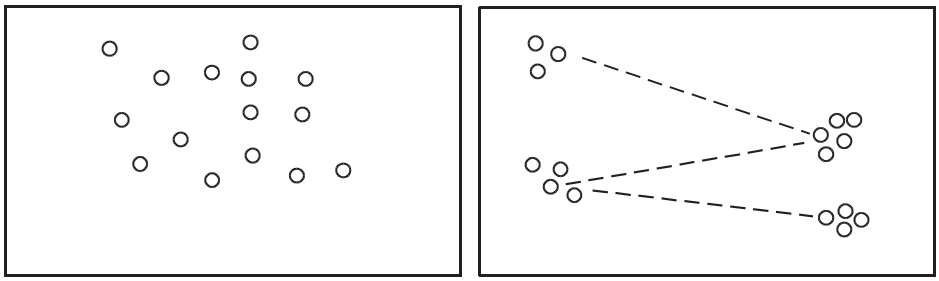


Obr.9 Hlavné kroky simulovaného žíhania. Vľavo prijatý stav, novovytvorený stav a nastavenie parametrov, vpravo tlačidlá pre hlavné kroky výpočtu a štatistika

Nastavovanie parametrov výpočtu. Parametre algoritmu simulovaného žíhania sa nedajú stanoviť nejakou všeobecnou metodikou, ale musia byť empiricky stanovované pre každý problém. Napriek tomu sa dajú použiť určité parametre, ktoré môžu byť postačujúce aspoň pre počiatočné spúšťanie:

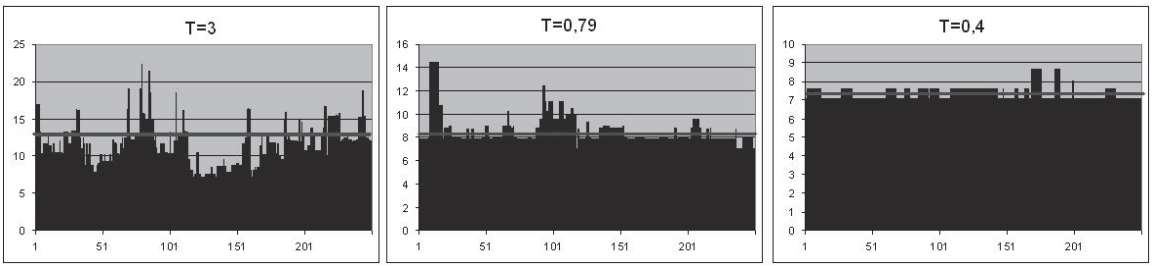
- ***plán chladenia - $T \rightarrow \alpha T$, $0 << \alpha < 1$***, v praxi sa ukazujú ako dobré hodnoty 0.8 až 0.99.
- ***T_{max}*** - počiatočná hodnota teploty sa obyčajne volí tak, aby bola Metropolisovým kritériom akceptovaná približne polovica porušených stavov (t.j. stavov s horšou hodnotou účelovej funkcie *f(x)*). Na Obr.9 je táto štatistika pre *T=3*. Z *250* vygenerovaných stavov malo *166* horšiu hodnotu účelovej funkcie, z toho *46%* bolo prijatých a *54%* bolo zamietnutých. Keď je T_{max} príliš veľké, dostávame stochastickú metódu hľadania globálneho optima ekvivalentnú so základnou metódou Monte Carlo, keď je T_{max} príliš malé, dostávame horolezecký algoritmus.

- ***kmax*** - počet opakovaní pre dané T, t.j. čas relaxácie – čas (či skôr počet vygenerovaných nových stavov) potrebný na nastavenie ekvilibria po zmene teploty, veľmi závisí na topografii účelovej funkcie, na momentálnej teplote T a komplikovaným spôsobom aj na spôsobe generovania porušených stavov. V literatúre občas možno nájsť odporučenie 10^4-10^6 pre dáta väčšieho rozsahu.



Obr.10 Dáta na obr. vľavo potrebujú inú Tmax, ako dáta na obr. vpravo, kde je v počiatočných fázach potrebné poprepájať správne dlhé úseky a Tmax musí byť na začiatku dostatočne veľké. Pokiaľ by sme vyšli z počiatočného stavu, ktorého prepojenia medzi dlhými úsekmi by boli ako sú na obr. vpravo naznačene čiarkovane, pri nedostatočnej Tmax by sa nepreklopili a nikdy by sme sa nepriblížili k optimu. Pri nízkej teplote sa už len dolaďujú prepojenia v rámci zhlukov bodov

Keďže vzorový príklad je malého rozsahu, pre jednotlivé teploty sa dá presne vypočítať stredná hodnota (obr.9 vpravo) a porovnať so strednou hodnotou dosiahnutou výpočtom pri danej teplote.



Obr.11 Graf hodnôt účelovej funkcie pre množinu vygenerovaných stavov pri rôznych teplotách. Hrubou vodorovnou čiarou je naznačená presná stredná hodnota

Na záver uvedieme len niekoľko príkladov. V tejto úlohe pre *N=12* trvalo autorovi nájdenie minimálnej cesty pomocou vygenerovania všetkých permutácií v jazyku C 18 minút. Na tom istom počítači, keď sme zvolili z 20 opakovaní algoritmus simulovaného žíhanie pri zvolení parametrov ako na Obr.9 našiel 5-krát presné riešenie, 14-krát blízke presnému riešeniu a jeden krát uviazol v lokálnom minime vzdialenom od globálneho. Bol použitý jednak štandardný generátor pseudonáhodných čísel, jednak generátor Mersen Twister. Nepotvrdil sa podstatný vplyv zvoleného generátora pseudonáhodných čísel. Pre N=6 sme robili experimenty s dvoma stochastickými operátormi poruchy. Pri zvolení operátora, ktorý vymieňal ľubovoľné dva vrcholy bolo globálne minimum dosiahnuté častejšie.

9. Praktické využitie

Na záver spomenieme niektoré typy úloh a oblasti, kde sa často využíva simulované žíhanie.

- Rôzne grafové úlohy:
  - výstavba a rozširovanie telekomunikačnej siete
  - riadenie dopravy
  - ekonomika, plánovanie
- Návrh počítačov, rozmiestňovanie súčiastok na doske plošných spojov
- Nukleárna fyzika, chémia
- Spracovanie obrazu (filtrácia zašumeného obrazu, segmentácia, nastavovanie parametrov)
- Vytváranie optimálnych triangulácií a tetrahedralizácií
- Molekulárna biológia
- Seizmografia – modelovanie seizmických vĺn
- Finančníctvo, ekonomika
- Vojenstvo – optimálne rozmiestňovanie striel zem - vzduch.

10. Záver

Dá sa ukázať, že simulované žíhanie nájde globálne minimum sa blíži k 1 pri rozsiahlom pláne chladenia. Tento teoretický výsledok ale nie je veľmi nápomocný, lebo čas, ktorý je potrebný na dosiahnutie významnej pravdepodobnosti k úspechu, obyčajne prekročí čas potrebný na úplné prehľadanie priestoru riešení.

Existuje veľa modifikácií simulovaného žíhania.

11. Literatúra

[1] KIRKPATRICK, S. – GELATTI, C. D. –VECCHI, M. P.: Optimization by simulating annealing. In: Science, Vol.220, No 4598, 1983, s. 671-680.

[2] ČERNÝ, V.**:** Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm. In Journal of Optimization Theory and Application:, Vol.45, No 1, 1985, s. 41-51.

[3] KVASNIČKA, V. – POSPÍCHAL, J. – TIŇO, P.: Evolučné algoritmy, STU Bratislava 2000, kap.6

[4] INGBER, L.: Simulated annealing: Practice versus theory, J Mathl. Comput. Modelling, Volume 18, No 11,1993, s. 29-57

[5] http://en.wikipedia.org/wiki/Simulated_annealing

Adresa autora:

Zuzana Krivá, doc. RNDr. PhD.

Stavebná fakulta STU

Radlinského 11, 813 68 Bratislava

kriva@math.sk

Adaptívna metóda konečných objemov na riešenie lineárnej difúznej rovnice na konzistentnej adaptívnej mriežke
Adaptive Finite Volume Method to Solve the Linear Diffusion Equation on a Consistent Quadtree Grid

Zuzana Krivá, Karol Mikula

Abstract: The paper deals with the solution to the linear heat equation on the nonuniform quadtree grid adapted for the finite volume method used for the space discretization. This grid, depending on the data, is built using the quadtree technique and is modified in such way, that the connection of representative points of two adjacent finite volumes is perpendicular to their common boundary. The paper presents the implicit finite volume numerical scheme, its EOC and shows some outputs of the adaptive algorithm for selected noisy data. The presented adaptive numerical scheme can become the stepping stone to solve the nonlinear diffusion equations on this type of grid.
Abstrakt: Tento článok sa zaoberá riešením rovnice vedenia tepla na nerovnonernej - adaptívnej mriežke, prispôsobenou pre metódu konečných objemov, ktorá je použitá pre priestorovú diskretizáciu. Táto mriežka je prispôsobená spracovávaným dátam pomocou techniky kvadrantových stromov a je zmodifikovaná tak, že spojnica stredov dvoch susedných elementov je kolmá na ich spoločnú hranicu. V článku ukazujeme odvodenie implicitnej konečno-objemovej schémy, popisujeme adaptívny algoritmus, pre ktorý vypočítavame experimentálny rád konvergencie a ukazujeme prácu algoritmu pre vybrané zašumené obrázky.

Key words: Image processing, the linear heat equation, finite volume method, adaptivity.

Kľúčové slová: Spracovanie obrazu, lineárna rovnica vedenia tepla, metóda konečných objemov, adaptivita.

JEL classification: C63, C67, C88.

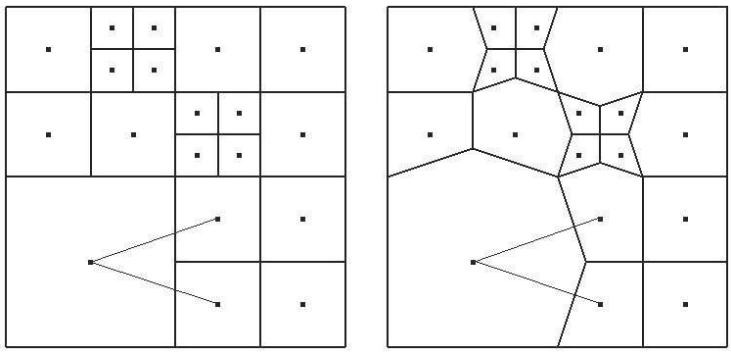
Introduction

In image processing many filtration algorithms are based on the nonlinear diffusion PDEs which modify the linear diffusion (heat) equation by slowing down the diffusion in the vicinity of edges. The linear heat equation is not only the base of these reliable and mathematically approved methods but it can be applied also when evaluating the so called "Gaussian gradient", i.e. the gradient of the image presmoothed by convolution with Gaussian function, which is often used as an edge detector also in other types of models (e.g. the modifications of level set -type PDEs models) [1,2,8].

For some of these models adaptive methods have been developed [1,2,4,5] - they use the fact that with the progress of diffusion removing the noise, more and more regions of homogenous image intensity are present in the image - the solution tends to be more flat with the increasing scale. For these regions we can use larger elements of the computational grid.

This paper deals with the numerical solution to the linear heat equation on the so called *consistent* adaptive grid built using the quadtree technique. This consistent grid posseses the important property, that the connection of two representative points of two adjacent finite volumes is perpendicular to their common boundary, what is imporatnt fact when we use the finite volume space discretization [3]. To make computations easier we use a quadtree grid with prescribed ratio of adjacent elements' sides: 1:1 or 1:2. This quadtree grid is procedurally adjusted to the consistent grid. Examples of basic quadtree and consistent grids are displayed in the fig. 1. We present the

geometric properties of the consistent grid and derive the implicit finite volume scheme.



Obr. 1. ***An example of the original quadtree grid together with the representative points of its elements (on the left). This grid is transformed into the consistent one (on the right).***

We solve the following problem:

$$\frac{\partial u(x,t)}{\partial t} - \Delta u(x,t) = 0 \quad \text{in } Q_T \equiv \Omega \times I, \tag{1}$$

$$\frac{\partial u(x,t)}{\partial n} = 0 \quad \forall x \in \partial\Omega \times I, \tag{2}$$

$$u(x,0) = u^0(x) \ \ \forall x \in \Omega. \tag{3}$$

Here, $u(x,t)$ is an unknown function defined in $\Omega \subset R^2$ and $I = [0,T]$ a time interval, n is a unit outer normal vector to $\partial\Omega$ and $u^0(x)$ is an initial condition.

1. The adaptive grid

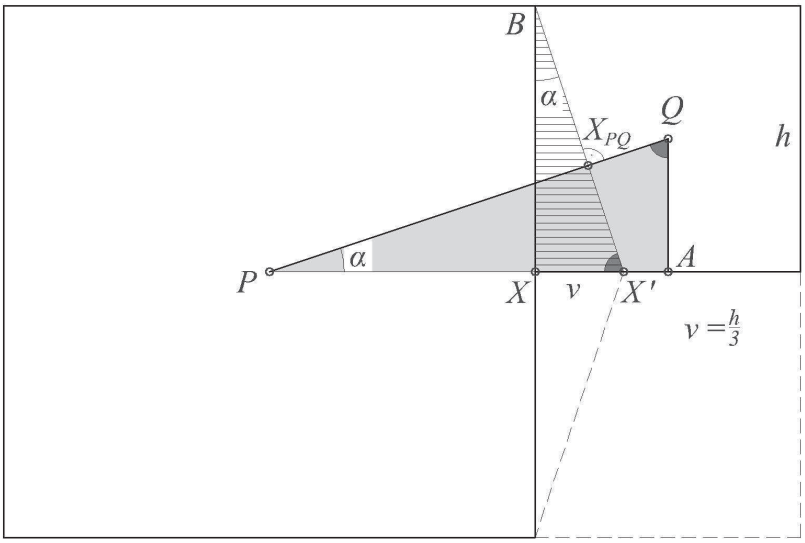
The initial image is given as a set of discrete grey (or RGB) values on cells of an initial regular - *nonadaptive* grid, which corresponds to the pixel structure of the image. At the beginning and especially with progress of smoothing algorithms, we can merge cells using some "coarsening" criterion and instead on the regular mesh, we will work on the irregular *adaptive* one. This idea was first used for solving the Perona-Malik equation, where the whole information about the image is contained in the initial grid and there is no spatial movement of the edges, no refinement is needed and the algorithm works just with grids, elements of which are obtained by merging of pixels. This process has been called *coarsening* [1,2].

To build our adaptive grid we used the ideas of [7,8], where adaptive strategies for rectangular 2D and 3D finite elements in large scale image processing had been elaborated. In our approach the adaptive grid is represented by leaves of the quadtree. However, instead of organizing resulting structure into a tree (which is known as being inconvenient when access to neighbors is needed) we use a procedural approach and maintain the field of *indicators* which enable us to find out easily whether a given cell is a leaf or not and what is the size of its neighbor. Traversing the quadtree, we stop on a higher level of hierarchy (i.e. on a coarse, nontrivial element) if some "stopping criterion" (described later) is fulfilled - this information is stored in an easily evaluated possition in the indicator field.

In order to simplify creating the linear system matrix, where access to neighbors is needed, we require that the ratio of sides of two adjacent squares is 1 : 1, 1 : 2 or 2 : 1. Later, such a structure will be called *balanced*. The technique of building the quadtree (and octree) adaptive

grids is described in [4,5]. It uses the following **coarsening criterion**: the cells are merged if a difference in their intensities is below a prescribed tolerance ε.

Adjustment to a consistent grid. The above grid is **inconsistent** in the sense, that we cannot find the unique representative points of the grid elements - finite volumes - such that the connection of representative points of two adjacent finite volumes is perpendicular to their common boundary. The adaptive grid fulfilling this condition is called *consistent*. The *consistent* grid will be encoded in the same way as the basic quadtree grid and it is adjusted to a consistent grid procedurally. We must adjust the shape and its area, if two adjacent finite volumes p and q are of the different size. If we denote the length of a common edge in the original quadtree by h, then if we shift the "hanging node" by $v = \frac{h}{3}$ (e.g. in fig. 2 we shift X to X') - then connection of representative points is perpendicular to the shifted common boundary. This fact (and also the fact, that $\frac{BX'}{PQ} = \frac{2}{3}$) follows from the similarity of triangles $\triangle AQP$ and $\triangle XX'B$, let us explore e.g. the angles in the quadrilateral $X'AQX_{PQ}$. The area of p is also evaluated procedurally - it depends on a configuration of its neighbors. If a finite volume p has two smaller neighbors, its original area P is enlarged by $\frac{1}{12}$ of P. If a finite volume p has a greater neighbor, its area P is reduced by $\frac{1}{6}$ of P, otherwise it is unchanged. These tests are performed on every edge of p.



Obr. 2. ***\|XX'\|=v=$\frac{1}{3}$h. XB=$\frac{2}{3}$PA, hence*** $\frac{BX'}{PQ} = \frac{2}{3}$**.**

2. Numerical scheme on a consistent adaptive grid

Let $\mathcal{T}_h$ be an adaptive grid with finite volumes p of measure $m(p)$ and let $N(p)$ be the set of neigbors $q \in \mathcal{T}_h$ for which common interface of p and q is a line segment e_{pq} with nonzero measure $m(e_{pq})$. Let every finite volume p have a representative point x_p lying in its center or in the center of the original square for an adjusted element of the consistent grid. Let u_p denote the solution value constant over p. Then $d_{pq} = |x_q - x_p|$ and n_{pq} is a normal vector to e_{pq} outward to p [3]. Having the grid, we can integrate the diffusion equation over a finite volume p and we use

the divergence theorem to obtain

$$\int_p \partial_t u \, dx - \int_{\partial p} \nabla u . n_p \, ds = 0, \tag{4}$$

where $n_p = (n_x, n_y)$ is the outward unit normal vector to ∂p.

We replace the time derivative by finite difference using the uniform time step $\tau = t^n - t^{n-1}$, where t^{n-1}, t^n are previous and current time steps, respectively. Let u^n be the solution in the n^{th} time step and u_p^n denotes the solution over the finite volume p in the n^{th} time step. Having the integral form of (4) for (1), let us denote by

$$F_{pq}^n = \int_{e_{pq}} \nabla u^n . n_{pq} \, ds \tag{5}$$

the implicit flux through boundary e_{pq} between p and its neighbor q. Then the implicit scheme can be rewritten in the following general form

$$(u_p^n - u_p^{n-1}) \, m(p) = \tau \sum_{q \in N(p)} F_{pq}^n, \tag{6}$$

where u_p^n is a representative value of approximated solution in the finite volume p at time step t^n.

The flux F_{pq}^n contains a normal derivative of a solution at the time step t^n evaluated on the boundary e_{pq} and we approximate it numerically by:

$$\nabla u^n \cdot n_{pq} \approx \frac{\left(u_q^n - u_p^n\right)}{d_{pq}}. \tag{7}$$

Let us denote by T_{pq} the term $\frac{m(e_{pq})}{d_{pq}}$. Now the flux can be approximated by

$$F_{pq}^n \approx T_{pq}(u_q^n - u_p^n), \tag{8}$$

where

- for q of the different size as p (in the original quadtree grid), $T_{pq} = \frac{2}{3}$.
- for q of the same size as p (in the original quadtree grid): if p and q have the same parent in the quadtree structure and a common larger neighbor, $T_{pq} = \frac{2}{3}$ (the length of such common edge is reduced, see e.g. fig.1). Otherwise $T_{pq} = 1$. In other words, if one edgepoint of the common edge is a "hanging node " in the original quadtreee then $T_{pq} = \frac{2}{3}$, otherwise $T_{pq} = 1$.

Now we are ready to write the **implicit finite volume scheme** for solving problem (1)-(3):
Let $0 = t_0 \leq t_1 \leq ... \leq t_{N_{\max}} = T$ *denote the time discretization with* $t_n = t_{n-1} + k$, *where* k *is the time step. For* $n = 1, ..., N_{\max}$ *we look for* u_p^n, $p \in \mathcal{T}_h$ *satisfying*

$$(u_p^n - u_p^{n-1}) m(p) = k \sum_{q \in N(p)} T_{pq} \left(u_q^n - u_p^n\right) \tag{9}$$

or rewritten

$$\left(\frac{m(p)}{k}+\sum_{q\in N(p)}T_{pq}\right)u_p^n-\sum_{q\in N(p)}T_{pq}u_q^n=\frac{m(p)}{k}u_p^{n-1} \tag{10}$$

The **adaptive algorithm** has three phases in every time step:

1. *We build the balanced quadtree adaptive grid and change it to a consistent one.*
2. *We compute the coefficients of the linear system during recursive traversal of the quadtree.*
3. *We solve the linear system (9).*
4. *Go to step 1.*

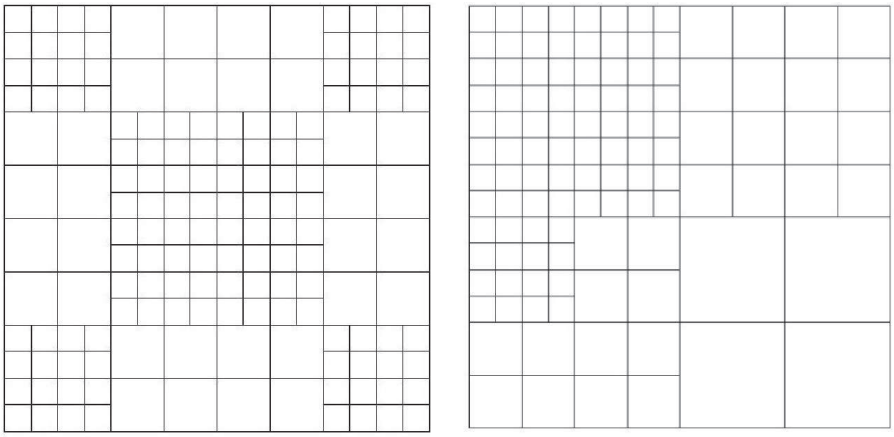
3. EOC for the consistent adaptive scheme

For numerical tests, let us have the linear diffusion equation with a right hand side

$$u_t-\Delta u=f, \tag{11}$$

where $u=u(x_1,x_2,t),\ f=f(x_1,x_2,t),\ (x_1,x_2)\in(0,1)\times(0,1)=\Omega$ and $t\in[T_1,T_2]$. For the function $f(x_1,x_2,t)=\cos(2\pi x_1)\cos(2\pi x_2)(1+8\pi^2 t)$, the exact solution is given by $u(x_1,x_2,t)=\cos(2\pi x_1)\cos(2\pi x_2)t$. We add numerically evaluated right hand side f to the scheme (9) and solve the problem in time interval $[0.5,0.6]$. We compute $L_2(I,L_2(\Omega))$ norm of the error by the formula

$$\sqrt{\sum_{n=1}^{N}\tau\sum_{p}(u(x_p,t^n)-u_p^n)^2m(p)}. \tag{12}$$

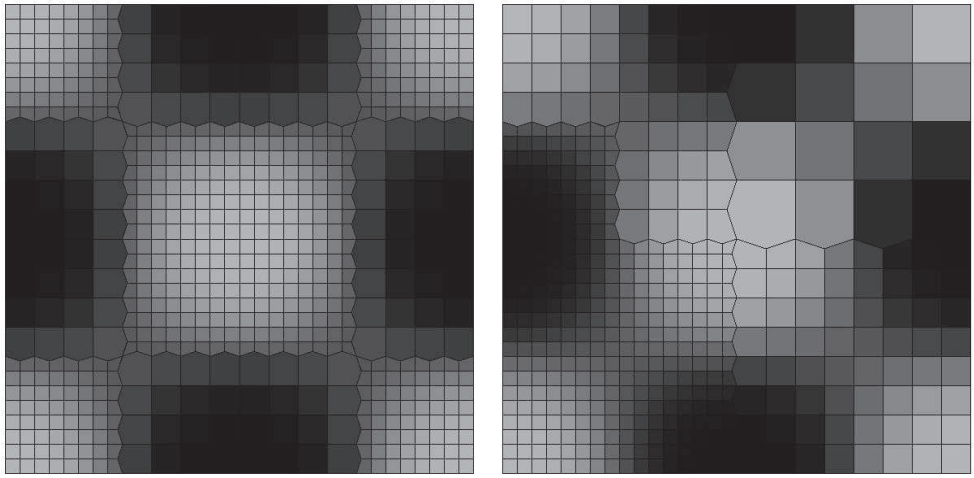


Obr. 3. ***The initial quadtree grids for EOC:grid a) on the left and grid b) on the right.***

To perform EOC experiment we take an initial nonuniform adaptive grid and solve the linear heat equation on this grid for a given time interval $T=[0.5,0.6]$. Then we refine the initial adaptive

grid by dividing every finite volume into four subvolumes and solve the linear heat equation again. Let us note that the grid is not changing in time, though we solve the time dependent problem. The initial uniform grid is of the size 16×16 ($h = \frac{1}{16}$). Over this grid we constructed two initial nonuniform grids depicted in fig. 3 and perform the experiments.

To study EOC, we refine each grid three times. We performed N time steps (the third column of Table 1) for each grid and evaluated the error given by (12). The error evaluations are shown in Table 1.



Obr. 4. ***Left: the consistent grid with*** $h = \frac{1}{32}$ ***obtained by refinement of the grid a) and the corresponding intensity function. Right: the consistent grid with*** $h = \frac{1}{32}$ ***obtained by refinement of the grid b) and the corresponding intensity function.***

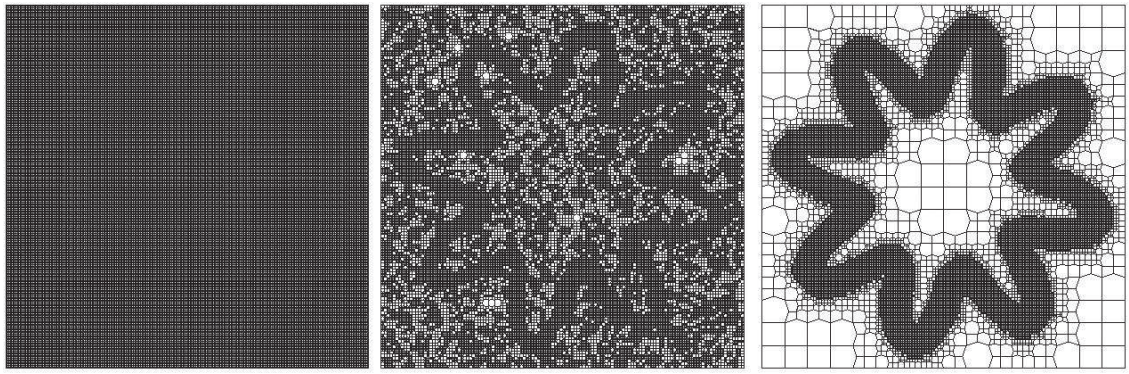
Tabul'ka 1. ***EOC calculations, from the left:*** h ***means the size of finite volumes in the initial grid if there would be no coarsening,*** τ ***is the time step,*** N ***is the number of time steps performed over*** $T = [0.5, 0.6]$***. Then the errors*** $E(h)$ ***and*** EOC ***for both Grid a) and Grid b) are given.***

h	τ	N	Grid a) $E(h)$	Grid a) EOC	Grid b $E(h)$	Grid b) EOC
$\frac{1}{16}$	0.003906	25	0.0188		0.01302	
$\frac{1}{32}$	0.000977	102	0.00465	2.015	0.003459	1.912
$\frac{1}{64}$	0.000244	409	0.001154	2.010	0.00089	1.958
$\frac{1}{128}$	0.000061	1638	0.000288	2.002	0.000109	1.996

4. Applying the adaptive algorithm to noisy data

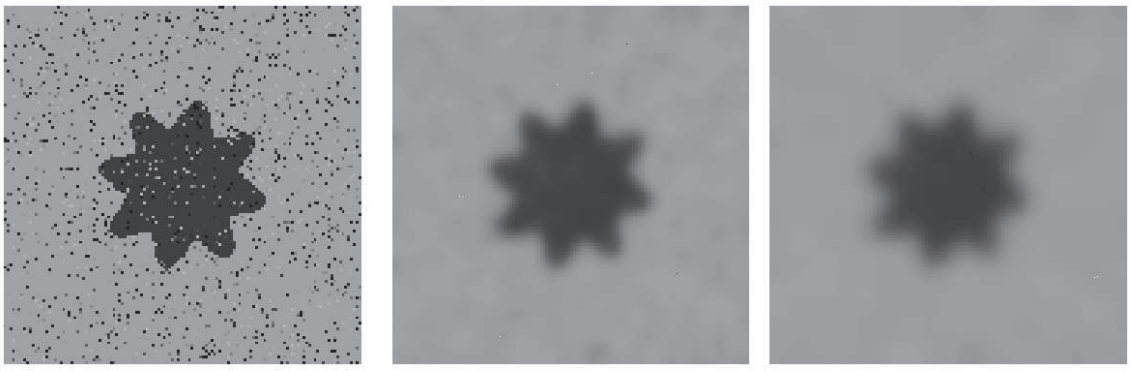
Experiment 1. The artificial data of the size 256×256 is disturbed by the additive uniform noise and is shown in the fig.5 on the left. We performed 20 scale steps with the size of the scale step $\tau = 1$ and the smallest grid element of the size $h = 1$. The results after 4 and 20 scale steps are shown in the same figure. At the beginning the grid is too dense, after 10 steps the initial number of elements 65 536 has fallen to 24332.

Obr. 5. ***On the left: the original noisy data. In the middle: the 4th scale step. On the right: the 20th scale step.***

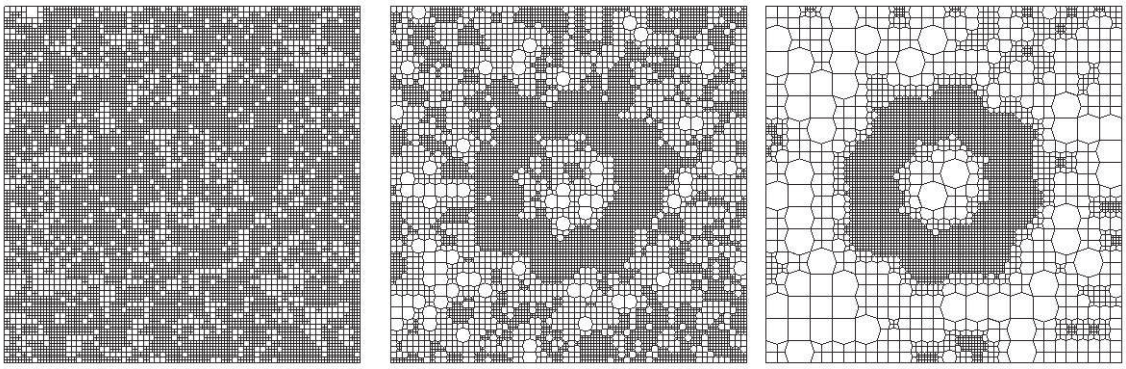


Obr. 6. ***On the left: the original regular grid (63 536 elements). In the middle: the grid in the 4th scale step (52195). On the right: the final grid(24322 elements).***

Experiment 2. This experiment was performed on the data of the size $128 \times 128 = 16384$ elements spoilt by 10% salt&pepper noise. The initial grid contained 12658 elements, after 5 scale steps 9835 elements and after 10 scale steps only 4615. The decrease of elements was caused by different type of noise and also by the fact, that the original image covered less part of the total image area than in the previous experiment. The smoothed data and corresponding grids are shown in the fig.7 and the fig.8.



Obr. 7. ***On the left: the original noisy data. In the middle: the 4th scale step. On the right: the 10th scale step.***



Obr. 8. ***On the left: the initial grid (12 658 elements). In the middle: the grid in the 4th scale step (9835). On the right: the final grid (the*** 10^{th} ***scale step with 4615 elements).***

5. Conclusion

In this paper we present a novel algorithm to solve the linear diffusion equation on the consistent adaptive grid. The algorithm proved the experimental order of convergence equal to two and we believe it can become a good starting point for deriving the consistent adaptive algorithms solving the nonlinear PDEs used in image processing. To be able to do this, we must evaluate the gradients on this type of grids, which will be the topic of the forthcoming paper.

6. Acknowledgements.

The work on this paper has been supported by the grant APVV-0184-10 and VEGA 1/1137/12.

7. References

[1] Bänsch, E. - Mikula, K., 1997. A coarsening finite element strategy in image selective smoothing. In Computing and Visualization in Science, Vol. 1 (1997), pp. 53-61.

[2] Bänsch, E. - Mikula, K., 2001. Adaptivity in 3D image processing. In Computing and Visualization in Science, Vol. 4 (2001), pp. 21-30.

[3] Eymard, R. - Gallouet, K. - Herbin, R., 2000. The finite volume method .In Handbook for Numerical Analysis, Elsevier, 2000.

[4] Krivá, Z. 2004. Adaptive finite volume methods in image processing. In Edícia vedeckých prác, Zošit č. 15, Vydavatelstvo STU, Bratislava 2004.

[5] Krivá, Z. - Mikula, K., 2002. An Adaptive Finite Volume Scheme for Solving Nonlinear Diffusion Equations in Image Processing. In Journal of Visual Communication and Image Representation, Vol. 13, Issues 1-2(2002), pp. 22-35.

[6] Krivá, Z. - Mikula, K. - Peyrièras, N. - Rizzi, B. - Sarti, A. - Stašová, O., 2010. 3D early embryogenesis image filtering by nonlinear partial differential equations. Med Image Anal, Vol. 14(4), 2010, pp. 510-526.

[7] Ohlberger, M. - Rumpf, M., 1998. Adaptive projection operators in multiresolutional scientific visualization. In IEEE Transactions on Visualization and Computer Graphics, Vol. 4, No.4 (1998), pp. 344-364.

[8] Preusser,T. - Rumpf, M. 1999. An Adaptive Finite Element Method for Large Scale Image Processing. In roceedings of ScaleSpace'99 (1999), pp. 223-234.

Adresa autora (autorov)

Zuzana Krivá, doc. RNDr. PhD.
KMaDG Svf STU
Radlinského 11, 813 68 Bratislava
kriva@math.sk

Karol Mikula, prof. RNDr. DrSc.
KMaDG Svf STU)
Radlinského 11 813 68 Bratislava
mikula@math.sk

Názory verejnosti na migrantov a ich integráciu v SR, VII. ako by ste pomohli imigrantom v SR? Public opinion on migrants and their integration in SR, VII. how would You help immigrants in SR?

Ján Luha, Lenka Berová, Martina Žáková

Abstract: We present results from public opinion research on imigrants and their integration in Slovak republic, seventh part – how would You help immigrants in Slovakia?
Abstrakt: V príspevku prezentujeme výsledky výskumu verejnej mienky názorov imigrantov na ich integráciu v Slovenske republike, siedma časť – ako by ste pomohli imigrantom v SR?
Key words: public opinion, immigrants, integration in SR, help immigrants.
Kľúčové slová: názory verejnosti, imigranti, integrácia v SR, pomoc imigrantom.
JEL Classification: C1, C12.

1. Úvod

V príspevku prezentujeme siedmu časť výsledkov vlastného výskumu názorov dospelej populácie Slovenskej republiky na aktuálne otázky a problémy spojené s problematikou migrácie, ktorý bol realizovaný v rámci vypracovania PhD. dizertačnej práce Berová L. (2012).
Dotazník obsahuje niekoľko oblastí, ktoré špeciálne prezentujeme v príspevkoch
prvá časť výsledkov bola uverejnená v práci Berová L., Luha J., Žáková M. (2012) - I. postoje k imigrantom prichádzajúcim do SR ,
druhá v práci Luha J., Berová L., Žáková M. (2012a) II. začleňovanie imigrantov do spoločnosti,
tretia v práci Luha J., Berová L., Žáková M. (2012b) - III. čím nás môžu imigranti obohatiť,
štvrtá časť v práci Luha J., Berová L., Žáková M. (2012c) IV. čo by Vám prekážalo, keby?
piata časť v práci Luha J., Berová L., Žáková M. (2013a) V. ako vnímajú občania imigrantov a
šiesta Luha J., Berová L., Žáková M. (2013b) VI. podarilo sa imigrantom prispôsobiť sa životu na Slovensku?

Základná charakterizácia výskumu je samozrejme rovnaká: Terénna fáza celoslovenského reprezentatívneho výskumu bola realizovaná v období od polovice novembra 2011 do konca januára 2012 poučenými dobrovoľnými anketármi.

Základný súbor tvorilo 4 405 673 dospelých obyvateľov SR, t.j. 81,06% z 5 435 273 všetkých obyvateľov SR k 31.12.2010, podľa údajov Štatistického úradu SR (Vekové zloženie obyvateľstva SR v roku 2010. Demografická a sociálna štatistika. ŠÚ SR Bratislava).

Výberový súbor o rozsahu 1120 respondentov bol reprezentatívny podľa kontrolovaných znakov pohlavie, vek, kraj a aj podľa nekontrolovaného znaku vzdelanie.

V tomto príspevku prezentujeme názory dospelej populácie SR na ďalšiu časť otázok o názoroch na cudzincov/migrantov, ktorí prichádzajú na Slovensko. Tieto otázky boli zamerané na ochotu respondentov pomáhať imigrantom na Slovensku. Tejto téme boli venované tri otázky.:

- Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - prispieť pre vás primeranou finančnou čiastkou.
- Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - prispieť obnoseným šatstvom a inými vecami do domácnosti
- Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - pracovať s nimi ako dobrovoľník

Škála odpovedí 1=áno; 2=skôr áno; 3= skôr nie; 4=nie umožňuje počítať priemerné skóre odpovedí. Stred škály je hodnota 2,5, čo značí, že hodnoty pod 2,5 ukazujú na priaznivý „vzťah“ ku pomoci imigrantom a hodnoty vyššie deklarujú nepriaznivú ochotu pomáhať imigrantom. Podiel neodpovedajúcich je pod 1,61% v otázkach výskumu, čiže výsledky považujeme za relevantné. Základné štatistické charakteristiky sú v tabuľke 1.

Tabuľka 1. Základné štatistické charakteristiky

Otázka	N	Mean	Std. Deviation
Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - prispieť pre vás primeranou finančnou čiastkou	1104	**2.95**	.877
Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - prispieť obnoseným šatstvom a inými vecami do domácnosti	1113	**1.74**	.853
Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - pracovať s nimi ako dobrovoľník	1102	**3.00**	.912

Najväčšiu ochotu pomáhať imigrantom vyjadrili respondenti pri možnosti „prispieť obnoseným šatstvom a inými vecami do domácnosti“. Respondenti sú menej ochotní pomáhať imigrantom finančne a pracovať s nimi ako dobrovoľníci.

2. Špecifiká názorov na imigrantov podľa demografických znakov

Diferenciáciu názorov respondentov skúmame podľa demografických znakov doplnených o dve otázky „Žili ste niekedy viac ako tri mesiace v zahraničí?“ (***zahraničie***) a „Poznáte vo svojom blízkom okolí migranta z iného štátu, ktorý sa rozhodol žiť v SR?“ (***imigran***t). V tabuľkách a grafoch ich uvádzame skrátene ***„zahraničie“ a „imigrant“.***

Najskôr uvádzame prehľadnú tabuľku priemerných hodnôt za tri skúmané otázky podľa kategórií demografických znakov. Podrobnejšiu analýzu realizujeme v troch podkapitolách, vrátane štatistického testovania a grafickej prezentácie.

Ako sme už uviedli, priemerné hodnoty menšie ako 2,5 ukazujú na väčšiu ochotu pomáhať imigrantom a priemerné hodnoty väčšie ako 2,5 signalizujú zase na menšiu ochotu pomáhať imigrantom, podľa skúmanej otázky a kategórie demografického znaku.

Tabuľka 2. Priemerné hodnoty ochoty pomáhať imigrantom podľa demografických znakov

pohlavie	boli by ste ochotní prispieť pre vás primeranou finančnou čiastkou?	boli by ste ochotní prispieť obnoseným šatstvom a inými vecami do domácnosti?	boli by ste ochotní pracovať s nimi ako dobrovoľník?	n
muž	3.03	1.86	3.14	535
žena	2.88	1.63	2.86	585
základne	3.13	1.98	3.19	181
stredoškolské bez maturity	3.05	1.77	3.10	292
stredoškolské s maturitou	2.92	1.72	2.87	423
vysokoškolské	2.76	1.57	2.95	224
v 18-24	2.92	1.86	2.87	144
v 25-29	2.86	1.68	2.81	118
v 30-39	3.09	1.81	2.89	226
v 40-49	2.94	1.72	2.97	188
v 50-59	2.87	1.62	3.02	197
v 60+	2.98	1.76	3.26	247
s vyznaním	2.92	1.75	3.01	834
bez vyznania	3.07	1.74	2.96	284
študent SŠ	2.90	2.04	2.90	51
študent VŠ	2.78	1.66	2.72	83
pracujúci	2.90	1.67	2.94	657
žena na MD, RD	2.98	1.75	2.96	48
nezamestnaný	3.20	1.99	3.14	93
dôchodca	3.10	1.82	3.28	188
Bratislavský	2.70	1.65	2.90	135
Trnavský	2.92	1.50	2.91	118
Trenčiansky	3.08	1.69	3.00	135
Nitriansky	2.91	1.76	3.07	149
Banskobystrický	3.02	1.76	2.98	135
Žilinský	3.18	1.98	3.13	142
Prešovský	2.91	1.85	3.03	157
Košický	2.92	1.71	2.94	149
viac ako 3mes. v zahraničí,áno	2.87	1.74	2.87	205
viac ako 3mes. v zahraničí,nie	2.97	1.74	3.02	908
imigrant v blízkom okolí,áno	2.81	1.64	2.86	462
imigrant v blízkom okolí,nie	3.06	1.82	3.10	656

2.1. Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - prispieť pre vás primeranou finančnou čiastkou

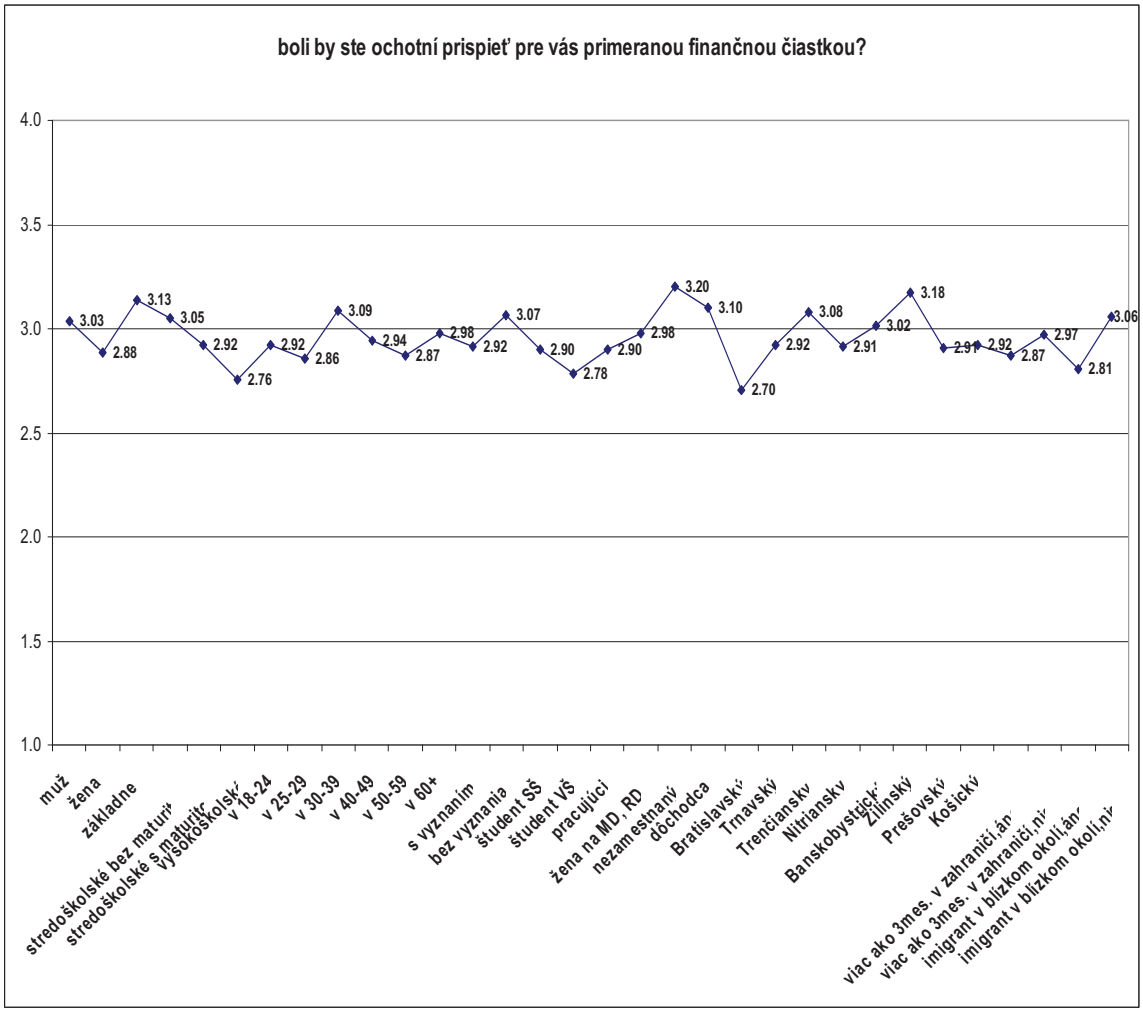
V tabuľke 3. sú P-hodnoty Chí-kvadrát testov, ktoré sme použili pri komparácii danej otázky podľa demografických znakov, rozšírených o hore uvedené dve otázky. Štatisticky signifikantné sú diferencie podľa pohlavia, vzdelania, viery, ekonomickej aktivity a podľa toho, či poznajú imigranta v blízkom okolí. Tí čo žili viac ako 3 mesiace v zahraničí sa v ochote pomôcť imigrantom finančným prispením nelíšia štatisticky signifikantne a diferenciácia podľa vekových kategórií nie je štatisticky signifikantná. Graf 1. tieto

výsledky vyjadruje plasticky. Miera ochoty prispieť finančnou čiastkou kolíše podľa kategórií demografických znakov okolo hodnoty 3, ktorá znamená „skôr nie“.

Tabuľka 3. P-hodnoty testov

	pohlavie	vzdelanie	vek	viera	ek.aktivita	kraj	zahraničie	imigrant
P-hodnoty	**0.023**	**0.003**	0.101	**0.037**	**0.003**	**0.003**	0.467	**0.000**

Graf1. Ochota prispieť imigrantom finančnou čiastkou podľa demografických znakov



2.2. Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - prispieť obnoseným šatstvom a inými vecami do domácnosti

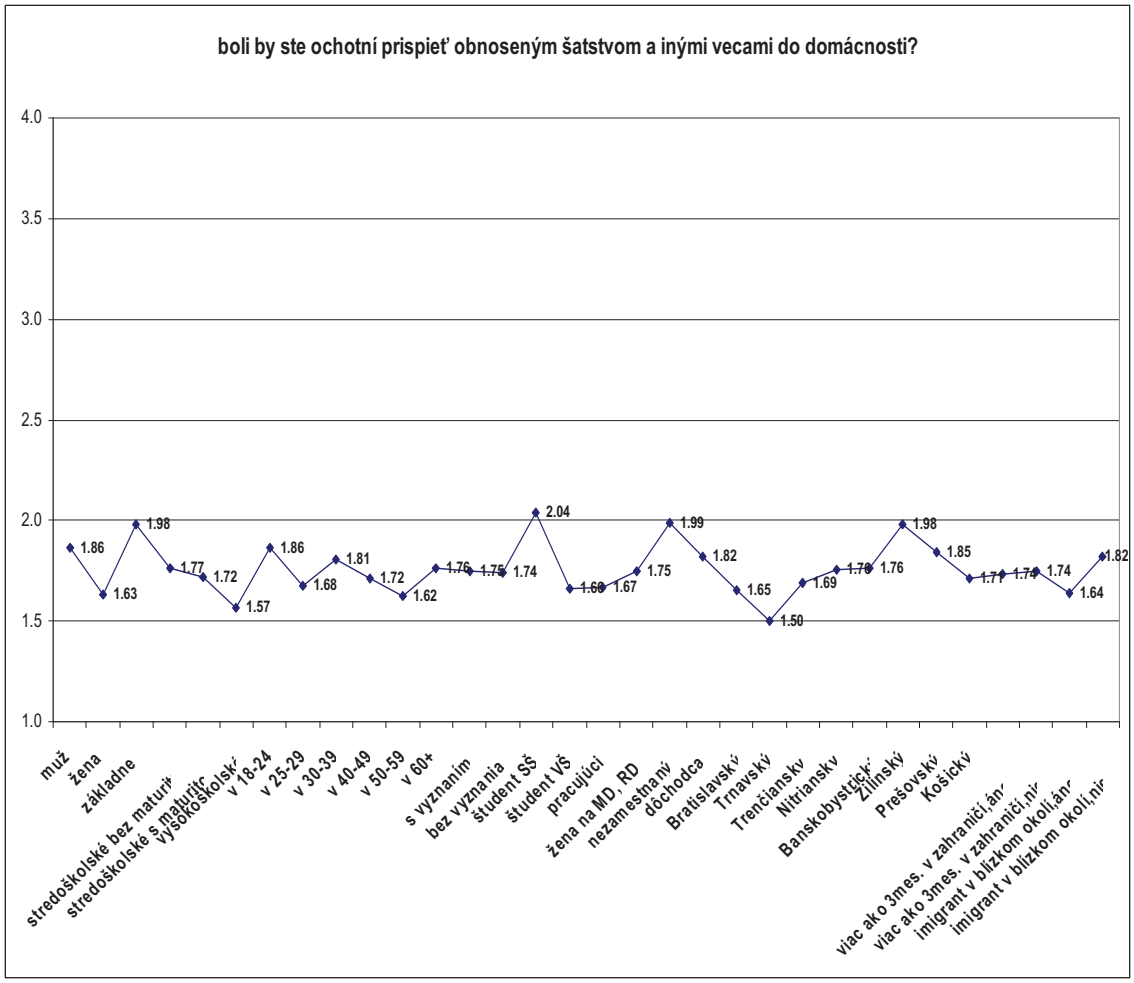
Výsledky testovania sú v tabuľke 4, ako P-hodnoty Chí-kvadrát testu skúmanej otázky podľa demografických znakov. Diferenciácia ochoty respondentov prispieť obnoseným šatstvom a inými vecami do domácnosti imigranta podľa demografických znakov je vidno v grafe 2. Štatisticky signifikantne sa líšia odpovede podľa pohlavia, vzdelania, ekonomickej aktivity, kraja a podľa toho či pozná v svojom okolí imigranta. Blízko hranice štatistickej signifikantnosti sú názory respondentov podľa veku a štatisticky nesignifikantne sa líšia

názory podľa vierovyznania a podľa toho či bol respondent dlhšie v zahraničí. Z grafu 2 zistíme, že miera ochoty prispieť obnoseným šatstvom a inými vecami do domácnosti imigrantov je väčšia, kolíše podľa kategórií demografických znakov približne v intervale od 1,5 až 2, kde 2 značí „skôr áno".

Tabuľka 4. P-hodnoty testov

	pohlavie	vzdelanie	vek	viera	ek.aktivita	kraj	zahraničie	imigrant
P-hodnoty	0.000	0.002	0.108	0.998	0.006	0.002	0.945	0.000

Graf 2. Ochota prispieť imigrantom obnoseným šatstvom a vecami do domácosti podľa demografických znakov



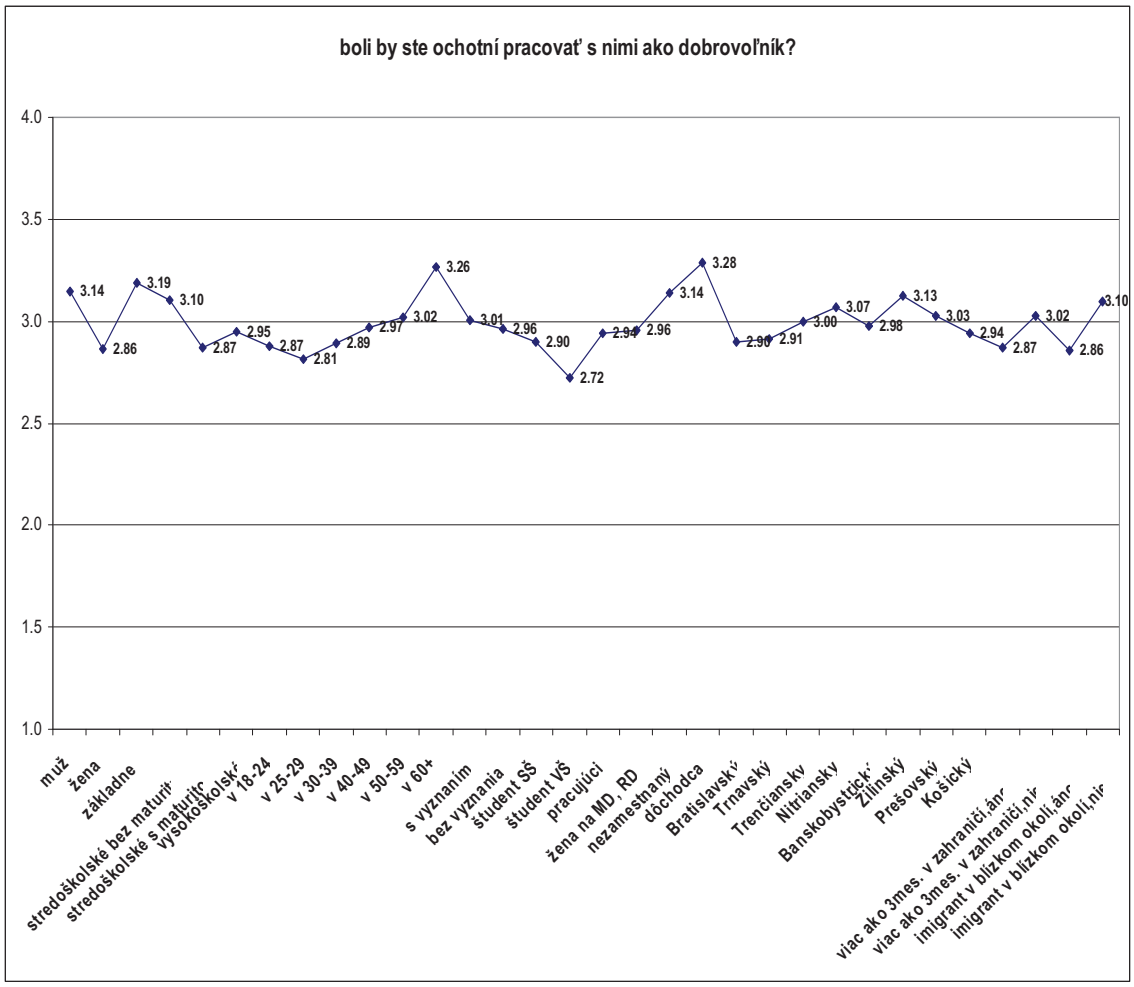
2.3. Ak by boli takéto činnosti zamerané na pomoc migrantom_cudzincom pri ich začleňovaní sa do spoločnosti, boli by ste ochotní? - pracovať s nimi ako dobrovoľník

Diferenciácia názorov respondentov pri ochote pracovať s imigrantmi ako dobrovoľník je podľa demografických znakov štatisticky signifikantná, okrem vierovyznania a kraja. Výsledky sú v tabuľke 5 a v grafe 3. Ochota pracovať s imigrantmi ako dobrovoľník je približne na rovnakej úrovni ako pri ochote prispieť finančnou čiastkou, dokonca je mierne nižšia. Taktiež kolíše podľa kategórií demografických znakov okolo hodnoty 3 „skôr nie".

Tabuľka 5. P-hodnoty testov

	pohlavie	vzdelanie	vek	viera	ek.aktivita	kraj	zahraničie	imigrant
P-hodnoty	**0.000**	**0.000**	**0.000**	0.209	**0.000**	0.423	**0.005**	**0.000**

Graf 3. Ochota pracovať s imigrantmi ako dobrovoľník podľa demografických znakov



3. Závery

Dôležitým aspektom v živote imigrantov je pomoc obyvateľov hosťujúcej krajiny. Názory „domácej populácie“ na imigrantov sú podmienené mnohými okolnosťami. V tomto príspevku sme sa zaoberali špecifickými pohľadmi na pomoc imigrantom zo strany domáceho obyvateľstva. Ochota pomáhať imigrantom má rôznu mieru, podľa oblasti pomoci. Z troch skúmaných „oblastí“ pomoci, bola miera ochoty respondentov pomáhať imigrantom pri pomoci darovaním obnoseného šatstva a inými vecami do domácnosti imigrantov. Ochota pomáhať finančným prispením a pracovať s nimi ako dobrovoľník sa dá charakterizovať „skôr nie“.

4. Literatúra

[1] Berová L. (2012): Názory verejnosti na migrantov a ich integráciu do spoločnosti. PhD. dizertačná práca. Trnavská univerzita v Trnave, Fakulta zdravotníctva a sociálnej práce. Trnava 2012.

[2]Berová L., Luha J., Žáková M. (2012): Názory verejnosti na migrantov a ich integráciu v SR: I. postoje k imigrantom prichádzajúcim do SR. FORUM STATISTICUM SLOVACUM 3/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[3]Kubanová, J.: Statistické metody pro ekonomickou a technickou praxi. Statis, Bratislava 2008. Vydání třetí – doplněné. ISBN 978- 80-85659-47-4. pp 245.

[4] Linda, B.: Pravděpodobnost. Monografie. Univerzita Pardubice, Pardubice 2010. ISBN 978-80-7395-303-4. pp 168.

[5] Luha, J. (1985): Testovanie štatistických hypotéz pri analýze súborov charakterizovaných kvalitatívnymi znakmi. STV Bratislava 1985.

[6] Luha J., Berová L., Žáková M. (2012a): Názory verejnosti na migrantov a ich integráciu v SR: II.začleňovanie imigrantov do spoločnosti. FORUM STATISTICUM SLOVACUM 4/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[7]Luha J., Berová L., Žáková M (2012b).: Názory verejnosti na migrantov a ich integráciu v SR: III. čím nás môžu imigranti obohatiť. FORUM STATISTICUM SLOVACUM 6/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[8]Luha J., Berová L., Žáková M. (2012c): Názory verejnosti na migrantov a ich integráciu v SR: IV. čo by Vám prekážalo, keby? FORUM STATISTICUM SLOVACUM 7/2012. SŠDS Bratislava 2012. ISSN 1336-7420.

[9]Luha J., Berová L., Žáková M. (2013a): Názory verejnosti na migrantov a ich integráciu v SR: V. ako vnímajú občania imigrantov. FORUM STATISTICUM SLOVACUM 3/2013. SŠDS Bratislava 2013. ISSN 1336-7420.

[10] Luha J., Berová L., Žáková M. (2013b) VI. podarilo sa imigrantom prispôsobiť sa životu na Slovensku? FORUM STATISTICUM SLOVACUM 4/2013. SŠDS Bratislava 2013. ISSN 1336-7420.

[11] Luha J. (2007): Kvótový výber. FORUM STATISTICUM SLOVACUM 1/2007. SŠDS Bratislava 2007. ISSN 1336-7420.

[12] Luha J. (2009): Matematicko-štatistické aspekty spracovania dotazníkových výskumov. FORUM STATISTICUM SLOVACUM 3/2009. SŠDS Bratislava 2009. ISSN 1336-7420.

[13] Luha J. (2010): Metodologické zásady záznamu dát z rozličných oblastí výskumu. FORUM STATISTICUM SLOVACUM 3/2010. SŠDS Bratislava 2010. ISSN 1336-7420.

[14]Pecáková I.(2008): Statistika v terénních průzkumech. Proffessional Publishing, Praha 2008. ISBN 978-80-86946-74-0.

[15]Řezanková A.(2007): Analýza dat z dotazníkových šetření. Proffessional Publishing, Praha 2007. ISBN 978-80-86946-49-8.

Adresy autorov:

Ján Luha, RNDr., CSc.
Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a UN Bratislava
jan.luha@fmed.uniba.sk

Lenka Berová, Ing.,PhD.
Katedra sociálnej práce
FZaSP, Trnavská univerzita
lenka.berova@gmail.com

Martina Žáková, doc. PhDr., PhD., Katedra sociálnej práce, FZaSP, Trnavská univerzita
martina.zakova@truni.sk

Měření podobnosti překladů básně Havran

Measuring the similarity of translations of poem Raven

Jaroslav Marek, Jan Šlahora

Abstract: We will take a look at possibilities of using statistical method and mathematical linguistics for measuring the similarity of texts. By help of linguistics characteristics we will measure the similarities of poems and we will try to recognize authors. For measurement of distances we will use methods of cluster analysis and method of principal components. Computation was realized on several czech translation of famous E.A.Poe's poem Raven.

Abstrakt: V článku upozorníme na možnosti využití statistických metod a matematické lingvistiky pro měření podobnosti textů. Pomocí lingvistických charakteristik budeme měřit podobnost básní a pokusíme se o rozpoznávání autorství. K měření podobnosti budou použity metody shlukové analýzy a metoda hlavních komponent, viz [6]. Výpočty budou realizovány na několika překladech slavné básně E. A. Poea Havran.

Keywords: mathematical linguistics, phonics variables, agrégation, alliteration, similarities of texts, cluster analysis, principal components analysis, confidence region, E.A.Poe, translations of poem Raven

Klíčová slova: matematická lingvistika, fónické veličiny, agregace, aliterace, podobnost textů, shluková analýza, metoda hlavních komponent, konfidenční oblast, E. A. Poe, překlady básně Havran

JEL classification: C02, C38,

1. ÚVOD

Analýza textů zahrnuje velké množství jevů, které je možné sledovat, a to na různých textových útvarech. Základní stavební kámen každého textu je písmeno. U písmen můžeme sledovat například jejich frekvenci výskytu v textu. Tato frekvence může být využita v dalších výpočtech u jiných jevů. Dále se zejména u poezie sledují fonické jevy, které zkoumají tzv. libozvuk – eufonii. Ten je dán uspořádáním a následným opakováním hlásek (především samohlásek) ve verši. V této práci budeme z fonických jevů měřit agregaci a aliteraci, viz [2, 3]. Dále se používá také asonance, se kterou ale pracovat nebudeme.

Můžeme si položit otázku: „Lze pomocí lingvistických jevů rozlišit autorství textu, resp. rozpoznat, jakým jazykem je text napsaný?"

Podobnost básní je diskutována v části fiktivního rozhovoru antických filosofů Sokrata a Platóna v knize Dialogy o matematice Alfreda Rényiho.

Sokrates: „...žádní dva básníci, kteří se neznají a navzájem o sobě nic nevědí, nenapíšou nikdy – ani o téže věci – stejnou báseň?"

Hippokrates: „ Nevím. To nedokážu vysvětlit. ... že by dva básníci napsali stejnou báseň, to jsem opravdu ještě nikdy neslyšel.“

Odpověď se pokusíme najít pomocí metod mnohorozměrné statistické analýzy aplikovaných na vypočtených lingvistických charakteristikách různých překladů básně Havran od E. A. Poea.

Měření a shlukování provedeme pro následující české překlady: 1) Šembera V. K., 2) Vrchlický J. 1890, 3) Mužík A. E., 4) Lutinov K. D., 5) Nezval V., 6) Babler O. F., 7) Taufer J., 8) Stoklas E., 9) Wagnerová D., 10) Havel R., 11) Čapek J. B., 12) Resler K., 13) Černý R., 14) Slavík I., 15) Kadlec S., 16) Bejblík A., 17) Vrchlický J. 1881, 18) Havel R. 1954.

2. AGREGACE A ALITERACE

V této části seznámíme čtenáře s agregací a aliterací.

2.1 Agregace

Agregace se pokouší změřit, zda věty (nebo v této práci verše) ležící blíže u sebe jsou si fonicky podobnější než ty, jejichž vzdálenost je větší.

Obvykle se hledají odhady neznámých parametrů *a* a *b* aproximační funkce ve tvaru $y = a \cdot x^{-b}$, která je řešením diferenciální rovnice

$$\frac{dy}{y} = -b \cdot \frac{dx}{x} \tag{2.1}$$

kde dx/x je relativní změna vzdálenosti, dy/y je relativní změna podobnosti, b je koeficient proporcionality. Platí $b > 0$.

Samotnou podobnost veršů vypočítáme tak, že vytvoříme pro dva verše dvě dvojice množin, množiny A_i, A_j a B_i, B_j, kde i je číslo prvního verše a j číslo druhého. V množině A_i budou abecedně uspořádané fonémy z prvního verše, v množině A_j pak fonémy verše dalšího. V množinách B_i a B_j budou uloženy dvojice po sobě jdoucích foném.

Chceme-li zkoumat agregaci například u prvních dvou veršů překladu Havrana od Augustina Eugena Mužíka

V půlnoc jednu pustou, tmavou,
když jsem se skloněnou hlavou

vytvoříme množinu $A_1 = \{a, c, d, e, j, l, m, n, n, o, p, s, t, u, ů, u, v\}$ a množinu $A_2 = \{a, d, e, ě, h, j, k, l, m, n, ň, o, s, u, v, y, ž\}$. Obě množiny mají po třinácti prvcích. Množina B_1 obsahuje tyto dvojice foném z prvního verše:
$B_1 = \{av, cj, dn, ed, je, ln, ma, no, nu, oc, ou, ou, pů, pu, st, to, tm, up, us, ůl, vo\}$ a

množina
$B_2 = \{av, dy, em, es, ěn, hl, js, kd, lo, la, ms, no, on, ou, ou, se, sk, uh, vo, yž, žj\}$.

První množiny obsahuje 15 a druhá množina obsahuje 16 dvojic foném.

Vzorec pro výpočet podobnosti veršů i a j je dán jako

$$S_{ij} = 100 \cdot \left(\frac{|A_i \cap A_j|^2}{|A_i| \cdot |A_j|} + \frac{|B_i \cap B_j|^2}{|B_i| \cdot |B_j|} \right) \tag{2.2}$$

kde $X_i \cap X_j$ – průnik množin X_i a X_j , $|X_i|$ je počet prvků v množině X_i.

Po dosazení do vzorce dostaneme $S_{1,2} = 100 \cdot \left(\frac{10^2}{15 \cdot 16} + \frac{3^2}{20 \cdot 20)} \right) \cong 43{,}92$.

Pro výpočet agregace je třeba vyhodnotit více párů veršů v různých vzdálenostech $k = 1,2, \dots, n$ a každý posun charakterizovat průměrem naměřených hodnotou $\overline{S_i}$. Získané hodnoty agregací pro různé vzdálenosti veršů se pak aproximují funkcí $y = a \cdot x^{-b}$.

2.2 Aliterace

Aliterace je literární technika, kdy se na začátku slov ve verši nebo na začátku samotných veršů opakuje stejná hláska. Příkladem mohou být verše z Bablerova a Tauferova překladu:

„Jistě je to tady," dím si, „zní to z okna, zní to z římsy,
***j**en v něm **j**edno slůvko tálo - **j**méno Lenořino, k níž*

Z pohledu analýzy textu není rozhodující, zda slova začínající stejným písmenem leží ve verši hned za sebou, či se mezi nimi nachází jedno nebo více slov začínajících na jiné písmeno. Díky tomu se aliterace vyskytuje v poezii poměrně často. Zajímavé je však sledovat, zda se jedná o náhodný jev nebo o nějakou charakteristiku básně nebo autora.
Základem pro výpočet aliterace jsou pravděpodobnosti výskytu jednotlivých hlásek p_i ($i = 1,2, \dots, k$, kde k je počet hlásek v textu). Odhady těchto pravděpodobností je možné získat buď z většího množství textu psaném v daném jazyce, případně je možno vypočítat pravděpodobnosti výskytu přímo ze zkoumaného textu. Pravděpodobnost toho, že ve verši najdeme x-násobnou aliteraci, kdy na hlásku i začíná x slov a všechna ostatní začínají na písmeno jiné, se vypočítá podle vzorce

$$P(X_i = x) = \binom{n}{x} \cdot p_i^x \cdot q_i^{n-x} \tag{2.3}$$

kde n – počet slov ve zkoumaném verši,

$\binom{n}{x}$ – kombinační číslo, které se vypočítá jako $\frac{n!}{n! \cdot (n-x)!}$,

p_i – pravděpodobnost výskytu hlásky i, $q_i = 1 - p_i$.

Pro výpočet aliterace však není důležitá pouze pravděpodobnost výskytu x-násobné aliterace v daném verši, ale i pravděpodobnosti výskytu extrémnějšího jevu, kdy $X_i \geq x$. Tato pravděpodobnost se vypočítá podle vztahu

$$P(X \geq x) = \sum_{r=x}^{n} \binom{n}{r} \cdot p_i^r \cdot q_i^{n-r} \tag{2.4}$$

Pro charakterizování aliterace v daném verši definujeme koeficient KA:

$$KA = \begin{cases} 100 \cdot [\alpha - P(X \geq x)], & pro\ \alpha > P(X \geq x) \\ 0, & pro\ \alpha \leq P(X \geq x) \end{cases} \tag{2.5}$$

Jiná situace by nastala v případě vícenásobné aliterace, to je v případě, že ve verši aliterují dvě (nebo více) hlásek, jako například v prvním verši originálu básně Havran: *„And my **s**oul **f**rom **o**ut that **s**hadow that lies **f**loating **o**n the **f**loor“*, kde je celkem $n = 13$ slov, z nichž dvě začínají hláskou „s“, dvě hláskou „o“ a tři hláskou „f“ (zbývajících šest potom libovolnými jinými hláskami). Pravděpodobnost výskytu trojnásobné aliterace je pomocí multinomického rozdělení vyjádřena jako

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = v) = \frac{n!}{x_1! \cdot x_2! \cdot x_3! \cdot v!} \cdot p_1^{x_1} \cdot p_2^{x_2} \cdot p_3^{x_3} \cdot (1 - p_1 - p_2 - p_3)^v \tag{2.6}$$

kde $v = n - x_1 - x_2 - x_3$.

Pro výpočet koeficientu KA je však potřeba, stejně jako v případě jednonásobné aliterace, vypočítat i pravděpodobnost výskytu $P(X_1 = y_1, X_2 = y_2, X_3 = w)$, kde $w = n - y_1 - y_2$, $x_1 \leq y_1 \leq n$, $x_2 \leq y_2 \leq n$ a $y_1 + y_2 \leq n$.

Pro porovnání aliterace u různých básní (například výběr básní jednoho autora) je možno vyjádřit aliterační charakter jako průměr všech KA určený pro počet n veršů v básni.

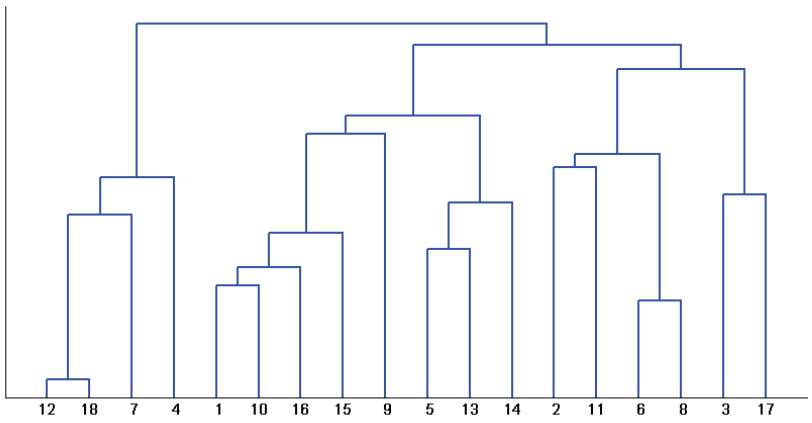
3. MĚŘENÍ CHARAKTERISTIK

Pomocí vytvořené počítačové aplikace viz [7] jsme určili u zvolených překladů resp. jejich strof průměrnou délku věty d_V, průměrnou délku slova d_S, hodnoty agregací při posunu jedna až pět, hodnotu aliterace *AL*. Změřené hodnoty u zvolených překladů básně Havran a u jednotlivých 18 slok určené pro uvedených k=12 lingvistických charakteristik jsou uvedeny na webové stránce [7]. V následující tabulce uvedeme příklad změřených hodnot pro první strofu.

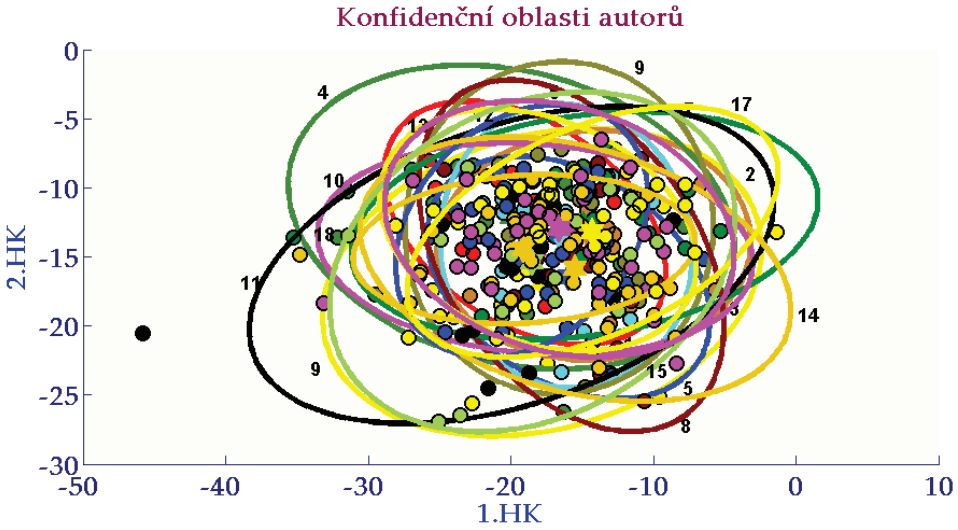
Tab. 1: Změřené charakteristiky 1. strofy překladů básně Havran

překlad	AG_1	AG_2	AG_3	AG_4	AG_5	AL	d_S	d_V
1	19,01	2,8	2,61	0,37	4,38	2,68	3,80	17,33
2	25,36	6,70	2,27	1,11	0,70	3,64	3,96	16,00
3	23,91	8,83	6,32	0,17	0,00	1,19	3,95	13,33
4	25,94	6,63	6,06	2,00	0,63	2,96	4,26	6,71
5	7,83	6,47	1,79	2,08	9,18	1,4	4,24	13,00

Samozřejmě nemůžeme získaná měření graficky zobrazit. Proto jsme provedli hierarchické shlukování s použitím metody nejvzdálenějšího souseda, viz dendrogram na obrázku 1.



Obr. 1: Dendrogram 18 překladů – hierarchické shlukování



Obr. 2: 1. a 2. hlavní komponenty 18 strof 18 překladů a konfidenční elipsy překladů

Na měření u vybraných strof byla dále aplikována metoda hlavních komponent, viz [6], která umožňuje redukci dimenze dat a transformaci měření do prostoru menší dimenze. Na obr. 2 jsou určeny ze všech 18 strof konfidenční elipsy, viz [4, 5] pro jejich první dvě hlavní komponenty. Je zřejmé, že konfidenční elipsy se výrazně neliší a básníky nejsme schopni rozlišit.

4. ZÁVĚR

V dendrogramu jsme mohli očekávat, že se budou shlukovat dva překlady od J. Vrchlického, resp. dva překlady R. Havla. Tento předpoklad se ale nenaplnil. Konfidenční elipsy zkonstruované u jednotlivých překladů z hodnot hlavních komponent jejich 18 strof se příliš neliší. Zdá se, že výsledky shlukování změřených zvolených charakteristik překladů neposkytují možnost rozlišit autorství.

5. LITERATURA

[1] POE, E.A.: Havran, šestnáct českých překladů, ODEON, Praha, 226 s. 1990

[2] WIMMER, G., ALTMANN, G., HŘEBÍČEK, L., ONDREJOVIČ, S., WIMMEROVÁ, S., Úvod do analýzy textov, VEDA, Bratislava, 345 s. 2003.

[3] ANDRES, J., KUBÁČEK, L., MACHALOVÁ, J., TUČKOVÁ, M.: Optimization of parameters in the Menzerath–Altmann law, Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica 51, 1 (2012), p. 5-27

[4] WIMMER, G. , PALENČÁR, R. , WITKOVSKÝ, V. Spracovanie a vyhodnocovanie meraní. Bratislava: VEDA, 2002. 187 s. ISBN 80-224-0734-8

[5] KUBÁČEK, L., KUBÁČKOVÁ, L. Statistika a metrologie. Vydavatelství Univerzity Palackého v Olomouci, Olomouc 2000.

[6] ŘEZÁNKOVÁ, H., HÚSEK, D., SNÁŠEL, V. Shluková analýza dat, PBtisk Příbram, 2007. 196 s. ISBN 978-80-86946-26-9

[7] MAREK, J.: Úložiště. http://jarekmarek.uloziste.sweb.cz/poe.html [online: 20. 06. 2013]

Adresa autorů:

Mgr. Jaroslav Marek, Ph.D.
Univerzita Pardubice,
Fakulta elektrotechniky a informatiky,
náměstí Čs. legií 565,
530 02 Pardubice,
email: jaroslav.marek@upce.cz

Bc. Jan Šlahora
Univerzita Pardubice,
Fakulta elektrotechniky a informatiky,
náměstí Čs. legií 565,
530 02 Pardubice,
email: jan@slahora.cz

Differential Equations of Selected Biological Structures
Diferenciálne rovnice vybraných biologických štruktúr

Mária Minárová, Jozef Sumec

Abstract: The aim of the paper is to present two different types of mathematical and mechanical models of biological processes. In the introduction part of the paper the general principles of mechanical and mathematical modeling are introduced. The mathematical elaboration includes the governing equation, boundary and initial conditions assessing, the appropriate method selecting and the computer implementation. In the paper two models governed by differential equations are introduced. The first one describes the mathematical modeling of the motion segment (two adjacent vertebrae and the intervertebral disc between) of the human lumbar spine. The elastic response of the biological domain (motion segment) is governed by the Lamé equations. The second one, the rheological model describes the behavior of material which is neither wholly elastic nor wholly plastic.

Abstrakt: V článku predstavujeme dva typy mechano – matematických modelov v biológii. V úvode článku sú uvedené všeobecné princípy mechano – matematického modelovania. Matematické rozpracovanie zahŕňa stanovenie riadiacich rovníc, určenie okrajových a začiatočných podmienok, výber a realizáciu metódy riešenia a počítačovú implementáciu. V článku sú predstavené modely dvoch biomechanických dejov.

Prvý popisuje správanie sa pohybového segmentu (skladba dvoch susedných stavcov a medzistavcovej platničky medzi nimi) ľudskej chrbtice. Mechanická odozva oblasti (pohybový segment) sa riadi diferenciálnymi rovnicami - Lamého rovnice.

Druhý model – reologický model – popisuje správanie sa materiálov, ktoré nikdy nie sú dokolnale pružné, ani dokolnale väzké.

Key words: elasticity equations, finite element modeling, motion segment, axisymmetry, rheology, viscoelastic behavior, creep, relaxation

Kľúčové slová: rovnice pružnosti, konečno prvkové modelovanie, pohybový segment, rotačná symetria, väzkopružnosť, dotvarovanie relaxácia

JEL classification: C60

Introduction – general modeling, principles

Biomechanics is the science dealing with the mechanical properties and problems as statics, kinematics, dynamics, biorheology, bioviscoelasticity, biotolerance and life ability of living systems based on the general methods of continuum mechanics. [4, 19].

Principal problems of human biomechanics are connected to the mechanics of healthy organism – motional apparatus, bone-muscular system, cardio-vascular system, ventilation, excretive apparatus, heat and mass transfer phenomenon, bioenergy, auricular and vestibular apparatus, peristaltic transport, joint tribology, mechanical properties of biomaterials, biocompatibility, sport and forensic biomechanics, etc.

The biomechanical modeling of mechanical properties of solid or liquid material at all, sources from thermodynamics laws of the open systems, which are complementary to the macroscopic deterministic event laws e.g. [18]:

- Mass balance law

$$\frac{d}{dt}\int_V \rho \, dV = 0\,, \tag{1}$$

t - time, V - volume, ρ - density

- Momentum balance law (1st Euler motion equation)

$$\frac{d}{dt}\int_V \rho\,\vec{v}\;dV = \vec{F} = \int_A \vec{t}\;dA + \int_V \vec{b}\,\rho\,dV \tag{2}$$

$\vec{v}$ velocity vector, $\vec{F}$ - resultant force vector, $\vec{t}$ - force per area unit vector, $\vec{b}$ - force per volume unit vector

- Angular momentum balance law (2nd Euler motion equation)

$$\frac{d}{dt}\int_V \left(\vec{p}\times\vec{v}\right)\rho\;dV = \vec{M} = \int_A \left(\vec{m} + \vec{p}\times\vec{t}\right)dA + \int_V \left(\vec{l} + \vec{p}\times\vec{b}\right)\rho\,dV \tag{3}$$

$\vec{p}$ - position vector, $\vec{M}$ - resultant vector of the angular momentums acting to the body, $\vec{m}$ - angular momentum per body surface unit vector, $\vec{l}$ - angular momentum per mass unit vector

- First thermo dynamical law – energy balance

$$\frac{d}{dt}K + U = W + \sum_\alpha H_\alpha \tag{4}$$

K - kinetic energy, U - inner energy, W - work of the outer forces per time unit, H_α other non-mechanical energy added to the body per time unit

- Second thermo dynamical law

$$\left(\frac{dS}{dt}\right)_V = \frac{dS}{dt} - \int_A \frac{q_i n_i}{T}\,dA - \int_V \frac{\rho h}{T}dV \geq 0 \tag{5}$$

S- global entropy of the body, $\left(\frac{dS}{dt}\right)_V$ - inner entropy production, n_i - unit normal component vector to the surface of the body, q_i heat flow vector component, h – flow coefficient of non-mechanical energy per body mass unit per time unit, T - temperature.

The law states that the entropy production is always nonnegative.

General theory of material constitutive equations of nonliving nature has to follow the physical principles [1]: causality axiom, deterministic axiom, uniform representation axiom, objectivity axiom, invariance axiom, surroundings axiom, memory axiom. The paper involves phenomenological description of mechanical phenomenon, which was elaborated by [2, 3, 17], for composite materials [7, 15], for poroelastic materials [11]. The common base for the phenomenological description is classical theory of continuum field, where we use for the description the arithmetic Euclidean space and continuum environment.

1. Degrees of freedom, stress, strain

Degree of freedom (DOF) of a mechanical system is the number of independent parameters that define its configuration. It is the number of parameters that determines the state of a physical system and is important to the analysis of systems of bodies; see Fig.1, [22]. Degrees of freedom is the number of independent motions and rotations that are allowed to the body or, in case of a mechanism made of several bodies, number of possible independent relative motions between the pieces of the mechanism.

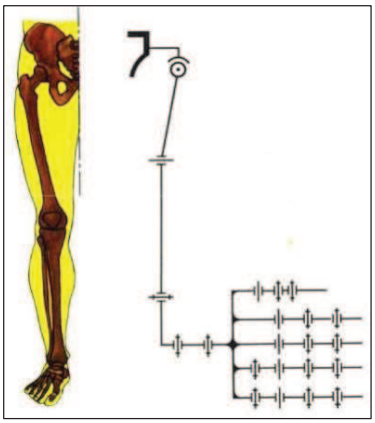


Fig. 1: Degrees of freedom in the human leg

The force that deforms a solid body can cause its compressing, stretching, squeezing, bending or twisting. *Stress* (force per area unit) - *strain* (prolongation/shortening per longitude unit) relationship can be represented by stress – strain diagram. The diagram is specific for each specific elastic material, see Fig.2.

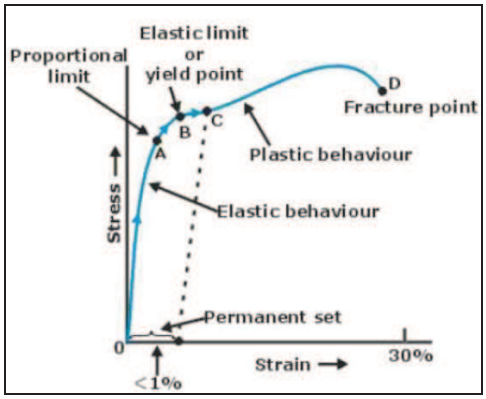


Fig. 2: Stress - strain diagram of a material [22]

Every elastic solid material stress – strain diagram has its linear part, i.e. small force causes the reversible displacement and both are proportional. Hooke's law, see Fig.3, describes this phenomenon:

$$\mathbf{F} = k\mathbf{x}\,, \quad \text{i.e.} \qquad \sigma = E\varepsilon\,, \tag{6}$$

where F[N] is an acting force, k[N/m], E[Pa] - elastic coefficients of a material, σ[Pa] stress, ε[-] strain (relative deformation), $\mathbf{x}$ – position vector

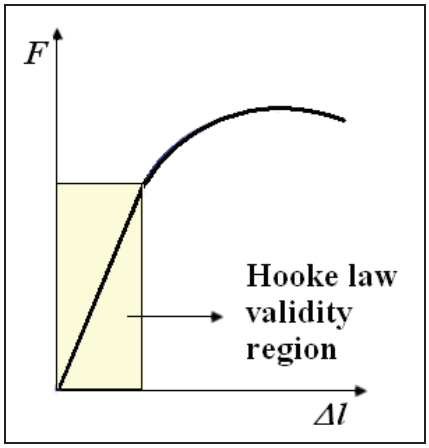


Fig. 3: Stress versus strain of a material, its linear part - elastic range – region of Hooke's law validity [22]

2. Elasticity equations

2.1. Stress tensor, volume forces, balance equation

Let us take an arbitrary point *P* in the rigid body and detach a volume element – prism with one vertex in *P* and dimensions dx_1, dx_2, dx_3. The rest of the body affects to the volume element by a force / stress. Denoting particular stress components as on the Fig.4, the components leading perpendicularly to the volume element planes $\sigma_{11}, \sigma_{22}, \sigma_{33}$ are called *normal* components of stress, whereas the components lying in the border planes of differential element $\sigma_{12}, \sigma_{21}, \sigma_{13}, \sigma_{31}, \sigma_{23}, \sigma_{32}$ are called *shear (tangential)* components of stress. (For the sake of depicting the situation synoptically, just three of six prism walls are drawn on the Fig.4)

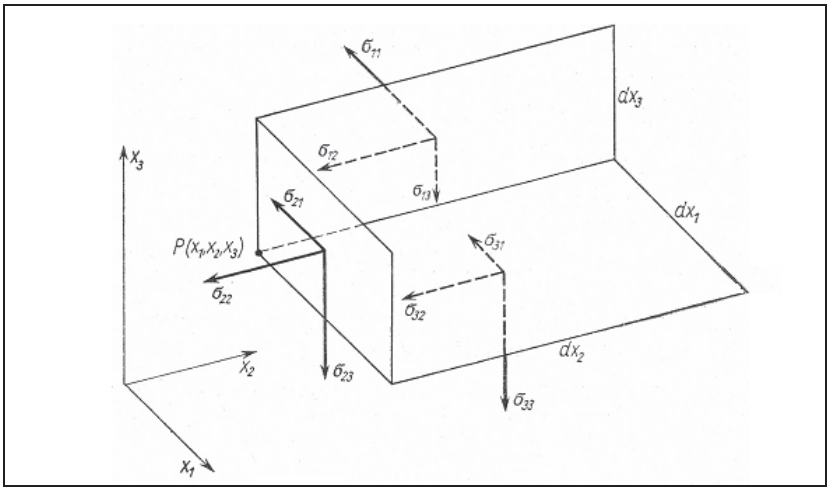


Fig. 4: Three planes (x=const, y=const and z=const) passing through point P and stress vector notation [8]

Hence, the entire element stress components can be collected in the stress tensor **s**:

$$\mathbf{s} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} = (\sigma_{ik}), \quad i,k = 1,2,3 \tag{7}$$

When denoting $\mathbf{K} = (X_1, X_2, X_3)$ volume force acting in centre of gravity, we can summarize all forces in the body in so-called balance equations:

$$\begin{aligned} \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{21}}{\partial x_2} + \frac{\partial \sigma_{31}}{\partial x_3} + X_1 = 0 \\ \frac{\partial \sigma_{12}}{\partial x_1} + \frac{\partial \sigma_{22}}{\partial x_2} + \frac{\partial \sigma_{32}}{\partial x_3} + X_2 = 0 \\ \frac{\partial \sigma_{13}}{\partial x_1} + \frac{\partial \sigma_{23}}{\partial x_2} + \frac{\partial \sigma_{33}}{\partial x_3} + X_3 = 0 \end{aligned} \tag{8}$$

2.2. Displacement, strain tensor, geometrical equations

Let us denote x_1, x_2, x_3 the coordinates of the arbitrary point of the observed solid body. Taking two nearby points x_i and $x_i + dx_i$, i=1,2,3 in this solid body with the distance *dl* between them before deformation and tracing their position after deformation – the point x_i moved to the new position x_i^* (the translation vector is $\vec{v}$), the point $x_i + dx_i$ moved to the new position $x_i^* + dx_i^*$ (the translation vector is $\vec{v} + d\vec{v}$), see Fig.5.

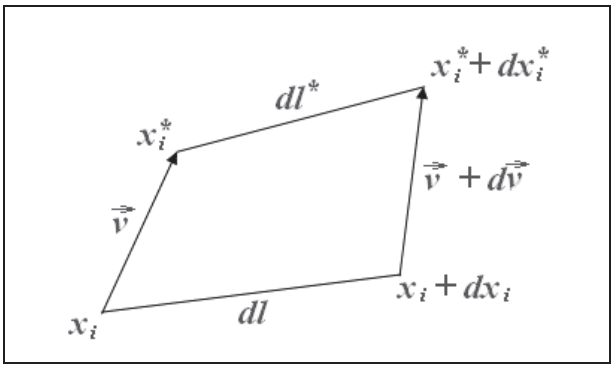


Fig. 5: Two points before and after deformation

The square of the distance between x_i and $x_i + dx_i$ is

$$dl^2 = dx_1^2 + dx_2^2 + dx_3^2 \tag{9}$$

and

$$dl^{*2} = dx_1^{*2} + dx_2^{*2} + dx_3^{*2} = \sum_{i=1}^{3} (dx_i + dv_i)^2 \tag{10}$$

By using the Taylor series expansion and neglecting the higher order derivatives as from the second order derivative we get the approximation of displacement vector increment $d\vec{v}$:

$$dv_i = \frac{\partial v_i}{\partial x_1} dx_1 + \frac{\partial v_i}{\partial x_2} dx_2 + \frac{\partial v_i}{\partial x_3} dx_3 , \quad i = 1,2,3 \ . \tag{11}$$

Then

$$dl^{*2} = \sum_{i=1}^{3} dx_i^2 + \sum_{i,k=1}^{3} \left(\frac{\partial v_i}{\partial x_k} + \frac{\partial v_k}{\partial x_i} \right) dx_k dx_i \tag{12}$$

By denoting the deformation quantity

$$\begin{aligned} \varepsilon_{11} &= \frac{\partial v_1}{\partial x_1}, \qquad \varepsilon_{12} = \varepsilon_{21} = \frac{1}{2}\left(\frac{\partial v_1}{\partial x_2} + \frac{\partial v_2}{\partial x_1} \right) \\ \varepsilon_{22} &= \frac{\partial v_2}{\partial x_2}, \qquad \varepsilon_{23} = \varepsilon_{32} = \frac{1}{2}\left(\frac{\partial v_2}{\partial x_3} + \frac{\partial v_3}{\partial x_2} \right) \\ \varepsilon_{33} &= \frac{\partial v_3}{\partial x_3}, \qquad \varepsilon_{31} = \varepsilon_{13} = \frac{1}{2}\left(\frac{\partial v_1}{\partial x_3} + \frac{\partial v_3}{\partial x_1} \right) \end{aligned} \tag{13}$$

we get (symmetric) deformation tensor **D**

$$\mathbf{D} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{pmatrix}, \tag{14}$$

that expresses dimensionless deformation of the body. When the deformation equals to zero, deformation does not occur, the distances between particular points stay unchanged even after the force acting, the effect of the acting force is only rotation or translation of the entire solid body.

As the volume of the element $dV = dx_1 dx_2 dx_3$ before deformation and $dV^* \approx dx_1 dx_2 dx_3 (1 + \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3})$ after deformation, by using the deformation term the relative volume dilatation can be expressed:

$$\varepsilon = \frac{dV^* - dV}{dV} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} = div\ \mathbf{v} = \varepsilon_{11} + \varepsilon_{22} + \varepsilon_{33}\ , \quad i = 1,2,3 \tag{15}$$

2.3. Stress – strain relationship - physical equations

In the following we consider modeled material being homogeneous, isotropic, and linear elastic.

When the volume element is loaded just by the normal stresses $\sigma_{11}, \sigma_{22}, \sigma_{33}$, the following relationship between the stress and strain goes into the consideration:

$$\begin{aligned} E\varepsilon_{11} &= \sigma_{11} - \mu(\sigma_{22} + \sigma_{33}) \\ E\varepsilon_{22} &= \sigma_{22} - \mu(\sigma_{11} + \sigma_{33})\ , \\ E\varepsilon_{33} &= \sigma_{33} - \mu(\sigma_{22} + \sigma_{11}) \end{aligned} \tag{16}$$

where E[Pa] is the Young elastic module of the material, μ[–] is the transversal deformation coefficient, ($1/\mu$ – Poisson constant)

When the volume element is loaded only by the tangential stress components $\sigma_{ik}, i,k = 1,2,3, i \neq k$, the stress and strain will be of the form:

$$\gamma_{ik} = 2\varepsilon_{ik} = \frac{2(1+\mu)}{E}\sigma_{ik} = \frac{1}{G}\sigma_{ik}, i \neq k, \tag{17}$$

where γ_{ik}[–] is a shear deformation, G[Pa] is a shear elastic module, $G = \frac{E}{2(1+\mu)}$.

Adding the equations (11) we get

$$\varepsilon = \frac{1-2\mu}{E}\mathbf{s} = \frac{1}{2G}\frac{1-2\mu}{1+\mu}\mathbf{s} \tag{18}$$

Followingly we can write

$$\varepsilon_{ik} = \frac{1}{2G}\left(\sigma_{ik} - \delta_{ik}\frac{\mu}{1+\mu}\mathbf{s}\right), \quad i,k = 1,2,3 \tag{19}$$

The basic elasticity equations in the sense of displacement (Lamé equations) we get by linking the balance equations, geometric and physical equation. In vector notation it has the following form.

$$\Delta\mathbf{v} + \frac{1}{1-2\mu}\nabla\left(\nabla.\mathbf{v}\right) + \frac{1}{G}\mathbf{K} = \mathbf{0} \tag{20}$$

3. Elasticity - biomechanical behavior of the motion segment of the human lumbar spine

As mentioned above, the mathematical and computational modeling of the biomechanical task includes the governing equation establishment, boundary and initial conditions assessing, the appropriate method selecting and the computer implementation.

3.1. Investigated domain, governing equations, boundary conditions

The domain geometry development of the model sometimes requires a lot of effort and simplifications are often inevitable. Our domain, motion segment of human lumbar spine is a composition of two neighbor vertebrae and intervertebral disc between them. The disc is connected to the upper and lower vertebra by thin layer called endplate [14]. The surface of this cartilage-like layer is hereby the contact surface which the essential part of load is transferred through. So in the following we concentrate our attention to two neighbor vertebral bodies joined together with the disc. And, we can consider even one simplification. As proven by the experiments [12, 13], we can neglect the oval shape of the upper and lower surface of the vertebral bodies and take circular shape. (Only the magnitude, not the exact shape of the surfaces is essential while computing the stress – strain field). So further, as a first approximation we use the rotation symmetry with the vertical axis of revolution within the motion segment. Using all symmetries, we cut the two dimensional domain (as shown on Fig.6), and using axisymmetric type of loads we get a fully *axisymmetric* task, see Fig.7.

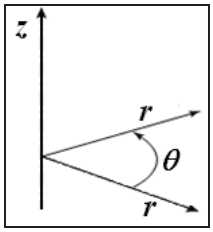


Fig. 6: Cylindrical system of coordinates

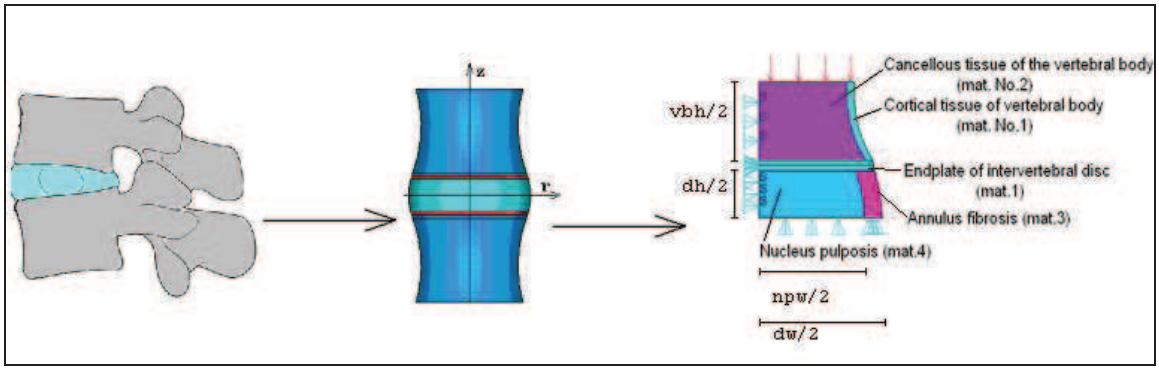


Fig. 7: Investigated domain, simplification, rotation symmetry

In the case of axisymmetry the displacement will consist of two components $\mathbf{v}(r,z)=\left(v_r(r,z),v_z(r,z)\right)^T$, corresponding *geometric equations* are of the form

$$\begin{aligned}\varepsilon_{rr}&=\frac{\partial v_r}{\partial r}\\ \varepsilon_{rr}&=\frac{\partial v_z}{\partial z}\\ \varepsilon_{\theta\theta}&=\frac{1}{r}v_r\\ \gamma_{rz}&=\frac{\partial v_r}{\partial z}+\frac{\partial v_z}{\partial r}\end{aligned}\quad . \tag{21}$$

The *physical equations* for the case of rotational symmetry (axisymmetry) case will be of the form

$$\begin{aligned}\sigma_{rr}&=2G\left(\varepsilon_{rr}+\frac{1}{1-2\mu}\varepsilon\right)\\ \sigma_{\varphi\varphi}&=2G\left(\varepsilon_{\varphi\varphi}+\frac{1}{1-2\mu}\varepsilon\right)\\ \sigma_{zz}&=2G\left(\varepsilon_{zz}+\frac{1}{1-2\mu}\varepsilon\right)\\ \sigma_{rz}&=2G\varepsilon_{rz}\end{aligned} \tag{22}$$

and *elasticity equations for the case of rotational symmetry* will be of the form

$$\begin{aligned}\Delta v_r-\frac{v_r}{r^2}+\frac{1}{1-2\mu}\frac{\partial\varepsilon}{\partial r}&=0\\ \Delta v_z+\frac{1}{1-2\mu}\frac{\partial\varepsilon}{\partial z}&=0\end{aligned} \tag{23}$$

Boundary conditions: On the upper border of the domain the uniform distributed load (traction) [Pa/m] is applied, see the downwards arrows on Fig. 7, and the lower border of the domain is fixed in vertical direction due to axisymmetry of the task, i.e. one degree of freedom is detracted.

3.2. Variational formulation and FEM modeling

Total potential energy functional minimization: Taking displacements as master field, the total potential energy functional is of the form:

$$\Pi[\mathbf{u}] = \mathcal{U}[\mathbf{u}] - \mathcal{W}[\mathbf{u}], \tag{24}$$

where $\mathcal{U}[\mathbf{u}]$ is the strain energy functional,

$$\mathcal{U}[\mathbf{u}] = \frac{1}{2}\int_V \boldsymbol{\sigma}^T \mathbf{e}\, dV = \frac{1}{2}\int_V \mathbf{e}^T \mathbf{E}\mathbf{e}\, dV\,, \tag{25}$$

$\mathcal{W}[\mathbf{u}]$ is the external work potential that can be written as a sum of contributions due to body force and contributions due to prescribed surface tractions,

$$\mathcal{W}[\mathbf{u}] = \mathcal{W}_b[\mathbf{u}] + \mathcal{W}_t[\mathbf{u}], \tag{26}$$

where

$$\mathcal{W}_b[\mathbf{u}] = \int_V \mathbf{b}^T \mathbf{u}\, dV\,, \tag{27}$$

$$\mathcal{W}_t[\mathbf{u}] = \int_{S_t} \mathbf{t}^T \mathbf{u}\, dS\,. \tag{28}$$

In the case of axisymmetry, the element of volume dV in (27) can be taken as a "ring element" $dV = 2\pi r dA$, where dA is an element of cross sectional area, and the element of surface $dS = 2\pi r ds$ in (28), ds being arc length element. Thus the *strain energy functional in the case of rotational symmetry* is of the form

$$\mathcal{U}[\mathbf{u}] = \frac{1}{2} 2\pi \int_A r\, \mathbf{e}^T \mathbf{E}\mathbf{e}\, dA \tag{29}$$

and external work potential of the form

$$\mathcal{W}[\mathbf{u}] = \mathcal{W}_b[\mathbf{u}] + \mathcal{W}_t[\mathbf{u}] = 2\pi \int_A r\, \mathbf{b}^T \mathbf{E}\mathbf{e}\, dA + 2\pi \int_{s_t} r\, \mathbf{t}^T \mathbf{u}\, ds \tag{30}$$

Thus we can use the ring element instead of 3D one, see Fig. 8.

In the Tab. 1, the material properties of all components involved in the model are provided.

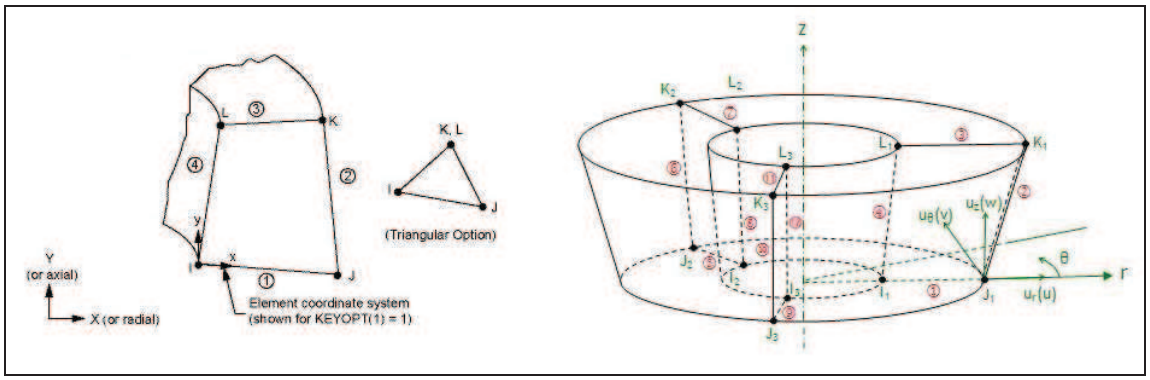


Fig. 8: Ring element and its node configuration [21]

Tab. 1: Material properties of the components[20]

	Cortical bone	Cancellous bone	Annulus fibrosis	Nucleus pulposus
Young's module [MPa]	16.1e9	75.8e6	24.3e6	0.013e6
Poisson ratio [-]	0.3	0.2	0.4	0.49999

On Fig. 9 there is an example of the result obtained by FEM software, the entire displacement field (3D representation) of the points of the domain relatively to the horizontal plane of symmetry of the domain.

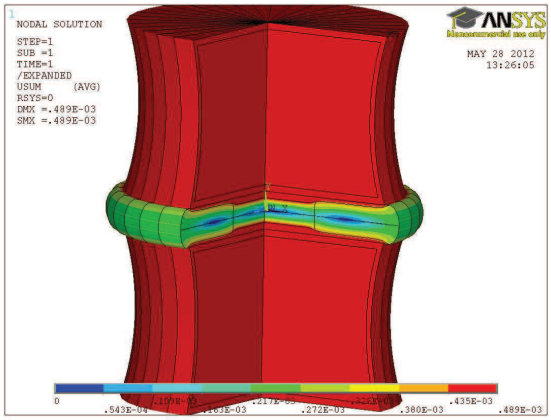


Fig. 9: Example of the result of finite element modeling obtained by the FEM software

Other biomechanical models are analyzed e.g. in [5, 6, 9].

4. Viscoelasticity in tissue behavior modeling

Rheology – the discipline that deals with the study of forces needed to achieve particular deformations or velocities, describes the tissues properly. It observes the materials which's mechanical behavior is neither strictly elastic (rigid body), nor viscous (liquid). The *tissues* of the organisms are just such a kind of materials. They are more or less soft, their mechanical behavior is viscoelastic. The simplest tissue that can be modeled by simple rheological model is skin tissue, see Maxwell rheological model on Fig. 11.

Elastic material obeys Hooke's law expressing the relationship between the stress and strain:

$$\varepsilon = \frac{\sigma}{E}, \tag{31}$$

where ε is proportional deformation, E Young elastic module, σ stress intensity.

Viscous material under the constant load flows with the velocity

$$\dot{\varepsilon} = \frac{\sigma}{\eta}, \tag{32}$$

where η is the Newton viscosity coefficient.

Whereas the (31) represents the *relation between force applied to a matter and its deformation,* (32) represents the *relation between force applied to a body and its deformation rate.*

Remark: In three-dimensional tasks E and η in (31), (32) are replaced by tensor operators, $\boldsymbol{\sigma} = \hat{\mathbf{H}}\boldsymbol{\varepsilon}$, i.e. $\mathbf{K}^{(r)}\boldsymbol{\sigma} = \hat{\mathbf{K}}_{(s)}\boldsymbol{\varepsilon}$, i.e. $\mathbf{Q}^{(r)}\boldsymbol{\varepsilon} = \hat{\mathbf{Q}}_{(s)}\boldsymbol{\sigma}$, where $\hat{\mathbf{H}}, \hat{\mathbf{K}}, \hat{\mathbf{Q}}$ are symmetric positive definite tensor operators $\mathbf{K}^{(r)}$ and $\mathbf{Q}^{(r)}$ are scalar operators.

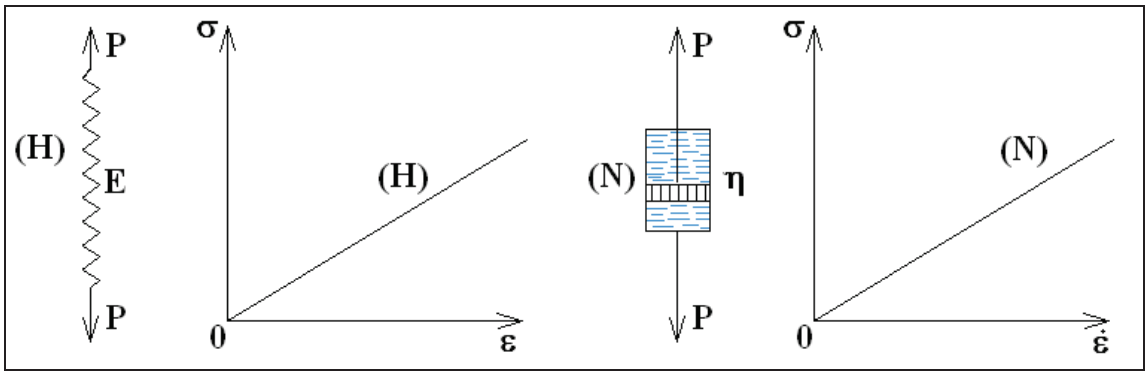


Fig. 10: Hooke's elastic matter (left) with constitutive equation (31) ***and Newton's viscous liquid with constitutive equation*** (32) ***and their stress – strain dependence***

When combining two basic rheological components: purely elastic material – Hooke's elastic material ***(H)*** and purely viscous material – Newton viscous liquid ***(N)***, see Fig. 10, we obtain various types of viscoelastic materials models. Below we introduce three of them:

Maxwell rheological model

By joining basic matters serially we get Maxwell rheological model ***(M)***; schematic formula is ***(M) = (H) – (N).***

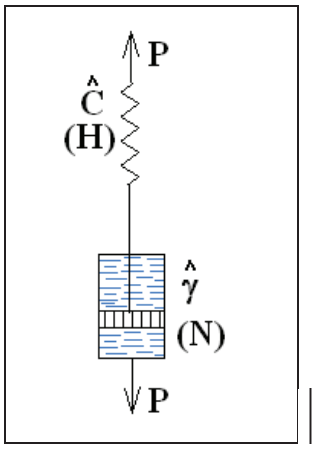


Fig. 11: Maxwell rheological model

Let a tensile force P act as on Fig. 11. The elastic component stretches immediately; the damping viscous component causes a delay. Let us denote ${}_H\boldsymbol{\sigma}, {}_H\boldsymbol{\varepsilon}$ stress, strain in Hook's mater, ${}_N\boldsymbol{\sigma}, {}_N\boldsymbol{\varepsilon}$ stress, strain in Newton's liquid. The stress in Maxwell model is distributed equally within the whole model,

$$\boldsymbol{\sigma} = {}_H\boldsymbol{\sigma} = {}_N\boldsymbol{\sigma}, \tag{33}$$

On the other hand the overall deformation is equal to the sum of deformations on the particular components

$$\boldsymbol{\varepsilon} = {}_H\boldsymbol{\varepsilon} + {}_N\boldsymbol{\varepsilon}\,. \tag{34}$$

The constitutive equations for particular components are

$$\begin{aligned} {}_H\boldsymbol{\varepsilon} &= \hat{\mathbf{C}}\boldsymbol{\sigma} \\ \frac{\partial}{\partial t}{}_N\boldsymbol{\varepsilon} &= \hat{\boldsymbol{\gamma}}\boldsymbol{\sigma} \end{aligned} \tag{35a,b}$$

Applying the time derivative operator to the equation (34) and (35a) and substituting to the and coupling the equations (33), (34), (35) we get the relation between stress and strain in the form of partial differential equation

$$\frac{\partial}{\partial t}\boldsymbol{\varepsilon} = \frac{\partial}{\partial t}\left({}_H\boldsymbol{\varepsilon} + {}_N\boldsymbol{\varepsilon}\right) = \hat{\mathbf{C}}\frac{\partial \boldsymbol{\sigma}}{\partial t} + \hat{\boldsymbol{\gamma}}\boldsymbol{\sigma}\,, \tag{36}$$

in one dimensional space we get the ordinary differential equation

$$\frac{d\boldsymbol{\varepsilon}}{dt} = \frac{1}{E}\frac{d\boldsymbol{\sigma}}{dt} + \frac{1}{\eta}\boldsymbol{\sigma} \tag{37}$$

Kelvin - Voigt rheological model

When we join basic matters parallelly by a stiff slab (parallel joining is denoted by | in the schematic formula), we get Kelvin - Voigt rheological model, ***(V)=(H)|(N),*** see Fig. 12. Now the deformation is the same in both matters, the stress is summed:

$$\boldsymbol{\sigma} = {}_H\boldsymbol{\sigma} + {}_N\boldsymbol{\sigma}, \quad {}_H\boldsymbol{\varepsilon} = {}_H\boldsymbol{\varepsilon} \tag{38}$$

By using constitutive equation for particular components we get the relationship between stress and strain in Kelvin – Voigt model:

$$\boldsymbol{\sigma} = \left(\hat{\mathbf{E}} + \hat{\boldsymbol{\eta}}\frac{\partial}{\partial t}\right)\boldsymbol{\varepsilon} \tag{39}$$

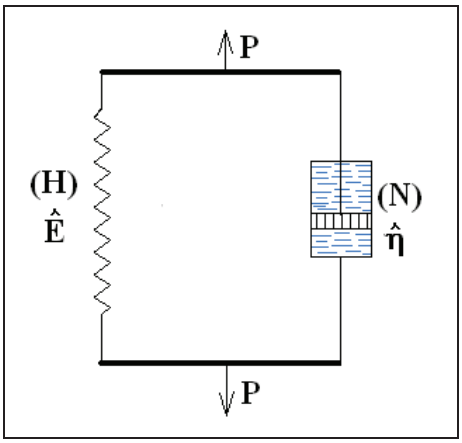


Fig. 12: Kelvin - Voigt rheological model

Zener rheological model

When we join Hook's matter to the Maxwell model parallelly, we get Zener rheological model, ***(Z)=(H₁)|[(H₂)-(N₂)]***, see Fig.13

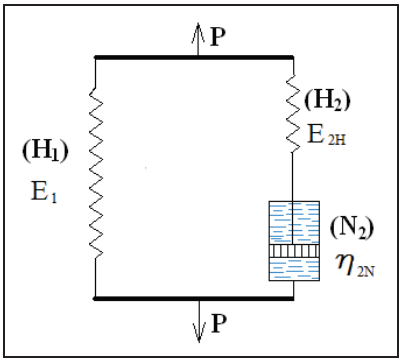


Fig. 13: Zener rheological model

In accordance with previous two models, we can write

$$\begin{aligned} \varepsilon_1 &= \varepsilon_2 , \\ \varepsilon_1 &= \varepsilon_{2N} + \varepsilon_{2H} , \\ \sigma &= \sigma_1 + \sigma_2 , \\ \sigma_1 &= E_1 \varepsilon_1 \\ \sigma_2 &= \eta_{2N} \dot{\varepsilon}_{2N} \end{aligned} \tag{40}$$

and because from (40) $\sigma_2 = \sigma - \sigma_1 = \sigma - E_1\varepsilon_1$ and $\dot{\varepsilon}_1 = \dot{\varepsilon}_{2N} + \dot{\varepsilon}_{2H} = \frac{\sigma_2}{\eta_{2N}} + \frac{\dot{\sigma}_2}{E_{2H}}$, then the constitutive equation of the Zener rheological model will be of the form

$$\dot{\varepsilon}_1 = \frac{\sigma - E_1\varepsilon_1}{\eta_{2N}} + \frac{\dot{\sigma} - E_1\dot{\varepsilon}_1}{E_{2H}} , \tag{41}$$

where dot over the entity represents its time derivative. Rearranging (36) to the form when stress and strain being on the opposite sides of the equation,

$$\dot{\varepsilon}_1\eta_{2N} + E_1\dot{\varepsilon}_1\frac{\eta_{2N}}{E_{2H}} + E_1\varepsilon_1 = \sigma + \frac{\eta_{2N}}{E_{2H}}\dot{\sigma} \tag{42}$$

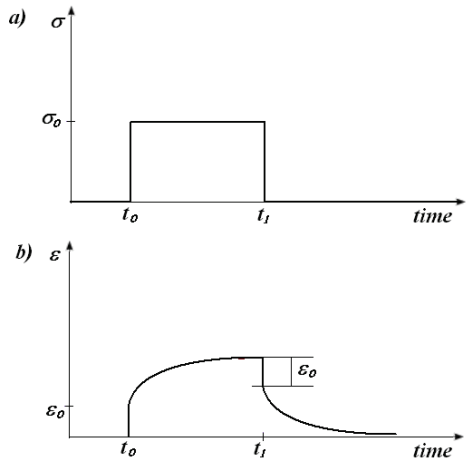


Fig. 14: Creep test of a material

5. Conclusion

Biomechanics is a transdisciplinary science. It collects the knowledge from various disciplines (technical mechanics, biology/anatomy, mathematics, biophysics, material science, etc.) and it provides its results for various branches (clinical medicine, sport medicine, natural sciences). It deals with mechanical structure, mechanical behavior and mechanical properties of the living organism, its parts and mechanical interactions between the organism and the surroundings.

The paper deals with physical and mathematical modeling in biomechanics. The basis of the modeling stands on the thermodynamical laws of open systems along with some deterministic events. By this attempt it is possible to assess the mechanical properties of analyzed biological structures. The adequate constitutive equations are essential as they are the physical-mathematical description of idealized material. The balance equations, physical equations and geometric equations for elastic contact problem vertebra – intervertebral disc – vertebra are derived, sourcing to the governing equation stipulation. The governing system - system of three partial differential equations expressing the balance in the sense of displacement (Lame equations) is finally completed by the boundary conditions. It is possible to show that the uniqueness of the solution is assured by the compatibility equation. The finite element analyze of the problem is derived from the global potential energy of the system.

An extra part of the article is devoted to the rheological models that are essential in biomechanical modeling of the tissues.

Acknowledgement: This work was supported by grant APVV – 0184 – 10.

References

[1] BEŇA, J., KOSSACZKÝ, E., (1981): Basis of Modeling Theory (in Slovak), Publishing house VEDA, Bratislava

[2] BRILLA, J., (1970): Variational Methods in Linear Anisotropic Viscoelasticity, ZAMM, Sonderheft

[3] BRILLA, J., (1958): Anisotropic Walls (in Slovak), Publishing house VEDA, Bratislava

[4] FUNK, Y.C.,(1990): Biomechanics, Flow, Stress, and Growth. Springer Verlag, New York – Berlin - Tokyo

[5] JIROUŠEK, O., JÍROVÁ, J., (2000): Construction of Mathematical Models for the Finite Element Method Using the Computer Tomography scan. In: 8th ANSYS Users meeting, Lednice, Czech Republic, pp. 1-5

[6] JÍROVÁ, J., (2002): Human mobility and implantate of the coxal joint. Habilitation thesis , ČVUT Prague, Fakulta of transport, (in Czech)

[7] KAFKA, V.R., (1981): On Mechanics of Two and Multi-Phased Solids. In: 4th National Congress on Theoretical and Applied Mechanics, Varna, Bulgaria

[8] KNESHKE, A., (1968): Using of differential equations in practice. (in Slovak) Publishing house ALFA, Bratislava, ISBN 63-078-69

[9] KULT, J., JÍROVÁ, J., (2002): Analysis of Finger Flexion Contractures Flexion. In: Proceedings of International Conference on Biomechanics of Man 2002, Charles University of Prague, Faculty of Physical Education and Sport

[10] NEMEC, J., (1985): Foreword to Book Biomechanics (in Czech)Academia, Prague,

[11] NOVOTNY, B., HANUŠKA, A., (1983): Theory of Layered Half Spaces. (in Slovak), Publishing house VEDA, Bratislava

[12] OGURKOWSKA, M.B et al., (2002): Interaction of the L4-L5 spinal segment by FEM analysis. Part I. Methods of geometrical data acquisition and validation. Acta of Bioengineering and Biomechanics, 13th Conference of the European Society of Biomechanics, Vol. 4, Suppl Wroclaw, Poland, pp.98-99.

[13] PANJABI, M.M. ET AL., (1992): Human Lumbar Vertebrae Quantitative Three Dimensional Anatomy, Spine, Vol.17, No.3, pp.299 – 306

[14] SINELNIKOV, R.D., (1980): Atlas of Human anatomy, Part I. (in Czech), Publishing house AVICENUM – Prague

[15] SOBOTKA, Z., (1981): Rheology of Mass and structures, (in Czech), Academia, Prague

[16] SPILKER, R.L:, (1982): A Simplified Hybrid-stress Finite Element Model of the Intervertebral Disc, Finite Elements in Biomechanics, University of Arizona, 14 chapter.

[17] SUMEC, J., SOKOL, M., JÍRA, J., (2003): Computer simulation of lumbar spine function. In|: Proceeding of International Conference Biomechanics, Rhodes, Greece

[18] SUMEC, J. SOKOL, M. AND WENDLOVÁ, J., (1996): Biomechanics of Human Spine and its Practical Using. In: 6. International Conference:" Biomechanics of Human '96, Tichonice '96, September 17 - 19, 191 - 194.

[19] VALENTA, J., KONVIČKOVÁ, S., (2006): Biomechanics of Human. Muscular Skeletal System. Part I, II (in Czech), Publishing house CVUT, Prague

[20] VALENTA, J., ed., (1993): Biomechanics. (in Czech), Publishing House Academia Prague

[21] ANSYS help

[22] www.substech.com/dokuwiki/doku.php?id%3Dtensile_test_and_strain-stress_diagram

Addresses of the authors:

Mária Minárová, RNDr., PhD.
Slovak University of Technology
Faculty of Civil Engineering
Dpt. of Mathematicsand Descr. Geometry
Radlinského 11, 813 68 Bratislava
minarova@math.sk

Jozef Sumec, Prof., Ing., RNDr., DrSc.
Slovak University of Technology
Faculty of Civil Engineering
Dpt. of Mechanical Engineering
Radlinského 11, 813 68 Bratislava
jozef.sumec@stuba.sk

Modelovanie viacrozmerných závislostí medzi svetovými menami počas finančnej krízy
Modelling multivariate dependencies between world currencies during financial crisis

Miroslav Sabo

Abstract: It is widely known that some world currencies depend on other much more than other. This dependence is in large part nonlinear and can not be described analytically. We therefore use support vector machines (SVM) regression model as statistical tool for modelling dependency and for predicting new values. The results show that SVM model is only able to describe dependencies between all currencies, not to make accurate predictions of future exchange rates values.

Abstrakt: Je známe, že niektoré svetové meny sú na iných závislé v omnoho väčšej miere ako iné. Z veľkej časti sú ale tieto závislosti nelineárne a nemožno ich popísať analyticky. V tejto práci používame metódu support vector machines modifikovanú na regresný problém za účelom modelovania závislostí a predikovania nových hodnôt pre vybrané svetové meny. Analýzu vzťahujeme na dáta (získané z Eurostatu) časovo spadajúce pod finančnú krízu. Výsledky práce ukázali, že metóda SVM dokáže dobre popísať nelineárne závislosti medzi menami, nedokáže však spoľahlivo predikovať budúce hodnoty mien.

Key words: financial crisis, support vector machines, dependency, prediction
Kľúčové slová: finančná kríza, support vector machines, závislosť, predikcia
JEL classification: C32, C45, C53, C58

Introduction

From all branches of science, prediction analysis in finance has special importance, since it can be very helpful in many areas. Another rarity is the fact that financial time series are not dependent only on past events but also on hidden factors that cannot be measured. World exchange rates analyzed in this article especially are dependent not only on other world currencies or past events, but mainly on psychological and political factors [12]. Therefore, many people are interested in modelling financial time series.

There are many philosophical views on predicting exchange rates. Some people say that they can be predictable, but there are also opinions that any prediction in finance is impossible, because financial time series are strictly stochastic (with strong nonlinear dependencies that can not be revealed) [8].

In the past, there were many attempts to predict true values of exchange rates. They can be divided into 4 main cathegories [12]:

a) technical analysis methods
b) fundamental analysis methods
c) traditional time series forecasting
d) machine learning methods

Traditional approaches [3] for financial time series are not able to handle multivariate data with strong nonlinear dependencies; therefore they are not widely used. On the contrary, modern methods include many machine learning procedures especially neural networks. In [12], prediction of univariate time series (national exchange rate) was performed with fuzzy interval neural networks.

Support vector machines are also a part of machine learning methods, but they were primarily developed for classification tasks and then they were successfully used for predictions in many applications in finance, for example [11] forecasted volatility using SVM model, [6] used SVM for bankruptcy prediction and [9] used SVM for detecting top management fraud.

There are also some multivariate parametric approaches for modelling multidimensional dependencies, widely used are mainly copulas. In [7], international stock market is analyzed via copula methods.

1. Data description and preprocessing

Data we analyze are available on Eurostat web page [1]. We selected all 35 available national currencies (but not all world currencies) described by average daily exchange rates of Euro against them since 1.1.2009 to 31.12.2009 (part of financial crisis). We removed all days with at least one missing value and whole two currencies (Bulgarian lev and Lithuanian litas), since their exchange rate during assumed period is monotone.

Afterwards we input data matrix (33 currencies described by 232 daily exchange rates from 1.1.2009 to 31.12.2009) into R language [10].

First, we displayed correlation matrix of all 33 variables (see Table 1). Due to big amount of values, it is better to see dependencies using special visualization, see Figure 1 [2]. Here, size of circles represents absolute value of correlation coefficient (moreover, black circles are negative and grey are positive dependencies). We can easily see that approximately one half of rows (currencies) has small circles in row, what means that there are only weak linear dependencies between these currencies and all the other ones. The rest of currencies have, on the contrary, big circles representing strong dependencies between them and other world currencies. To be more precise, we computed average correlation coefficient of each currency with all other currencies (omitting diagonal values). After sorting these values in ascending order, we can separate aforementioned two groups of variables (see Figure 2). We removed all currencies with average absolute value of correlation coefficient less than 0.6 to omit variables that have small correlations with other ones.

Rest 17 variables were following: Czech koruna (denoted in article as V14), Danish krone (V10), Hungarian forint (V15), Denar (of the former Yugoslav republic of Macedonia) (V17), South African rand (V13), Indian rupee (V16), Brazilian real (V11), Icelandic krona (V7), Singapore dollar (V12), Thai baht (V8), Australian dollar (V5), Malaysian ringgit (V9), New Zealand dollar (V1), Hong Kong dollar (V3), US dollar (V2), Philippine peso (V6) and Renminbi yuan (V4).

We selected three currencies for further analysis. They have the highest value of average absolute correlation coefficient (US dollar, Renminbi yuan and Hong Kong dollar). Average correlation coefficient of US dollar with all other 16 variables was 0.911, what

indicates very strong correlation of this currency with other assumed non omitted world currencies.

The task now is to model exchange rates of aforementioned three the most dependent currencies and to make one-step predictions (with three past values).

Tab. 1: Correlation matrix of all 33 assumed world currencies

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	V32	V33
V1	1	-1	-1	-1	1	-0.9	-0.9	-0.9	-0.9	0.9	0.9	-0.9	0.9	0.9	0.8	-0.8	0.9	0.8	-0.8	0.8	0.8	0.8	-0.6	0.7	0.6	0.6	-0.5	-0.4	-0.2	0.2	0.3	0.1	0
V2	-1	1	1	1	-0.9	1	0.9	1	0.9	-0.9	-0.9	0.9	-0.8	-0.8	-0.9	0.9	-0.8	-0.8	0.7	-0.8	-0.7	-0.7	0.7	-0.6	-0.6	-0.5	0.5	0.6	0.3	-0.1	-0.3	0	0
V3	-1	1	1	1	-0.9	1	0.9	1	0.9	-0.9	-0.9	0.9	-0.8	-0.8	-0.9	0.9	-0.8	-0.8	0.7	-0.8	-0.7	-0.7	0.7	-0.6	-0.6	-0.5	0.5	0.6	0.3	-0.1	-0.3	0	0
V4	-1	1	1	1	-0.9	1	0.9	1	0.9	-0.9	-0.9	0.9	-0.8	-0.8	-0.9	0.9	-0.8	-0.8	0.7	-0.8	-0.7	-0.7	0.7	-0.6	-0.6	-0.5	0.5	0.6	0.3	-0.1	-0.3	0	0
V5	1	-0.9	-0.9	-0.9	1	-0.9	-0.9	-0.9	-0.9	0.9	1	-0.8	0.9	0.8	0.8	-0.8	0.8	0.8	-0.8	0.7	0.8	0.7	-0.7	0.7	0.7	0.7	-0.4	-0.4	-0.2	0.2	0.3	0.1	0
V6	-0.9	1	1	1	-0.9	1	0.9	1	1	-0.9	-0.9	1	-0.8	-0.9	-0.8	0.9	-0.9	-0.8	0.8	-0.7	-0.6	-0.7	0.7	-0.5	-0.5	-0.5	0.5	0.6	0.3	-0.2	-0.2	0	-0.1
V7	-0.9	0.9	0.9	0.9	-0.9	0.9	1	0.8	0.8	-0.9	-0.8	0.8	-0.9	-0.9	-0.9	0.8	-0.8	-0.8	0.7	-0.8	-0.8	-0.7	0.6	-0.7	-0.4	-0.6	0.5	0.3	0.1	-0.3	-0.4	-0.2	-0.1
V8	-0.9	1	1	1	-0.9	1	0.8	1	1	-0.9	-0.8	1	-0.8	-0.9	-0.8	0.9	-0.8	-0.7	0.8	-0.7	-0.5	-0.7	0.7	-0.5	-0.6	-0.4	0.5	0.6	0.4	0	-0.3	0	0.1
V9	-0.9	0.9	0.9	0.9	-0.9	1	0.8	1	1	-0.9	-0.8	1	-0.8	-0.9	-0.8	0.9	-0.8	-0.7	0.8	-0.7	-0.5	-0.6	0.6	-0.4	-0.5	-0.4	0.6	0.6	0.4	-0.2	-0.1	0	-0.1
V10	0.9	-0.9	-0.9	-0.9	0.9	-0.9	-0.9	-0.9	-0.9	1	0.9	-0.8	0.9	0.8	0.7	-0.8	0.8	0.7	-0.7	0.7	0.7	0.7	-0.7	0.6	0.6	0.6	-0.3	-0.5	-0.3	0.2	0.2	0.1	0
V11	0.9	-0.9	-0.9	-0.9	1	-0.9	-0.8	-0.8	-0.8	0.9	1	-0.8	0.9	0.8	0.8	-0.7	0.9	0.8	-0.7	0.7	0.8	0.7	-0.6	0.6	0.7	0.7	-0.4	-0.4	-0.2	0.3	0.1	0	0.1
V12	-0.9	0.9	0.9	0.9	-0.8	1	0.8	1	1	-0.8	-0.8	1	-0.7	-0.9	-0.8	0.9	-0.8	-0.7	0.8	-0.7	-0.5	-0.6	0.7	-0.4	-0.5	-0.3	0.6	0.6	0.4	-0.1	-0.2	0	-0.1
V13	0.9	-0.8	-0.8	-0.8	0.9	-0.8	-0.9	-0.8	-0.8	0.9	0.9	-0.7	1	0.8	0.8	-0.7	0.8	0.7	-0.8	0.6	0.8	0.6	-0.7	0.7	0.6	0.7	-0.5	-0.2	0	0.4	0.2	0.3	0.1
V14	0.9	-0.8	-0.8	-0.8	0.8	-0.9	-0.9	-0.9	-0.9	0.8	0.8	-0.9	0.8	1	0.8	-0.9	0.8	0.7	-0.8	0.8	0.6	0.7	-0.6	0.5	0.4	0.4	-0.6	-0.4	-0.1	0.2	0.2	0.2	0.2
V15	0.8	-0.9	-0.9	-0.9	0.8	-0.8	-0.9	-0.8	-0.8	0.7	0.8	-0.8	0.8	0.8	1	-0.7	0.8	0.8	-0.6	0.9	0.7	0.8	-0.4	0.6	0.4	0.6	-0.4	-0.3	0.1	0.2	0.4	0.2	0.3
V16	-0.8	0.9	0.9	0.9	-0.8	0.9	0.8	0.9	0.9	-0.8	-0.7	0.9	-0.7	-0.9	-0.7	1	-0.8	-0.6	0.8	-0.7	-0.5	-0.6	0.7	-0.3	-0.5	-0.3	0.6	0.6	0.4	0	-0.1	-0.1	0
V17	0.9	-0.8	-0.8	-0.8	0.8	-0.9	-0.8	-0.8	-0.8	0.8	0.9	-0.8	0.8	0.8	0.8	-0.8	1	0.7	-0.6	0.7	0.6	0.6	-0.4	0.4	0.5	0.5	-0.4	-0.4	-0.2	0.3	0.1	-0.1	0.3
V18	0.8	-0.8	-0.8	-0.8	0.8	-0.8	-0.8	-0.7	-0.7	0.7	0.8	-0.7	0.7	0.7	0.8	-0.6	0.7	1	-0.6	0.6	0.7	0.6	-0.4	0.6	0.4	0.6	-0.2	-0.4	-0.1	0.2	0.4	0.1	0.2
V19	-0.8	0.7	0.7	0.7	-0.8	0.8	0.7	0.8	0.8	-0.7	-0.7	0.8	-0.8	-0.8	-0.6	0.8	-0.6	-0.6	1	-0.4	-0.5	-0.4	0.8	-0.4	-0.4	-0.4	0.7	0.2	0.2	-0.4	-0.2	-0.4	0
V20	0.8	-0.8	-0.8	-0.8	0.7	-0.7	-0.8	-0.7	-0.7	0.7	0.7	-0.7	0.6	0.8	0.9	-0.7	0.7	0.6	-0.4	1	0.6	0.8	-0.3	0.6	0.4	0.5	-0.4	-0.2	0	0	0.4	0.1	0.1
V21	0.8	-0.7	-0.7	-0.7	0.8	-0.6	-0.8	-0.5	-0.5	0.7	0.8	-0.5	0.8	0.6	0.7	-0.5	0.6	0.7	-0.5	0.6	1	0.6	-0.5	0.8	0.4	0.7	-0.2	0	0.1	0.3	0.4	0.2	0.1
V22	0.8	-0.7	-0.7	-0.7	0.7	-0.7	-0.7	-0.7	-0.6	0.7	0.7	-0.6	0.6	0.7	0.8	-0.6	0.6	0.6	-0.4	0.8	0.6	1	-0.3	0.7	0.6	0.6	-0.4	-0.3	0.1	0	0.1	0.1	0
V23	-0.6	0.7	0.7	0.7	-0.7	0.7	0.6	0.7	0.6	-0.7	-0.6	0.7	-0.7	-0.6	-0.4	0.7	-0.4	-0.4	0.8	-0.3	-0.5	-0.3	1	-0.5	-0.5	-0.2	0.5	0.3	0.4	-0.1	-0.1	-0.2	0.3
V24	0.7	-0.6	-0.6	-0.6	0.7	-0.5	-0.7	-0.5	-0.4	0.6	0.6	-0.4	0.7	0.5	0.6	-0.3	0.4	0.6	-0.4	0.6	0.8	0.7	-0.5	1	0.4	0.6	-0.3	0.1	0.2	0.1	0.4	0.3	-0.1
V25	0.6	-0.6	-0.6	-0.6	0.7	-0.5	-0.4	-0.6	-0.5	0.6	0.7	-0.5	0.6	0.4	0.4	-0.5	0.5	0.4	-0.4	0.4	0.4	0.6	-0.5	0.4	1	0.5	-0.2	-0.4	-0.4	-0.2	-0.1	-0.3	-0.4
V26	0.6	-0.5	-0.5	-0.5	0.7	-0.5	-0.6	-0.4	-0.4	0.6	0.7	-0.3	0.7	0.4	0.6	-0.3	0.5	0.6	-0.4	0.5	0.7	0.6	-0.2	0.6	0.5	1	-0.1	-0.1	0.3	0.4	0	0.1	0.1
V27	-0.5	0.5	0.5	0.5	-0.4	0.5	0.5	0.5	0.6	-0.3	-0.4	0.6	-0.5	-0.6	-0.4	0.6	-0.4	-0.2	0.7	-0.4	-0.2	-0.4	0.5	-0.3	-0.2	-0.1	1	0	0	-0.2	0.1	-0.3	-0.2
V28	-0.4	0.6	0.6	0.6	-0.4	0.6	0.3	0.6	0.6	-0.5	-0.4	0.6	-0.2	-0.4	-0.3	0.6	-0.4	-0.4	0.2	-0.2	0	-0.3	0.3	0.1	-0.4	-0.1	0	1	0.6	0.2	0	0.5	0.1
V29	-0.2	0.3	0.3	0.3	-0.2	0.3	0.1	0.4	0.4	-0.3	-0.2	0.4	0	-0.1	0.1	0.4	-0.2	-0.1	0.2	0	0.1	0.1	0.4	0.2	-0.4	0.3	0	0.6	1	0.4	0	0.4	0.3
V30	0.2	-0.1	-0.1	-0.1	0.2	-0.2	-0.3	0	-0.2	0.2	0.3	-0.1	0.4	0.2	0.2	0	0.3	0.2	-0.4	0	0.3	0	-0.1	0.1	-0.2	0.4	-0.2	0.2	0.4	1	-0.2	0.3	0.6
V31	0.3	-0.3	-0.3	-0.3	0.3	-0.2	-0.4	-0.3	-0.1	0.2	0.1	-0.2	0.2	0.2	0.4	-0.1	0.1	0.4	-0.2	0.4	0.4	0.1	-0.1	0.4	-0.1	0	0.1	0	0	-0.2	1	0.2	-0.2
V32	0.1	0	0	0	0.1	0	-0.2	0	0	0.1	0	0	0.3	0.2	0.2	-0.1	-0.1	0.1	-0.4	0.1	0.2	0.1	-0.2	0.3	-0.3	0.1	-0.3	0.5	0.4	0.3	0.2	1	0.1
V33	0	0	0	0	0	-0.1	-0.1	0.1	-0.1	0	0.1	-0.1	0.1	0.2	0.3	0	0.3	0.2	0	0.1	0.1	0	0.3	-0.1	-0.4	0.1	-0.2	0.1	0.3	0.6	-0.2	0.1	1

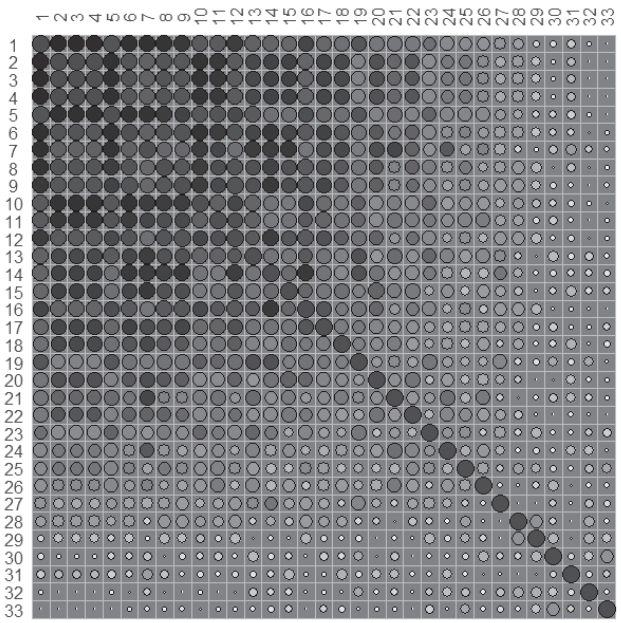


Fig. 1: Visualization of correlation matrix

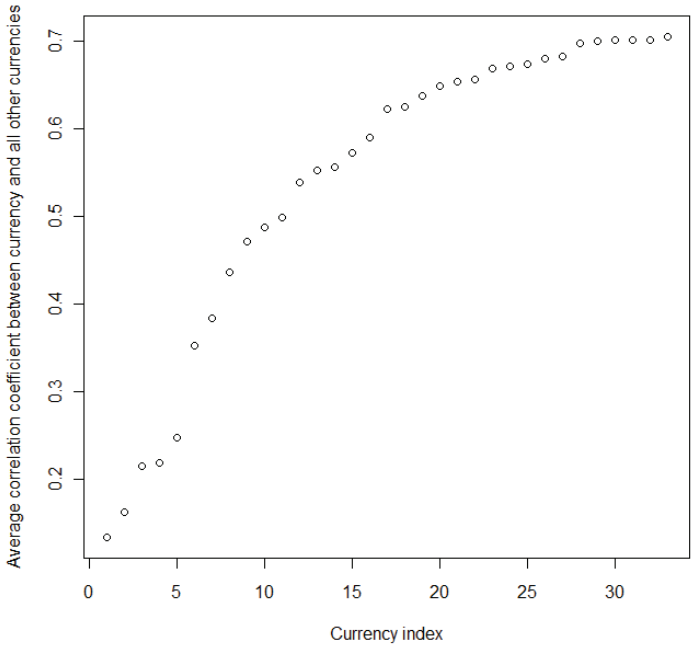


Fig. 2: Average absolute correlations of each currency with all other currencies

2. Support vector machines

Support vector machines [5] is technique used for supervised learning (term for classification and regression). As in all supervised learning methods, we have to know in advance the class (or the value) that each object belongs to. Then the task is to use new data and to predict which class it belongs to. SVM can solve especially following problems

- General classification (non-linear)
- Regression (also non-linear)
- Novelty detection

It has also some advantages when comparing with neural networks. In detail, SVM produce more precise classification and prediction, SVM do not deal with several local minima problem and also SVM performs much better in terms of time complexity than training neural networks (neural networks use all training data during training process, but SVM use at the end of the procedure only small part of input vectors; therefore their training is not so time-consuming as in the case of neural networks).

One can define supervised learning problem as follows: We are given set of data $D = \{(x_i, y_i)\}_{i=1}^{n}$ and the task is to fit a function $g(x)$ which approximates data points as much as possible. Then we often want to predict value of $g()$ for some new data points. Many regression algorithms work with so called loss function $L(y, g(x))$ that measures how the estimated function deviates from the true one. In literature, many different loss functions have been used, especially linear, quadratic, exponential and $\varepsilon-$insensitive. For example, quadratic loss fuction is used in least squares method and $\varepsilon-$insensitive loss function is commonly used in SVM regression (this one is also called Vapnik loss function).

$\varepsilon-$insensitive loss function is defined as

$$L(y, g(x)) = \max\{0, |y - g(x)| - \varepsilon\},$$

where ε is positive constant. When looking at formula, one can immediately see that if y is sufficiently close to $g(x)$, then value of loss function is zero. If not, the value is proportional to the difference of y and $g(x)$. Roughly speaking, this loss function accepts small inaccuracies (i.e. if two values are very close, then their bias is supposed to be zero). For more details about SVM, see [5].

3. Modelling dependecies between currencies

So far we have inspected only pairwise dependencies between currencies. The aim of this section is to model exchange rate of three selected currencies (dependent variables of model) with 16 rest world currencies (independent variables of model) using SVM regression approach.

After loading e1071 package [4] into R language we first divided all observations (232 days) to two groups - training (150) and testing (82) data. Then we tuned best parameters (C=100, Gamma=0.001) of SVM regression procudure using svm.tune() command to get the best predictions of dependent variable.

As a measure of accuracy we calculated for each model mean relative error of prediction. The resulting values for US dollar, Renminbi yuan and Hong Kong dollar are 0.917 %, 0.723 % and 0.770 % respectively.

4. Predicting exchange rates

The task of this part is to find out how accurate are one-step future predictions of assumed exchange rates. In this model, we consider 51 independent variables (17 exchange rates of all currencies during three succeeding time periods) and one dependent variable (for example, for the first model it is exchange rate of US dollar in the next time period). We can expect significantly worse predictions as before, since now we do not model how one currency depends on the other ones, but we predict future value of fixed exchange rate with three past values of 17 exchange rates.

Analogously as before, we divided initial data into two parts - training (150) and testing (79). Total number of observations is not equal to 232 as before, since we had to omit first three predictions, since our model predicts “fourth value from first three values“.

We calculated again mean relative errors as before. The results were 1.237 %, 1.654 % and 1.750 % for US dollar, Renminbi yuan and Hong Kong dollar respectively.

Moreover, we were interested in how accurate are increase-decrease predictions, i.e. we wanted to know, if SVM regression is able to predict increase or decrease of currency only with knowledge of three past values of all currencies. From 78 predictions for US dollar, 37 were correct; for Renminbi yuan the same number and for Hong Kong dollar only 34 values. This corresponds to following relative successes of binary (increase - decrease) predictions:

47.436 %, 47.436 % and 43.590 %. These predictions are similar to simple tossing a coin; therefore they are not assumed as successful.

5. Conclusion

Possible reasons of our bad results might be some "hidden variables" that also influence behavior of currencies.

As a special contribution of article we confirmed that US dollar is key world currency, since its dependencies on other world currencies were the strongest from all available currencies.

Acknowledgments: This work was supported by VEGA 1/0143/11.

6. References

[1] http:/appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ert_bil_eur_d&lang=en (available on 3.8.2011)

[2] http://addictedtor.free.fr/graphiques/graphcode.php?graph=152 (Wei T., code available on 3.8.2011)

[3] BOX, G. E. P. – JENKINS, G. M. 1976. Time series analysis: forecasting and control. Holden-day.

[4] DIMITRIADOU, E. – HORNIK, K. – LEISCH, F. – MEYER, D. – WEINGESSEL, A. 2011. e1071: Misc Functions of the Department of Statistics (e1071). TU Wien. R package version 1.5-26. http://CRAN.R-project.org/package=e1071.

[5] DRUCKER, H. – BURGES, Ch. J. C. – KAUFMAN, L. – SMOLA, A. – VAPNIK, V. 1997. Support Vector Regression Machines. In: *Advances* in Neural Information Processing Systems 9, NIPS 1996, MIT Press, s. 155 - 161.

[6] CHAUDHURI, A. – DE, K. 2011. Fuzzy Support Vector Machine for bankruptcy prediction. In: Applied Soft Computing, Volume 11, Issue 2, The Impact of Soft Computing for the Progress of Artificial Intelligence, s. 2472 – 2486.

[7] JONDEAU, E. – ROCKINGER, M. 2006. The Copula-GARCH model of conditional dependencies: An international stock market application. In: Journal of International Money and Finance, Volume 25, Issue 5, s. 827 – 853.

[8] MARKEI, B. G. 1999. A random walk down wall street. W.W. Norton and Company, New York, London.

[9] PAI, P. F. – HSU, M. F. – WANG, M. Ch. 2011. A support vector machine-based model for detecting top management fraud. In: Knowledge-Based Systems, Volume 24, Issue 2, s. 314 - 321.

[10] R DEVELOPMENT CORE TEAM (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[11] TANG, L. B. – TANG, L. X. - SHENG H.Y. 2009. Forecasting volatility based on wavelet support vector machine. In: Expert Systems with Applications, Volume 36, Issue 2, Part 2, s. 2901 – 2909.

[12] ZHANG, Y.Q. - WAN, X. 2007. Statistical fuzzy interval neural networks for currency exchange rate time series prediction. In: Applied Soft Computing, Volume 7, Issue 4, Soft Computing for Time Series Prediction, s. 1149 – 1156.

Adresa autora:
Miroslav Sabo, Mgr.
FCE STU in Bratislava
Radlinského 11, 813 68, Bratislava
miro165sabo@gmail.com

Grafická identifikácia prahovej hodnoty v SETAR modeloch

Grafical identification of the threshold value in SETAR models

Danuša Szőkeová

Abstract: This paper proposes some grafical methods of the threshold identification in two regime SETAR models. Threshold identification is the fundamental problem in model specification, parameter estimation and model fitting for multiregime models. We can use the statistical data handling, as histogram creations and clusters analysis but also grafical represention of autoregressive parameters to be estimated in the ordered autoregression. The results of grafical data analysis aimed to identify the threshold value are illustrated for one simulated data set and one real data set.

Abstrakt: V článku sú popísané spôsoby grafickej identifikácie prahovej hodnoty v dvojrežimových SETAR modeloch. Stanovenie prahovej hodnoty predstavuje základný problém pri špecifikácii, odhade parametrov a konštrukcii viacrežimových modelov. Okrem možností založených na štatistickom spracovaní údajov, ako sú tvorba histogramov a zhluková analýza dát, je možné použiť grafickú reprezentáciu autoregresných parametrov odhadnutých v preusporiadanej autoregresii. Výsledky grafickej analýzy údajov zameranej na identifikáciu prahovej hodnoty sú ilustrované na jednej množine simulovaných dát a jednej množine reálnych dát.

Key words: Nonlinear multiregime models, SETAR models, threshold value, ordered autoregression.

Kľúčové slová: Nelineárne viacrežimové modely, SETAR modely, prahová hodnota, preusporiadaná autoregresia.

JEL classification: C22, C24,C52

1. Úvod

V posledných dvoch desaťročiach sa objavilo veľa prác, ktoré sa zaoberajú analýzou časových radov zameranou na detekciu nelineárnych štruktúr v týchto radoch, obzvlášť na existenciu viacerých režimov. V najjednoduchšom prípade sa jedná o nelineárne modely typu SETAR (Self Exciting Threshold Autoregressive) so skokovým prepínaním medzi režimami na základe posunutej hodnoty analyzovaného radu, ktorá sa porovnáva s určitou prahovou hodnotou.

So špecifikáciou a odhadom viacrežimových modelov súvisí veľa nových problémov. V prvom rade je potrebné ukázať opodstatnenosť aplikácie týchto modelov pre konkrétne dáta. V literatúre sa objavuje veľa štatistických testov, v ktorých sa testuje hypotéza H_0 o existencii jedného režimu, voči hypotéze H_1 o existencii viacerých režimov v danom procese. V niektorých testoch sa predpokladá znalosť prahovej hodnoty, na základe ktorej sa prepínajú jednotlivé režimy (Chan, 1990), v iných prípadoch sa testujú rezíduá preusporiadanej autoregresie (Petrucelli a Davis, 1996) a znalosť prahovej hodnoty nie je

potrebná. Diagnostické testy možno využiť nielen na potvrdenie SETAR nelinearity v časovom rade, ale tiež na špecifikáciu prahovej hodnoty a parametra posunutia.

V procese odhadu parametrov a konštrukcie hodnôt SETAR modelu je stanovenie vhodnej prahovej hodnoty, prípadne intervalu, v ktorom leží prahová hodnota, základnou úlohou. Okrem prostriedkov štatistického spracovania údajov, ako sú histogramy a zhluková analýza dát (clusters), je možné odhadnúť prahovú hodnotu aj pomocou preusporiadanej autoregresie. Takto sa označuje každá regresia, v ktorej sú regresory a závislé hodnoty preusporiadané na základe nejakého kritéria. V nasledujúcej kapitole popíšeme preusporiadanú regresiu autoregresných parametrov lineárneho modelu, na základe ktorej možno odhadnúť prahovú hodnotu dvojrežimového SETAR modelu. V závere uvedieme dvojrežimové SETAR modely odhadnuté pre dva rady dát s prahovou hodnotou určenou pomocou preusporiadanej regresie.

2. Dvojrežimový SETAR model

Pri analýze časových radov sa často stáva, že pomocou jednoduchého lineárneho modelu nie je možné zachytiť vlastnosti pozorovaného časového radu. V takom prípade sa uvažuje o aplikácii nelineárnych modelov, napríklad viacrežimových typu SETAR (Tong, 1978).

Uvažujme jednorozmerný časový rad $\{y_t\}$, ktorý reprezentuje pozorovania v čase $t = 1,2,\ldots,n$. Vo všeobecnosti možno dvojrežimový model SETAR(c,p,d) popísať nasledovne

$$y_t = (\phi_0^1 + \phi_1^1 y_{t-1} + \ldots + \phi_p^1 y_{t-p} + \varepsilon_t^1) I[y_{t-d} \le c] + (\phi_0^2 + \phi_1^2 y_{t-1} + \ldots + \phi_p^2 y_{t-p} + \varepsilon_t^2) I[y_{t-d} > c] \quad (1)$$

kde

- c je prahová hodnota (reálne číslo, $|c| < \infty$),
- $\boldsymbol{\Phi}_i = (\phi_0^i, \ldots, \phi_p^i)$, i=1,2, sú autoregresné parametre dvoch režimov, $p = \max(p_1, p_2)$,
- I[A] je indikačná funkcia (I[A] = 1, ak A je pravdivý výraz, I[A] = 0 v opačnom prípade),
- d je parameter posunutia (kladné celé číslo, $d \ge 1$),
- $\varepsilon_t^i \approx$ i.i.d $N(0, \sigma_i^2)$, i=1,2, je proces bieleho šumu ($\sigma_i^2 < \infty$ je rozptyl chybovej zložky jednotlivých režimov).

V tomto spôsobe reprezentácie modelu prepínanie medzi jednotlivými režimami závisí od hodnoty indikačnej funkcie. Uvedený popis SETAR modelu možno jednoducho zovšeobecniť pre viac ako dva režimy.

Pomocou prepínania medzi rôznymi autoregresnými modelmi v závislosti od hodnoty parametra c možno pomocou SETAR modelov efektívne zachytiť skoky, cykly a iné nepravidelnosti v dátach.

3. Určenie prahovej hodnoty v SETAR modeloch

Nech n je počet pozorovaní, p je rád autoregresných polynómov v jednotlivých režimoch, $p = \max(p_1, p_2)$. Pri odhade dvojrežimového SETAR(c,p,d) modelu možno postupovať tak, že prvým krokom je určenie prahovej hodnoty c, nasleduje odhad vektorov autoregresných parametrov $\boldsymbol{\Phi}_i$, i=1,2, a napokon sa stanoví parameter posunutia d. Množinu

relevantných hodnôt, z ktorej sa vyberie konkrétna hodnota c, možno zistiť na základe regresie parametrov založenej na preusporiadaní regresorov v rastúcom alebo klesajúcom poradí vzhľadom na posunuté hodnoty radu y_{t-d} (Pekár, 2004).

Zvoľme pevné p a d, potom procedúra pozostáva z nasledujúcich krokov:

1. vytvoria sa vektory regresorov a závislej hodnoty y_t, $(y_t,y_{t-1},\ldots,y_{t-p})$, $t = h+d,\ldots,n$, $h = \max(1, p-d+1)$,
2. vektory pre regresiu sa preusporiadajú tak, aby v preusporiadanej postupnosti vektorov platilo $y_{(t-d)} \leq y_{(t-d+1)}$, $(t-d) = h,\ldots,n-d$, kde (t) sú indexy v preusporiadanej postupnosti vektorov,
3. zvolí sa hodnota $m > d+h-1$ (t.j. $m > d$, ak $d \geq p$; $m > p$, ak $p > d$),
4. urobí sa autoregresia prvých m preusporiadaných vektorov a odhadnú sa autoregresné parametre $\phi_0,\phi_1,....,\phi_p$:

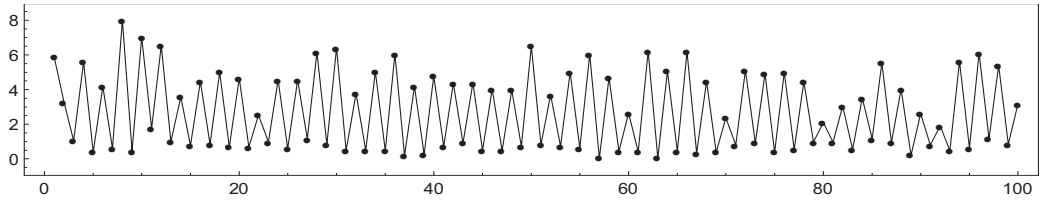
$$y_{(t+d)} = \phi_0+\phi_1 y_{(t+d-1)} + \ldots+\phi_p y_{(t+d-p)} ,$$

5. preusporiadaná regresia sa vykoná postupne pre všetky m, $m=d+h,\ldots,n$,
6. zobrazia sa postupnosti hodnôt autoregresných parametrov $\phi_0,\phi_1,....,\phi_p$ odhadnutých postupne pre všetky m a vyhodnotením grafu sa odhadne prahová hodnota c, prípadne interval, z ktorého sa prahová hodnota vyberie v procese odhadu ďalších parametrov modelu.

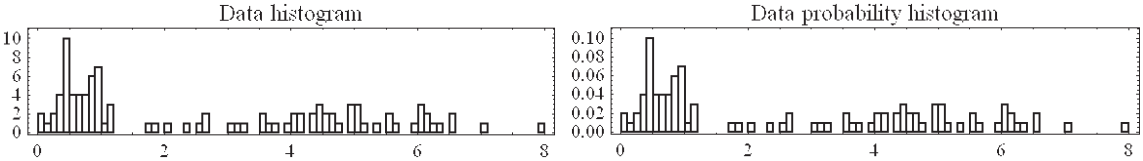
4. Výsledky grafickej identifikácie prahovej hodnoty na dvoch množinách dát

Uvedený algoritmus sme testovali na jednej množine simulovaných dát s dĺžkou radu n=100 a na jednej množine reálnych dát, ktorá reprezentuje priemerné mesačné prietoky na toku Váh, stanica Liptovský Mikuláš, pozorované v období rokov 1991 až 2000, n=120. Všetky výpočty boli realizované pomocou výpočtového systému Mathematica 8.

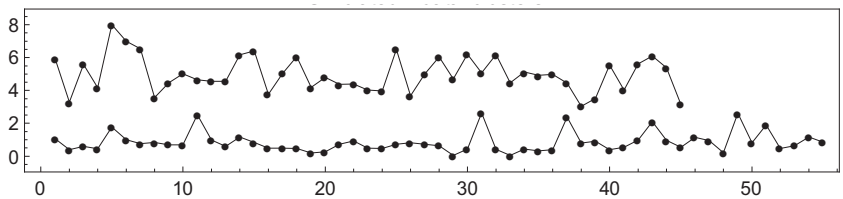
Na nasledujúcich obrázkoch 1 až 6 uvádzame grafickú reprezentáciu údajov ako aj histogramy a zhluky (clustre) stanovené v oboch množinách dát.



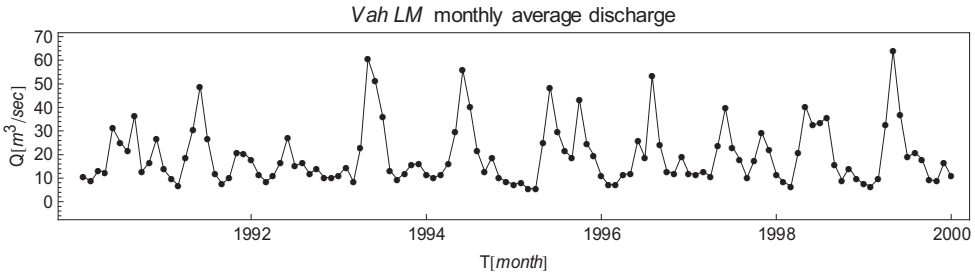
Obr.1 Zobrazenie postupnosti simulovaných dát, n=100



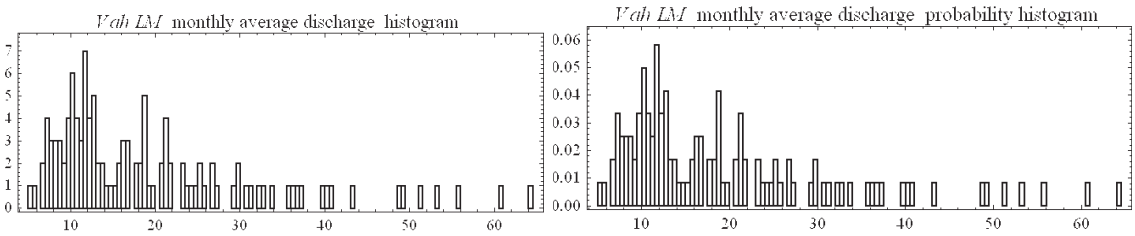
Obr.2 Histogramy simulovanej množiny dát



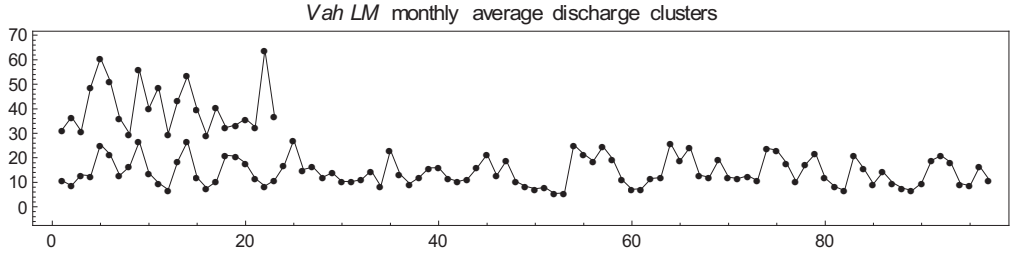
Obr.3 Zobrazenie hodnôt dvoch zhlukov stanovených v množine simulovaných dát, n_1=55, n_2=45



Obr.4 Zobrazenie postupnosti hodnôt priemerných mesačných prietokov na toku Váh, stanica Liptovský Mikuláš, obdobie 1991–2000, n=120



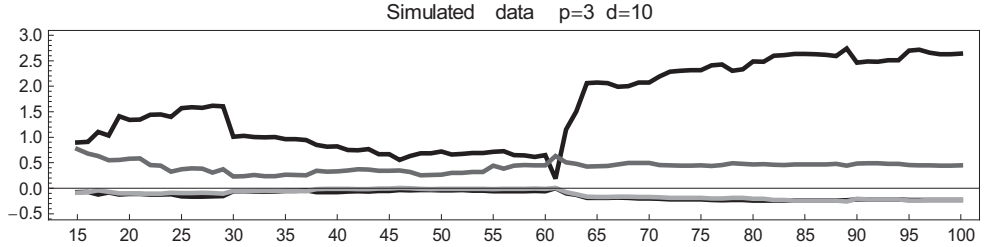
Obr.5 Histogramy množiny priemerných mesačných prietokov, tok Váh, stanica Liptovský Mikuláš, obdobie 1991–2000

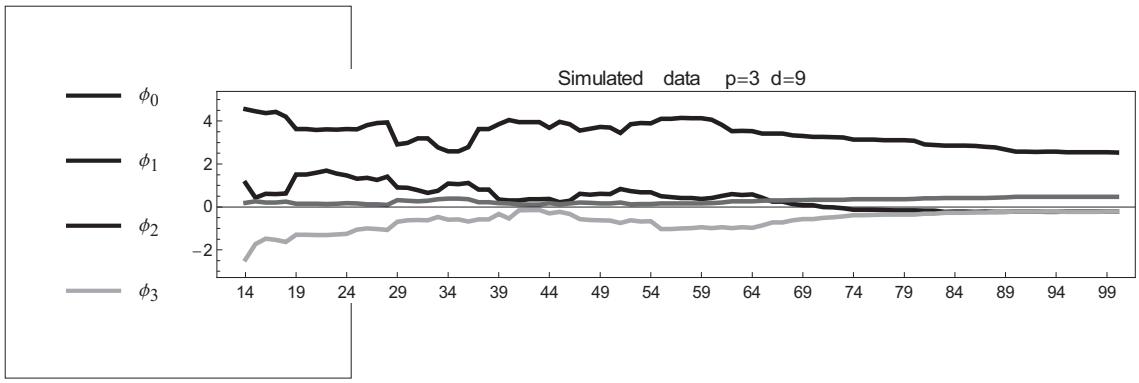


Obr.6 Zobrazenie hodnôt dvoch zhlukov stanovených v množine priemerných mesačných prietokov na toku Váh, stanica Liptovský Mikuláš, n_1=97, n_2=23

Jednoduchou analýzou histogramov a zhlukov je možné hľadať prahovú hodnotu c v prípade simulovaných dát približne v intervale (2,4), v prípade prietokov v intervale (14,17).

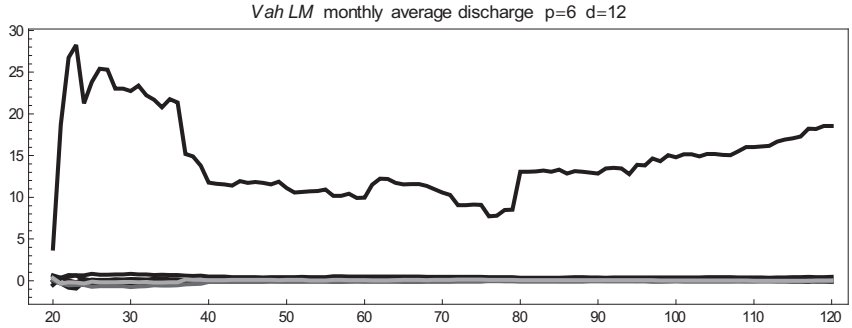
Okrem tejto jednoduchej grafickej analýzy sme podľa algoritmu popísaného v časti 3 zostavili program pre preusporiadanú regresiu autoregresných parametrov. Grafy odhadnutých parametrov $\phi_0,\phi_1,\phi_2,\phi_3$ pre simulované dáta, p=3, sú na obrázku 7. Z grafickej reprezentácie je zrejmé, že významný skok v hodnotách parametrov nastáva v prípade posunutia d=10 približne pre počet regresovaných vektorov m=60, čo odpovedá posunutej hodnote preusporiadaného radu $y_{(m-d)}$=3.038 (odhad prahovej hodnoty c). Pre iné hodnoty posunutia d v hodnotách autoregresných parametrov významný skok nenastal, čo je zrejmé na príklade grafov pre d=9.



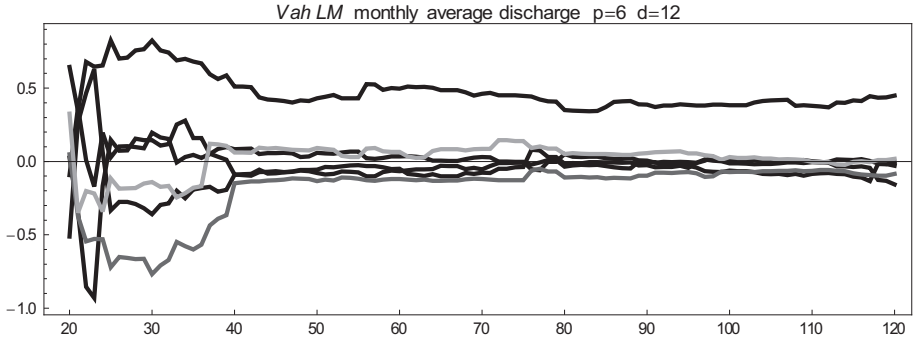


Obr.7 Zobrazenie postupností pre autoregresné parametre $\phi_0, \phi_1, \phi_2, \phi_3$ vyčíslených v preusporiadanej regresii, p=3, d=10,9, m=d+h,…,n, h = max(1, p–d+1)

Podobné grafy autoregresných parametrov $\phi_0,\phi_1,...,\phi_6$ vyčíslených v preusporiadanej regresii, p=6, pre priemerné mesačné prietoky sú na obrázku 8. Skok v hodnotách parametrov nie je taký výrazný a jednoznačný ako v prípade simulovaných dát. V prípade posunutia d=12 skok nastáva približne pre m=40, čo zodpovedá posunutej hodnote $y_{(m-d)}$=16.1 (odhad prahovej hodnoty c).



a) grafy pre parametre $\phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6$



b) grafy pre parametre $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6$

Obr.8 Zobrazenie postupností pre autoregresné parametre $\phi_0, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6$ vyčíslených v preusporiadanej regresii, p=6, d=12, m=d+h,…,n, h = max(1, p–d+1)

5. Odhad a konštrukcia SETAR modelu

Proces tvorby stochastického modelu má niekoľko etáp. Špecifikácia modelu znamená určenie typu modelu, ktorý sa pre dané dáta bude aplikovať. Ďalšou etapou je odhad parametrov špecifikovaného modelu a potom nasleduje zostavenie modelu (hodnoty radu). Na záver by sa mala urobiť verifikácia modelu. V prípade SETAR modelu to znamená:

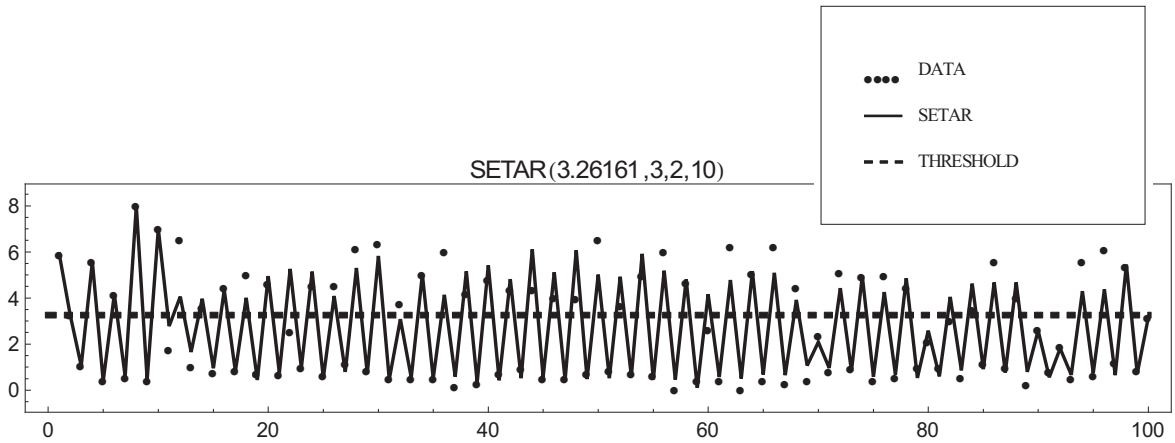
a) špecifikovať
– počet režimov r ,
– maximálny rád p autoregresných AR(p) modelov v jednotlivých režimoch,
– interval prahových hodnôt c,
b) pre každé r, p, c odhadnúť
– množiny autoregresných parametrov $\mathbf{\Phi_1},\ldots,\mathbf{\Phi_r}$ pre každý režim,
– parameter posunutia d,
c) vybrať konkrétny model a generovať hodnoty radu (deterministického bez chybovej zložky alebo stochastického s chybovou zložkou),
d) overiť model pomocou
– štatistických testov reziduálneho radu,
– tvorby predpovedí.

Po vykonaní týchto krokov sa podľa potreby model upraví (zmení sa špecifikácia, prepočítajú sa parametre).

Po vykonaní uvedených krokov sme pre množinu simulovaných dát vybrali dvojrežimový SETAR(c_1,p_1,p_2,d) model :

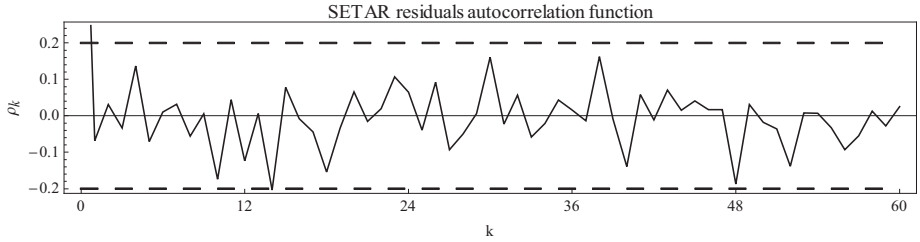
$$\textbf{SETAR(3.26,3,2,10)=}\begin{cases} 0.647-0.056\, y_{t-1}+0.447\, y_{t-2}-0.009\, y_{t-3}, & \text{ak } y_{t-10} \le 3.26,\ \sigma_1=0.175 \\ 4.592-0.329\, y_{t-1}-0.009\, y_{t-2}, & \text{ak } y_{t-10} > 3.26,\ \sigma_2=0.527 \end{cases}$$

Na obrázku 9 sú znázornené pôvodné hodnoty a rad zostavený na základe uvedeného SETAR modelu spolu s prahovou hodnotou c.



Obr.9 Zobrazenie hodnôt generovaných pomocou dvojrežimového SETAR modelu simulovanej množiny dát

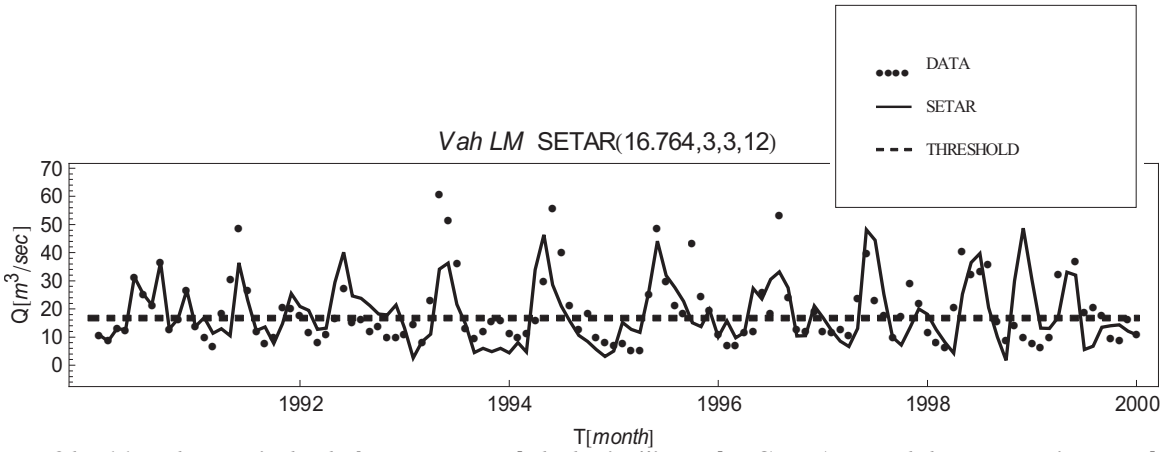
Na obrázku 10 je znázornená autokorelačná funkcia reziduálneho radu, t.j. radu, ktorý reprezentuje rozdiely medzi modelovanými dátami a hodnotami generovaných modelom. Ak je autokorelačná funkcia od nejakej hodnoty k_0 štatisticky nulová (interval vyznačený na obrázku vodorovnými prerušovanými čiarami), možno daný typ modelu považovať za vhodný. Okrem takejto jednoduchej verifikácie modelu sa adekvátnosť overuje aj inými spôsobmi, ako sú štatistické testy rezíduí a predpovede generovanými modelom, čím sa v rámci tohoto článku nebudeme zaoberať.



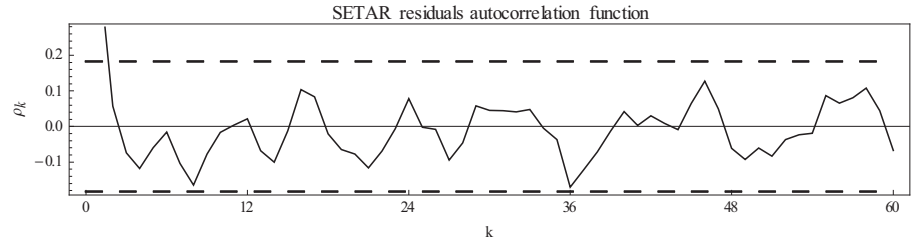
Obr.10 Zobrazenie hodnôt autokorelačnej funkcie reziduálneho radu SETAR modelu pre simulovanú množinu dát

Dvojrežimový SETAR(c_1,p_1,p_2,d) model odhadnutý pre množinu priemerných mesačných prietokov na toku Váh má nasledovné parametre:

$$\textbf{SETAR(16.754,3,3,12)} = \begin{cases} 4.476+0.679\, y_{t-1}-0.129\, y_{t-2}+0.038\, y_{t-3}, \text{ ak } y_{t-12} \le 16.764, \sigma_1=3.46 \\ 23.647+0.578\, y_{t-1}-0.319\, y_{t-2}-0.315\, y_{t-3}, \text{ ak } y_{t-12} > 16.764, \sigma_2=6.42 \end{cases}$$



Obr.11 Zobrazenie hodnôt generovaných dvojrežimovým SETAR modelom pre priemerné mesačné prietoky na toku Váh, obdobie 1991–2000



Obr.12 Zobrazenie hodnôt autokorelačnej funkcie reziduálneho radu SETAR modelu pre priemerné mesačné prietoky na toku Váh, obdobie 1991–2000

Adekvátnosť modelu možno zistiť predovšetkým jednoduchým posúdením štatistických charakteristík modelovaných hodnôt a hodnôt generovaných modelom. V tabuľke 1 sú uvedené stredné hodnoty a štandardné odchýlky pre obidve modelované množiny, pre rady generované dvojrežimovými SETAR modelmi a pre porovnanie aj radov generovaných jednorežimovými lineárnymi AR(p) modelmi. V prípade SETAR modelu a AR modelu sme vyčíslili aj koeficienty determinácie R^2. Posúdením štandardnej odchýlky rezíduí σ_{rez} (čím menšia, tým lepší model) a koeficientu determinácie R^2 (čím väčší, tým adekvátnejší model) možno zistiť, že dvojrežimový model v oboch prípadoch má v testovaných radoch lepšie popisné vlastnosti.

Tab.1 Štatistické charakteristiky dát a hodnôt generovaných lineárnym modelom AR(p) a dvojrežimovým modelom SETAR(c,p1,p2,d)

Množina dát	**Dáta**		**One regime model AR(3)**				**Two regimes SETAR(c,p_1,p_2,d)**			
	μ	**σ**	**μ**	**σ**	**σ_{rez}**	**R^2**	**μ**	**σ**	**σ_{rez}**	**R^2**
Simulované	2.71	2.24	2.68	1.24	1.73	0.39	2.75	2.20	0.72	0.89
Váh LM	19.47	12.53	19.48	5.79	13.86	–0.22	18.37	10.93	10.51	0.28

6. Záver

V článku sme sa zaoberali problematikou aplikácie viacrežimových SETAR modelov na jednej množine simulovaných dát a na množine priemerných mesačných prietokov. Ukázalo sa, že v oboch prípadoch môžu byť viacrežimové modely vhodnejšie ako jednoduché lineárne autoregresné modely. Okrem toho, že majú lepšie popisné vlastosti, prinášajú aj dodatočnú infomáciu o režimoch, ktoré sa prejavujú v modelovaných procesoch. Výhodou viacrežimových modelov typu SETAR je, že proces odhadu parametrov a generovania modelovaných hodnôt je v porovnaní s inými typmi viacrežimových modelov, ako sú napríklad STAR modely (Smooth Transition AutoRegressive) alebo MSW modely (Markov Switching), pomerne jednoduchý a oveľa menej časovo náročný.

Hlavnú pozornosť sme venovali predovšetkým efektívnym spôsobom odhadu prahovej hodnoty c. Poukázali sme na to, ako na základe jednoduchej grafickej reprezentácie dát možno zistiť existenciu viacerých režimov v modelovaných procesoch a ako určiť vhodnú prahovú hodnotu. Tento krok môže významne zjednodušiť a urýchliť jednak špecifikáciu modelu a jednak odhad parametrov SETAR modelu.

Popísané metódy sme ilustrovali na dvoch konkrétnych príkladoch. Identifikácia prahovej hodnoty a odhadnuté autoregresné parametre v jednotlivých režimoch môže predstavovať pre odborníkov významnú informáciu o modelovaných dátach. Okrem toho sme sa na konkrétnych príkladoch presvedčili, že v porovnaní s jednorežimovým lineárnym modelom môže byť viacrežimový SETAR model oveľa vhodnejší a spoľahlivejší nástroj pri popise procesov.

Prahové premenné, na základe ktorých sa realizuje prepínanie medzi jednotlivými režimami, môžu byť dané nielen posunutými hodnotami daného procesu, ale aj hodnotami váženého priemeru niekoľkých posunutých hodnôt, prípadne hodnotami exogénnych premenných. Takéto varianty základného SETAR modelu predstavujú nové problémy,

ktorými sa chceme zaoberať. Je taktiež zrejmé, že proces môže mať viac ako dva režimy. Odhad modelov s viac ako dvoma režimami je v podstate podobný, narastá iba časová náročnosť výpočtov. Z našich doterajších skúseností vyplýva, že aplikácia trojrežimového modelu môže v niektorých prípadoch priniesť ďalšie skvalitnenie popisu dát pomocou SETAR modelov.

7. Literatúra

[1] Chan,K.S. 1990 *Testing for threshold autoregression*, Annals of Statistics 18, 1886–94

[2] Petrucelli,J., Davis,N. 1986 A Portmanteau Test for Self-Exciting Threshold Autoregressive Type Nonlinearity, Biometrica,73, 687-694

[3] Hansen,B.E. 2000. *Sample splitting and threshold estimation*, Econometrica 68, No.3, 575-603

[4] Komorníková,M., Szőkeová,D. 2008. *Analysis of annual average river flows based on application of various classes of modeling procedures*, Anniversary conference FCI STU

[5] Pekár,J. 2004. *Gross domestic product autoregressive models of Slovakia* (*Autoregresné modely hrubého domáceho produktu Slovenska)* (skriptum), Comenium university, Bratislava

[6] Tsay,R.S. 1991. *Detecting and modeling nonlinearity in univariate time series analysis*, Statistica Sinica 1

[7] Tong, H 1978. *On a threshold model*. C. H. Chen (ed.), Pattern recognition and Signal Processing, Amsterdam, 101–141

[5] Wolfram Research: *Time Series Pack* / Reference and User's Guide, Published by Wolfram Research, Illinois

Poďakovanie Táto práca bola podporená grantom VEGA 1/0143/11

Adresa autora :
Danuše Szőkeová, RNDr.
Katedra matematiky a deskriptívnej geometrie
Stavebná fakulta, STU Bratislava
Radlinského 11, 813 68 Bratislava
szoke@math.sk

Zo života SŠDS

From live of SSDS

Nitrianske štatistické dni 2013
Statistical days in Nitra 2013

Ondrej Šedivý

Katedra matematiky Fakulty prírodných vied Univerzity Konštantína Filozofa v Nitre a Slovenská štatistická demografická spoločnosť v dňoch 9. a 10. mája 2013 usporiadali vedeckú konferenciu „Nitrianske štatistické dni 2013“. Konferencia sa konala pod záštitou dekana Fakulty prírodných vied prof. RNDr. Ľubomíra Zelenického, CSc. pri príležitosti 20. výročia konštituovania Fakulty prírodných vied UKF v Nitre.

Konferenciu otvoril dekan FPV UKF prof. RNDr. Ľubomír Zelenický, CSc. a vedecký tajomník Slovenskej štatistickej a demografickej spoločnosti RNDr. Ján Luha, CSc.

Na pôde Katedry matematiky FPV UKF účastníkov konferencie privítal a pozdravil vedúci Katedry matematiky RNDr. Dušan Vallo, PhD.

Konferencia bola zameraná najmä na tieto tematické okruhy: Regionálna štatistika, Aplikácie štatistických metód v biológii, ekológii a environmentalistike, Metodológia a prax zberu štatistických údajov, Matematická štatistika a pravdepodobnosť, Štatistický softvér.

Úvodnú prednášku v pléne predniesol prof. Ing. Mirko Navara, DrSc. z Katedry kybernetiky z Fakulty elektrotechnickej Českého vysokého učení technického v Prahe na tému: Center of Machine Perception.

V prednáške podrobne charakterizoval úlohu štatistiky ako vedy. Pripomenul, že štatistika je v oblasti vzdelávania často krát podceňovaná. Vzhľadom na to, že so štatistikou sa v reálnom živote stretávame denne, bolo by potrebné, aby vyučovaniu štatistiky bola venovaná adekvátna pozornosť. Uviedol viacero dôležitých argumentov prečo je treba učiť štatistiku. V ďalšej časti prednášky sa venoval neštandardným metódam štatistiky, kedy pre popísanie daného problému nestačia klasické štatistické metódy, ale je potrebné daný problém nahradiť modelom kvantových logík a fuzzy logík. Tieto metódy ilustroval na príkladoch zo života.

Po plenárnej prednáške nasledovali prihlásené referáty účastníkov konferencie. Súčasťou konferencie bol workshop pre študentov doktorandského štúdia z Teórie vyučovania matematiky na tému: ***Implikačná analýza a jej využitie v teórii vyučovania matematiky.***

Konferenciu garantoval organizačný výbor v zložení: Prof. RNDr. Anna Tirpáková, CSc. – predseda, FPV UKF Nitra, Doc. Ing. Jozef Chajdiak, CSc., STU Bratislava, Doc. PaedDr. Jana Kubanová, CSc., FES UPC Pardubice, doc. RNDr. Bohdan Linda, CSc., FES UPCE Pardubice, RNDr. Ján Luha, CSc. – LF UK a UN Bratislava, RNDr. Jitka Poměnková, PhD., PEF MU Brno, Prof. RNDr. Ondrej Šedivý, CSc., FPV UKF Nitra, Doc. RNDr. Marta Vrábelová, CSc., FPV UKF Nitra, Doc. RNDr. Dagmar Markechová, CSc., FPV UKF Nitra.

Organizačne konferenciu pripravil Organizačný výbor, ktorého predsedom bola PaedDr. Janka Melušová, PhD., FPV UKF v Nitre.

Prof. RNDr. Ondrej Šedivý, CSc.
Katedra matematiky FPV UKF v Nitre

Konferenciu slávnostne zahájil dekan Fakulty prírodných vied UKF v Nitre prof. RNDr. Ľubomír Zelenický, CSc.

prof. Ing. Mirko Navara, DrSc. počas svojej prednášky

Pohľad do pléna

Škola štatistiky Ekomstat 2013

School of statistics Ekomstat 2013

Jozef Chajdiak, Ján Luha

V dňoch 19. - 24.5.2013, v priestoroch Domova Speváckeho zboru slovenských učiteľov v Trenčianskych Tepliciach, Slovenská štatistická a demografická spoločnosť zorganizovala akciu EKOMSTAT 2013. Ekomstat (EKOnoMická ŠTATistika) sa prvý krát uskutočnil v roku 1988 a prebieha každoročne na uvedenom mieste s podnázvom akcie „škola štatistiky“.

Organizačný a programový výbor Ekomstat-u 2013: Doc. Ing. Jozef Chajdiak, CSc. – predseda; RNDr. Ján Luha, CSc. – tajomník; členovia: RNDr. Viliam Páleník, PhD. h. doc., Ing. Marek Radvanský, PhD., Doc. Ing. Beáta Gavurová, PhD., MBA, RNDr. Samuel Koróny, PhD, Ing. Vladimír Velikanič. Ekomstatu 2013 sa zúčastnilo 32 účastníkov z vysokých škôl, SAV, výskumných organizácií, hospodárskej praxe, rezortu štatistiky.

Akcia sleduje súčasne niekoľko cieľov. Prvým je konferenčný aspekt akcie. Účastníci prezentujú svoje príspevky. V hlavných prednáškach ide prakticky o výučbové prednášky, v ktorých autori (prednášatelia) prednášajú výsledky svojej vedeckovýskumnej práce (postup riešenia úlohy, výsledky riešenia úlohy, forma prezentácie, prehľad problematiky). U mladších autorov ide o aspekt vyrovnania sa s prípadnou trémou u prednášateľa, pestovanie schopnosti pružne reagovať na otázky v rámci diskusie. Dôležitou je tiež schopnosť formulovať otázky k prednášanej problematike.

Prednášky sformalizované do vedeckého alebo odborného článku sú publikované v Zborníku akcie alebo v niektorom z čísiel nášho časopisu Forum Statisticum Slovacum, konkrétne príspevky na Ekomstat 2013 v FSS 4/2013.

Ďalším cieľom je rozšírenie vedecko-odborného obzoru účastníka. Príspevky autorov sú zvyčajne orientované na témy z oblasti štatistických metód, analýzy kvalitatívnych znakov, z ekonomickej štatistiky, makroekonomických analýz, z riadiacich metód a postupov a prípadne iné témy.

Súčasťou EKOMSTATU sú aj vzájomné formálne aj neformálne diskusie, besedy, konzultácie účastníkov.

Na EKOMSTAT-e 2013 skupinu hlavných prednášajúcich a ich prednášky tvorili:

Luha, J.: Intervaly spoľahlivosti pre podiely.

Páleník, V.: Uskutočniteľnosť zavedenia stabilizačných dlhopisov v podmienkach EU - kvalitatívna analýza.

Haluška,J.: Rýchle odhady vývoja zamestnanosti.

Kaščáková, A. – Nedelová, G.: Metódy prípravy a spracovania dotazníkov.

Radvanský, M.: Odhad budúcich potrieb a nákladov zdravotnej starostlivosti.

Chajdiak, J.: Zadanie a výsledky štatistických pre zvyšovanie efektívnosti medzinárodnej spolupráce malých a stredných podnikov v oblasti inovácií.

Frankovič, B.: Kalibrácia váh štatistických zisťovaní

Aktívne na Ekomstate tiež vystúpili:

Beáta Gavurová, B. - Klepáková, A.: Vybrané aspekty adaptability modelu Flexicurity na Slovensku

Korónу, S.: Relatívne ukazovatele

Ďurechová, M.: Zvyšovanie konkurencieschopnosti a výkonnosti EÚ prostredníctvom rozvoja vedy, techniky a inovácií

Chajdiak, J.: Ekonomický informačný systém Slovensko - analýza aktuálneho vývoja vybraných makroekonomických ukazovateľov do apríla 2013.

Lichner, I.: Starnutie v poľnohospodárstve

König **,** B.:Analýza zahraničného obchodu v SR: ECM prístup

Páleník M.: Regionálny rozmer dlhodobej nezamestnanosti na Slovensku

Zajko, M.: Stav a možnosti rozvoja medzinárodnej spolupráce slovenských malých a stredných podnikov v inováciách

Lichner, I.: Zmena odvodového zaťaženia práce na dohodu

Miklošovič, T.: Porovnávanie efektivity výučby na fakultách slovenských verejných vysokých škôl s využitím DEA analýzy

Štefánik, M.: Spracovanie individuálnych údajov z Európskeho zisťovania o pracovných podmienkach zamerané na faktory pracovnej schopnosti starších pracovníkov

Horvát, P.: Komplikácie pri modelovaní vstupu práce: Prípad DSGE

Petríková, K.: Modelovanie striebornej ekonomiky

Majerník, M.: Due diligence - kde sa s ním môžeme stretnúť

Plchová, J.: Mapa úspešnosti ako nástroj merania konkurencieschopnosti firiem

Fuksová , N.: Zvyšovanie konkurenčnej spôsobilosti malých a stredných podnikov v SR v rámci medzinárodnej spolupráce

Chodasová , Z.: Ukazovatele konkurencieschopnosti

Tekulová, Z.: Controlling kapitálovej štruktúry v podniku

Mišota, B.: Systém na Správu Inovačných Informácií - výber vhodného technologického riešenia

Potančok, M.: Závislosť a príčina – filozofická reflexia

Chajdiak, J.: Objem tržieb – východiskový cieľ ekonomiky firmy

29. škola štatistiky EKOMSTAT 2014 sa uskutoční 25.-30.5.2014 v Trenčianskych Tepliciach v Domove Speváckeho zboru slovenských učiteľov.

Adresy autorov:

Jozef Chajdiak, Doc., Ing., CSc.
Ústav manažmentu STU
Vazovova 5, Bratislava
jozef.chajdiak@stuba.sk

Ján Luha, RNDr., CSc.
Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a UN
Sasinkova 4, Bratislava
jan.luha@fmed.uniba.sk

Prastan a Stakan – z histórie
PRASTAN and STAKAN – from the history

Martin Kalina, Oľga Nanásivá, Ján Luha

Prvá konferencia PRASTAN, vtedy pod názvom ***Výučba matematickej štatistiky, pravdepodobnosti a numerickej matematiky*** sa konala v roku 1996 v Kočovciach. Bola zameraná na výmenu skúseností vysokoškolských pedagógov. Prvý ročník konferencie bol bez zborníka. Počnúc rokom 1997 konferencie mali odbornú náplň, kde si pedagógovia vypočuli nové poznatky z matematickej štatistiky aj numerickej matematiky. V roku 1997 bol prvýkrát publikovaný aj konferenčný zborník.

Spoločná akcia Slovenskej štatistickej a demografickej spoločnosti a Českej štatistickej spoločnosti je spravidla organizovaná každé dva roky, striedavo v Čechách a na Slovenku (Čechy - Stakan 1999, Slovensko - Prastan 2001, Čechy - Stakan 2003, Slovensko - Prastan 2005, Čechy – Stakan 2007, Slovensko – Prastan 2009, Slovensko - 2013). Cieľom tejto akcie je neprerušiť a upevňovať dobré vzťahy a vzájomné poznanie sa medzi štatistikmi z ČR a SR.

Konferencia je určená hlavne vysokoškolským učiteľom, zaoberajúcim sa problematikou výučby a aplikácii štatistiky a príbuzných disciplín. Akcie na Slovensku majú medzi témami aj numerickú matematiku. Nezávisle na tejto spolupráci sa konferencia PRASTAN organizovala aj samostatne s tematickým zameraním na pravdepodobnosť, štatistiku a numerickú matematiku.

Od roku 2001 nesie konferencia meno PRASTAN. V roku 2003 sa konferencia nekonala.

Od roku 2005 sa konferenčné príspevky z PRASTANu publikujú vo vedeckom recenzovanom časopise FORUM STATISTIUCUM SLOVACUM, ktorý vydáva Slovenská štatistická a demografická spoločnosť. Zatiaľ posledný ročník konferencie bol v roku 2013.

Chronologický prehľad konferencií:

1. **Kočovce. 1996**, 8.6.-10.6. asi.,organizátori: SvF STU Bratislava, FM UK
2. **Kálnica 1997**, 16.6.-19.6, organizátori: SvF STU Bratislava, FM UK
3. **Kálnica 1998**, 1.6.-5.6., organizátori: SvF STU Bratislava, FM UK
4. **Kočovce 1999,** 14.6-18.6., organizátori: SvF STU Bratislava, FM UK
5. **Bezovec 2000**, 22.5-26.5., organizátori: SvF STU, Bratislava, FM UK, TEMPUS AC_JEP-13425-98

PRASTAN

6. **Kočovce 2001**, 12.9. - 14.9., organizátori: SvF STU, Bratislava, FM UK, FMFI UK, FHI UK SŠDS a ČŠS
7. **Bratislava 2002,** 23.10. – 24.10., organizátori SvF STU, FM UK, ŠSDS
8. **Kočovce 2004**, 17.5. - 21.5. organizátori: SvF STU, Bratislava, FM UK, VA Liptovský Mikuláš, SŠDS
9. **Tajov 2005,** hotel Lesák, 10.6. - 15.6., organizátori: SvF STU, Bratislava, FPV UMB, Banská Bystrica, VA Liptovský Mikuláš, FM UK, SŠDS
10. **Selce 2006,** 12.6.-16.6., organizátori: SvF STU, Bratislava, FPV UMB, Banská Bystrica, SŠDS
11. **Banská Bystrica 2007,** organizátori: SvF a SjF STU Bratislava, FPV UMB, Banská Bystrica, SŠDS, FM UK
12. **Kočovce 2009,** 10.6. - 12.6. organizátori: SvF STU Bratislava, SŠDS
13. **Prastan 2013,** 27.5. -29.5. 2013, organizátori: SvF STU Bratislava, SŠDS, JSMF

Hlavným organizátorom konferencií bola od prvého ročníka Stavebná fakulta STU v Bratislave. Od roku 2001 sa konferencie konajú pod patronátom Slovenskej štatistickej a demografickej spoločnosti.

Adresy autorov

Kalina Martin, Doc. RNDr. PhD.
STU V Bratislave, Stavebná fakulta
Katedra matematiky a deskr. geometrie
Radlinského 11;813 68 Bratislava
martin.kalina@stuba.sk

Nánásiová Oľga, doc. RNDr. PhD.
STU V Bratislave, Fakulta elektrotechniky a informatiky
Ústav informatiky a matematiky
Ilkovičova 3;812 19 Bratislava
olga.nanasiova@stuba.sk

Ján Luha, RNDr., CSc.
Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a UN
Sasinkova 4, 811 08 Bratislava
jan.luha@fmed.uniba.sk

OBSAH
CONTENTS

Pokyny pre autorov

Jednotlivé čísla vedeckého recenzovaného časopisu FORUM STATISTICUM SLOVACUM sú prevažne tematicky zamerané zhodne s tematickým zameraním akcií SŠDS. Príspevky v elektronickej podobe prijíma zástupca redakčnej rady na elektronickej adrese uvedenej v pozvánke na konkrétne odborné podujatie Slovenskej štatistickej a demografickej spoločnosti. Akceptujeme príspevky v slovenčine, češtine, angličtine, nemčine, ruštine a výnimočne po schválení redakčnou radou aj inom jazyku. Názov word-súboru uvádzajte a posielajte v tvare: ***priezvisko_nazovakcie.doc resp. docx***

Forma: Príspevky písané výlučne len v textovom editore MS WORD, verzia 6 a vyššia, písmo Times New Roman CE 12, riadkovanie jednoduché (1), formát strany A4, všetky okraje 2,5 cm, strany nečíslovať. Tabuľky a grafy v čierno-bielom prevedení zaradiť priamo do textu článku a označiť podľa šablóny. Bibliografické odkazy uvádzať v súlade s normou STN ISO 690 a v súlade s medzinárodnými štandardami. Citácie s poradovým číslom z bibliografického zoznamu uvádzať priamo v texte.

Rozsah: Maximálny rozsah príspevku je 6 strán.

Príspevky sú recenzované. Redakčná rada zabezpečí posúdenie príspevku oponentom.

Príspevky nie sú honorované, poplatok za uverejnenie akceptovaného príspevku je minimálne 30 €. Za každú stranu naviac je poplatok 5 €.

Štruktúra príspevku: ***(Pri písaní príspevku využite elektronickú šablónu: http://www.ssds.sk/ v časti Vedecký časopis, Pokyny pre autorov.). Časti v angličtine sú povinné!***

Názov príspevku v slovenskom jazyku (štýl Názov: Time New Roman 14, Bold, centrovať)
Názov príspevku v anglickom jazyku (štýl Názov: Time New Roman 14, Bold, centrovať)
Vynechať riadok
Meno1 Priezvisko1, Meno2 Priezvisko2 (štýl normálny: Time New Roman 12, centrovať)
Vynechať riadok

Abstrakt: Text abstraktu v slovenskom jazyku, max. 10 riadkov (štýl normálny: Time New Roman 12).

Abstract: Text abstraktu v anglickom jazyku, max. 10 riadkov (štýl normálny: Time New Roman 12).

Kľúčové slová: Kľúčové slová v slovenskom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

Key words: Kľúčové slová v anglickom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

JEL classification: Uviesť kódy klasifikácie podľa pokynov v:
<http://www.aeaweb.org/journal/jel_class_system.php>

Vynechať riadok a nastaviť si medzery odseku pre nadpisy takto: medzera pred 12 pt a po 3 pt. Nasleduje vlastný text príspevku v členení:

1. **Úvod** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať,)*
2. **Názov časti 1** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*
3. **Názov časti 1. . .**
4. **Záver** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

Vlastný text jednotlivých častí je písaný štýlom Normal: písmo Time New Roman 12, prvý riadok odseku je odsadený vždy na 1 cm, odsek je zarovnaný s pevným okrajom. Riadky medzi časťami a odsekmi nevynechávajte. Nastavte si medzi odsekmi medzeru pred 0 pt a po 3 pt.

5. **Literatúra** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

[1] Písať podľa normy STN ISO 690

[2] GRANGER, C.W. – NEWBOLD, P. 1974. Spurious Regression in Econometrics. In: Journal of Econometrics, č. 2, 1974, s. 111 – 120.

Adresa autora (-ov): ***Uveďte svoju pracovnú adresu!!!*** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, adresy vpísať do tabuľky bez orámovania s potrebným počtom stĺpcov a s 1 riadkom):*

Meno1 Priezvisko1, tituly1 (študenti ročník)
Pracovisko1 (študenti škola1)
Ulica1, 970 00 Mesto1
meno1.priezvisko1@mail.sk

Meno2 Priezvisko2 , tituly2 (študenti ročník)
Pracovisko2 (študenti škola2)
Ulica2, 970 00 Mesto2
meno2.priezvisko2@mail.sk

FORUM STATISTICUM SLOVACUM

vedecký recenzovaný časopis Slovenskej štatistickej a demografickej spoločnosti

Vydavateľ:
Slovenská štatistická a demografická spoločnosť
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Redakcia:
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Fax: 02/39004009

e-mail:
chajdiak@statis.biz
jan.luha@fmed.uniba.sk

Dátum vydania: august 2013

Registráciu vykonalo:
Ministerstvo kultúry Slovenskej republiky

Dátum registrácie: 22. 7. 2005

Evidenčné číslo: EV 3287/09

Tematická skupina: B1

Periodicita vydávania:
minimálne 2 krát ročne

Objednávky:
Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava 24
Slovenská republika

IČO: 178764
DIČ: 2021504276
Číslo účtu: 0011469672/0900

ISSN 1336-7420

Ročník: IX.

Číslo: 5/2013

Cena výtlačku: 30 EUR
Ročné predplatné: 120 EUR