INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 24, číslo 3–4, prosinec 2013

ROBUST 2012
KLIMATEXT 2012
REQUEST 2012

Vážené kolegyně, vážení kolegové,

toto číslo Bulletinu České statistické společnosti přináší vybrané práce 17. letní školy JČMF ROBUST 2012, kterou ve dnech 9.-14. září 2012 uspořádala v Němčičkách skupina pro výpočetní statistiku JČMF za podpory ČStS, KPMS MFF UK a KAP PF TUL, workshopu KLIMATEXT zaměřeného na na modelování extrémů v klimatologii, kerý se uskutečnil 9. září 2012 v Němčičkách. Součástí čísla jsou i příspěvky 7. konference Request 2012, pořádané 29. listopadu 2012 Centrem pro jakost a spolehlivost výroby a ČStS za podpory projektu OPVK Praxe pro praxi.

Toho číslo bylo vytištěno s laskavou podporou JČMF, konference ROBUST a s podporou projektu OPVK Klimatext č. CZ.1.07/2.3. 00/20.0086.

Jaromír Antoch, Gejza Dohnal a Jan Picek

evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Obsah

USING MIXED MODELS THEORY IN PROBLEMATICS OF OBSERVATIONAL STUDIES

Silvie Bělašková, Jan Janoušek

Keywords: Mixed models, REML, random effect, observational studies, ventricular pacing, Pacing site

Abstract:
Medicine is an intensively studied discipline. Not all phenomena can be studies solely by experimentation due to obvious ethical or logistical restrictions. Observational studies play an important role in investigating treatment oeffects. Many clinical studies are organized as multi-centre trials. It is usually because there is no adequate number of suitable patients in any single centre. The data within a given centre are assumed to be correlated. One of possible statistical approaches to analyze this kind of data is analysis using mixed models approach. In this report we deal with mixed model approach and REML method for estimation of parameters. As an illustration, we present an application of the mixed model approach to determine the effects of the site of ventricular pacing on left ventricular ejection fraction.

1. Introduction

Analysis of variance (ANOVA) and regression analysis are for many decades the mainstay of statistical data analysis. The basic assumption for this type of models is that the error terms for different sampling units are independently distributed. When this assumption is not justified a different approach is needed. Inclusion of random effects in ANOVA or regression models allows us to relax the independence assumption and take into account possible dependence between observations. Linear or generalized linear mixed model are statistical models for outcome variables in which the error terms may not be independent or have constant variance. Linear and generalized linear mixed models are suitable for modeling responses resulting from clustered, longitudinal, or repeated-measures designs. these models include fixed-effect parameters associated with one or more continuous or categorical covariates and random effects. The fixed-effect parameters describe the relationships of the covariates to the dependent variable for an entire population and the random effects are specific to the sampling unit. The inference usually includes estimation of model parameters as well as hypotheses tests about population parameters.

The matrix form of a linear mixed model in general:.

$$Y = X\beta + ZU + \varepsilon, \tag{1}$$

where Y is an n dimensional response vector and β is an m-vector of unknown fixed-effects parameters. The $n \times m$-matrix X and the $n \times p$-matrix Z are known design matrices for fixed and random effects, respectively.

The random effects vector U and the error vector ε are jointly normally distributed as

$$\begin{pmatrix} U \\ \varepsilon \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G & 0 \\ 0 & H \end{pmatrix}\right), \tag{2}$$

Where the matrices G and H represent the covariance matrices of U and ε and usually depend on unknown parameters (variance components).

The covariance matrix of Y is therefore $V = ZGZ' + H$. Under assumption of normality of the response Y, maximum likelihood and restricted maximum likelihood (REML) approach are generally adopted for estimation of variance components. Our approach for estimate of variance components was use REML. Corresponding log-likelihood function is following.

$$l_R(G, H) = -\frac{1}{2}\log|V| - \frac{1}{2}|X'V^{-1}X| - \frac{1}{2}r'V^{-1}r - \frac{n-p}{2}\log(2\pi) \tag{3}$$

where $r = Y - X(X'V^{-1}X)^{-}X'V^{-1}Y$ and p is the rank of X. REML provide estimates of G and H, which are denoted $\widehat{G}$ and $\widehat{H}$ respectively. This procedure estimated covariance matrix which is denoted $\widehat{V}$, which is then used to obtain estimates of β using the standard generalized least squares method.

$$\widehat{\beta} = (X'\widehat{V}^{-1}X)^{-1}X'\widehat{V}^{-1}Y \tag{4}$$

and with the estimated covariance matrix

$$var(\widehat{\beta}) = (X'\widehat{V}^{-1}X)^{-1}. \tag{5}$$

2. **Testing linear hypotheses**

The linear hypothesis to be tested is usually formulated as

$$H_0 : H'\beta = 0 \text{ vs. } H'\beta \neq 0, \tag{6}$$

where H is a known matrix of contrasts. The modified F statistic for testing the null hypothesis is calculated from the statistic which is given by

$$F = \frac{1}{DF}(H'\widehat{\beta})'(H'(X'\widehat{V}^{-1}X)^{-}H)^{-}(H'\widehat{\beta}). \tag{7}$$

In mixed models exact F-tests do not always exist for testing significance of all the variance components. The way how to test significance is to use approximate F-tests. One approach to approximate inference about fixed effects is the method of Kenward and Roger[4] for determination of DF.

3. **Application to real data**

The general approach of mixed model analysis has been to analysis of data resulting from a multicentre study described in details in paper Pacing Site: A Multi-Center Study.[3] In his study a cohort of 178 patients with slow heart rate was treated by seven treatment modalities (one modality per each patient) with a potential for differing adverse effects on cardiac function. Right

ventricular (RV) pacing is associated with asynchronous left ventricular (LV) activation, which can lead to deleterious pathological remodeling and LV failure. Several recent studies have demonstrated that increased percentage of RV apical correlates with morbidity and mortality from heart failure in adults The main aim of this study was to evaluate the effects of the site of ventricular pacing on left ventricular ejection fraction (LVEF) in children requiring permanent pacing. The response in this model is the left ventricular ejection fraction (LVEF). In cardiovascular physiology, ejection fraction represents the volumetric fraction of blood pumped out of the ventricle (heart) with each heart beat or cardiac cycle. Design matrix of this model includes the main factor tested and the set of additive covariates like age at implantation, pacing duration, and QRS duration which are continuous covariates. QRS duration is the name for the combination of three of the graphical deflections seen on a typical electrocardiogram. It is usually the central and most visually obvious part of the tracing. The dichotomous covariates were gender, presence of maternal antibodies, presence of congenital block, and DDD pacing. The main treatment factor was the pacing site (Figure 2) with seven levels or a combination of specific pacing sites: free wall of the RV outflow tract (RVOT), lateral RV wall (RVLat), RV apex (RVA), RV septum (RVS) (any position), LV apex (LVA), lateral LV wall (LVLat), and LV base (LVB)(Figure 3). As an additive random effect was included the variable "Contributing center" (Figure 1). For the random "Center" effect a simple covariance structure was assumed. The statistical test of main treatment effect was an Kenward-Roger's adjusted F tests which is available in SAS and widely used. For the "Pacing site" main effect, multiple comparisons were performed using the Tukey-Kramer adjustment. REML is used as the estimation methods for the covariance parameters in mixed model. Statistical software package SAS Version 9.3 (SAS Institute, NC, USA) and R version 2.15.1 were used for all statistical analysis. Significance was accepted at 0.05 level.

4. **Conclusion**

Age at implantation, pre-implantation LV size and function, duration of pacing, DDD mode, QRS duration, and presence of maternal auto-antibodies had no significant impact on LVEF. The site of ventricular pacing has a major impact on LV efficiency in children that require life-long pacing. LVA/LVLat pacing allows for optimal prevention of pacing-induced heart failure. The original data are included in a manuscript submitted to Circulation[3].

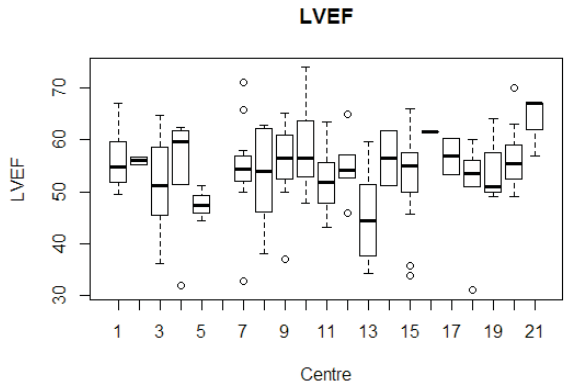


FIGURE 1. Box Plot classified by "Centre".

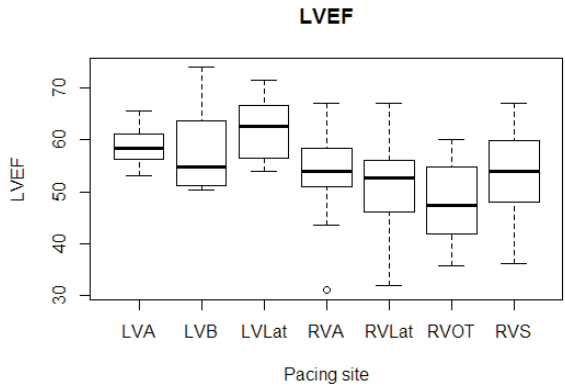


FIGURE 2. Box Plot classified by "Pacing site".

Literature

[1] Ramon C. Littell, George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, Oliver, Ph.D. Schabenberber: SAS for Mixed Models. Second Edition. SAS Inst. 2006.

[2] Roman A. Gebauer, Viktor Tomek, Petr Kubuš, Vít Rázek, Tomáš Matějka, Aida Salameh, Martin Kostelka, and Jan Janoušek: Differential effects of the site of permanent epicardial pacing on left ventricular synchrony and function in the young: implications for lead placement. Europace (2009) 11, 1654–1659

[3] Janousek J, van Geldorp IE, Krupicková S, Rosenthal E, Nugent K, Tomaske M, Früh A, Elders J, Hiippala A, Kerst G, Gebauer RA, Kubus P, Frias P, Gabbarini F, Clur SA, Nagel B, Ganame J, Papagiannis J, Marek J, Tisma-Dupanovic S, Tsao S, Nürnberg JH, Wren C, Friedberg M, de Guillebon M, Volaufova J, Prinzen FW, Delhaas T. Permanent

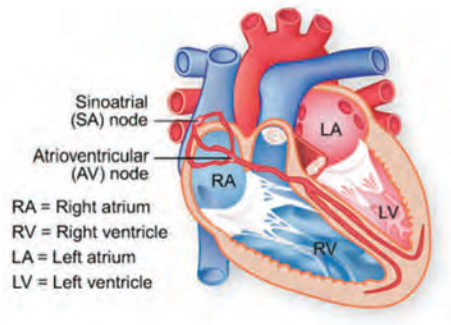


FIGURE 3. Diagram of the electrical conduction system of the heart [2]

Cardiac Pacing in Children - Choosing the Optimal Pacing Site: A Multi-Center Study. Circulation. 2012 Dec 30. [Epub ahead of print] PubMed PMID: 23275383.

[4] Kenward, M. G. and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," Biometrics 53: 983-997.

Acknowledgement: Supported by the grant PřF-2012-017 of the Internal Grant Agency of Palacky University Olomouc.

Address: 1 Tomas Bata University in Zlin, Czech Republic;
2 Children's Heart Centre, University Hospital Motol, Prague, Czech Republic

E-mail: belaskova@fai.utb.cz

INTERVALOVÁ DATA, ALGORITMY A VÝPOČETNÍ SLOŽITOST

Michal Černý[1] **and Milan Hladík**[2]

Klíčová slova: Intervalová data, algoritmy, výpočetní složitost, regrese.

Keywords: Interval data, algorithms, computational complexity, linear regression.

Abstrakt: Tato přehledová práce shrnuje některé algoritmické výsledky o problémech, které vznikají při analýze intervalových dat. Zkoumá algoritmické vlastnosti množiny všech možných hodnot min-norm estimátoru lineárního regresního modelu, jehož data probíhají dané intervaly. (Min-norm estimátorem se pro účely tohoto textu rozumí minimalizace $\|y - X\beta\|$, kde (X, y) jsou data a $\|\cdot\|$ je L_1, L_2 nebo L_∞-norma.) Text se také věnuje zkoumání intervalu, ve kterém se pohybuje reziduální součet čtverců (a analogické statistiky při užití L_1 či L_∞-normy). Text se také zabývá případy, kdy výpočet hodnot i banálních statistik nad jednorozměrnými intervalovými daty, jakými je např. výběrový rozptyl, t-poměr či F-poměr, je algoritmicky obtížný (***NP***-těžký).

1 Úvod

Intervalová data. V tomto textu rozumíme pojmu *intervalová data* $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$ jedním z následujících způsobů.

- **Způsob první.** Existuje skutečná (teoreticky pozorovatelná) hodnota x, kterou ovšem neznáme; známe jen interval $[\underline{x}, \overline{x}]$, o kterém s jistotou víme, že v něm hodnota x leží. Přitom na x nepohlížíme jako na náhodnou veličinu s nosičem $[\underline{x}, \overline{x}]$. (V tomto textu je důvod prozaický: žádný předpoklad o distribuci x nad intervalem $[\underline{x}, \overline{x}]$ nepotřebujeme. Nicméně prakticky zajímavá je situace, kdy pro přijetí předpokladu o konkrétní distribuci x nad $[\underline{x}, \overline{x}]$ nemáme dostatek informací.)
- **Způsob druhý.** Pozorovaná data jsou ze své podstaty intervalová. Z jejich smyslu vyplývá, že předpoklad distribuce na pozorovaném intervalu nemá dobrou interpretaci. Příkladem je informace, že jistá významná událost se konala od $\underline{x} =$ *9. září 2012* do $\overline{x} =$ *14. září 2012*.

Příklady. Intervalová data vznikají v řadě reálných situací. Uveďme několik příkladů.

[1]Katedra ekonometrie, Fakulta informatiky a statistiky, Vysoká škola ekonomická Praha. Nám. W. Churchilla 4, 13067 Praha 3. E-mail: cernym@vse.cz. Autor byl podpořen grantem GAČR P402/12/G097.

[2]Katedra aplikované matematiky, Matematicko-fyzikální fakulta, Univerzita Karlova Praha. Malostranské nám. 25, 18000 Praha 1. E-mail: milan.hladik@matfyz.cz. Autor byl podpořen grantem GAČR P403/12/1947.

- *Zaokrouhlování a reprezentace dat pomocí datových typů s omezeným počtem desetinných míst.* Reprezentujeme-li například číslo $x = 0.8156$ jen na jedno desetinné místo, tj. ve tvaru $\widetilde{x} = 0.8$, ztrácíme jistou informaci: o čísle x pak s jistotou víme jen to, že leží v intervalu $[\widetilde{x} - 0,05; \widetilde{x} + 0,05]$.

- *Obecněji: při ztrátě informace.* Ke ztrátě informace může také docházet při kategorizaci dat, při utajování individuálních hodnot či při diskretizaci spojitých dat. Například po kategorizaci dat $x_1, \ldots, x_n \in \mathbb{R}$ máme pro každou kategorii k dispozici jen meze $\underline{x}, \overline{x}$ a bez dalších informací může být obtížné na intervalu $[\underline{x}, \overline{x}]$ předpokládat distribuci individuálních hodnot, jež do kategorie spadají.

- *Nestabilita dat.* Stává se, že „konstanty" nejsou ve skutečnosti konstanty (např. některé fyzikální „konstanty" se mohou mírně měnit s polohou); pak může být vhodné takovou „konstantu" nahradit intervalem. Jiným příkladem je situace, kdy máme k dispozici pozorování jisté veličiny v diskrétních časových obdobích; uvnitř období ovšem veličina stabilní (konstantní) není.

- *Predikce.* Predikce budoucích hodnot ekonomických, klimatologických či podobných veličin bývají často intervalové. Uvažme pro příklad situaci, kdy jeden model generuje intervalovou predikci budoucí inflace. (Tím modelem může být ekonometrický model, ale také třeba panel expertů.) Předpokládejme, že tato predikce utváří inflační očekávání. Druhý model, např. model spotřební či model kapitálových výdajů, má inflační očekávání jako jeden z regresorů. Pak se druhý model musí vypořádat se situací, kdy mezi regresory figurují intervalová data.

Pravděpodobnostní pohled na intervalová data. Ve statistice je obvyklé na intervalová data $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_t, \overline{x}_t], \ldots, [\underline{x}_n, \overline{x}_n]$ nahlížet pravděpodobnostně: často tak, že data vznikla realizací dvojice procesů $v_t \in \mathbb{R}$, $w_t \geq 0$ takových, že v_t generuje středy $\frac{1}{2}(\overline{x}_t + \underline{x}_t)$ a w_t generuje poloměry $\frac{1}{2}(\overline{x}_t - \underline{x}_t)$. Odtud je pak možné začít budovat teorii (např. předpokládat vhodná rozdělení v_t, w_t, konstruovat estimátory jejich parametrů, konstruovat testy apod.).

Algoritmický pohled na intervalová data. V tomto textu se zabýváme jiným pohledem na data $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$. Hlavní otázkou je, jaké vlastnosti mají *algoritmy*, které intervalová data zpracovávají. Půjde nám především o vlastnosti zajímavé z hlediska teoretické informatiky, tj. o otázky z oblasti rozhodnutelnosti a výpočetní složitosti.

Pro algoritmus jsou data vždy jen *pevné konstanty*, se kterými algoritmus provádí jisté manipulace; jediné, co budeme předpokládat, je, že jde o racionální čísla. (Iracionální čísla totiž nelze na počítači reprezentovat už z toho prostého důvodu, že jich je nespočetně.)

2 Značení

Intervalové matice. Intervaly, intervalové vektory a intervalové matice značíme tučně. *Intervalová matice* $\boldsymbol{X} = [\underline{X}, \overline{X}]$ *rozměru* $n \times p$ je systém reálných matic

$$\{X \in \mathbb{R}^{n\times p} : \underline{X} \leq X \leq \overline{X}\},$$

kde relace $\leq$ se rozumí po složkách. Systém všech intervalových matic rozměru $n \times p$ značíme $\mathbb{IR}^{n\times p}$.

Střed a poloměr. *Středem* intervalové matice $\boldsymbol{X}$ rozumíme matici

$$X^c := \tfrac{1}{2}(\overline{X} + \underline{X})$$

a *poloměrem* intervalové matice $\boldsymbol{X}$ rozumíme matici

$$X^\Delta := \tfrac{1}{2}(\overline{X} - \underline{X}).$$

Okamžitě je vidět, že matice X^Δ je nezáporná.

Analogické pojmy jako pro intervalové matice užíváme také pro intervalové vektory. Není-li řečeno jinak, vektory chápeme jako sloupcové.

3 Množiny B_k

Je-li dána funkce $f(x) : \Psi \longrightarrow \mathbb{R}$, pak symbolem $\operatorname{argmin}_{x\in\Psi} f(x)$ rozumíme množinu všech $x \in \Psi$, ve kterých funkce f nabývá minima na Ψ.

Množina B s obecnou normou [9]. Nech» $\|\cdot\|$ značí některou vektorovou normu. Budeme se zabývat algoritmickými vlastnostmi množiny

$$\begin{aligned} B(\boldsymbol{X}, \boldsymbol{y}) &:= \{b \in \mathbb{R}^p : \quad b \in \operatorname{argmin}_{b'\in\mathbb{R}^p} \|y - Xb'\|, \quad X \in \boldsymbol{X}, \quad y \in \boldsymbol{y}\} \\ &= \bigcup_{X\in\boldsymbol{X}, y\in\boldsymbol{y}} \operatorname{argmin}_{b'\in\mathbb{R}^p} \|y - Xb'\|, \end{aligned} \tag{1}$$

je-li dána matice $\boldsymbol{X} \in \mathbb{IR}^{n\times p}$ a vektor $\boldsymbol{y} \in \mathbb{IR}^n$.

Vezmeme-li za $\|\cdot\|$ například L_2 normu $\|x\|_2 = \sqrt{x^{\mathrm{T}}x}$, obdržíme množinu

$$B_2(\boldsymbol{X}, \boldsymbol{y}) := \{b \in \mathbb{R}^p : X^{\mathrm{T}}Xb = X^{\mathrm{T}}y \text{ pro některé } X \in \boldsymbol{X} \text{ a } y \in \boldsymbol{y}\}. \tag{2}$$

Význam množiny B_2 [3, 4]. Množinu B_2 můžeme chápat jako množinu *všech možných odhadů*, které lze získat metodou nejmenších čtverců, jestliže v lineárním regresním modelu $y = X\beta + \varepsilon$ necháme data (X, y) probíhat intervaly $(\boldsymbol{X}, \boldsymbol{y})$. (Množinu B_2 záměrně definujeme algebraicky vztahem (2), abychom se v tomto textu, který je zaměřen na algoritmické problémy, mohli vyhnout [nesnadným] otázkám, jak přesně rozumět lineárnímu regresnímu modelu „$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$“ s intervalovými daty, jak přesně rozumět disturbanci ε apod.).

Množina B_2 má také tento význam. Interpretujme data $(\boldsymbol{X}, \boldsymbol{y})$ tak, že intervaly $(\boldsymbol{X}, \boldsymbol{y})$ obsahují nám neznámé hodnoty (X, y), které by ovšem teoreticky bylo možné změřit. Pak množina B_2 jistě obsahuje hodnotu $\widehat{\beta} =$

$(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$, která nás zajímá. Množina B_2 tedy poskytuje meze, ve kterých se nám nedostupná hodnota $\widehat{\beta}$ jistě nachází, jestliže namísto pozorování (X, y) máme k dispozici jen intervaly $(\boldsymbol{X}, \boldsymbol{y})$. Při znalosti množiny B_2 si můžeme začít klást např. otázku, zdali je (v nějakém smyslu) „velká“ či „malá“: je-li „malá“, můžeme neformálně říci, že ztráta informace způsobená tím, že namísto hodnot (X, y) máme k dispozici jen intervaly $(\boldsymbol{X}, \boldsymbol{y})$, má na znalost hodnoty estimátoru „malý vliv“. Může se také stát, že množina B_2 je v některém směru „úzká“ a v některém směru „široká“ — pak můžeme neformálně říci, že ztráta informace způsobená tím, že namísto hodnot (X, y) máme k dispozici jen intervaly $(\boldsymbol{X}, \boldsymbol{y})$, má na znalost odhadu jednoho regresního parametru „malý vliv“, zatímco na znalost odhadu jiného regresního parametru může mít „velký vliv“. Buďme přesnější: jestliže například zjistíme, že

$$B_2 \subseteq \boldsymbol{b}, \tag{3}$$

kde $\boldsymbol{b}$ je intervalový vektor (box) takový, že $\overline{b}_i - \underline{b}_i = \delta_i$, kde $i = 1, \ldots, p$, pak můžeme říci, že hodnotu odhadu i-tého regresního parametru známe s chybou nanejvýš δ_i.

Množinou B_2 se budeme zabývat v části 4.

Množina B s jinými normami [9]. Zvolíme-li v (1) jinou normu $\|\cdot\|$, obdržíme množinu všech možných hodnot estimátoru $\operatorname{argmin}_{b'} \|y - Xb'\|$, probíhají-li data (X, y) intervaly $(\boldsymbol{X}, \boldsymbol{y})$. (Nezabýváme se zde otázkou, za jaké situace je vhodné volit ten či onen estimátor.) Například:

- L_1-*norma*: B je množina všech možných hodnot odhadů metodou LAD (Least Absolute Deviations), která má velký význam v robustní statistice;
- L_∞-*norma*: B je množina všech možných hodnot odhadů pomocí čebyševovské aproximace;
- Ω-*norma* ($\|x\|_\Omega = \sqrt{x^{\mathrm{T}}\Omega^{-1}x}$, kde Ω je positivně definitní): B je množina všech možných hodnot odhadů metodou zobecněných nejmenších čtverců s kovarianční maticí Ω.

Nadále budeme užívat značení

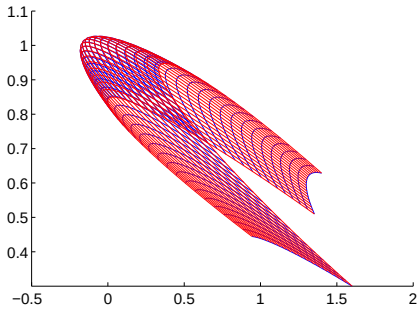
$$B_k := \text{množina } B \text{ definovaná v (1) s } L_k\text{-normou.}$$

Algoritmickým vlastnostem množiny B_2 věnujeme část 4. Množinami B_k s $k = 1$ a $k = \infty$ se budeme zabývat v části 5.

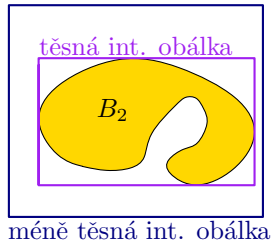
4 Algoritmické vlastnosti množiny B_2

4.1 Obálka množiny B_2

Množina B_2 je obecně složitá; nemusí být ani konvexní. Je proto obtížné podat jiný popis množiny B_2, než je její definice (2).



Obrázek 1: Množina B_2 s daty (4).



Obrázek 2: Intervalové obálky.

Příklad. Uvažme

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 \\ 1 & [0;5] \\ 1 & [2;4] \\ 1 & 4 \end{pmatrix}, \qquad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}. \tag{4}$$

Množina $B_2(\boldsymbol{X}, y)$ je zachycena na obrázku 1.

Intervalová obálka. Motivaci k následujícímu postupu jsme podali již v (3): chceme-li „nahlédnout“, jak složitá množina B_2 přibližně vypadá, je přirozené se pokusit sestrojit její aproximaci pomocí jednoduchého geometrického objektu, například elipsy či boxu (intervalového vektoru, tzv. *intervalové obálky* – viz obrázek 2). Zde se omezíme jen na intervalové obálky. Položme si otázku, jak je možné zkonstruovat — pokud možno — *těsnou* intervalovou obálku, tj. intervalový vektor $\boldsymbol{b}$ takový, že platí $\boldsymbol{b} \supseteq B_2$, ale pro žádný intervalový vektor $\boldsymbol{b}' \subsetneqq \boldsymbol{b}$ už neplatí $\boldsymbol{b}' \supseteq B_2$?

Omezenost množiny B_2. Aby vůbec mělo smysl se do konstrukce intervalové (ale také elipsoidové či jiné) obálky pouštět, musíme být nejprve schopni zodpovědět otázku: *je množina B_2 omezená*? Jinými slovy: každý algoritmus, který (korektně) zkonstruuje jakoukoliv intervalovou (či elipsoidovou či jinou) obálku množiny B_2, musí také být schopen rozhodnout, zdali

množina B_2 je omezená.

Cílem následujícího textu je ukázat, že tento problém není z algoritmického hlediska vůbec snadný.

4.2 Složitost rozhodování o omezenosti množiny B_2 — krok první: rozhodnutelnost

Je problém rozhodnutelný? Mějme data $(\boldsymbol{X}, \boldsymbol{y})$ a pokusme se navrhnout algoritmus, který rozhodne, zdali množina $B_2(\boldsymbol{X}, \boldsymbol{y})$ je omezená. Na první pohled není zřejmé, zdali takový algoritmus vůbec existuje: mohlo by se stát, že tento problém je nerozhodnutelný. Připomeňme, že matematika zná řadu problémů, které jsou algoritmicky nerozhodnutelné. Například:

- *Nulové body funkcí.*
  - *Zadání:* funkce $f(x) : \mathbb{R} \longrightarrow \mathbb{R}$ složená z konstant, $+, -, \times, \sin(\cdot)$.
  - *Úkol:* rozhodnout, zdali existuje $x_0 \in \mathbb{R}$ splňující $f(x_0) = 0$.
- *Konvergence integrálu.*
  - *Zadání:* funkce $f(x) : \mathbb{R} \longrightarrow \mathbb{R}$ složená z konstant, $+, -, \times, \div$, $\sin(\cdot)$.
  - *Úkol:* rozhodnout, zdali $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$ konverguje.
- *Diofantické rovnice* [tzv. Matijasevičova věta, také tzv. desátý Hilbertův problém, [12]].
  - *Zadání:* polynom $p(x_1, \ldots, x_9)$ s celočíselnými koeficienty.
  - *Úkol:* rozhodnout, zdali existují $x_1^*, \ldots, x_9^* \in \mathbb{Z}$ splňující $p(x_1^*, \ldots, x_9^*) = 0$.
- *Maticová smrt.*
  - *Zadání:* celočíselné čtvercové matice $A_1, \ldots, A_n$.
  - *Úkol:* rozhodnout, zdali lze z matic $A_1, \ldots, A_n$ násobením (v libovolném pořadí, opakování povoleno) obdržet nulovou matici.
- *Celočíselné programování s kvadratickými omezeními.*
  - Celočíselné lineární programování je optimalizační problém $\max\{c^{\mathrm{T}}x : \ Ax \leq b, \ \ x \in \mathbb{Z}^p\}$. Připustíme-li vedle lineárních omezení $Ax \leq b$ také kvadratická omezení (tj. připustíme-li součiny dvou proměnných), je otázka „je úloha přípustná?“ nerozhodnutelná.
- *Dokazatelnost* [tzv. Gödelova věta, viz [2]].
  - *Zadání:* tvrzení (= uzavřená formule množinového jazyka) φ.
  - *Úkol:* rozhodnout, zdali tvrzení φ je dokazatelné v teorii množin.

- *Halting problem.*
  - *Zadání:* text programu P a data x.
  - *Úkol:* rozhodnout, zdali výpočet programu P nad daty x skončí, anebo zdali počítá věčně.
- *Isomorfismus prezentovaných grup.*
  - *Zadání:* dvě prezentace grup (G_1, R_1) a (G_2, R_2), kde G_i je množina generátorů a R_i je systém relací, které generátory splňují (např. cyklická grupa řádu n má prezentaci $(G = \{a\}, R = \{a^n = 1\})$).
  - *Úkol:* rozhodnout, zdali grupy s prezentacemi (G_1, R_1) a (G_2, R_2) jsou isomorfní.
- *Náhodnost hodů mincí.*
  - *Zadání:* konečná posloupnost γ nul a jedniček a číslo K.
  - *Úkol:* rozhodnout, zdali posloupnost γ má kolmogorovskou složitost $\geq K$.

Následující věta ukazuje, že naše situace není tak špatná: problém omezenosti množiny B_2 seznam nerozhodnutelných problémů neprodlouží.

Věta 4.1 ([3])**.** *Nechť data $(\boldsymbol{X}, \boldsymbol{y})$ jsou racionální. Rozhodnout, zdali množina $B_2(\boldsymbol{X}, \boldsymbol{y})$ je omezená, lze algoritmicky.*

Idea důkazu. Snadno se nahlédne, že platí: množina B_2 je neomezená, právě když existuje $X \in \boldsymbol{X}$ neplné sloupcové hodnosti, a to nastává právě tehdy, když

$$(\exists X \in \mathbb{R}^{n\times p})[\underline{X} \leq X \leq \overline{X} \ \ \& \ \ \det X^{\mathrm{T}} X = 0]. \tag{5}$$

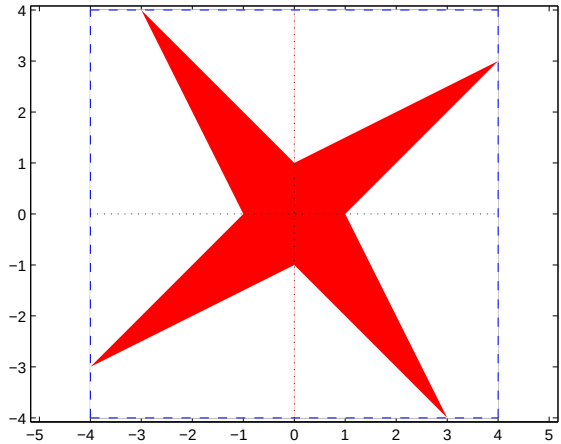
Výraz (5) je formule jazyka teorie těles (formuli (5) lze skutečně zapsat jen pomocí logických operací a pomocí operací $\leq$, $+$, $-$, $\times$, protože determinant je polynom). Podle Tarského věty je teorie reálně uzavřených těles rozhodnutelná pomocí eliminace kvantifikátorů (blíže viz [2, 15, 16, 17, 19]). Eliminaci kvantifikátorů tudíž můžeme použít i na rozhodnutí o (5). □

Efektivita algoritmů. Algoritmus z důkazu věty 4.1 ovšem není prakticky užitečný: pracuje totiž extrémně dlouho, obecně až v čase $\approx 2^{2^N}$, kde N je velikost vstupu $(\boldsymbol{X}, \boldsymbol{y})$. Otvírá se přirozená otázka, zdali je možné najít rychlejší algoritmus. K tomu potřebujeme základní věty z intervalové analýzy.

4.3 Lineární rovnice s intervalovými koeficienty

Řešení lineárních rovnic s intervalovými koeficienty. Následující definice říká, jak pro účely tohoto textu rozumíme pojmu *řešení* systému lineárních rovnic s intervalovými koeficienty.

Definice 4.1. *Nechť $\boldsymbol{A} \in \mathbb{IR}^{n\times p}$ a $\boldsymbol{b} \in \mathbb{IR}^n$. Vektor $z_0 \in \mathbb{R}^p$ je **řešením systému** $\boldsymbol{A}z = \boldsymbol{b}$, jestliže existují $A \in \boldsymbol{A}$ a $b \in \boldsymbol{b}$ takové, že $Az_0 = b$.*



Obrázek 3: Množina řešení systému rovnic (6).

Příklad (tzv. Barthův-Nudingův systém)**.** Množinu řešení soustavy rovnic

$$\begin{pmatrix} [2,4] & [-2,1] \\ [-1,2] & [2,4] \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} [-2,2] \\ [-2,2] \end{pmatrix} \tag{6}$$

ukazuje obrázek 3.

Oettliho-Pragerova věta [6]. Obrázek 3 naznačuje, že množina řešení systému intervalových rovnic má jistou pozoruhodnou strukturu: omezíme-li se na jeden (libovolně zvolený) orthant, vypadá jako konvexní polyedr. To skutečně není náhoda.

Věta 4.2. *Vektor $z \in \mathbb{R}^p$ je řešením systému $\boldsymbol{A}z = \boldsymbol{b}$, právě když platí*

$$|A^c z - b^c| \leq A^\Delta |z| + b^\Delta,$$

kde absolutní hodnota vektoru se rozumí po složkách. □

Důsledek 4.1. *Nechť $s \in \{\pm 1\}^p$. Nechť $\mathbb{R}^p_s$ označuje orthant $\{x \in \mathbb{R}^p : \operatorname{diag}(s)x \geq 0\}$. Množinu řešení systému $\boldsymbol{A}z = \boldsymbol{b}$ ležících v orthantu $\mathbb{R}^p_s$ tvoří polyedr*

$$\begin{aligned} \{z \in \mathbb{R}^p : \ & (A^c - A^\Delta \operatorname{diag}(s))z \leq \underline{b}, \\ & (-A^c - A^\Delta \operatorname{diag}(s))z \leq -\overline{b}, \ \operatorname{diag}(s)z \geq 0\}. \end{aligned} \tag{7}$$

□

4.4 Složitost rozhodování o omezenosti množiny B_2 — krok druhý: exponenciální algoritmus

Pozorování. Snadno nahlédneme, že platí:

množina B_2 je neomezená

$\longleftrightarrow$ existuje $X \in [\underline{X}, \overline{X}]$ neplné sloupcové hodnosti

$\longleftrightarrow$ intervalový systém $\boldsymbol{X}z = 0$ má nenulové řešení.

Algoritmus [3, 9]. Důsledek 4.1 ukazuje, že test existence nenulového řešení systému $\boldsymbol{X}z = 0$ je možné realizovat tak, že projdeme všech 2^p orthantů; v každém pak stačí řešit jeden lineární program. Vystačíme tedy s výpočetním časem

$$2^p \times (\text{polynomiální čas pro lineární programování}), \tag{8}$$

protože lineární programování má polynomiální algoritmy [2, 18]. Výpočetní čas (8) je podstatně lepší, než byl výpočetní čas eliminace kvantifikátorů z důkazu věty 4.1.

Složitostní výsledek. Z důsledku 4.1 vyplývá ještě jedno pozoruhodné zjištění.

Důsledek 4.2. *Otázka „je množina B_2 omezená?“ je v co-**NP**.* □

Důsledek 4.2 vyplývá z toho, že pozitivní odpověď na komplement — otázku „je množina B_2 neomezená?“ — lze doložit ***NP***-certifikátem. Certifikátem je orthant s, ve kterém má polyedr (7) nenulový bod.

4.5 Složitost rozhodování o omezenosti množiny B_2 — krok třetí: *co-NP*-úplnost

Přirozeně se otvírá otázka, zdali lze rozhodovat o omezenosti množiny B_2 v polynomiálním čase. Odpověď je záporná (leda by platilo $\boldsymbol{P} = \boldsymbol{NP}$):

Věta 4.3 ([3]). *Otázka „je množina B_2 omezená?“ je co-**NP**-úplná.* □

Připomeňme, že neformálně můžeme přijmout tuto definici: rozhodovací problém A je *co-**NP***-úplný, je-li $A \in$ *co-**NP*** a platí-li: jestliže problém A lze řešit v polynomiálním čase, pak lze v polynomiálním čase řešit libovolný problém v *co-**NP***. (Přesněji: $A \in$ *co-**NP*** a každý problém z *co-**NP*** je v polynomiálním čase převoditelný na problém A.)

Věta 4.3 ukazuje, že patrně nemůžeme očekávat existenci rychlých (subexponenciálních, nebo dokonce polynomiálních) algoritmů rozhodujících, zdali množina B_2 je omezená. A tudíž nemůžeme ani očekávat existenci rychlých algoritmů pro problém, který nás skutečně zajímá — totiž pro konstrukci intervalových (či elipsoidových či jiných) obálek množiny B_2. A tím spíše nemůžeme očekávat existenci rychlých algoritmů pro těsné intervalové obálky.

Otázka o omezenosti množiny B_2 se tak zařazuje do široké třídy známých *co-**NP***-úplných problémů; připomeňme zde například

- problém rozhodnout, zdali daná výroková formule je tautologie;
- problém rozhodnout, zdali je pravda, že barevnost daného grafu je větší než předepsané číslo k (tento problém zůstává *co-**NP***-úplný dokonce i při fixaci čísla k na libovolnou pevnou hodnotu ≥ 3);
- problém rozhodnout, zdali je pravda, že klikové číslo daného grafu je menší než předepsané číslo k;
- problém rozhodnout, zdali je pravda, že přípustný prostor daného celočíselného lineárního programu $\max\{c^{\mathrm{T}}x : Ax \leq b, x \in \mathbb{Z}^p\}$ je prázdný.

Komentář k větě 4.3. Věta 4.3 ukazuje, že patrně neexistuje *jeden obecný* rychlý algoritmus pro rozhodování o omezenosti, a tudíž i pro konstrukci intervalových obálek množiny B_2, který by fungoval v polynomiálním čase, a» mu předložíme jakákoliv data $(\boldsymbol{X}, \boldsymbol{y})$, tj. a» je počet pozorování n a dimenze p prostoru parametrů libovolná. Nicméně to neznamená, že by problém musel být nutně beznadějný ve speciálních případech. A zde se ukazuje, že situace je lepší v některých speciálních případech, které se při analýze dat často vyskytují. Negativní výsledek věty 4.3 lze také chápat jako motivaci ke studiu těchto speciálních případů. Těmi se budeme zabývat v následujících částech.

4.6 Dva speciální případy

Speciální případ I: modely s nezápornými hodnotami regresních parametrů [3]. Někdy se stává, že z věcného (fyzikálního, ekonomického apod.) hlediska mají smysl pouze nezáporné hodnoty regresních parametrů. V takovém případě se stačí omezit jen na nezáporný orthant v prostoru parametrů. Pak se výpočetní čas (8) sníží; stačí totiž prozkoumat jediný orthant, a tak stačí řešit jediný lineární program.

Důsledek 4.3. *O omezenosti množiny $B_2 \cap \{b : b \geq 0\}$ lze rozhodovat v polynomiálním čase.* □

Speciální případ II: modely s nízkým počtem regresních parametrů [3]. Již speciální případ I ukázal, že by» je výpočetní čas (8) exponenciální, situace není tak špatná: výpočetní čas (8) je exponenciální *v dimenzi prostoru parametrů, nikoliv ovšem v počtu pozorování.* (Skutečně špatnou zprávou by bylo, kdyby se v (8) vyskytoval faktor 2^n namísto faktoru 2^p.) Omezíme-li se na třídu modelů s nanejvýš p_0 regresními parametry, kde p_0 je pevná konstanta (pro praxi postačí např. $p_0 = 10$), pak 2^{p_0} je kontanta. A tudíž:

Důsledek 4.4. *Ve třídě modelů s nanejvýš p_0 regresními parametry lze o omezenosti množiny B_2 rozhodovat v polynomiálním čase.* □

Užijme přesnější formulace, aby lépe vynikl rozdíl mezi tvrzením věty 4.3 a tvrzením důsledku 4.4.

Věta 4.3 říká, že množina

$$\{(\boldsymbol{X}, \boldsymbol{y}) : \text{množina } B_2(\boldsymbol{X}, \boldsymbol{y}) \text{ je omezená}\}$$

je *co-**NP***-úplná. Naproti tomu důsledek 4.4 říká, že pro každé p_0 platí

$$A_{p_0} \in \boldsymbol{P},$$

kde

$$A_{p_0} := \{(\boldsymbol{X}, \boldsymbol{y}) : \text{matice } \boldsymbol{X} \text{ má} \le p_0 \text{ sloupců a množina } B_2(\boldsymbol{X}, \boldsymbol{y}) \text{ je omezená}\}.$$

Tvrzení důsledku 4.4 také můžeme formulovat tak, že problém je rozhodnutelný v polynomiálním čase v případě, kdy počet regresních parametrů je pevný a počet pozorování roste přes všechny meze.

Silnější formulace speciálního případu II. Snadno se nahlédne, že můžeme vyslovit i silnější formulaci: stačí totiž, aby počet regresních parametrů p nerostl vzhledem k počtu pozorování n příliš rychle. Je-li totiž $p = O(\log n)$, pak pro jisté pevné k jest $2^p \approx n^k$ a výpočetní čas (8) získá polynomiální tvar

$$n^k \cdot (\text{polynomiální čas pro lineární programování}).$$

4.7 Ještě jeden speciální případ: intervaly jen v pozorování vysvětlované proměnné

Nyní se soustředíme na případ, kdy $\underline{X} = \overline{X} =: X$. Data tedy mají tvar $(X, \boldsymbol{y})$. Rozhodování o omezenosti množiny B_2 je nyní triviální — stačí zjistit, zdali X má plnou sloupcovou hodnost.

Nadále předpokládejme, že matice ,X má plnou sloupcovou hodnost.
O množině B_2 se v tomto případě dá říci leccos zajímavého. Množinu B_2 nyní můžeme popsat ve tvaru

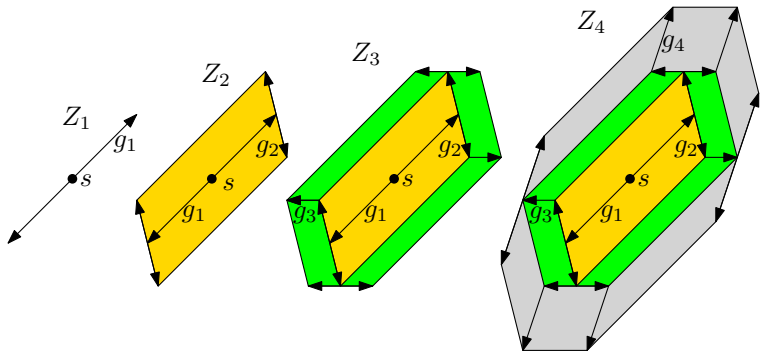
$$B_2(X, \boldsymbol{y}) = \{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y : y \in \boldsymbol{y}\}.$$

Odtud obdržíme zajímavou geometrickou charakteristiku: množina B_2 vznikne zobrazením n-dimenzionálního boxu $\boldsymbol{y}$ do prostoru parametrů při lineárním zobrazení

$$v \mapsto Gv, \quad \text{kde} \quad G := (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}. \tag{9}$$

Můžeme neformálně říci, že množina B_2 je obraz „krychle“ vysoké dimenze (totiž dimenze n = počet pozorování) v prostoru nízké dimenze (totiž v prostoru parametrů) při lineárním zobrazení (9). To speciálně znamená:

Pozorování 4.1. *Množina $B_2(X, \boldsymbol{y})$ je omezený konvexní polyedr v prostoru parametrů.* □



Obrázek 4: Příklady zonotopů $Z_1 = \{s\} +_M g_1$; $Z_2 = \{s\} +_M g_1 +_M g_2$; $Z_3 = \{s\} +_M g_1 +_M g_2 +_M g_3$; $Z_4 = \{s\} +_M g_1 +_M g_2 +_M g_3 +_M g_4$.

Přesnější geometrická charakterizace množiny $B_2(X, \boldsymbol{y})$ [3, 4]. *Zonotopy* jsou speciální konvexní polyedry, jejichž konstrukci stručně popíšeme.

Minkowského součet množiny A s vektorem g je definován jako množina

$$A +_M g := \{a + \lambda g : \ a \in A, \lambda \in [-1, 1]\}.$$

Nech» je dán

- vektor $s \in \mathbb{R}^p$, zvaný *posunutí*;
- systém vektorů $g_1, \ldots, g_n$, zvaných *generátory*.

Zonotop určený posunutím s a generátory $g_1, \ldots, g_n$ je pak množina

$$\{s\} +_M g_1 +_M \cdots +_M g_n;$$

viz též obrázek 4.

Zonotopy mají řadu pozoruhodných strukturálních vlastností; z nejvýznamnějších uveďme jen to, že jsou středově symetrické.

Věta 4.4. *Množina $B_2(X, \boldsymbol{y})$ je zonotop v prostoru parametrů určený posunutím $s = Gy^c$ a generátory*

$$g_i = y_i^{\Delta} \cdot G_i, \quad i = 1, \ldots, n,$$

kde G_i je i-tý sloupec matice G z (9). □

Těsná intervalová obálka. Spočítat těsnou intervalovou obálku zonotopu je snadné: stačí vyhodnotit výraz $G\boldsymbol{y}$ pomocí tzv. *intervalové aritmetiky* [13, 14]. Je definována přirozeně: pro dva intervaly $\boldsymbol{u} = [\underline{u}, \overline{u}]$ a $\boldsymbol{v} = [\underline{v}, \overline{v}]$ jest

$$\boldsymbol{u} + \boldsymbol{v} = [\underline{u} + \underline{v}, \overline{u} + \overline{v}], \tag{10}$$

$$\boldsymbol{u} \cdot \boldsymbol{v} = [\min\{\underline{u} \cdot \underline{v}, \underline{u} \cdot \overline{v}, \overline{u} \cdot \underline{v}, \overline{u} \cdot \overline{v}\}, \max\{\underline{u} \cdot \underline{v}, \underline{u} \cdot \overline{v}, \overline{u} \cdot \underline{v}, \overline{u} \cdot \overline{v}\}]. \tag{11}$$

To speciálně znamená, že těsnou obálku množiny $B_2(X, \boldsymbol{y})$ lze zkonstruovat v polynomiálním čase.

K zonotopům lze konstruovat také (více či méně) těsné elipsoidové obálky, viz [1].

5 Obálka pro množinu hodnot odhadů regresních parametrů

Víme již, že rozhodnout, zda B_2 je omezená množina, je *co-**NP**-úplný* problém. Tím spíše platí, že najít nejtěsnější možnou intervalovou obálku B_2 je také výpočetně těžký problém. Podobný negativní výsledek máme pro každou L_k-normu.

Věta 5.1 ([9])**.** *Pro libovolnou L_k-normu jsou následující problémy **NP**-těžké:*

- *nalézt intervalovou obálku množiny B_k s danou relativní či absolutní chybou,*
- *rozhodnout zda B_k je neomezená.* □

Přestože věta říká, že (pokud $\boldsymbol{P} \neq \boldsymbol{NP}$) neexistuje algoritmus, který by v polynomiálním čase našel *vždy* těsnou obálku B_k, neznamená to ještě, že neexistují metody, které by pro většinu vstupních dat nespočítaly rozumně těsnou obálku.

5.1 Obálka pro B_2

Připomeňme, že $B_2(\boldsymbol{X}, \boldsymbol{y})$ je množinou všech řešení soustavy normálních rovnic

$$X^{\mathrm{T}} X b = X^{\mathrm{T}} y,$$

když X probíhá $\boldsymbol{X}$ a y probíhá $\boldsymbol{y}$. Pokud spočítáme $\boldsymbol{M} := \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, $\boldsymbol{m} := \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ pomocí intervalové aritmetiky (10), (11), pak každé $b \in B_2(\boldsymbol{X}, \boldsymbol{y})$ je zároveň řešením soustavy

$$Mx = m \quad \text{pro jisté } M \in \boldsymbol{M},\ m \in \boldsymbol{m}.$$

Naopak implikace neplatí, protože výpočtem intervalovou aritmetikou jsme ztratili informaci o závislosti matice X (jako kdyby tři instance matice X v soustavě byly najednou nezávislé). Zde se jedná o standardní soustavu intervalových lineárních rovnic. Víme sice již, že najít těsnou obálku je výpočetně těžké i pro tento případ, nicméně existuje celá řada metod, které v krátkém čase spočítají většinou dostatečně těsnou obálku, viz [3].

Jinou možností (viz [3]) je přepsat normální rovnice do tvaru

$$\begin{pmatrix} 0 & X^T \\ X & I_n \end{pmatrix} \begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix}, \tag{12}$$

kde $X \in \boldsymbol{X}$ a $y \in \boldsymbol{y}$. Je-li (b, c) řešení této soustavy, pak b řeší původní soustavu normálních rovnic; a naopak, každé řešení soustavy normálních rovnic jde rozšířit o složku c, abychom dostali řešení soustavy (12). Soustava (12) opět představuje intervalové lineární rovnice a obálku množiny řešení můžeme najít nějakou standardní metodou. Tato soustava je větší, má rozměr $(n + p) \times (n + p)$ oproti původnímu $p \times p$. Dá se dokázat, že její množina řešení je podmnožinou množiny řešení původní intervalové soustavy, a tudíž vypočítaná obálka bude vždy těsnější (nebo alespoň stejně těsná).

5.2 Obálka pro B_1 a B_∞

Při odhadu pomocí L_∞-normy je třeba řešit optimalizační problém $\min_{b' \in \mathbb{R}^p} \|Xb' - y\|_\infty$. To vede na úlohu lineárního programování

$$\min t \quad \text{za omezení} \quad Xb' - y \leq te, \ -Xb' + y \leq te, \ t \geq 0. \tag{13}$$

A podobně, výpočet hodnoty L_1-estimátoru $\min_{b' \in \mathbb{R}^p} \|Xb' - y\|_1$ vede na lineární program

$$\min e^{\mathrm{T}} w \quad \text{za omezení} \quad Xb' - y \leq w, \ -Xb' + y \leq w, \ w \geq 0. \tag{14}$$

Obě úlohy můžeme zapsat jednotně jako

$$\min c^{\mathrm{T}} u \quad \text{za omezení} \quad Au \leq d \tag{15}$$

pro jistou matici $A \in \mathbb{R}^{r \times s}$ a vektory $d \in \mathbb{R}^r, c \in \mathbb{R}^s$. Pokud nás zajímá množina B_1 (resp. B_∞) všech odhadů v L_1 (resp. L_∞) normě, pak nás to vede přímo na třídu lineárních programů

$$(15), \quad A \in \boldsymbol{A}, \ d \in \boldsymbol{d}, \tag{16}$$

kde $\boldsymbol{A} \in \mathbb{IR}^{r \times s}$, $\boldsymbol{d} \in \mathbb{IR}^r$ jsou intervalová matice a intervalový vektor sestavené z původních dat. Třídě úloh (16) se říká *intervalové lineární programování* a jejímu studiu se věnují např. publikace [6, 7].

5.2.1 Bazická stabilita Připomeňme, že *báze* lineárního programu (15) je indexová množina $B \subseteq \{1, \ldots, r\}$ velikosti s taková, že řádky matice A indexované množinou B jsou lineárně nezávislé. Pro jednoduchost budeme symbolem A_B značit podmatici A sestavenou z řádků indexovaných B. Báze B je *přípustná*, pokud lineární soustava

$$A_B u = d_B, \ A_N u \leq d_N$$

má řešení, kde $N := \{1, \ldots, n\} \setminus B$ značí nebazické indexy. V tomto případě je řešení soustavy $u = A_B^{-1} d_B$ jednoznačné a odpovídá vrcholu odpovídající konvexní polyedrické množiny.

Říkáme, že intervalový lineární program (16) je *bazicky stabilní*, pokud existuje báze B, která je optimální bází úlohy (15) pro každé $A \in \boldsymbol{A}$, $d \in \boldsymbol{d}$.

Nejprve pesimistická zpráva.

Věta 5.2 ([9]). *Rozhodnout, zda (16) je bazicky stabilní pro danou bázi B, je co-**NP**-těžký problém.* □

Platí dokonce silnější tvrzení, že *co-**NP***-těžké je rozhodovat o bazické stabilitě nejen pro obecný intervalový lineární program, nýbrž i pro speciální lineární programy (13) a (14). Na druhou stranu ovšem existují silné postačující podmínky zaručující bazickou stabilitu, viz [8].

Nyní optimističtější zpráva. Pokud B je stabilní báze intervalového lineárního programu (16), pak množina všech optimálních řešení je rovna množině řešení intervalové soustavy rovnic

$$A_B u = d_B, \quad A_B \in \boldsymbol{A}_B,\ d_B \in \boldsymbol{d}_B.$$

Tudíž můžeme použít řadu metod na nalezení obálky množiny řešení. Pokud víme navíc, že proměnné u jsou nezáporné, pak dle důsledku 4.1 je množina řešení přímo rovna konvexní polyedrické množině popsané nerovnostmi

$$\underline{A}_B u \le \overline{d}_B,\ \overline{A}_B u \ge \underline{d}_B,\ u \ge 0.$$

Speciální případ: matice X je neintervalová [9]. Je-li matice X neintervalová, je i matice A neintervalová. Bazická stabilita pak pro bázi B nastává právě tehdy, když platí dvě podmínky

- B je optimální báze pro nějaké $A \in \boldsymbol{A}$, $d \in \boldsymbol{d}$,
- $\overline{(A_N A_B^{-1})\boldsymbol{d}_B} \le \underline{d}_N$.

V tomto případě je bazická stabilita rozhodnutelná v polynomiálním čase (a také velmi rychle v praktickém slova smyslu). První podmínku ověříme jednoduše volbou libovolných hodnot z intervalů (např. středů). V druhé podmínce vyhodnotíme $(A_N A_B^{-1})\boldsymbol{d}_B$ pomocí intervalové aritmetiky (10), (11) a otestujeme, zda horní kraj je menší nebo roven dolnímu kraji intervalového vektoru $\boldsymbol{d}_N$.

6 Residuální hodnoty lineárních regresních modelů s intervalovými daty

Vyjděme opět z lineárního regresního model $y = X\beta + \varepsilon$ a předpokládejme, že pozorovaná data (X, y) probíhají dané intervaly $(\boldsymbol{X}, \boldsymbol{y})$. Budeme se zabývat otázkou, jakých hodnot může nabývat residuální součet čtverců. Analogickou otázku si položme i v případě, užijeme-li namísto L_2-estimátoru $\operatorname{argmin}_b \|y - Xb\|_2$ estimátor $\operatorname{argmin}_b \|y - Xb\|_k$ s jinou normou $\|\cdot\|_k$. Omezíme se zde na případ s $k = 1$ a $k = \infty$.

Přesněji: pro danou L_k-normu $\|\cdot\|_k$ definujme

$$\underline{R}_k := \min_{X \in \boldsymbol{X}, y \in \boldsymbol{y}} \min_{b \in \mathbb{R}^p} \|y - Xb\|_k,$$

$$\overline{R}_k := \max_{X \in \boldsymbol{X}, y \in \boldsymbol{y}} \min_{b \in \mathbb{R}^p} \|y - Xb\|_k.$$

Připomeňme, že (neformálně) můžeme přijmout tuto definici: problém A je ***NP**-těžký*, jestliže platí: je-li problém A řešitelný v polynomiálním čase, pak $\boldsymbol{P} = \boldsymbol{NP}$.

6.1 Obecný případ

Ukazuje se, že obecně je výpočet hodnot $\underline{R}_k$ i $\overline{R}_k$ algoritmicky obtížný, a to v případě všech norem L_1, L_2, L_k.

Věta 6.1 ([9]). *(a) Spočítat hodnotu $\overline{R}_k$ je **NP**-těžký problém pro libovolné $k \in \{1, 2, \infty\}$.*
*(b) Spočítat hodnotu $\underline{R}_k$ je **NP**-těžký problém pro libovolné $k \in \{1, 2, \infty\}$.* □

Podobně jako v částech 4.5 – 4.7, i zde negativní výsledek věty 6.1 podává motivaci ke zkoumání speciálních případů. Podává také motivaci ke studiu těsných horních a dolních mezí pro $\underline{R}_k$ a $\overline{R}_k$, viz [9].

6.2 Speciální případy

Speciální případ I: prostor parametrů je omezen na nezáporný orthant [9]. Předpokládejme, že a priori víme, že regresní parametry jsou nezáporné (například proto, že záporné hodnoty nemají fyzikální či ekonomický smysl). Pak získá definice statistik $\underline{R}_k$ a $\overline{R}_k$ tvar

$$\underline{R}_k^+ := \min_{X \in \boldsymbol{X}, y \in \boldsymbol{y}} \min_{b \geq 0} \|y - Xb\|_k,$$
$$\overline{R}_k^+ := \max_{X \in \boldsymbol{X}, y \in \boldsymbol{y}} \min_{b \geq 0} \|y - Xb\|_k.$$

Ukazuje se, že předpoklad nezápornosti situaci částečně zlepší.

Věta 6.2. *(a) Spočítat hodnotu $\overline{R}_k^+$ je **NP**-těžký problém pro libovolné $k \in \{1, 2, \infty\}$.*
(b) Spočítat hodnotu $\underline{R}_k^+$ lze v polynomiálním čase pro každé $k \in \{1, 2, \infty\}$. □

Speciální případ II: počet regresních parametrů je pevný [9]. Nyní prozkoumáme třídu modelů, kde prostorem parametrů je $\mathbb{R}^p$, kde $p \leq p_0$ a p_0 je pevně zvolená konstanta. Ukazuje se, že v tomto případě je situace optimističtější než v obecném případě popsaném větou 6.1.

Věta 6.3. *Nechť p_0 je pevné. Omezíme-li se na třídu modelů s nanejvýš p_0 regresními parametry, pak platí následující tvrzení.*

*(a) Spočítat hodnotu $\overline{R}_2$ je **NP**-těžký problém.*
(b) Spočítat hodnoty $\overline{R}_1$ a $\overline{R}_\infty$ lze v polynomiálním čase.

(c) Spočítat hodnoty $\underline{R}_k$ lze v polynomiálním čase pro každé $k \in \{1, 2, \infty\}$. □

Tvrzení věty 6.3(b) je ovšem třeba číst opatrně: prokazují jej totiž algoritmy, jejichž výpočetní čas může dosahovat

$$(4n)^{p_0+1} \times \text{(polynomiální čas na lineární programování)}.$$

To jen ukazuje, že výraz „polynomiální čas“ nemusí nutně znamenat „opravdu rychlý algoritmus“, zejména je-li p_0 velké.

Analogicky, tvrzení (c) prokazují algoritmy, jejichž výpočetní čas činí

$$2^{p_0} \times \text{(polynomiální čas na lineární programování)}$$

v případě L_1 a L_∞-normy a

$$2^{p_0} \times \text{(polynomiální čas na konvexní kvadratické programování)}$$

v případě L_2-normy.

Speciální případ III: matice X je neintervalová [9]. Nyní vyšetříme případ $\underline{X} = \overline{X}$; intervaly se tedy vyskytují jen ve vektoru $\boldsymbol{y}$. Připomeňme, že v tomto případě z části 4.7 víme, že množina B_2 je zonotop v prostoru parametrů. Počet regresních parametrů je obecný (tj. nepředpokládáme omezení na $\leq p_0$ regresních parametrů).

Věta 6.4. *Nechť $\underline{X} = \overline{X}$. Pak platí následující tvrzení.*

*(a) Spočítat hodnotu $\overline{R}_2$ je **NP**-těžký problém.*
(b) Spočítat hodnoty $\overline{R}_1$ a $\overline{R}_\infty$ lze v polynomiálním čase.
(c) Spočítat hodnoty $\underline{R}_k$ lze v polynomiálním čase pro každé $k \in \{1, 2, \infty\}$. □

Poznamenejme, že pro libovolné $k \in \mathbb{N} \cup \{\infty\}$ se hodnota $\underline{R}_k$ dá spočítat pomocí optimalizační úlohy

$$\underline{R}_k = \min\{\|w\|_k : Xb - \overline{y} \leq w,\ -Xb + \underline{y} \leq w,\ w \geq 0\}.$$

Pro $k \in \{1, \infty\}$ se úloha jednoduše zredukuje na úlohu lineárního programování a pro $k = 2$ zase na konvexní kvadratický program. Ve všech třech případech tedy umíme $\underline{R}_k$ vypočítat nejen polynomiálně (tedy: teoreticky „rychle“), ale i „rychle“ z praktického hlediska.

Speciální případ IV: matice X je neintervalová a počet regresních parametrů je pevný [9]. Je zřejmé, že pozitivní výsledky věty (6.4) zůstávají v platnosti: hodnoty $\overline{R}_1, \overline{R}_\infty, \underline{R}_1, \underline{R}_2, \underline{R}_\infty$ lze vyčíslit v polynomiálním čase. Nicméně ukazuje se, že ani předpoklad na omezení počtu regresních parametrů neumožní spočítat hodnotu $\overline{R}_2$ efektivně.

Věta 6.5. *Nechť $\underline{X} = \overline{X}$ a nechť je p_0 pevné. Spočítat hodnotu $\overline{R}_2$ v modelu, který má nanejvýš p_0 regresních parametrů, je **NP**-těžký problém.* □

Shrnutí. Výsledky vět 6.1, 6.2, 6.3, 6.4 a 6.5 přehledně shrnuje tabulka 1.

věta	6.1	6.2	6.3	6.4	6.5
p	obecné	obecné	pevné	obecné	pevné
X	intervalové	intervalové	intervalové	reálné	reálné
y	intervalové	intervalové	intervalové	intervalové	intervalové
β	obecné	nezáporné	obecné	obecné	obecné
$\overline{R}_1$	***NPH***	***NPH***	***P***	***P***	***P***
$\underline{R}_1$	***NPH***	***P***	***P***	***P***	***P***
$\overline{R}_2$	***NPH***	***NPH***	***NPH***	***NPH***	***NPH***
$\underline{R}_2$	***NPH***	***P***	***P***	***P***	***P***
$\overline{R}_\infty$	***NPH***	***NPH***	***P***	***P***	***P***
$\underline{R}_\infty$	***NPH***	***P***	***P***	***P***	***P***

Tabulka 1: Přehled výsledků částí 6.1 a 6.2 podle [9]. Symbol ***NPH*** značí ***NP***-těžký problém, symbol ***P*** značí problém řešitelný v polynomiálním čase.

7 Složitost výpočtu některých statistik nad intervalovými daty

Na závěr ukažme ještě několik příkladů složitostních výsledků, které mají význam v analýze intervalových dat. Omezíme se tentokrát jen na jednorozměrná intervalová data

$$\boldsymbol{x}_1 = [\underline{x}_1, \overline{x}_1], \ldots, \boldsymbol{x}_n = [\underline{x}_n, \overline{x}_n].$$

Půjde nám zejména o to ukázat, že i při analýze takto jednoduchých dat je řada problémů algoritmicky obtížná — a tím spíše jsou analogické problémy ještě obtížnější v případě analýzy složitějších dat pomocí složitějších modelů.

Nech» S je nějaká statistika. Z algebraického pohledu je S prostě funkcí dat $x_1, \ldots, x_n$; pak píšeme $S(x_1, \ldots, x_n)$. Je zřejmé, že platí:

(a) nahlížíme-li na data $x_1, \ldots, x_n$ jako na pevné konstanty, je $S(x_1, \ldots, x_n)$ pevná hodnota;
(b) nahlížíme-li na data $x_1, \ldots, x_n$ jako na náhodné veličiny, je $S(x_1, \ldots, x_n)$ náhodná veličina;
(c) nahlížíme-li na data $x_1, \ldots, x_n$ jako na intervaly, je $S(x_1, \ldots, x_n)$ interval (za předpokladu spojitosti S).

Zabývejme se nyní bodem (c) a položme si otázku, jak je z algoritmického pohledu snadné či obtížné vyčíslit hodnoty

$$\overline{S} = \overline{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sup\{S(x_1, \ldots, x_n) : x_1 \in \boldsymbol{x}_1, \ldots, x_n \in \boldsymbol{x}_n\},$$
$$\underline{S} = \underline{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \inf\{S(x_1, \ldots, x_n) : x_1 \in \boldsymbol{x}_1, \ldots, x_n \in \boldsymbol{x}_n\}.$$

Hodnotám $\overline{S}$ a $\underline{S}$ říkáme *horní hodnota* statistiky S a *dolní hodnota* statistiky S.

7.1 Obecný případ

První výsledek je negativního typu.

Věta 7.1 ([2, 12]). *(a) Hodnotu $\overline{S}$ nelze vyčíslit algoritmicky.*
(b) Hodnotu $\underline{S}$ nelze vyčíslit algoritmicky. □

Negativní výsledek věty 7.1 je důsledkem toho, že jsme statistiku definovali jako obecnou funkci dat. Tvrzení věty (a) je třeba rozumět takto: neexistuje žádný algoritmus, který na vstup obdrží formuli pro statistiku S (řekněme: zapsanou jako elementární funkci) a číslo $\epsilon > 0$ a spočte racionální číslo $\overline{S}^*$, které se od hodnoty $\overline{S}$ liší nanejvýš od ϵ. (Tato opatrná formulace je nezbytná z toho důvodu, aby nevznikl problém s vyčíslováním odmocnin a dalších iracionálních čísel.) A analogicky je třeba rozumět tvrzení (b).

7.2 Konkrétní statistiky

Negativní výsledek věty 7.1 ovšem neznamená, že situace je nutně špatná v případě *konkrétních* statistik. Vezmeme-li za příklad výběrový průměr

$$\widehat{\mu} = \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i,$$

je zřejmé, že jeho horní a dolní hodnotu lze vyčíslit v polynomiálním čase snadno:

$$\overline{\widehat{\mu}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \overline{x}_i, \qquad \underline{\widehat{\mu}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \underline{x}_i.$$

Následující věty ukazují, že už pro základní, často užívané statistiky situace takto snadná není.

Rozptyl. Uvažme nyní výběrový rozptyl

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (x_i - \widehat{\mu})^2 = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} \left(x_i - \frac{1}{n} \sum\nolimits_{j=1}^{n} x_j \right)^2 .$$

Věta 7.2 ([5]). *(a) Hodnotu $\underline{\widehat{\sigma^2}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ lze vyčíslit v polynomiálním čase.*
*(b) Výpočet hodnoty $\overline{\widehat{\sigma^2}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ je **NP**-těžký problém.* □

Tvrzení věty 7.2(a) plyne z toho, že výpočet dolní hodnoty rozptylu lze vyjádřit jako konvexní kvadratický program

$$\min \left\{ \frac{1}{n-1} \sum\nolimits_{i=1}^{n} \left(x_i - \frac{1}{n} \sum\nolimits_{j=1}^{n} x_j \right)^2 : x_1 \in \boldsymbol{x}_1, \ldots, x_n \in \boldsymbol{x}_n \right\}.$$

Alespoň neformálně nahlédněme tvrzení (b), by≫ tento náhled *není* důkazem ***NP***-těžkosti. Nekonvexní optimalizační problém

$$\max \left\{ \frac{1}{n-1} \sum\nolimits_{i=1}^{n} \left(x_i - \frac{1}{n} \sum\nolimits_{j=1}^{n} x_j \right)^2 : x_1 \in \boldsymbol{x}_1, \ldots, x_n \in \boldsymbol{x}_n \right\}$$

má tu vlastnost, že optima se nabývá v některém vrcholu krychle $\boldsymbol{x}_1 \times \cdots \times \boldsymbol{x}_n$; problém je algoritmicky obtížný z toho důvodu, že je třeba obecně prozkoumat všech 2^n vrcholů. (Tvrzení věty 7.2(b) ukazuje, že nelze očekávat existenci nějaké výrazně rychlejší algoritmické metody, než je právě popsaná metoda hrubé síly.) Viz také [20].

Všimněme si, že věta 6.5 je důsledkem věty 7.2(b).

Výpočet rozptylu, je-li střední hodnota známá. Je pozoruhodné si všimnout, že situace s výpočtem horní hodnoty výběrového rozptylu je odlišná v případě, kdy je střední hodnota μ známá. Pak pracujeme se statistikou

$$\widehat{\sigma^2_{\mu \text{ známé}}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{17}$$

Pozorování 7.1. *Hodnotu* $\overline{\widehat{\sigma^2_{\mu \text{ známé}}}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *lze vyčíslit v polynomiálním čase.* □

Dokonce je metoda jejího výpočtu velice jednoduchá: stačí totiž výraz (17) vyhodnotit pomocí intervalové aritmetiky (10), (11).

***t*-statistika.** Uvažme t-statistiku ve tvaru

$$t = \sqrt{n} \cdot \frac{|\widehat{\mu}|}{\sqrt{\widehat{\sigma^2}}}.$$

Pro ni platí následující věta.

Věta 7.3 ([9])**.** *(a) Výpočet hodnoty* $\underline{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *je* ***NP****-těžký problém.*
(b) Hodnotu $\overline{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *lze vyčíslit v polynomiálním čase.* □

***F*-statistika.** Uvažme F-statistiku ve tvaru

$$F = \frac{\sum_{i=1}^{n/2}\left(x_i - \frac{1}{n/2}\sum_{j=1}^{n/2} x_j\right)^2}{\sum_{i=1+(n/2)}^{n}\left(x_i - \frac{1}{n/2}\sum_{j=1+(n/2)}^{n} x_j\right)^2}.$$

Platí následující věta.

Věta 7.4 ([9])**.** *(a) Výpočet hodnoty* $\underline{F}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *je* ***NP****-těžký problém.*
(b) Výpočet hodnoty $\overline{F}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *je* ***NP****-těžký problém.* □

Není překvapující, že tvrzení vět 7.3(a), 7.4(a) a 7.4(b) lze prokázat jako důsledky věty 7.2(b). Zatímco v případě vět 7.4(a) a 7.4(b) je důkaz snadný, důkaz věty 7.3(a) je obtížný.

Literatura

[1] M. Černý: Goffin's algorithm for zonotopes. *Kybernetika* 48 **(5)**, 2012, 890–906.

[2] M. Černý: *Výpočty.* Svazky I – III. Professional Publishing, Praha, 2011, 2012.
[3] M. Černý, J. Antoch and M. Hladík: On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data. V recenzním řízení. Preprint: Technická zpráva Katedry ekonometrie V©E Praha, 2011. `http://nb.vse.cz/~cernym/plr.pdf`.
[4] M. Černý, M. Rada: On the possibilistic approach to linear regression with rounded or interval-censored data. *Measurement Science Review* 11 **(2)**, 2011, 34–40.
[5] S. Ferson, L. Ginzburg, V. Krejnovič, L. Longpré, M. Aviles: Exact bounds on finite populations of interval data. *Reliable Computing* 11 **(3)**, 2005, 207–233.
[6] M. Fiedler, J. Nedoma, J. Ramík, J. Rohn, K. Zimmermann: *Linear optimization problems with inexact data.* Springer, 2006.
[7] M. Hladík: *Interval linear programming: A survey.* In: Z. Á. Mann (ed.): *Linear Programming – New Frontiers in Theory and Applications.* Nova Science Publishers, New York, 2012, 85–120.
[8] M. Hladík: How to determine basis stability in interval linear programming. *Optimization Letters*, přijato k publikaci.
[9] M. Hladík, M. Černý: Interval data and complexity of statistics based on minimum norm estimators. V recenzním řízení. Preprint: Technická zpráva Katedry ekonometrie V©E Praha, 2013. `http://nb.vse.cz/~cernym/minnorm.pdf`.
[10] V. Krejnovič, A. Lakajev, J. Rohn, P. Kahl: *Computational complexity and feasibility of data processing and interval computations.* Kluwer, 1998.
[11] V. Krejnovič, G. Xiang: Fast algorithms for computing statistics under interval uncertainty: An overview. In: V.-N. Huynh et al. (eds.): *Interval/probabilistic uncertainty and non-classical logics.* Springer, Berlin, 2008, 19–31.
[12] Ju. Matijasevič: *Hilbert's Tenth Problem.* MIT Press, 1993.
[13] R. E. Moore, R. B. Kearfott, M. J. Cloud: *Introduction to interval analysis.* SIAM, Philadelphia, 2009.
[14] A. Neumaier: *Interval methods for systems of equations.* Cambridge University Press, 1990.
[15] J. Renegar: On the computational complexity and geometry of the first-order theory of the reals. Part I: Introduction. Preliminaries. The geometry of semi-algebraic sets. The decision problem for the existential theory of the reals. *Journal of Symbolic Computation* 13 **(3)**, 1992, 255–299.
[16] J. Renegar: On the computational complexity and geometry of the first-order theory of the reals. Part II: The general decision problem. Preliminaries for quantifier elimination. *Journal of Symbolic Computation* 13 **(3)**, 1992, 301–327.
[17] J. Renegar: On the computational complexity and geometry of the first-order theory of the reals. Part III: Quantifier elimination. *Journal of Symbolic Computation* 13 **(3)**, 1992, 329–352.
[18] A. Schrijver: *Theory of linear and integer programming.* Wiley, 1988.
[19] V. ©vejdar: *Logika: neúplnost, složitost a nutnost.* Academia, Praha, 2002.
[20] S. A. Vavasis: *Nonlinear Optimization: Complexity Issues.* Oxford University Press, 1991.

RESUSCITACE MOMENTOVÉ METODY

Zdeněk Fabián, Ústav informatiky AV ČR, Praha

Klíčová slova: skórová funkce, charakteristiky rozdělení, charakteristiky dat

Abstrakt: V článku definuji skórovou funkci rozdělení (s libovolným intervalovým nosičem), vysvětluji její vztah k verzím publikovaným v předcházejících sbornících ROBUST a doporučuji její použití v (obecné) metodě momentů.

Abstrakt: A score function of distribution (with arbitrary interval support) is introduced, its relation to versions presented on the past ROBUST schools are explained and its use in (a general) moment method is recommended.

1 Úvod

Buď $\mathcal{X} \subseteq \mathbb{R}$ otevřený interval a $\mathbf{x} = (x_1, ..., x_n)$ náhodný výběr z rozdělení F. Předpokládáme že F je členem parametrické rodiny $\{F_\theta, \theta \in \Theta\}$ kde $\Theta \subseteq \mathbb{R}_m$, s nosičem $\mathcal{X}$ a hustotami $f(x;\theta)$. Úlohu nalézt na základě $\mathbf{x}$ takové $\hat{\theta} \in \Theta$, aby $F_{\hat{\theta}}$ co nejlépe aproximovala F řeší (m.j.) tradiční metody: momentová metoda a metoda maximální věrohodnosti.

a) Buď S "šikovná" funkce. Pro $k \in \mathcal{N}$, k-tý moment náhodné veličiny $S(X)$ je hodnota

$$ES^k(X) = \int_{\mathcal{X}} S^k(x) f(x)\, dx. \tag{1}$$

Odhad θ momentovou metodou spočívá v řešení rovnic

$$\frac{1}{n}\sum_{i=1}^{n} S^k(x_i;\theta) = ES^k(\theta), \qquad k = 1, ..., m, \tag{2}$$

představujících konečnou aproximaci parametrického tvaru vzorce (1) na základě pozorovaného $\mathbf{x}$. Ve vzorcích se odedávna používá funkce $S(x;\theta) = x$, což má za následek, že pro řadu rozdělení integrály (1) nekonvergují a metodu nelze použít. Což je škoda, protože kromě odhadu parametrů poskytuje názorné charakteristiky dat: výběrové momenty.

Metoda maximální věrohodnosti zavádí obecně vektorovou (Fisherovu) skórovou funkci

$$U(\theta) = \frac{\partial}{\partial\theta} \log L(\theta) \tag{3}$$

jejíž složky jsou parciální skóry pro jednotlivé parametry. Pro odhady parametrů máme rovnice

$$\frac{1}{n}\sum_{i=1}^{n} U_{\theta_k}(x_i;\theta) = 0 \qquad k = 1, ..., m$$

poskytující odhady v dobře známém smyslu optimální.

Překvapivě, pro rozdělení s nosičem $\mathbb{R}$ a hustotou tvaru $f(x-\mu)$, kde μ je parametrem polohy, platí že

$$U_\mu(x-\mu) = \frac{\partial}{\partial\mu}\log f(x-\mu) = -\frac{f'(x-\mu)}{f(x-\mu)} \equiv S(x-\mu), \tag{4}$$

a to znamená, že (Fisherovu) skórovou funkci pro parametr μ můžeme najít v tomto případě i derivováním podle proměnné.

Položíme-li na pravé straně (4) $\mu = 0$, máme na $\mathbb{R}$ funkci

$$S(x) = -\frac{f'(x)}{f(x)} \tag{5}$$

charakterizující rozdělení F, shodnou se skórovou funkcí pro nejdůležitější parametr, kterou Hampel [13] a Jurečková [15] nazývají skórovou funkcí rozdělení F. Za skórovou funkci rozdělení $F_\theta(x)$ pak můžeme považovat funkci

$$S(x;\theta) = -\frac{1}{f(x;\theta)}\frac{d}{dx}f(x;\theta) \tag{6}$$

a použít ji v (2), momenty (1) pak budou existovat, budou často vyjádřeny jako jednoduché funkce parametrů a $ES^2(\theta)$ bude Fisherova infomace rozdělení $F(x;\theta)$ [7]. Zobecnění (6) zahrnuje i rozdělení bez parametru polohy, jako je F s hustotou

$$f(x;p,q) = \frac{1}{B(p,q)}\frac{e^{px}}{(1+e^x)^{p+q}}$$

a skórovou funkcí rozdělení $S(x;p,q) = \frac{qe^x-p}{e^x+1}$. Pro všechna unimodální rozdělení s nosičem R pak řešení x^* rovnice $S(x;\theta) = 0$ představuje jeho pěknou centrální charakteristiku: mód.

Bohužel, funkce (5) ztrácí smysl pro rozdělení s nosičem $\mathcal{X} \neq R$, stačí uvážit exponenciální či rovnoměrné rozdělení. Zobecnění (5) po rozdělení s obecným intervalovým nosičem uvádím v následující kapitole.

2 Skórová funkce spojitého rozdělení

Zobecnění (5) pro rozdělení s nosičem $\mathcal{X} = (0,\infty)$ je

$$S(x) = -\frac{1}{f(x)}\frac{d}{dx}[xf(x)]. \tag{7}$$

Důvodem je, že pro rozdělení s hustotou $f(x/\tau) = \frac{1}{\tau}h(x/\tau)$ je

$$U_\tau(x;\tau) = \frac{\partial}{\partial\tau}\log[\frac{1}{\tau}h(x/\tau)] = \frac{1}{\tau^2}[1 - \frac{x}{\tau}\frac{h'(x/\tau)}{h(x/\tau)}],$$

$$T(x;\tau) = -\frac{1}{\frac{1}{\tau}h(x/\tau)}\frac{d}{dx}[x\frac{1}{\tau}h(x/\tau)] = \frac{1}{\tau}[1 - \frac{x}{\tau}\frac{h'(x/\tau)}{h(x/\tau)}].$$

Položíme-li

$$S(x;\tau) = \frac{1}{\tau}T(x;\tau), \tag{8}$$

je $S(x) \equiv S(x;1)$ identické s (Fisherovou) skórovou funkcí pro $\tau = 1$.

Veličinu v lomené závorce vzorce (7) jsem si vysvětlil následovně: Představíme-li si, že F je transformované rozdělení $F(x) = G(\log x)$, kde "prototyp" G má nosič $\mathbb{R}$, je to hustota dělená Jacobianem té inverzní transfromace. Pro libovolný interval $\mathcal{X}$ to zobecníme následovně:

Definice 1. Buď F s nosičem $\mathcal{X}$ a spojitě diferencovatelnou hustotou $f(x)$, buď $\eta : \mathcal{X} \to R$ spojité rostoucí zobrazení a $G(y) = F(\eta^{-1}(y))$ unimodální. Položme

$$T(x) = -\frac{1}{f(x)}\frac{d}{dx}\left(\frac{1}{\eta'(x)}f(x)\right) \tag{9}$$

a označme x^* řešení rovnice $T(x) = 0$. Pak

$$S(x) = \eta'(x^*)T(x) \tag{10}$$

je $\eta-$skór rozdělení F.

Definice 2. Skórová funkce rozdělení F je jeho matematicky nejjednodušší $\eta-$skór.

Poslední definice vypadá sice vágně, ale není. Nejjednodušší $\eta-$skór bude výsledkem (virtuální) transfrormace, jejíž Jacobian je součástí vzorce pro transformovanou hustotu

$$f(x) = g(\eta(x))\eta'(x). \tag{11}$$

Pak totiž podle (9)

$$S(x) = -\eta'(x^*)\frac{1}{f(x)}\frac{d}{dx}g(\eta(x)).$$

Rozdělení s nosičem $\mathbb{R}$ mají zpravidla skórovou funkci (6). Ale nemusí. Rozdělení s hustotou

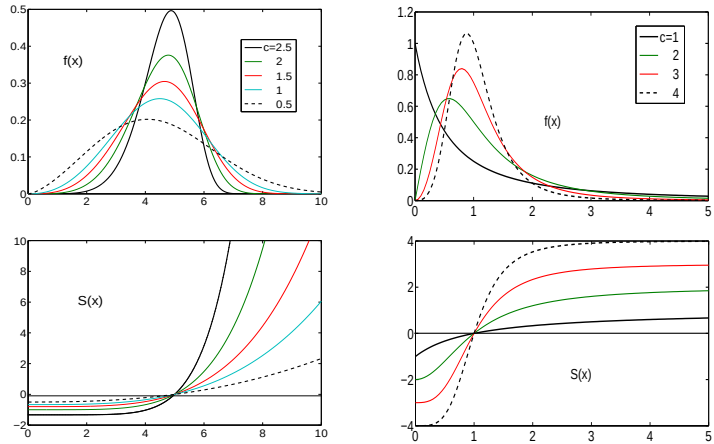
$$f(x) = \frac{1}{\sqrt{(1+x^2)}}\frac{e^{\sinh^{-1}x}}{(1+e^{\sinh^{-1}x})^2} \tag{12}$$

je vhodné považovat za transformované pomocí $\eta(x) = \sinh^{-1}x$, protože $\eta'(x) = \frac{1}{\sqrt{(1+x^2)}}$. Skórová́ funkce rozdělení je pak

$$S(x) = \frac{e^{\sinh^{-1}x} - 1}{e^{\sinh^{-1}x} + 1}$$

(prototyp rozdělení F má zřejmě hustotu $g(y) = e^y/(1+e^y)^2$).

Vzorce pro hustoty rozdělení s nosičem $\mathcal{X} = (0, \infty)$ většinou obsahují člen $1/x$ (který lze případně snadno zavést, viz exponenciální rozdělení s hustotou $f(x) = \frac{1}{x}xe^{-x}$). Na obr.1 jsou znázorněny hustoty a skórové funkce dvou rozdělení s tímto nosičem (příslušné vzorce viz Tabulka I).



obr. 1. Hustoty a skórové funkce rozdělení: vlevo Weibullova ($\tau = 5$), vpravo log-logistického ($\tau = 1$) pro různá c (skórová funkce pro $c = 4$ je Kovanicova funkce h, viz kap. 3).

Pro jiné nosiče už S tak jednoduše určit nelze, tak na $\mathcal{X} = (1, \infty)$ jsou nejméně dvě "šikovné" transformace, $\eta(x) = \log(x-1)$ a $\eta(x) = \log\log x$. Třeba příslušné η-skóry Paretova rozdělení s hustotou $f(x;c) = c/x^{c+1}$ jsou

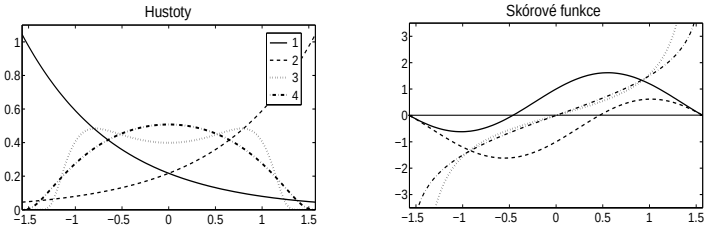
$$T_1(x;c) = -\frac{1}{f(x)}\frac{d}{dx}[(x-1)f(x)] = c - \frac{c+1}{x},$$

tedy $x^* = \frac{c+1}{c}$ a $S_1(x;c) = c(1/x^* - 1/x)$, a

$$T_2(x;c) = -\frac{1}{f(x)}\frac{d}{dx}[x^{-c}\log x] = c\log x - 1,$$

kdy $x^* = e^{1/c}$ a $S_2(x;c) = e^{-1/c}(c\log x - 1)$, což je (Fisherova) skórová funkce pro c . Zde je asi nutno dát přednost omezené inferenční funkci $S_1(x;c)$.

Vzorce pro hustoty rozdělení s nosičem v podobě konečného intervalu mohou obsahovat řadu různých Jacobianů, tak v případě $\mathcal{X} = (0,1)$ kromě $\eta(x) = \log\frac{x}{1-x}$, $\eta'(x) = \frac{1}{x(1-x)}$ kterou obsahuje vzorec pro hustotu beta rozdělení (Tabulka I), i třeba $\eta(x) = -\log(-\log x)$, $\eta'(x) = \frac{1}{x\log x}$. Rozdělení s nosičem $\mathcal{X} = (-\pi/2, \pi/2)$ je často výhodné považovat za virtuálně transformované funkcí $\eta(x) = \tanh^{-1}(x)$, $\eta'(x) = 1/\cos^2 x$ (obr. 2).



obr. 2. Hustoty a skórové funkce rozdělení s nosičem $(-\pi/2, \pi/2)$.
1 $f(x) = e^{-x}/\kappa$, 2 $f(x) = e^{x}/\kappa$, 3 $f(x) = \frac{1}{\sqrt{2\pi}\cos^2 x} e^{-\frac{1}{2}\tan^2 x}$
4 $f(x) = \sqrt{\frac{\pi}{2}} \frac{1}{(x+\pi/2)(\pi/2-x)} e^{-\frac{1}{2}\log^2 \frac{\pi/2-x}{x+\pi/2}}$.

3 Průvodce po džungli mých článků z minulých ROBUSTů

Pro účastníky škol ROBUST a čtenáře minulých sborníků prací ROBUST připojuji pár poznámek.

V soupisu literatury na konci článku uvádím hlavně práce, které byly otištěny ve sbornících ROBUST a Informačním Bulletinu ČSS. Tyto práce ve všech případech předcházely rozšířeným anglickým verzím, které byly, často se značným zpožděním, publikovány v mezinárodních časopisech.

Hledat skórovou funkci rozdělení s nosičem $\mathcal{X} = (0, \infty)$ mě napadlo po "rozšifrování" Kovanicovy t.zv. gnostické teorie [16] zpracování dat "bez pomoci statistiky". Kovanic odhaduje "datový parametr", řekněme δ, jako průměr hodnot jisté inferenční funkce $h(x; \delta)$, odvozené spolu s funkcí, kterou jsem detekoval jako nenormovanou hustotu rozdělení log-logistického typu na $\mathcal{X} = (0, \infty)$. Na jiné rozdělení však jeho postup odvození inferenční funkce nelze aplikovat.

Funkce $T(x)$, (9), libovolného rozdělení s intervalovým nosičem, odvozené od skórové funce rozdělení s nosičem $\mathbb{R}$ ("prototypu") byla zavedena v [1] pod názvem geometrická funkce rozdělení. V [2]-[5] jí říkám core funkce (vzniklo jako "vnitřek score funkce"), v [6] jí říkám t-skór (z transformation-based score). Skórovou funkci rozdělení (10) jsem představil v [4] pod názvem Johnsonova skórová funkce, později [6] jako věrohodnostní skór pro těžiště a na přednáškách jsem jí říkal skalární skór nebo skalární skórová funkce. Všechna tato (nepříliš smyslná) pojmenování byla vedena snahou vše vyjasnit (odlišit novou funkci od skórových funkcí klasické a robustní statistiky) a docílila ovšem pravého opaku. K současnému termínu, který je jistě nejvhodnější, mě přivedla přednáška [15] prof. Jurečkové.

Definice skórové funkce rozdělení jako nejjednoduššího $\eta-$skóru je nová. Už v [2] sice říkám, "že obecně lze za core funkci rozdělení F považovat ten nejjednodušší smysluplný popis rozdělení pomocí funkce obsahující člen $-f'/f$, o který si matematický tvar hustoty řekne", ale nečta pozorně starší ročníky ROBUSTu, byl jsem pak veden mylným úsilím o jedno-jednoznačný

vztah mezi f a S (s cílem uspořádat pravděpodobnostní modely podle chování skórových funkcí rozdělení) tak, že jsem fixoval virtuální transformaci inspirovanou Johnsonem [14] (odtud Johnsonův skór) jako

$$\eta(x) = \begin{cases} x & \text{if} \quad \mathcal{X} = \mathbb{R} \\ \log(x-a) & \text{if} \quad \mathcal{X} = (a, \infty) \\ \log \dfrac{(x-a)}{(b-x)} & \text{if} \quad \mathcal{X} = (a, b). \end{cases} \tag{13}$$

Důležitější ovšem je, jak jsem si konečně uvědomil, aby inferenční funkce byla co nejjednodušší, a její momenty v rovnicích (1) byly vyjádřeny jednoduchými funkcemi parametrů. To je samozřejmě možné jen když si S a F "padnou do noty".

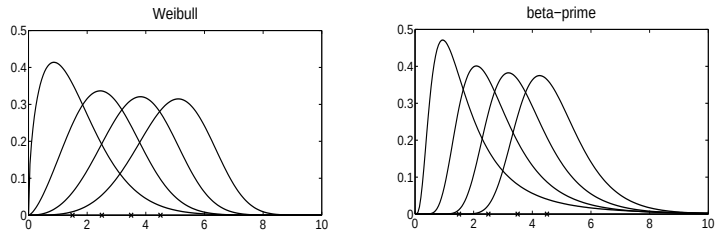
4 Momenty

Momenty (1), kde S je skórová funkce rozdělení F, nazveme skórové momenty. Oproti momentům náhodné veličiny X, skórové momenty existují. Za obvyklého předpokladu $ET^2(X) < \infty$ existuje i $ET^k(X)$, důkaz viz [2], a $\eta'(x^*)$ v (10) je konstanta. Dodejme, že důležitá, představuje zobecnění vztahu (8) pro obecná rozdělení.

$ES = 0$. Těžiště, t.j. řešení x^* rovnice $T(x;\theta) = 0$, považuji za typickou hodnotu rozdělení F (je to "obraz módu jeho prototypu"). ES^2 je Fisherova informace rozdělení F [3], její reciprokou hodnotu

$$\omega^2 = 1/ES^2$$

kterou zde nazveme s-variance (v [5] a [6] "Johnsonova disperze"), je vhodná míra variability rozdělení [5]. To naznačují i hustoty s jednotkovou variabilitou $\omega^2 = 1$ na obr. 3.



Obr. 3. Hustoty Weibullových a beta-prime (neboli beta II) rozdělení (viz Tabulka 1) s $\omega^2 = 1$. Křížky na ose x jsou vyznačena těžiště.

ES^3 je jistou mírou nesymetrie rozdělení (na $\mathcal{X} = (0, \infty)$ mírou odchylky od základního nesymetrického tvaru) a ES^4 charakterizuje "plochost" (opak špičatosti): položíme-li $\tilde{\gamma}_2 = ES^4/(ES^2)^2$, pro rozdělení s hustotou $f(x) \sim e^{-x^4/4}$ je $\tilde{\gamma}_2 = 10.9$, pro normální $\tilde{\gamma}_2 = 3$, pro logistické $\tilde{\gamma}_2 = 1.8$ a pro Laplaceovo $\gamma_2 = 1$.

Pro představu uvádím skórovou funkci rozdělení, těžiště a skórovou variancí některých známých rozdělení v Tabulce I. Písmenem m v tabulce označuji střední hodnotu, která, jak je patrné, pro některá rozdělení (s těžkým koncem) existuje jen v omezeném oboru parametru. To platí i pro log-log. (log-logistické) rozdělení, vzorec se mi do tabulky nevešel.

Tabulka I. Skórová funkce, těžiště a s-variance některých rozdělení

F	$f(x)$	$S(x)$	m	x^*	ω^2
exp.	$\frac{1}{\tau}e^{-x/\tau}$	$\frac{1}{\tau}(\frac{x}{\tau}-1)$	τ	τ	τ^2
lognorm.	$\frac{c}{\sqrt{2\pi}x}e^{-\frac{1}{2}\ln^2(\frac{x}{\tau})^c}$	$\frac{c}{\tau}\ln(\frac{x}{\tau})^c$	$\tau(e^{\frac{1}{c^2}})^{1/2}$	τ	τ^2/c^2
Weibull	$\frac{c}{x}(\frac{x}{\tau})^c e^{-(\frac{x}{\tau})^c}$	$\frac{c}{\tau}((\frac{x}{\tau})^c-1)$	$\tau\Gamma(\frac{1}{c}+1)$	τ	τ^2/c^2
Fréchet	$\frac{c}{x}(\frac{\tau}{x})^c e^{-(\frac{\tau}{x})^c}$	$\frac{c}{\tau}[1-\left(\frac{\tau}{x}\right)^c]$	$\tau\Gamma(1-\frac{1}{c})$	τ	τ^2/c^2
log-log.	$\frac{c}{x}\frac{(x/\tau)^c}{((x/\tau)^c+1)^2}$	$\frac{c}{\tau}\frac{(x/\tau)^c-1}{(x/\tau)^c+1}$		τ	$\frac{3\tau^2}{c^2}$
gamma	$\frac{\gamma^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\gamma x}$	$\frac{\gamma^2}{\alpha}(x-x^*)$	$\frac{\alpha}{\gamma}$	$\frac{\alpha}{\gamma}$	$\frac{\alpha}{\gamma^2}$
beta II	$\frac{1}{B(p,q)}\frac{x^{p-1}}{(x+1)^{p+q}}$	$\frac{q^2}{p}\frac{x-x^*}{x+1}$	$\frac{p}{q-1}$	$\frac{p}{q}$	$\frac{p(p+q+1)}{q^3}$
Pareto	c/x^{c+1}	$\frac{c^2}{c+1}(1-\frac{x^*}{x})$	$\frac{c}{c-1}$	$\frac{c+1}{c}$	$\frac{c+2}{c^3}$
beta	$\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$	$(p+q)x-p$	$\frac{p}{p+q}$	$\frac{p}{p+q}$	$\frac{pq(p+q+1)}{(p+q)^4}$

5 Metoda skórových momentů

5.1. Estimátor daný rovnicemi (2), v nichž S značí skórovou funkci rozdělení F, je specielní forma $M-$estimátoru s inferenční funkcí

$$\mathbf{\Psi}(x;\theta) = [S(x;\theta), S^2(x,\theta) - ES^2(\theta), ..., S^m(x;\theta) - ES^m(\theta)].$$

Pokud jsou S a ES^2 spojitě diferencovatelné podle jednotlivých parametrů, jsou skórové momentové odhady na základě obecných vět robustní statistiky konzistentní a asymptoticky normální s kovarianční maticí odvozenou v [9]. Skórové momenty jsou často jednoduché funkce parametrů a momentové rovnice jsou zejména v případě více parametrů mnohem jednodušší než rovnice (3). Odhady nejsou eficientní, ale zato při omezené S robustní pro všechny parametry. Řadu příkladů obsahují práce [1] (zde jim říkám "geometrické momenty"), [3] ("core momenty"), [6] ("zobecněné momenty") a [12], ve vzorcích zde vystupuje T místo S, což je ale vzhledem k (10) jedno.

5.2. x^* rozdělení s jedním parametrem (někdy i se dvěma) lze často jednoduše vyjádřit jako $x^*(\theta)$. První z rovnic (2) pak má tvar

$$\sum_{i=1}^{n} S(x_i; x^*) = 0$$

a její řešení $\hat{x}^*$, výběrovou typickou hodnotu, lze nazvat skórovým průměrem. Tak pro rozdělení s nosičem $\mathbb{R}$, hustotou $g(x) = \frac{\gamma^\alpha}{\Gamma(\alpha)}e^{\alpha x}e^{-\gamma e^x}$ a skórovou

funkcí rozdělení $S(x) = \gamma e^x - \alpha$ je $x^* = \log(\alpha/\gamma)$ a skórový průměr je

$$\hat{x}^* = \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{x_i}\right).$$

Podle [3], skórový průměr normálního, gamma a beta rozdělení je aritmetický průměr, lognormálního geometrický průměr, Paretova a Fréchetova harmonický průměr a Weibullova (pro pevné c) $\hat{x}^* = (\frac{1}{n}\sum x_i^c)^{1/c}$. Rozdělení všech skórových průměrů je asymptoticky normální s asymptotickým rozptylem [11] $\sigma_{as}^2 = ES^2/[\frac{\partial}{\partial x^*}S(x;x^*)]^2$. Obecnou metodou je ovšem odhadnout θ a položit $\hat{x}^* = x^*(\hat{\theta}_n)$. Podobně lze najít výběrovou s-varianci jako $\hat{\omega}^2 = \omega^2(\hat{\theta}_n)$, jinou možností je spočíst

$$\hat{\omega}^2 = \frac{n}{\sum_{i=1}^{n} S^2(x_i;\hat{\theta}_n)}. \tag{14}$$

5.3. Rozdělení s neomezenou sórovou funkcí rozdělení je možno použít i v případě kontaminovaných dat, ošetříme-li ji metodami, které nabízí robustní statistika. Protože je S jediná funkce i v případě vektorového parametru, je situace principiálně jednoduchá.

Uvažujme Weibullovo rozdělení s hustotou f_W (Tabulka I) s neomezenou skórovou funkci rozdělení pro $x \to \infty$. Zvolme za inferenční funkci "useknutý t-skór"

$$\Psi(x;\tau,c) = \begin{cases} (x/\tau)^c - 1 & \text{když} \quad x \le v \\ q & \text{když} \quad x > v \end{cases} \tag{15}$$

kde $r = (v/\tau)^c - 1$. $\theta = (\tau, c)$ odhadneme z rovnic

$$\sum_{i=1}^{n} \psi(x_i;\theta) = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} \psi^2(x_i;\theta) = E\psi^2(\theta)$$

kde $\psi = \Psi - E\Psi$. Položíme-li

$$v = \tau_0 + r\omega_0 = \tau_0(1 + r/c_0)$$

kde r je "ladící" konstanta a τ_0, ω_0 jsou nějaké počáteční odhady, např. medián a MADN(x) (to ale bohužel není u rozdělení na polopřímce vždycky vhodná volba), lze odvodit rovnice

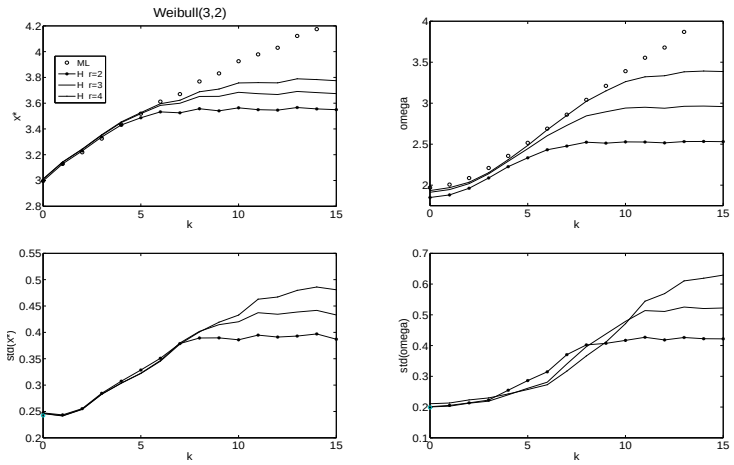
$$\begin{aligned} \tau^c &= \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^c}{1 - e^{-w}} \\ \tau^{2c} &= \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^{2c}}{2[1 - (w+1)e^{-w}]} \end{aligned}$$

kde $w = (1 + r/c_0)^{c_0}$ a $c_0 = \tau_0/\omega_0$.

Při simulacích jsem generoval kontaminované Weibullovo rozdělení jako

$$f_c(x^*,\omega) = (1-\epsilon)f_W(x^*,\omega) + \epsilon f_W(x^* + k, \omega) \tag{16}$$

kde $\epsilon = 0.1$. Průměrné hodnoty a standardní odchylky těchto "huberizovaných" odhadů s rostoucím k a pro několik hodnot "ladící" konstanty r jsou znázorněny na obr. 4. Zatímco maximálně věrohodné ML a jejich standardní odchylky (nevyznačeno) rostou lineárně s rostoucím k, "huberizované" odhady se "zastaví" na určité hladině, která ale neodpovídá hodnotě těžiště a odmocnině z s-variance nekontaminovaného souboru.



obr. 4. Závislost maximálně věrohodných a "huberizovaných" (s různou hodnotou "ladící" konstanty) odhadů x^* a ω Weibullova rozdělení na rostoucí kontaminaci.

6 Charakteristiky dat

Obecně lze samozřejmě vyjádřit výběrovou typickou hodnotu a výběrovou s-varianci pomocí odhadnutého θ_n, t.j. $\hat{x}^* = x^*(\hat{\theta}_n)$ a $\hat{\omega}^2 = \omega^2(\hat{\theta}_n)$ (jinou možností je spočíst $\hat{\omega}^2$ podle (14). Tak je možné převést odhady parametrů různým způsobem parametrizovaných rozdělení na společný jmenovatel.

7 Závěr

Domnívám se, že takto oživená momentová metoda by mohla fungovat v situacích, kdy máme dobrou představu o modelu odlišném od normálního a očekáváme, že data budou značně kontaminovaná.

Reference

[1] Fabián Z. (1997) *Geometrické momenty.* Sb. Robust'96, 49 – 62.

[2] Fabián Z. (2001) *MM-odhady.* Sb. Robust'2000, 33 – 41.

[3] Fabián Z. (2003) *Informace ve výběru z rozdělení.* Sb. Robust'2002, 95 – 106.

[4] Fabián Z. (2005) *Modifikovaná Raova vzdálenost.* Sb. Robust'2004, 459 – 466.

[5] Fabián Z. (2007) *Nové charakteristiky rozdělení a výběrů z rozdělení.* Sb. Robust'2006, 459 – 466.

[6] Fabián Z. (2009) *O rozděleních s těžkými chvosty.* Statistický bulletin, 459 – 466.

[7] Cover, T.M. and Thomas, J.A. (1991). *Elements of Information Theory.* Wiley, New York.

[8] Fabián, Z. (1997). *On the relation between gnostical and probability theories.* Kybernetika 33, 3, 259 – 270.

[9] Fabián, Z. (2001). *Induced cores and their use in robust parametric estimation.* Comm. Statist. Theory Methods 30, 537 – 556.

[10] Fabián, Z. (2007). *Estimation of simple characteristics of samples from skewed and heavy-tailed distribution.* In Skiadas, C. (ed.), Recent Advances in Stochastic Modeling and Data Analysis, Singapore, World Scientific, 43 – 50.

[11] Fabián, Z. (2009). *Confidence intervals for a new characteristic of central tendency of distributions.* Comm. Statist. Theory Methods 38, 1804 – 1814.

[12] Fabián, Z. (2010). *Score moment estimators.* In Proc. of conference COMPSTAT, Physica-Verlag, Springer.

[13] Hampel F. R., Rousseeuw P. J., Ronchetti E. M. and Stahel W. A. (1986) *Robust Statistic. The Approach Based on Influence Functions*, Wiley, New York.

[14] Johnson, N.L. (1949). *Systems of frequency curves generated by methods of translations.* Biometrika 36, 149 – 176.

[15] Jurečková J. (2012). *Score functions of distributions and their role.* Přednáška na workshopu Recent advances in math. statistics, Praha, konaném na počest prof. M. Huškové.

[16] Kovanic, P. (1986). *A new theoretical and algorithmical tool for estimation, identification and control.* Automatica 22, 6, 657 – 674.

Poděkování: Tato práce měla instituciální podporu RVO:67985807.

Adresa: Ústav Informatiky AV ČR, Pod vodárenskou věží 2, 182 07 Praha 8

TESTOVÁNÍ NORMALITY ZE ZAOKROUHLENÝCH DAT

Michal Friesl

Klíčová slova: Test dobré shody, normální rozdělení, seskupená data.

Abstrakt: Stojíme před úkolem provést test dobré shody s normálním rozdělením na základě pozorování seskupených do intervalů — pozorována jsou např. zaokrouhlená data, v našem případě celočíselná. Rozdělení může mít malý rozptyl (směrodatnou odchylku srovnatelnou s šíří intervalů), rozsah výběru může být malý a pozorování cenzorovaná. Pomocí simulací zvážíme použití variant Kolmogorovova-Smirnovova testu a chí-kvadrát testu dobré shody a navrhneme randomizovanou variantu.

Abstract: We face a task to test the goodness of fit with the normal distribution based on observations grouped into intervals — we observe rounded data, in our case integer. Distribution may have a small dispersion (standard deviation comparable to the width of intervals), the sample size may be small and data censored. Using simulations, we consider appropriateness of variants of Kolmogorov-Smirnov test and the chi-square test of goodness of fit and propose a randomized variant.

1. Úvod

Uvažujme náhodný výběr $X_1, \dots, X_n$ ze spojitého rozdělení s distribuční funkcí F, pozorujeme však jen data seskupená do intervalů. Konkrétně pozorujeme zaokrouhlené hodnoty $X_1^d, \dots, X_n^d$, pracujeme tedy s náhodným výběrem z diskretizovaného rozdělení F^d. Zaokrouhlením míníme v našem případě zaokrouhlení na celá čísla, výsledky pro jiný než jednotkový krok diskretizace by se dostaly přenásobením příslušnou konstantou.

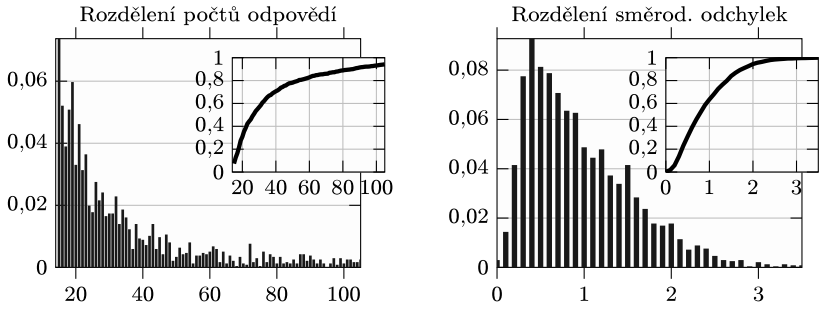
Budeme se zabývat testováním dobré shody s normálním rozdělením ze zaokrouhlených a případně i cenzorovaných dat. Chceme testovat hypotézu

$$H_0\colon F = \mathrm{N}(\mu_0, \sigma_0^2), \qquad \text{resp.} \qquad H_0'\colon F \in \{\mathrm{N}(\mu, \sigma^2),\ \mu \in \mathbf{R}, \sigma > 0\},$$

a to s ohledem na situace, kdy šířka intervalů není zanedbatelná vzhledem k rozptylu a zároveň kdy rozsah výběru n může být malý. Uveďme dva konkrétní příklady takových dat z prostředí Západočeské univerzity.

Příklad 1. Data ze studentského hodnocení kvality výuky. Studenti u každého předmětu posuzují tři až pět otázek, a to tak, že vyjadřují míru souhlasu s předloženými tvrzením, kterými jsou např. *přednáška byla srozumitelná* či *cvičení bylo vedeno dobře*. Zaznamenané hodnoty jsou z množiny $\{0, \dots, 5\}$, kde 0 značí naprostý nesouhlas a 5 naprostý souhlas. Úkolem je u každého předmětu zjistit, zda rozložení číslených odpovědí u jednotlivých tvrzení je normální. Pozorovanou hodnotu chápeme jako zaokrouhlenou hodnotu

skutečné míry souhlasu studenta (zaokrouhluje se na celá čísla). V případech, kdy student chce souhlasit více než naprosto (nebo silněji než naprosto nesouhlasí) dostáváme jen hodnotu cenzorovanou — pětkou zprava (nebo 0 zleva). Vezměme data z ankety za zimní semestr 2011/2012 dostupná z [3] a omezme se jen na 2361 tvrzení, kde odpovědělo aspoň 15 studentů, viz obr. 1. Počet respondentů byl v průměru 41, u poloviny tvrzení byl menší než 27,



Obrázek 1. Rozdělení počtů respondentů a směrodatných odchylek hodnocení u otázek jednotlivých předmětů. V malém obrázku kumulované relativní četnosti.

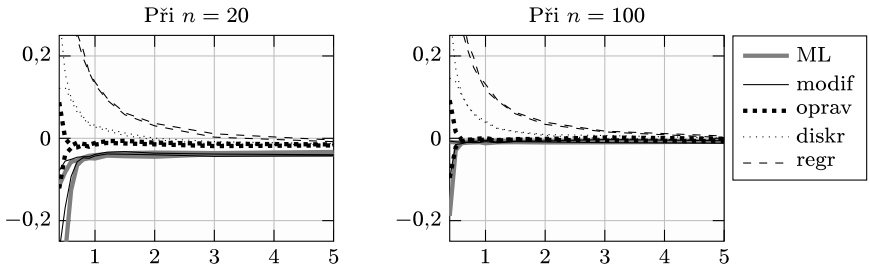
průměrné hodnocení 4,05. Směrodatná odchylka odpovědí u jednoho tvrzení vychází v průměru 0,95, u 50 % tvrzení vyšla menší než 0,85.

Příklad 2. Data ze vstupního testu znalostí studentů 1. ročníku z matematiky. Jde o test, který je pořádán od roku 2006/2007. Student formou výběru ze 4 až 5 odpovědí řeší 10 jednoduchých příkladů ze středoškolské matematiky. Výsledkem je počet správně zodpovězených otázek, tedy celé číslo mezi 0 a 10 (které považujeme za aproximaci spojité míry znalostí studenta). U výsledků z roku 2011/2012 (základní údaje viz [5]) pozorujeme v závislosti na fakultě či oboru směrodatné odchylky kolem 1,9, průměr je kolem 4 až 6 bodů.

Při úvahách o možnostech testování normality se omezíme jen na dva typy testů: chí-kvadrát test dobré shody a Kolmogorovův-Smirnovův test. V části 2 nejprve připomeneme některá fakta o diskretizovaném rozdělení a odhadech parametrů. V části 3 se budeme věnovat necenzorovaným pozorováním, pomocí simulací vyšetříme použitelnost testů a navrhneme randomizovanou variantu Kolmogorovova-Smirnovova testu. V části 4 pak uvážíme i cenzorovaná data a uvedeme souhrnné výsledky u výše zmíněných příkladů.

2. Diskretizované normální rozdělení

Uvažme rozdělení veličiny X^d, která vznikne celočíselným zaokrouhlením hodnot veličiny s normálním rozdělením $\mathrm{N}(\mu, \sigma^2)$. Tvar diskretizovaného rozdělení X^d závisí nejen na parametru σ, ale i na střední hodnotě μ (hlavně



Obrázek 2. Vychýlení odhadů σ (relativně k σ) v závislosti na σ při $n = 20$ a $n = 100$.

při menších σ). Se zaokrouhlením se změní i momenty, tj. střední hodnota a směrodatná odchylka veličiny X^d již nejsou rovny parametrům μ a σ. Zatímco u střední hodnoty je odchylka od μ relativně malá, u směrodatné odchylky může být rozdíl oproti σ větší. Při větších σ se odchylky zmenšují.

Při testování shody budeme potřebovat odhadnout parametry μ, σ z diskretizovaných dat. Výběrové momenty z původních spojitých dat jsou nedostupné a výběrové momenty získané ze zaokrouhlených dat konvergují k momentům diskretizovaného rozdělení, budou tedy i limitně různé od hledaných parametrů. Asymptoticky nestranné odhady získáme některým z obecných přístupů, ať už metodu maximální věrohodnosti (s použitím věrohodností funkce diskretizované veličiny), či metodou minimálního chí-kvadrát, resp. modifikovanou metodou minimálního chí-kvadrát. Při těchto přístupech jsou odhady řešením soustavy rovnic a prakticky je lze najít pouze numericky.

Můžeme se také uchýlit k některému jednoduchému přístupu, např. odhadnout parametry μ, σ jako odhad parametrů regresní přímky proložené metodou nejmenších čtverců body kvantilového grafu. V případě, že by data nebyla cenzorovaná, nabízí se také jednoduše vyjít z výběrových momentů z diskrétních dat a v případě výběrové směrodatné odchylky ji „opravit“. Tím myslíme nikoli odečíst $1/12$ diskretizačního kroku jako je tomu u Sheppardovy korekce odhadu rozptylu pořízeného z diskretizovaných dat, ale zvolit za odhad takové σ, při kterém je teoretická směrodatná odchylka veličiny X^d rovna pozorované. Obr. 2 ilustruje střední hodnoty odchylky diskutovaných odhadů od skutečného σ v závislosti na $\sigma \in \langle 0{,}4; 5\rangle$. Při každé hodnotě σ bylo pro $\mu = 0, 0{,}25, 0{,}5$ simulováno 2000krát a střední hodnota odchylek odhadnuta výběrovým průměrem. Vynesena jsou pro každé σ nejmenší a největší výchylení. Analogickým způsobem vznikly i další grafy v textu.

3. Testování z necenzorovaných dat

Tradičním testem pro seskupená data je Pearsonův chí-kvadrát test dobré shody. U rozdělení s malým rozptylem však při dané šířce třídy narazíme na problém s malým počtem tříd. K testu hypotézy H_0', kdy odhadujeme

2 parametry rozdělení, potřebujeme pozorování aspoň ve 4 třídách. Toto ale např. při 20 (resp. 100) pozorováních z $\mathrm{N}(\mu, \sigma^2)$ se $\sigma = 0{,}6$ nebude v závislosti na μ splněno s pravděpodobností 20–80 % (resp. až 30 %), při $\sigma = 0{,}8$ ve 20–30 % případů. Pokud se pozorovaná data nacházejí nejvýše ve dvou sousedních třídách, zřejmě mohou přesně odpovídat normálnímu rozdělení s nějakým malým rozptylem a vhodnou střední hodnotou. V tomto případě proto hypotézu H_0' nezamítneme. Stejně postupujeme, když počet tříd je 3 a počet stupňů volnosti vychází 0.

K tomu přistupuje problém, že test pracuje s asymptotickou hladinou významnosti při $n \to \infty$. Při malém rozsahu výběru tak používaná kritéria, deklarující přijatelnost asymptotiky prostřednictvím dostatečných očekávaných četností, zůstanou často i po sdružení tříd nenaplněna. Skutečná hladina významnosti tak může být odlišná (křivka CH v obr. 3).

Kolmogorovův-Smirnovův test pro spojitá rozdělení vychází ze vzdálenosti mezi teoretickou distribuční funkcí F_0 (za platnosti nulové hypotézy) a empirickou distribuční funkcí F_n, a pracuje s testovou statistikou

$$KS = \sup_{-\infty<x<\infty} \big|F_0(x) - F_n(x)\big| = \max_{1\leqq i\leqq n} \big|F_0(X_{(i)}) - F_n(X_{(i)}\pm)\big|,$$

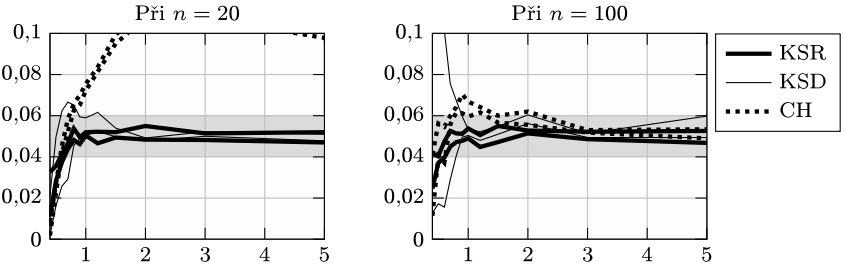
kde $X_{(i)}$ jsou pořádkové statistiky. Jiným způsobem poměřují odlišnost mezi F_n a F_0 test Cramérův-von Misesův, Andersenův-Darlingův a další, my se ale omezíme na test Kolmogorovův-Smirnovův. V základní variantě při testu jednoduché hypotézy H_0 rozdělení statistiky KS za platnosti H_0 nezávisí na F_0, z limitního rozdělení procesu $\sqrt{n}(F_n - F_0)$ vyplyne kritická hodnota pro velké rozsahy výběru a i v případě menších rozsahů výběru jsou kritické hodnoty tabelovány. Při testu složené hypotézy H_0' se v předpisu pro KS místo neznámého konkrétního tvaru F_0 použije distribuční funkce rozdělení s parametry rovnými odhadům. Tím se změní rozdělení KS, ale např. když šlo o parametry měřítka či polohy a když se použijí odhady invariantní ke změně měřítka či polohy, kritické hodnoty sice budou pro různé typy rozdělení a různé odhady různé, ale nezávislé na parametrech rozdělení (detailněji viz [1]). To umožňuje testovat složenou hypotézu, pro obvyklá rozdělení jsou kritické hodnoty tohoto Lillieforsova testu tabelovány.

V případě zaokrouhlených dat empirickou distribuční funkci F_n původního spojitého rozdělení neznáme, můžeme ale počítat s empirickou distribuční funkcí F_n^d určenou z pozorovaných diskrétních hodnot. Kolmogorovova-Smirnovova statistika KS^d sestavená za pomoci této distribuční funkce však může vykazovat odlišné chování než statistika KS ze spojitého případu. Zapíšeme-li

$$KS^d = F_n^d - F_0 = (F_n^d - F_n) + (F_n - F) = D_n + E_n, \tag{1}$$

kde $|D_n|$ se řídí binomickým rozdělením

$$|D_n(x)| = |F_n^d(x) - F(x)| \sim \frac{1}{n}\,\mathrm{Bi}(n, p(x)), \qquad p(x) = |F^d(x) - F(x)|,$$



Obrázek 3. Skutečné hladiny významnosti testů H_0' jako funkce σ při nominální hladině 5 % a při $n = 20$ a $n = 100$.

a E_n odpovídá rozdílu ze spojitého případu. Asymptoticky při $n \to \infty$ pak při pevné šířce diskretizace rozhodne D_n; pokud by se diskretizace s rostoucím n zjemňovala tak, že šířky intervalů by byly řádově menší než $n^{-1/2}$, limitní rozdělení by odpovídalo případu spojitému; limitní rozdělení při šířce řádově stejné jako $n^{-1/2}$ je odvozeno v [4]. Alternativou k tomuto postupu je neporovnávat hodnoty distribuční funkce F_0 v bodech pozorování s F_n, která tam má skok $1/n$, ale s průměrem limity F_n zleva a zprava v daném bodě. Dosáhneme tak sice snížení příslušné statistiky, ale limitně rozdělení ze spojitého případu odpovídat nebude.

Další možností je Kolmogorovův-Smirnovův test pro diskrétní rozdělení (v grafech značeno KSD), který využívá toho, že ve skutečnosti známe přírůstky F_n přes intervaly, jejichž krajní body d_j definovaly zaokrouhlené hodnoty. Tato diskrétní verze Kolmogorovova-Smirnovova testu pak porovnává F_n a F_0 v krajních bodech intervalů, tj. kumulované relativní četnosti a pravděpodobnosti tříd. V našem případě zaokrouhlování na celá čísla tedy

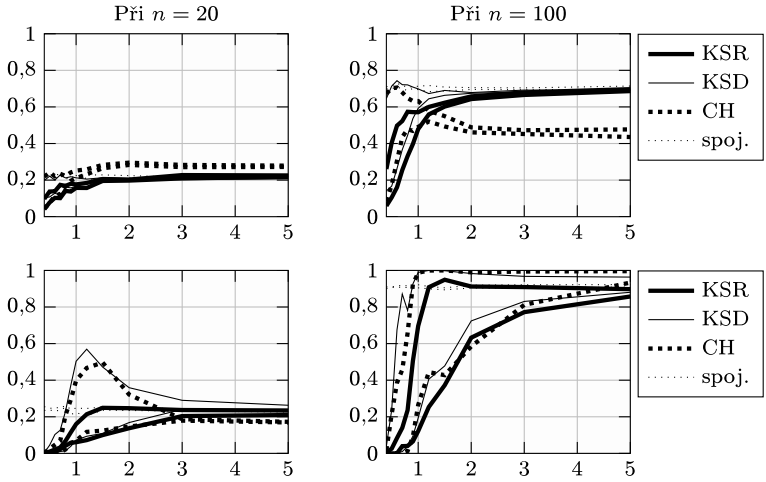
$$KSD = \sup_j \left|F_n(d_{(j)}) - F_0(d_{(j)})\right| = \max_i \left|F_n\Big(X_{(i)} \pm \frac{1}{2}\Big) - F_0\Big(X_{(i)} \pm \frac{1}{2}\Big)\right|.$$

Oproti verzi pro spojitá rozdělení probíhá porovnání empirické a teoretické distribuční funkce v méně bodech, proto kritické hodnoty testu budou nutně jiné. Na rozdíl od spojitého případu se budou bohužel lišit pro různá rozdělení F_0 (a jeho parametry) a pro každé dělení. Určením kritické hodnoty či síly tohoto testu se věnuje např. [2]. Aplikovat tento přístup pro test složené hypotézy H_0' by znamenalo dosadit do (1) místo neznámých parametrů jejich odhady a např. naivně použít kritickou hodnotu odpovídající testu jednoduché hypotézy s hypotetickou hodnotou parametru rovnou hodnotě odhadnuté z dat. Takto ale nelze očekávat dodržení stanovené hladiny významnosti.

Jako alternativu navrhujeme randomizovaný test (v grafu KSR), který hypotézu H_0 bude při daných pozorováních zamítat s pravděpodobností

$$\mathrm{P}\big(KS(X_1, \ldots, X_n) > KS_{\mathrm{krit}} \mid X_1^d, \ldots, X_n^d\big), \tag{2}$$

kde pravděpodobnost se myslí za platnosti H_0 a KS_{krit} je tradiční kritická hodnota Kolmogorova-Smirnovova testu pro spojitá rozdělení při zvolené

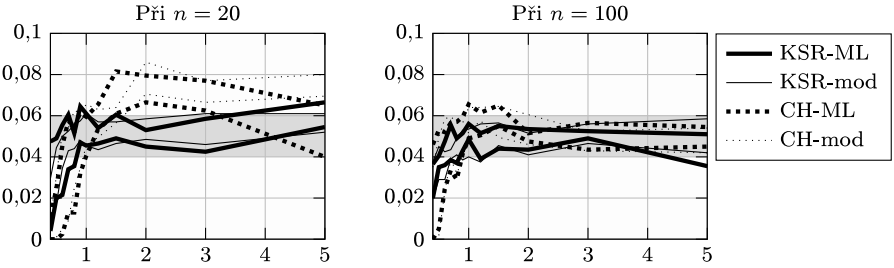


Obrázek 4. Síla testů hypotézy H_0' na hladině významnosti 5 % proti alternativám Laplaceovo (nahoře) a useknuté normální (dole) jako funkce parametru σ, tečkovaně KS ze spojitých dat. Při $n = 20$ (vlevo) a $n = 100$ (vpravo).

hladině významnosti. Jak by se (podmíněná) pravděpodobnost (2) spočetla, není potřeba zkoumat, protože při praktické realizaci testu pouze potřebujeme s pravděpodobností (2) na základě daných pozorování vygenerovat rozhodnutí o zamítnutí nebo nezamítnutí hypotézy H_0. Jelikož známe počty pozorování v jednotlivých intervalech, můžeme v každém dle rozdělení F_0 dogenerovat příslušný počet konkrétních hodnot a na dogenerovaná pozorování (která za H_0 tvoří náhodný výběr z F_0) aplikovat statistiku KS ze spojitého případu. Tudíž tento test přesně dosáhne předem zvolené hladiny významnosti.

V případě složené hypotézy H_0' už situace není takto příznivá. Místo parametrů v předpisu KS můžeme použít odpovídající odhady a na místě kritické hodnoty v (2) Lillieforsovu kritickou hodnotu, avšak neumíme výše uvedeným postupem bez znalosti skutečných hodnot parametrů dogenerovat náhodný výběr z F_0. Přesto, nabízí se příslušný počet hodnot v každém intervalu rozprostřít do tohoto intervalu nikoli podle rozdělení F_0, které neznáme, ale podle rozdělení, kde místo skutečných hodnot parametrů použijeme jejich odhady spočtené některou metodou z pozorovaných hodnot diskretizovaného rozdělení. Zde můžeme jen doufat, že skutečná hladina významnosti bude blízká nominální.

Obrázek 3 zachycuje chování testů za platnosti nulové hypotézy na základě 5000 simulací, ve všech případech se parametry odhadovaly modifikovanou metodou minimálního chí-kvadrát. Zkoumali jsme sílu proti některým alternativám. Na obr. 4 vidíme sílu testů normality (2000 simulací), když data



Obrázek 5. Skutečné hladiny významnosti testů z cenzorovaných dat (nominální 5 %) při $n = 20$ a $n = 100$.

pocházejí ve skutečnosti z Laplaceova rozdělení s $\mathrm{E}\, X = \mu$ a $\mathrm{E}\,|X - \mu| = \sigma$ a useknutého normálního $\max(\mu, \mathrm{N}(\mu, \sigma^2))$. Analogické rozdíly mezi testy lze pozorovat, když pocházejí např. ze směsi normálních rozdělení či z t-rozdělení.

Zkonstruovat nerandomizovanou verzi testu KSR můžeme např. tak, že zamítneme H_0 při daných pozorováních, když $\mathrm{E}(KS \mid X_1^d, \ldots, X_n^d) > \mathrm{E}_{\mathrm{krit}}$. Podmíněnou střední hodnotu na levé straně, stejně jako kritickou hodnotu $\mathrm{E}_{\mathrm{krit}}$ pro ni, by bylo potřeba odhadnout např. simulacemi. Podobně bychom mohli zamítnout H_0, když podmíněný kvantil statistiky KS při daných pozorováních překročí kritickou hodnotu rozdělení tohoto kvantilu.

4. **Cenzorovaná data**

V příkladech uvedených v úvodu byla pozorována zaokrouhlená data s tím, že v krajních třídách byla zahrnuta i pozorování se skutečnými hodnotami většími (v pravé krajní třídě) i menšími (v levé třídě). Tj. pozorování jsou cenzorována zprava (levým krajním bodem nejvyšší třídy) a zleva (pravým krajním bodem nejnižší třídy) a cenzorující veličiny jsou nenáhodné. V případě chí-kvadrát testu žádný podstatný problém navíc nevzniká, cenzorovaná data přispívají do krajních tříd a při testu složené hypotézy je jen nutno příslušně upravit odhady parametrů.

U Kolmogorovova-Smirnovova testu a jeho variant můžeme postupovat tak, že porovnáme empirickou a teoretickou distribuční funkci jen v části definičního oboru — bez krajů určených cenzorujícími veličinami. Je třeba vzít ale v úvahu, že už v případě Kolmogorovova-Smirnovova testu bude kritická hodnota záviset na cenzorujících veličinách, pro každou cenzorující hodnotu by tedy bylo třeba vytvořit tabulky (nebo kritickou hodnotu určit simulačně). U nastíněných variant pak přistupuje ještě závislost na parametrech skutečného rozdělení. Na ukázku v obr. 5 uvádíme skutečně dosažené hladiny významnosti (zjištěné z 2000 simulací) pro model ankety, kdy pracujeme se zaokrouhlenými pozorováními náhodného výběru z $\mathrm{N}(\mu, \sigma^2)$ cenzorovaného zleva 0 zprava 5. Použita byla $\mu = 3{,}5, 3{,}75, \ldots, 4{,}5$

Nakonec porovnejme výsledky testování na skutečných datech. Testování probíhalo na hladině významnosti 5 % a použit byl chí-kvadrát test a randomizovaný Kolmogorovův-Smirnovův test, vždy jednou s odhadem očekávaných četností resp. parametrů rozdělení na základě odhadu maximální věrohodnosti a jednou dle modifikované metody minimálního chí-kvadrát.

Na datech z příkladu 1 byla testována normalita u 241 otázek o předmětech jedné fakulty. Chí-kvadarát test zamítl normalitu v 45 případech (19 %), randomizovaný test ve 21, resp. 7 případech. Neshoda mezi závěry některých testů nastala u 40 otázek, z čehož ve 20 případech bylo rozhodnutí těsné, resp. ve 21 se odlišoval výsledek jednoho z testů (z provedené čtveřice).

V příkladu 2 při testu normality po fakultách vedly všechny testy vždy ke stejnému rozhodnutí. Tímto rozhodnutím bylo v případě fakult s větším počtem účastníků (170–650) zamítnutí normality, u fakult s menším počtem (20–60) se normalita nezamítla. Dále byl proveden test normality u 24 studijních programů, k zamítnutí normality došlo u 6 až 8 oborů, v závislosti na použitém testu. Neshoda mezi testy se projevila u 7 oborů, z nich u 3 bylo rozhodnutí těsné a u dalších 2 vyšel odlišně 1 test.

5. Závěr

Pro pozorování s danými specifickými vlastnostmi není chí-kvadrát test dobré shody při tak malém rozsahu výběru vhodný a i při větším rozsahu může mít v případě menších σ vyšší než stanovenou pravděpodobnost chyby prvního druhu. S tím je třeba počítat při pohledu na jeho někdy zdánlivě vyšší sílu. U navržené randomizované varianty Kolmogorovova-Smirnovova testu skutečná hladina významnosti odpovídá stanovené řekněme už od $\sigma = 0{,}6$, jeho síla se přiblíží síle testu ze spojitých pozorování až při větším rozptylu. V případě necenzorovaných dat lze jako odhad směrodatné odchylky použít výpočetně jednoduchou metodu „opravy“ výběrové směrodatné odchylky pořízené z diskretizovaných pozorování.

Reference

[1] David F. N. and Johnson N. L. (1948) *The probability integral transformation when parameters are estimated from the sample.* Biometrika **35**, 182–190.

[2] Gleser L. J. (1985) *Exact power of goodness-of-fit test of Kolmogorov type for discontinuous distributions.* J. Amer. Stataist. Assoc. **80** (no. 392), 954–958.

[3] *Hodnocení kvality výuky.* Portál ZČU, Informační systém Západočeské univerzity v Plzni, [cit. 2012-12-20], dostupné z http://portal.zcu.cz/wps/portal/kvalita-vyuky/.

[4] O'Reilly F. J., Rueda R. and Garza-Jinich M. (2003) *How important is the effect of rounding in goodness-of-fit.* Comm. Statist. Simulation Comput. **32** (no. 3), 953–976.

[5] *Výsledky vstupních testů 2011/2012.* Statistici na KMA, [cit. 2012-12-20], dostupné z http://stat.kma.zcu.cz/testik11/.

Poděkování: Tato práce byla podpořena projektem CZ.1.07/2.2.00/15.0377.

Adresa: FAV ZČU, KMA, Univerzitní 22, 306 14 Plzeň

E-mail: friesl@kma.zcu.cz

REGIONÁLNÍ ANALÝZA SRÁŽKOVÝCH EXTRÉMŮ V SIMULACÍCH REGIONÁLNÍCH KLIMATICKÝCH MODELŮ

Martin Hanel (1,2), T. Adri Buishand (3)

Abstrakt: Článek informuje o metodice a výsledcích analýzy srážkových extrémů a jejich změn v simulacích regionálních klimatických modelů. Stručně je popsán statistický model vycházející z klasické regionální frekvenční analýzy, který je upraven za účelem využití v nestacionárních podmínkách. Tento model byl aplikován při analýze simulovaných srážkových extrémů pro Nizozemí, povodí Rýna a Českou republiku. Hlavním cílem představených analýz bylo zejména posouzení schopnosti regionálních klimatických modelů simulovat srážkové extrémy a vyhodnocení změn těchto extrémů v možném budoucím klimatu. Denní srážkové extrémy jsou v simulacích regionálních klimatických modelů relativně dobře reprezentovány, naopak charakteristiky hodinových srážkových extrémů v simulacích se výrazně liší od charakteristik odvozených z pozorovaných dat. Regionální klimatické modely projektují pro druhou polovinu jedenadvacátého století obecně spíše zvyšování srážkových extrémů, nicméně v případě vícedenních srážkových extrémů může docházet ke stagnaci nebo poklesu, zejména v zimním období. Výsledky pro jednotlivé simulace se nicméně poměrně značně liší.

Klíčová slova: srážkové extrémy; regionání frekvenční analýza; změna klimatu

Abstract: The present paper brings information on development and application of non-stationary index-flood model for precipitation extremes in regional climate model simulations. The model has been applied for the assessment of precipitaion extremes in Netherlands, Rhine basin and the Czech Republic. The objective of these analyses was to evaluate the skill of the climate models in simulating precipitation extremes and assessment of changes of these extremes. It turned out that daily precipitation extremes are reasonably well reproduced by in the climate model simulations, however extremes for shorter (hour) and longer (10 days and more) durations deviate from the observed precipitation extremes. Regional climate model indicate increase of daily and subdaily precipitation extremes, for multi-day precipitation extremes, however, the increase is weak, some models even indicate a decrease.

Keywords: precipitation extremes; regional frequency analysis; climate change

1. Úvod

Srážky jsou klíčovým faktorem ovlivňujícím výskyt extrémních hydrologických jevů. Studiu jejich změn v podmínkách klimatické změny se proto v posledních letech věnuje značná pozornost. Dle simulací klimatických modelů

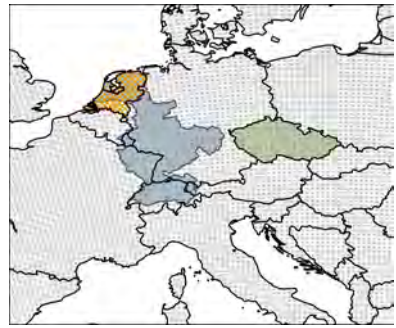
leží Česká republika v oblasti vykazující jen minimální změny průměrných ročních srážkových úhrnů. Nicméně s velkou pravděpodobností se dají očekávat změny sezónního rozložení srážek (viz např. [14]), jenž může do značné míry ovlivnit hydrologickou bilanci i při stagnaci ročních průměrných srážek. Pro extrémní hydrologické jevy, zejména povodně, jsou však spíše než změny průměrných srážkových úhrnů podstatnější změny srážkových extrémů. Tyto změny mohou být často značně odlišné od změn průměrných úhrnů [3, 1].

Detekovat systematické změny v rozdělení srážkových extrémů pro jednu stanici (výpočetní bod klimatického modelu) je velmi obtížné, zejména kvůli značné přirozené variabilitě klimatického systému. Navíc extrémy jsou ze své podstaty jevy nevyskytující se často, tj. pro odhady charakteristik srážkových extrémů je k dispozici pouze omezený počet pozorování, což vede k značné nejistotě odhadů. Tyto nejistoty je možno snížit použitím metod regionální analýzy, jež předpokládá, že nejvíce nejisté parametry rozdělení extrémních hodnot jsou v definovaných homogenních oblastech konstantní (viz např. [10]).

V předkládaném příspěvku prezentujeme výsledky analýzy extrémních srážek a jejich změn dle simulací regionálních klimatických modelů provedených v rámci evropského projektu ENSEMBLES [9]. Cílem této analýzy byla jednak validace srážkových extrémů v klimatických modelech pro kontrolní období (1961-1990), jednak vyhodnocení změn kvantilů extrémních srážek (do 50ti leté srážky) mezi kontrolním a scénářovým obdobím (2070-2099). Za účelem této analýzy byl vyvinut statistický model umožňující regionální analýzu srážkových extrémů v případě nestacionárních dat [8]. Statistický model je stručně popsán v následující kapitole. Dále předkládáme přehled prezentovaných případových studií spolu s použitými daty. Následuje přehled výsledků tří případových studií demonstrující využití statistického modelu a shrnující hlavní poznatky o reprezentaci srážkových extrémů v simulacích regionálních klimatických modelů a jejich změnách v možném budoucím klimatu.

2. Statistický model

K analýze srážkových extrémů je využit nestacionární regionální model [8]. Předpokladem modelu je, že srážkové extrémy v každém výpočetním bodě regionálního klimatického modelu předem definované oblasti mohou být normovány tak, že rozdělení těchto normovaných srážkových extrémů je v dané oblasti stejné. Normovací faktor, který je určen pro jednotlivé regionální klimatické modely (RCM) výpočetní body, je zpravidla označován „index-flood“, stejně jako tato metoda[10]. Pro srážkové extrémy v případě pozorovaných srážek i srážek simulovaných regionálním klimatickým modelem je často uvažováno GEV rozdělení [12, 11, 2, 4], které má tři parametry (ξ - parametr polohy, α - parametr měřítka, κ - parametr tvaru) a kombinuje tři limitní rozdělení extrémů (Gumbelovo, Fréchetovo a obrácené Weibullovo). Použité předpoklady implikují, že parametr κ a koeficient $\gamma = \alpha/\xi$ jsou v uvažované oblasti konstantní.



Obrázek 1. Lokalizace jednotlivých případových studií - Česká republika (zeleně), povodí Rýna (modře) a Nizozemí (oranžová). Výpočetní síť společná pro většinu regionálních klimatických modelů z projektu ENSEMBLES je zobrazena šedou barvou v pozadí.

Jako normovací faktor bývá často použit průměr nebo medián rozdělení extrémů na jednotlivém výpočetním bodě/stanici, nicméně v případě nestacionárního modelu je výhodné normovat rozdělení extrémů pomocí GEV parametru ξ. Nestacionární regionální model je pak definován tak, že parametry γ a κ jsou v ploše konstantní, ale mění se v čase (s trendem společným pro celou oblast) a parametr ξ je různý pro jednotlivé výpočetní body, ale v čase se mění s trendem společným pro celou oblast. Všechny parametry jsou pak vztaženy k ukazateli času, např. roku, teplotě (rostoucí v důsledku globální změny klimatu), atp. Parametry jsou odhadnuty metodou maximální věrohodnosti a umožňují odvození odhadů libovolných kvantilů rozdělení extrémních srážek a jejich změn. K posouzení nejistot odhadu parametrů je využit bootstrap. Další podrobnosti, testování modelu apod. viz [8].

Zpravidla není možné najít oblasti, které by byly homogenní pro všechny RCM simulace, roční období a časové agregace srážkových extrémů. Volba homogenních oblastí je tedy vždy do jisté míry subjektivní a homogenita oblastí může být v jednotlivých případech porušena. Nicméně je známo, že porušení předpokladů homogenity oblastí nevede k zásadním chybám v odhadech kvantilů srážkových extrémů ani v případě odhadů jejich změn. V aplikacích zmiňovaných v tomto příspěvku byly homogenní oblasti definovány zpravidla na základě map parametru γ a změn parametrů γ a ξ pro jednotlivé simulace, roční období, případně časové agregace.

3. **Vybrané studie**

V této kapitole předkládáme výsledky tří případových studií (viz Obrázek 1): validace hodinových a denních ročních srážkových extrémů pro Nizozemí [5], vyhodnocení změn 1denních letních a 5denních zimních srážkových extrémů

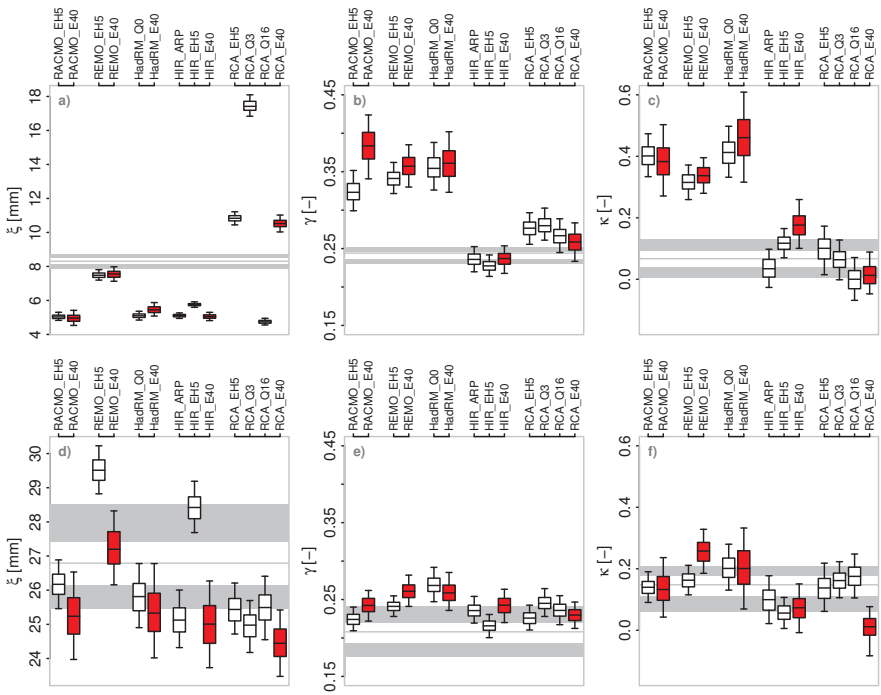
TABULKA 1. Přehled simulací regionálních klimatických modelů použitých v jednotlivých studiích.

akronym	model	Nizozemí	povodí Rýna	Česká republika
řídící model ECHAM5				
RACMO_EH5	RACMO2.1	×	×	×
REMO_EH5	REMO5.7	×	×	×
RCA_EH5	RCA3.0	×	×	×
RegCM_EH5	RegCM3		×	×
HIR_EH5	HIRHAM5	×		×
řídící model HadCM3Q0, HadCM3Q3, HadCM3Q16				
HadRM_Q0	HadRM3.0	×	×	×
CLM_Q0	CLM2.4.6		×	×
HadRM_Q3	HadRM3.0		×	×
RCA_Q3	RCA3.0	×	×	×
HadRM_Q16	HadRM3.0		×	×
RCA_Q16	RCA3.0	×	×	×
řídící model ARPEGE4.5				
HIR_ARP	HIRHAM5	×	×	×
CNRM_ARP	CNRM-RM4.5		×	
CNRM5_ARP	CNRM-RM5.1			×
řídící model BCM2.0				
HIR_BCM	HIRHAM2	×		
RCA_BCM	RCA3.0		×	×
řídící model CGCM3.1				
CRCM_CCC	CRCM4.2.1		×	
řízeno reanalýzou ERA-40				
RACMO_E40	RACMO2.1	×		
REMO_E40	REMO5.7	×		
HadRM_E40	HadRM3.0	×		
HIR_E40	HIRHAM5	×		
RCA_E40	RCA3.0	×		

pro povodí Rýna [6] a odhad změn sezónních srážkových extrémů pro různé časové agregace pro Českou republiku [7].

Pro všechny případové studie byly využity simulace regionálních klimatických modelů z projektu ENSEMBLES (viz Tab. 1). Regionální klimatické modely byly řízeny různými simulacemi globálních klimatických modelů dle emisního scénáře SRES A1B (v rámci rodiny SRES scénářů jde o scénář předpokládající zhruba průměrné zvyšování emisí skleníkových plynů), případně reanalýzou pozorovaného klimatu ERA-40. Většina simulací byla provedena na stejné výpočetní síti a všechny simulace mají rozlišení 25 km × 25 km.

Pro všechny uvedené simulace byly k dispozici denní řady srážek pro období 1961-2099 (většina simulací začíná již 1950). Pro simulace použité k validaci hodinových srážkových extrémů jsou dostupná i denní maxima hodinových srážek. Pro jednotlivé simulace regionálních klimatických modelů byla extrahována maxima pro relevantní výpočetní body: pro validaci simulací pro Nizozemí roční hodinová a denní, pro povodí Rýna 1denní letní a 5denní

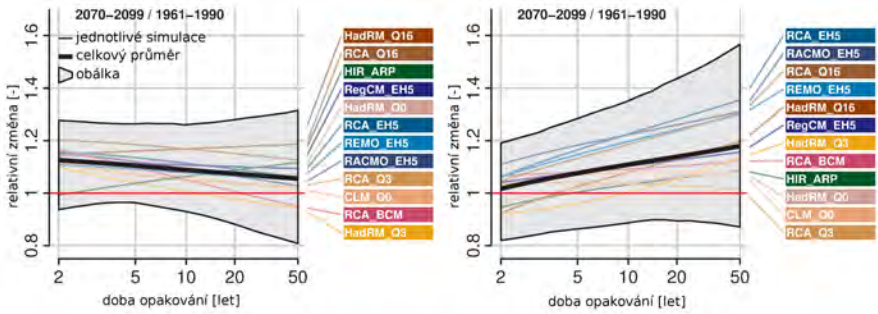


Obrázek 2. Odhady parametrů GEV modelu pro hodinová (a-c) a denní (d-f) roční srážková maxima v Nizozemí pro kontrolní období. Prázdné boxy odpovídají simulacím řízeným globálním klimatickým modelem, červené pak simulace řízené reanalýzou ERA-40. V pozadí je odhad parametrů z radarových dat. Nejistota byla odhadnuta pomocí bootstrap resamplingu, indikován je 5, 25, 50, 75 a 95% kvantil rozdělení.

zimní a v případě České republiky 1, 3, 5, 7, 10, 15, 30denní srážková maxima pro všechna roční období.

3.1. Validace hodinových srážkových extrémů v RCM simulacích

Pro území Nizozemí jsou k dispozici odhady parametrů GEV rozdělení ročních srážkových extrémů pocházející z radarových dat pro různé časové agregace a různě velké plochy [13]. Tyto odhady pro hodinová a denní srážková maxima pro plochu 25 km × 25 km bylo možno přímo porovnat s odhady parametrů nestacionárního GEV modelu založenými na simulacích regionálních klimatických modelů. Výsledky udává Obázek 2.



Obrázek 3. Změny 1denních letních (vlevo) a 5denních zimních (vpravo) v simulacích regionálních klimatických modelů pro povodí Rýna. Šedě je vyznačena obálka (5-95 %) z 500 bootstrap samplů.

Parametry rozdělení hodinových maxim v simulacích regionálních klimatických modelů se značně liší od parametrů maxim odvozených z radarových dat (na obrázku vyznačeny šedou barvou v pozadí). Ve většině simulací je rozdělení hodinových srážkových extrémů posunuto k nižším hodnotám (Obrázek 2a), nicméně toto posunutí je u části modelů kompenzováno výrazně těžším pravým chvostem (parametry γ a κ, viz Obrázek 2b-c). Pro většinu modelů jsou tedy kvantily do cca 10leté hodinové srážky podhodnoceny a nad tuto dobu opakování nadhodnoceny, s tím že chyba se s rostoucí dobou opakování zvětšuje.

Jednodenní maxima jsou ve většině simulací naopak reprezentovány víceméně adekvátně. Většina simulací rozdělení extrémů posunuje mírně k nižším hodnotám(viz Obrázek 2d), nicméně tento posun je částečně kompenzován vyšší variabilitou extrémů (Obrázek 2e). Obecně jsou kvantily denních srážkových extrémů pro delší doby opakování (např. 50 let) spíše nadhodnoceny, nicméně rozdíly od odhadů z radarových dat jsou podstatně menší než v případě hodinových dat.

Zároveň je možno konstatovat, že charakter srážkových extrémů je do značné míry určen regionálním klimatickým modelem, tj. simulace řízené reanalýzou ERA-40 se významně neliší od simulací řízených globálním klimatickým modelem.

3.2. Změny srážkových extrémů v simulacích regionálních klimatických modelů pro povodí Rýna

Pro povodí Rýna (viz Obrázek 1) byly analyzovány 1denní letní a 5denní zimní srážkové extrémy v 15 simulacích regionálních klimatických modelů. Vícedenní zimní extrémy byly vyhodnoceny zejména proto, že zimní povodně v této oblasti jsou způsobeny spíše vícedenními srážkami.

Změny letních srážkových extrémů jsou ovlivněny zejména zvyšováním relativní variability extrémů (parametr γ), jenž je patrné pro většinu simulací. Ostatní parametry spíše stagnují. Výsledkem je stagnace kvantilů srážkových extrémů pro kratší doby opakování (např. 2 roky) a růst srážkových extrémů s prodlužováním doby opakování až o 20 % pro 50letou srážku (viz Obrázek 3). Rozdíly v odhadech změn dle jednotlivých simulací jsou pro 50letou srážku řádově ± 15 %, nicméně změny jsou v souboru modelů konzistentní, tj. stagnace kvantilů pro kratší doby opakování a jejich růst s prodlužující se dobou opakování můžeme konstatovat pro téměř všechny modely.

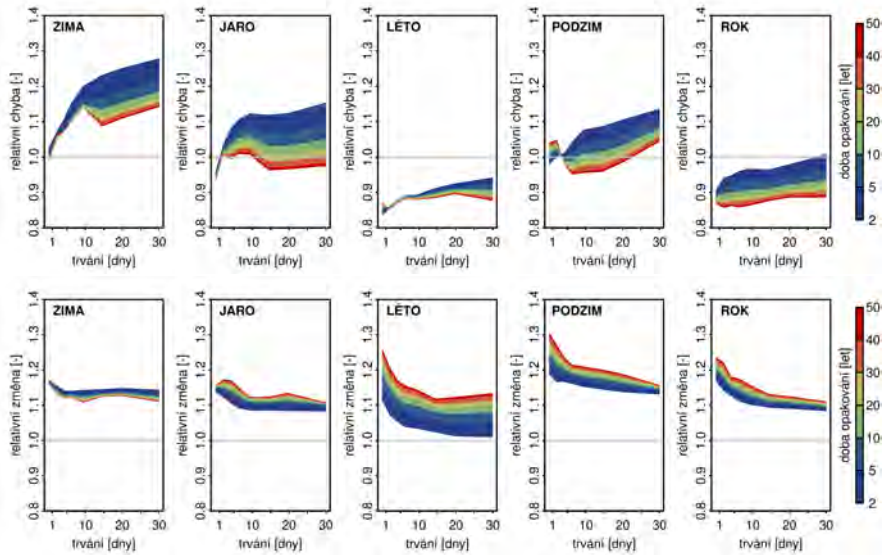
Změny 5denních zimních srážkových extrémů jsou ovlivněny zejména posunem celého rozdělení k vyšším hodnotám (růst parametru ξ) a poklesem relativní variability extrémů a odlehčování pravého chvostu rozdělení (pokles parametrů γ a κ). To vede k růstu kvantilů srážkových extrémů pro kratší doby opakování a jejich stagnaci, pro některé simulace i k poklesu pro dobu opakování 50 let (viz Obrázek 3). Rozdíly mezi jednotlivými simulacemi jsou řádově podobné rozdílu pro letní extrémy. Nicméně charakter změn je v souboru modelů méně konzistentní.

Dále jsme analyzovali nakolik porušení předpokladů o homogenitě oblasti ovlivňuje výsledný odhad změn kvantilů rozdělení srážkových extrémů. K tomuto účelu byl upraven statistický model tak, aby umožňoval prostorovou variabilitu parametru γ, tedy relativní variabilita srážkových extrémů se mění v ploše oblasti. To vedlo k podstatnému zlepšení testů homogenity oblasti, nicméně výsledné změny kvantilů srážkových extrémů byly téměř identické jako výsledky vycházející z původního modelu.

3.3. Vyhodnocení systematických chyb a projektovaných změn 1 až 30denních srážkových extrémů pro Českou republiku

Pro Českou republiku byla provedena rozsáhlá studie zaměřená na validaci a vyhodnocení změn 1 až 30denních srážkových extrémů pro všechna roční období. Pro porovnání s pozorovanými srážkovými extrémy byla využita datová sada pozorování interpolovaných na výpočetní síť regionálních klimatických modelů [15]. V dalším textu diskutujeme průměrné systematické chyby a změny ze souboru klimatických modelů. Výsledky pro jednotlivé modely jsou do jisté míry odlišné.

Pro podzimní, zimní a jarní srážkové extrémy platí, že lépe jsou simulovány extrémy pro kratší (denní) agregace (viz Obrázek 4). S prodlužováním agregace až na 30 dní roste systematická chyba, zejména pro nižší doby opakování (až na 10 - 20 %). Paradoxně, extrémnější události jsou simulovány lépe, což je způsobeno nadhodnocením polohy rozdělení, které je pro vyšší kvantily kompenzováno lehčím chvostem rozdělení. Naproti tomu letní extrémy jsou podhodnoceny (cca 10 - 15 %), přičemž systematická chyba je mírně nižší pro delší agregace. Jelikož roční extrémy (zejména pro vyšší doby opakování) jsou z velké části tvořeny extrémy letními, systematická chyba v ročních srážkových extrémech je do značné míry srovnatelná s letním obdobím, nicméně



Obrázek 4. Průměrná systematická chyba v kvantilech rozdělení sezónních a ročních srážkových maxim pro Českou republiku v období 1961-1990 (nahoře). Průměrná změna kvantilů rozdělení sezónních a ročních srážkových maxim pro Českou republiku mezi obdobími 2070-2099 a 1961-1990 (dole).

pro nižší doby opakování (2 roky) je podhodnocení obecně o cca 10 % nižší než pro vyšší doby opakování (50 let).

Pro všechna roční období projektují klimatické modely zvýšení srážkových extrémů. Pro vícedenní extrémy je toto zvýšení podobné pro všechna roční období (řádově 10-20 %). V létě a na podzim je patrný vyšší růst srážkových extrémů pro kratší agregace, v létě navíc srážkové extrémy pro kratší doby opakování rostou podstatně méně než pro delší doby opakování (o cca 10 %). Obecně nejvyšší růst modely projektují pro podzimní období (až 30 % pro 1denní 50letou srážku). Změny ročních extrémů jsou kombinací změn podzimních a letních extrémů s vyšším, cca 20% růstem pro krátké agregace, cca 10% růstem pro dlouhé agregace a s malým rozdílem ve změnách pro různé doby opakování. Příčiny změn srážkových extrémů v letním a zimním období jsou podobné jako v případě povodí Rýna, tedy v zimě zejména posun rozdělení k vyšším hodnotám a v letním období růst relativní variability extrémů.

4. Závěr

Cílem tohoto příspěvku nebylo podat vyčerpávající informace o provedených, výše zmíněných analýzách, ale spíše demonstrovat možnosti využití statistického modelování srážkových extrémů pro validaci klimatických modelů a sumarizaci systematických chyb i projektovaných změn. Použitý statistický model poskytuje tuto sumarizaci přirozeně ve formě odchylek/změn parametrů GEV modelu a tím zároveň zprostředkovává vhled do jejich podstaty. Regionální přístup navíc omezuje vliv přirozené variability srážkových extrémů a umožňuje odhadovat parametry jejich rozdělení s větší přesností.

Regionální klimatické modely nejsou zcela schopny reprezentovat krátkodobé (hodinové) srážkové extrémy a i simulace vícedenních extrémů většinou vykazují podstatné systematické chyby (kromě léta jsou extrémy nadhodnoceny). Denní srážkové extrémy jsou zachyceny relativně věrně. Nicméně v porovnání s chybou simulovaných průměrných srážek, která se pohybuje během podzimu-jara v rozmezí 20-40 %, je chyba srážkových extrémů zpravidla nižší.

V podmínkách změny klimatu regionální klimatické modely obecně projektují spíše růst srážkových extrémů, nicméně v zimním období pro některé oblasti indikují stagnaci či pokles vícedenních srážkových maxim. Přestože se absolutní hodnoty změn mezi jednotlivými modely liší, jejich podstata je často obdobná. Neurčitost scénářů změn srážkových extrémů ve střední Evropě je přesto poměrně velká, zejména pokud se vezmou do úvahy i další faktory, na které jsme se v předložené práci nezaměřili (např. různé scénáře budoucího vývoje koncentrací skleníkových plynů a lidské společnosti obecně).

Přestože v posledních cca deseti letech byla provedena celá řada analýz extrémních srážek v simulacích klimatických modelů, nelze tvrdit, že by problematika byla uzavřena - jednak dochází k dalšímu vývoji a zlepšování klimatických modelů (vyšší rozlišení, lepší zachycení orografie, lepší reprezentace procesů vedoucích ke vzniku srážek) a realizaci nových simulací klimatických modelů (v současnosti zejména v souvislosti s připravovanou pátou hodnotící zprávou Mezivládního panelu pro klimatickou změnu) - jednak i po teoretické stránce je problematika (prostorového) modelování nestacionárních extrémů stále živá, příkladem může být rychlý vývoj POT (peak over threshold) přístupů k nestacionárním prostorovým extrémům, dalším aktuálním tématem je například využití neparametrických modelů trendu, případně adaptace alternativních metod regionální frekvenční analýzy (např. ROI - region-of-influence) pro nestacionární podmínky.

Literatura

[1] Boberg F., Berg P., Thejll P., Gutowski W.J., Christensen J.H. (2009) *Improved confidence in climate change projections of precipitation evaluated using daily statistics from the PRUDENCE ensemble.* Climate Dynamics, **32**, 1097 – 1106.

[2] Buonomo E., Jones R., Huntingford C., Hannaford J. (2007) *On the robustness of changes in extreme precipitation over Europe from two high resolution climate change simulations.* Quarterly Journal of the Royal Meteorological Society, **133(622)**, 65 – 81.

[3] Christensen O.B., Christensen J.H. (2004) *Intensification of extreme European summer precipitation in a warmer climate.* Global and Planetary Change, **44**, 107 – 117.

[4] Goubanova K., Li L. (2007) *Extremes in temperature and precipitation around the Mediterranean basin in an ensemble of future climate scenario simulations.* Global and Planetary Change, **57(1–2)**, 27 – 42.

[5] Hanel M., Buishand T.A. (2010) *On the value of hourly precipitation extremes in regional climate model simulations.* Journal of Hydrology, **393(3–4)**, 265 – 273.

[6] Hanel M., Buishand T.A. (2011) *Analysis of precipitation extremes in an ensemble of transient regional climate model simulations for the Rhine basin.* Climate Dynamics, **36**, 1135 – 1153.

[7] Hanel M., Buishand T.A. (2012) *Multi-model analysis of RCM simulated 1-day to 30-day seasonal precipitation extremes in the Czech Republic.* Journal of Hydrology, **412–413(0)**, 141 – 150.

[8] Hanel M., Buishand T.A., Ferro C.A.T. (2009) *A nonstationary index flood model for precipitation extremes in transient regional climate model simulations.* Journal of Geophysical Research, **114(D15)**.

[9] Hewitt C.D., Griggs D.J. (2004) *Ensembles-based Predictions of Climate Changes and their Impacts.* Eos, **85**, 566.

[10] Hosking J.R.M., Wallis J.R. (1997) *Regional frequency analysis: an approach based on L-moments.* Cambridge University Press.

[11] Kyselý J., Beranová R. (2009) *Climate-change effects on extreme precipitation in central Europe: uncertainties of scenarios based on regional climate models.* Theoretical and Applied Climatology, **95**, 361 – 374.10.1007/s00704-008-0014-8.

[12] Kyselý J., Picek J. (2007) *Regional growth curves and improved design value estimates of extreme precipitation events in the Czech Republic.* Climate Research, **33(3)**, 243 – 255.

[13] Overeem A., Buishand T.A., Holleman I., Uijlenhoet R. (2010) *Extreme value modeling of areal rainfall from weather radar.* Water Resources Research, **46(W09514)**.

[14] Solomon S., Qin D., Manning M., et al., eds. (2007) *Climate Change 2007: The Physical Science Basis.* Cambridge University Press, Cambridge, UK and New York, NY, USA.

[15] Štěpánek P., Zahradníček P., Huth R. (2011) *Interpolation techniques used for data quality control and calculation of technical series: an example of a Central European daily time series.* IDŐJÁRÁS - Quarterly Journal of the Hungarian Meteorological Service, **115(1–2)**, 87 — 98.

Poděkování: Tento příspěvek vznikl s podporou projektu ESF „Zapojení týmu KLIMATEXT do mezinárodní spolupráce (reg.č. č. CZ.1.07./2.3.00/20.0086)“. Využitá data byla poskytnuta v rámci projektů EU ENSEMBLES (No.505539) a VaV „Zpřesnění dosavadních odhadů dopadů klimatické změny v sektorech vodního hospodářství, zemědělství a lesnictví a návrhy adaptačních opatření“ (SP/1a6/108/07).

Adresa: (1) Technická univerzita v Liberci, Studentská 2, 461 17 Liberec (2) Česká zemědělská univerzita v Praze, Kamýcká 129, 165 21 Praha (3) Královský nizozemský meteorologický institut, PO Box 201, 3730 AE De Bilt, Nizozemí

E-mail: `hanel@fzp.czu.cz`

SIMULTÁNNE TESTOVANIE STREDNEJ HODNOTY A VARIANCIE NORMÁLNEHO ROZDELENIA

Martina Chvosteková

Kľúčové slová: Test, oblasť spoľahlivosti, silofunkcia.

Abstrakt: V príspevku sa zaoberáme simultánnym testovaním strednej hodnoty a variancie jednorozmerného normálneho rozdelenia. Navrhnutý je test $H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ proti $H_1 : (\mu, \sigma^2) \neq (\mu_0, \sigma_0^2)$. Silu nového testu sme porovnávali so silami známych presných testov, konkrétne s Moodovym testom (Mood, 1950) a s testom pomerom vierohodnosti (Choudhari–Kundu–Misra, 2001). Porovnané sú aj veľkosti oblastí spoľahlivosti zároveň pre oba parametre prisluchajúce k uvažovaným testom.

Abstract: We deal with the simultaneous testing of the mean and the variance of a univariate normal distribution. The test of null hypothesis $H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ against the alternative $H_1 : (\mu, \sigma^2) \neq (\mu_0, \sigma_0^2)$ is suggested. The power function of the presented test is compared with the power functions of the known exact tests, concrete with the Mood test (Mood, 1950) and with the likelihood ratio test (Choudhari–Kundu–Misra, 2001). We also determined the sizes of the confidence regions corresponding to the tests.

1. Úvod

Simultánne testovanie parametrov normálneho rozdelenia $N(\mu, \sigma^2)$ je prirodzenou nadstavbou testov pre parametre normálneho rozdelenia nachádzajúcich sa v každej základnej knihe o štatistike. Nech Θ označuje parametrický priestor, teda $\Theta = \{\theta = (\mu, \sigma^2) : \ -\infty < \mu < \infty, \sigma^2 > 0\}$. Úloha testovať

$$H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2) \in \Theta_0 \quad \text{proti} \quad H_1 : (\mu, \sigma^2) \in \Theta - \Theta_0 = \Theta_1,$$

nie je nová, už však Kendall a Sturat (1961) ukázali, že neexistuje rovnomerne najsilnejší test.

Štatistické testovanie parametrov je silne prepojené s intervalovými odhadmi a dá sa povedať, že každému testu odpovedá oblasť spoľahlivosti a naopak (Casella a Berger, 1990). Nech α, $\alpha \in (0, 1)$ označuje hladinu významnosti testu. Pod presnou $100(1-\alpha)\%$-nou oblasťou spoľahlivosti zároveň pre oba parametre normálneho rozdelenia rozumieme takú množinu $\mathcal{R}$, pre ktorú platí $P((\mu, \sigma^2) \in \mathcal{R}) = 1 - \alpha$. Presnú $100(1-\alpha)\%$-nú oblasť spoľahlivosti ako prvý skonštruoval Mood (1950). V literatúre možno nájsť množstvo oblastí spoľahlivosti, avšak s približne $1 - \alpha$ spoľahlivosťou (pozri napr. Meeker a Escobar 1995, Arnold a Shavelle 1998).

Presným testom (veľkosti α) budeme rozumieť test, pre ktorý platí rovnosť $P(H_0 \text{ nezamietame} | (\mu, \sigma^2) = (\mu_0, \sigma_0^2)) = 1 - \alpha$. Navrhli sme test $H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ proti $H_1 : (\mu, \sigma^2) \neq (\mu_0, \sigma_0^2)$ a porovnali ho so známymi

presnými testmi, konkrétne s testom pomerom vierohodnosti (Choudhari–Kundu–Misra, 2001) a s Moodovym testom (Mood, 1950) prisluchajúcim k Moodovej presne oblasti spoľahlivosti pre parametre (μ, σ^2). Zaujímala nás sila jednotlivých testov a veľkosti oblastí spoľahlivosti odpovedajúcich k testom.

2. **Testy, silofunkcie, oblasti spoľahlivosti**

Majme náhodný výber $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ rozsahu n z normálneho rozdelenia s neznámou strednou hodnotou μ a s neznámou varianciou σ^2. Označme $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ výberový priemer a $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ výberový rozptyl. Náhodné premenné $\bar{X}$, S^2 sú nezávislé a nech $v = n - 1$, potom platí

$$Y_1 = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \sim N(0,1), \quad Y_2 = \frac{vS^2}{\sigma^2} \sim \chi^2_v, \tag{1}$$

kde χ^2_v označuje chi-kvadrát rozdelenie s v stupňami voľnosti. Pri zostavení vhodného testu sa stačí obmedziť na využitie postačujúcich štatistík. Vieme, že $(\bar{X}, S^2)$ je postačujúca štatistika pre (μ, σ^2). Navyše rozdelenie náhodných premenných Y_1, Y_2 nezávisí od neznámych parametrov, čo využijeme pri stanovení presných kritických hodnôt v odvodených testoch.

Uvažujme teraz asymptotické rozdelenie $\sqrt{2\chi^2_v} \approx N(\sqrt{2v-1}, 1)$ (Fisher, 1928) a nech $k = \sqrt{2v-1}$. Pre navrhnutú testovaciu štatistiku H_0 proti H_1 v tvare

$$G = \frac{n(\bar{X} - \mu_0)^2}{\sigma_0^2} + \left(\sqrt{\frac{2vS^2}{\sigma_0^2}} - k\right)^2 \tag{2}$$

za platnosti nulovej hypotézy platí $G \approx \chi^2_2$. Teda test zamietnúť H_0 ak $G > \chi^2_2(1-\alpha)$ a ináč nezamietnúť proti alternatíve H_1 je len približne veľkosti α. Invertovaním presnej distribučnej funkcie náhodnej premennej G môžeme však vypočítať presné hodnoty kvantilov, ozn. $g_n(1-\alpha)$, pre ktoré platí

$$\begin{aligned} 1 - \alpha &= P(G \le g_n(1-\alpha)) = \\ &= P\left(Y_1^2 + \left(\sqrt{2Y_2} - k\right)^2 \le g_n(1-\alpha)\right) = \\ &= \int_0^\infty P_{Y_1^2}\left(g_n(1-\alpha) - \left(\sqrt{2x} - k\right)^2\right) f_{Y_2}(x)dx, \end{aligned}$$

kde $Y_1^2 \sim \chi^2_1$, $P_{Y_1^2}(\cdot)$ označuje distribučnú Y_1^2 a $f_{Y_2}(\cdot)$ označuje hustotu náhodnej premennej Y_2. Oblasť spoľahlivosti pre (μ, σ^2) skonštruovaná na základe presného testu pomocou štatistiky G má tvar

$$\mathcal{R} = \left\{(\mu, \sigma^2) : \frac{n(\bar{X} - \mu)^2}{\sigma^2} + \left(\sqrt{\frac{2vS^2}{\sigma^2}} - k\right)^2 \le g_n(1-\alpha)\right\}.$$

Obsah oblasti $\mathcal{R}$ je daný

$$\text{obsah } = \left[2 \int_{m_1}^{m_2} \sqrt{\frac{m g_n(1-\alpha) - (\sqrt{2v} - k\sqrt{m})^2}{n}} dm \right] \times S^3,$$

$$\text{kde } \; m_1, m_2 = \frac{2v}{(k \pm \sqrt{g_n(1-\alpha)})^2}.$$

Nech $(\mu_1, \sigma_1^2) \in \Theta$, silofunkcia navrhnutého testu je vyjadrená vzťahom

$$\begin{aligned}
\beta(\mu_1, \sigma_1^2) &= \\
&= 1 - P\left(\frac{n(\bar{X} \pm \mu_1 - \mu_0)^2}{\sigma_1^2} \frac{\sigma_1^2}{\sigma_0^2} + \left(\sqrt{\frac{2vS^2}{\sigma_1^2} \frac{\sigma_1^2}{\sigma_0^2}} - k \right)^2 \le g_n(1-\alpha) \right) = \\
&= 1 - \int_0^\infty P_{Y_3}\left(g_n(1-\alpha) \frac{\sigma_0^2}{\sigma_1^2} - \left(\sqrt{2x} - k \frac{\sigma_0}{\sigma_1} \right)^2 \right) f_{Y_2}(x) dx,
\end{aligned}$$

kde $Y_3 = n(\bar{X} - \mu_0)/\sigma_1^2$ má chi-kvadrát rozdelenie s 1 stupňom voľnosti a s parametrom necentrality $n(\mu_1 - \mu_0)^2/\sigma_1^2$.

Moodov presný test na hladine významnosti α nezamietne H_0 proti H_1, ak

$$\frac{|\bar{X} - \mu_0|}{\sigma_0} \sqrt{n} \le u\left(1 - \frac{\alpha_1}{2}\right) \quad \wedge \quad \chi_v^2(\delta) \le \frac{vS^2}{\sigma_0^2} \le \chi_v^2(1 - \alpha_2 + \delta),$$

kde $u(1-\alpha_1/2)$ je $(1-\alpha_1/2)$-kvantil štandardného normálneho rozdelenia, $\chi_v^2(\delta)$ je δ-kvantil chi-kvadrát rozdelenia s v stupňami voľnosti a platí $1 - \alpha = (1-\alpha_1)(1-\alpha_2)$, $\delta \ge 0$. Z Bonferroni nerovnosti vyplýva, že oblasť spoľahlivosti pre (μ, σ^2) skonštruovaná množinovým súčin $100(1-\alpha_1)\%$-ného intervalu spoľahlivosti pre strednú hodnotu a $100(1-\alpha_2)\%$-ného intervalu spoľahlivosti pre varianciu je presne $1-\alpha$ spoľahlivá. V najjednoduchšom prípade môžeme predpokladať $1 - \alpha_1 = 1 - \alpha_2 = \sqrt{1-\alpha}$. Hodnoty α_1, α_2 však môžeme voliť s ohľadom na veľkosť prisluchajúcej oblasti spoľahlivosti pre (μ, σ^2). Pre rôzne veľkosti výberov n, hladiny významnosti α sú uvedené optimálne hodnoty $\alpha_1, \alpha_2, \delta$ na dosiahnutie minimálnej plochy oblasti spoľahlivosti ako aj vzťah na výpočet obsahu uvedené v Arnold a Shavelle (1998). Tvar silofunkcie je odvodený v Choudhari–Kundu–Misra (2001).

Presný test pomerom vierohodnosti (LRT) nezamietne H_0 proti H_1 na hladine významnosti α, ak platí

$$\lambda = \frac{n(\bar{X} - \mu_0)^2}{\sigma_0^2} + \frac{vS^2}{\sigma_0^2} - n \ln\left(\frac{vS^2}{\sigma_0^2} \right) \le d_n(1-\alpha), \tag{3}$$

kde $d_n(1-\alpha)$ je $(1-\alpha)$-kvantil rozdelenia testovacej štatistiky λ. V Choudhari–Kundu–Misra (2001) je prezentovaný vzťah pre silofunkciu testu a pre vybrané hodnoty n, $1-\alpha$ sú kvantily vypočítané invertovaním distribučnej

funkcie λ uvedené v priložených tabuľkách. Obsah oblasti spoľahlivosti prisluchajúcej k LRT je daný vzťahom

$$\text{obsah } = \left[2\int\sqrt{\frac{yd_n(1-\alpha)}{n}+y\ln\left(\frac{v}{y}\right)-1}dy\right]\times S^3,$$

kde hranice integrovania určené definičným oborom integrovaného výrazu sú numericky počítané.

V literatúre sa uvádza aj Wilksov test (Wilks, 1962), ktorý na hladine významnosti α nezamieta H_0 ak

$$\frac{n(\bar{X}-\mu_0)^2}{\sigma_0^2}+\frac{vS^2}{\sigma_0^2}\leq\chi_n^2(1-\alpha). \tag{4}$$

Tento test však nie je vhodné použiť pri alterantíve H_1, ale pri $H_1^*:\mu\neq\mu_0\ \vee\ \sigma^2>\sigma_0^2$ už áno. Silofunkcia testu je daná vzťahom

$$\beta(\mu_1,\sigma_1^2)=1-P\left(\frac{n(\bar{X}\pm\mu_1-\mu_0)^2}{\sigma_1^2}\frac{\sigma_1^2}{\sigma_0^2}+\frac{vS^2}{\sigma_1^2}\frac{\sigma_1^2}{\sigma_0^2}\leq\chi_n^2(1-\alpha)\right)=$$
$$=1-\int_0^\infty P_{Y_3}\left(\chi_n^2(1-\alpha)\frac{\sigma_0^2}{\sigma_1^2}-x\right)f_{Y_2}(x)dx.$$

Silu Wilksovho testu porovnáme v následujúcej časti so silou upraveného Moodovho test, t.j. modifikovaný Moodov test na hladine významnosti α nezamietne H_0 proti H_1^*, ak platí

$$\frac{|\bar{X}-\mu_0|}{\sigma_0}\sqrt{n}\leq u\left(1-\frac{\alpha_1}{2}\right)\quad\wedge\quad\frac{vS^2}{\sigma_0^2}\leq\chi_v^2(1-\alpha_2).$$

Odvodenie výpočtu silofunkcie testu je analogické k odvodeniu výpočtov pre pôvodný tvar Moodovho testu, preto ho tu nebudeme uvádzať. Obsahy oblastí spoľahlivosti prisluchajúce k Wilksovmu a modifikovanému Moodovmu testu vzhľadom na neohraničenú množinu hodnôt σ^2 nebudeme porovnávať.

Poznamenajme, že test pomerom vierohodnosti obsahuje v sebe Wilksov test a teda zrejme testovacia štatistika v tvare $L=\frac{n(\bar{X}-\mu_0)^2}{\sigma_0^2}-n\ln\left(\frac{vS^2}{\sigma_0^2}\right)$ je vhodná na testovanie H_0 proti $H_1^{**}:\mu\neq\mu_0\ \vee\ \sigma^2<\sigma_0^2$, pričom presné kvantily pre uvedený test môžu byť dopočítané opäť invertovaním distribučnej funkcie L. Navrhnutý test H_0 proti H_1^{**} však nebudeme numericky analyzovať. V následujúcej časti porovnáme presné testy H_0 proti H_1 a presné testy H_0 proti H_1^* veľkosti α.

3. Numerické výsledky

V tejto časti sa zameriame na porovnanie síl testov a oblastí spoľahlivostí odpovedajúcich k uvedeným presným testom. Vzájomne najprv porovnáme testy pre $H_0:\mu=\mu_0\wedge\sigma^2=\sigma_0^2$ proti $H_1:\mu\neq\mu_0\vee\sigma^2\neq\sigma_0^2$ a to navrhnutý test (nový), Moodov test (Mood) a test pomerom vierohodnosti (LRT). Pre testovanie $H_0:\mu=\mu_0\ \wedge\ \sigma^2=\sigma_0^2$ proti $H_1^*:\mu\neq\mu_0\ \vee\ \sigma^2>\sigma_0^2$ budeme

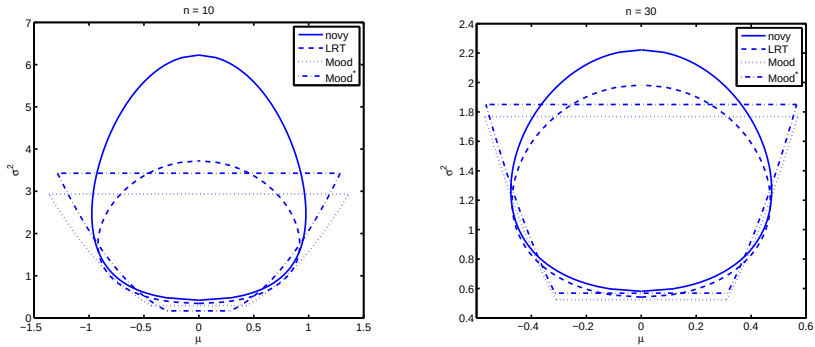
$\sigma_1^2 \downarrow$	$\mu_1 \rightarrow$	0.0	0.1	0.2	0.3	0.4
	nový	*0.0301*	0.0368	0.0584	0.0992	0.1642
0.8	LRT	0.0642	0.0704	0.0903	**0.1273**	0.1856
	Mood	**0.0824**	**0.0849**	**0.0942**	0.1144	0.1521
	Mood*	0.0613	0.0671	0.0868	0.1255	**0.1893**
		0.0365	0.0437	0.0667	0.1088	0.1739
0.9		0.0533	0.0596	0.0795	0.1163	0.1734
		0.0616	**0.0650**	0.0766	0.1005	0.1422
		0.0524	0.0593	**0.0817**	**0.1238**	**0.1900**
		0.0500	**0.0578**	**0.0824**	0.1261	0.1915
1.0		0.0500	0.0565	0.0771	0.1145	0.1715
		0.0500	0.0542	0.0679	0.0949	0.1396
		0.0500	0.0578	0.0823	**0.1269**	**0.1943**
		0.0708	**0.0792**	**0.1052**	**0.1502**	**0.2157**
1.1		0.0531	0.0601	0.0817	0.1201	0.1775
		0.0447	0.0495	0.0651	0.0945	0.1415
		0.0517	0.0601	0.0863	0.1324	0.2005
		0.0987	**0.1075**	**0.1343**	**0.1801**	**0.2451**
1.2		0.0624	0.0697	0.0924	0.1319	0.1898
		0.0440	0.0494	0.0665	0.0979	0.1465
		0.0561	0.0650	0.0923	0.1394	0.2075

TABUĽKA 1. Sily presných testov H_0 proti alternatíve H_1 pre $n = 10$ a $\alpha = 0.05$.

porovnávať modifikovaný Moodov test (mMood) a Wilksov test (Wilks). Silofunkcie testov závisia na hodnotách $(\mu_0 - \mu_1)^2$ a σ_1^2/σ_0^2. Bez straty na všeobecnosti vo výpočtoch predpokladáme $\mu_0 = 0$, $\sigma_0^2 = 1$. V tabuľkách 1-2 sú hodnoty sily testov pre kombinácie hodnôt $\mu_1 = \{0, 0.1, 0.2, 0.3, 0.4\}$, $\sigma_1^2 = \{0.8, 0.9, 1, 1.1, 1.2\}$ pri zvolenom $\alpha = 0.05$ a $n = \{10, 30\}$. Pre každú dvojicu (μ_1, σ_1^2) je uvedená štvorica čísel, pričom prvé číslo je sila navrhnutého testu, druhé číslo je sila testu pomeru vierohodnosti, tretie číslo je sila Moodovho testu. Hodnoty $\alpha_1, \alpha_2, \delta$ boli zvolené tak, aby Moodova oblasť spoľahlivosti mala minimálny obsah. Konkrétne $\alpha_1 = 0.0117$, $\alpha_2 = 0.0388$, $\delta = 0.0384$ pre $n = 10$, $\alpha_1 = 0.0190$, $\alpha_2 = 0.0316$, $\delta = 0.0293$ pre $n = 30$. V stĺpci $\mu_1 = 0$ sú odlišené hodnoty, kde je sila navrhnutého testu a Moodovho testu menšia ako α. Dopočítali sme optimálne hodnoty $\alpha_1, \alpha_2, \delta$ tak, aby silofunkcia pre Moodov test neklesla pod zvolenú hodnotu α, pričom bola rovná práve α jedine v $\mu_1 = 0$ pre $\sigma_1^2 = 1$. Konkrétne $\alpha_1 = 0.0279$, $\alpha_2 = 0.0227$, $\delta = 0.0227$ pre $n = 10$, $\alpha_1 = 0.0227$, $\alpha_2 = 0.0279$, $\delta = 0.0209$ pre $n = 30$. Silofunkcia Moodovho testu s takto stanovenými hodnotami $\alpha_1, \alpha_2, \delta$ je označená ako Mood* a je udávaná ako štvrtá hodnota. Zvýraznená je najväčšia sila pre každú kombináciu (μ_1, σ_1^2). Vidíme, že pre hodnoty $\sigma_1^2 = \{1.0, 1.1, 1.2\}$ dosahuje navrhnutý test najvyššie hodnoty avšak pre $\sigma_1^2 = \{0.8, 0.9\}$ pre

$\sigma_1^2 \downarrow$	$\mu_1 \rightarrow$	0.0	0.1	0.2	0.3	0.4
	nový	0.0616	0.0852	0.1662	0.3194	0.5291
0.8	LRT	0.1007	0.1248	**0.2052**	**0.3529**	**0.5518**
	Mood	**0.1310**	**0.1433**	0.1945	0.3129	0.5016
	Mood*	0.1047	0.1196	0.1791	0.3110	0.5121
		0.0441	0.0666	0.1437	0.2899	0.4931
0.9		0.0613	0.0834	**0.1586**	**0.3001**	**0.4969**
		0.0734	**0.0892**	0.1491	0.2765	0.4697
		0.0617	0.0800	0.1476	0.2861	0.4880
		0.0500	**0.0735**	**0.1516**	**0.2945**	**0.4890**
1.0		0.0500	0.0722	0.1467	0.2848	0.4759
		0.0500	0.0684	0.1342	0.2653	0.4564
		0.0500	0.0708	0.1434	0.2832	0.4800
		0.0783	**0.1037**	**0.1846**	**0.3248**	**0.5085**
1.1		0.0605	0.0842	0.1609	0.2974	0.4809
		0.0475	0.0679	0.1373	0.2687	0.4542
		0.0603	0.0827	0.1576	0.2953	0.4835
		0.1295	**0.1564**	**0.2386**	**0.3741**	**0.5444**
1.2		0.0920	0.1174	0.1966	0.3309	0.5049
		0.0614	0.0830	0.1541	0.2832	0.4607
		0.0907	0.1138	0.1885	0.3208	0.4975

TABUĽKA 2. Sily presných testov H_0 proti alternatíve H_1 pre $n = 30$ a $\alpha = 0.05$.



OBR. 1. Hranice oblastí spoľahlivosti pre (μ, σ^2) prisluchajúce k presným testom H_0 proti H_1 vykreslené pre $\alpha = 0.05$ a $n = \{10, 30\}$.

niektoré μ_1 je sila testu menšia ako zvolená α a podobný jav pozorujeme aj pre Moodov test. Pre kombinácie $(0, 0.8), (0.1, 0.8), (0.0, 0.9), (0.1, 0.9)$ je najsilnejší Moodov test avšak pri kombinácii $\mu_1 = \{0.0, 0.1\}$ s hodnotami

	$\alpha = 0.10$		$\alpha = 0.05$		$\alpha = 0.01$	
	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 10$	$n = 30$
nový	5.1218	0.8801	8.8844	1.2236	32.5699	2.2175
LRT	**3.2079**	**0.7819**	**4.8910**	**1.0668**	**10.9590**	**1.8329**
Mood	3.2702	0.8087	5.1747	1.1256	13.4743	2.2458
Mood*	3.8056	0.8332	5.8177	1.1587	14.2137	2.2778

TABUĽKA 3. Obsahy oblastí spoľahlivosti pre (μ, σ^2) prislúchajúce k presným testom H_0 proti H_1 pre $\alpha = \{0.10, 0.05, 0.01\}$ a $n = \{10, 30\}$.

$\sigma_1^2 = \{1.1, 1.2\}$ je sila testu nižšia ako $\alpha = 0.05$. Silnejší spomedzi testov, pre ktoré sila neklesla pod hodnotu α je test pomerom vierohodnosti. Na obr. 1 sú vykreslené hranice oblastí spoľahlivosti prislúchajúce k uvažovaným testom H_0 proti H_1 pre zvolené $\alpha = 0.05$ a $n = \{10, 30\}$. V tabuľke 3 sú uvedené obsahy jednotlivých oblastí pre kombinácie vybraných hodnôt $n = \{10, 30\}$ a $\alpha = \{0.1, 0.05, 0.01\}$. Pre každú kombináciu je zvýraznená najnižšia hodnota. Vzhľadom na voľbu $\alpha_1, \alpha_2, \delta$ je zrejme, že obsah Moodovej oblasti je vždy menší ako oblasť skonštruovaná Mood* testom. Pre všetky kombinácie má oblasť skonštruovaná LRT testom najnižší obsah. Vzhľadom na numerické porovonanie presných testov odporúčame na testovanie H_0 proti H_1 použiť test pomerom vierohodnosti. V tabuľke 4-5

$\sigma_1^2 \downarrow$	$\mu_1 \rightarrow$	0.0	0.2	0.4	0.6	0.8
	Wilks	0.0000	0.0000	0.0000	0.0000	0.0005
0.2	mMood	0.0000	**0.0002**	**0.0185**	**0.2507**	**0.7709**
		0.0000	0.0000	0.0001	0.0026	0.0281
0.4		**0.0005**	**0.0067**	**0.0700**	**0.3172**	**0.7001**
		0.0007	0.0014	0.0062	0.0295	0.1133
0.6		**0.0047**	**0.0219**	**0.1143**	**0.3491**	**0.6658**
		0.0112	0.0161	0.0374	0.0973	0.2280
0.8		**0.0180**	**0.0447**	**0.1519**	**0.3709**	**0.6461**
		0.0500	0.0619	0.1045	0.1946	0.3453
1.0		0.0500	**0.0823**	**0.1943**	**0.3959**	**0.6384**
		0.1230	0.1408	0.1983	0.3027	0.4523
1.2		0.1069	**0.1411**	**0.2501**	**0.4315**	**0.6438**

TABUĽKA 4. Sily presných testov H_0 proti alternatíve H_1^* pre $n = 10$ a $\alpha = 0.05$.

sú sily Wilksovho testu a modifikovaného Moodovho testu pre kombinácie hodnôt $\mu_1 = \{0.0, 0.2, 0.4, 0.6, 0.8\}$ a $\sigma_1^2 = \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$. Pre

$\sigma_1^2 \downarrow$	$\mu_1 \rightarrow$	0.0	0.2	0.4	0.6	0.8
	Wilks	0.0000	0.0000	0.0000	0.0000	0.0001
0.2	mMood	0.0000	**0.0041**	**0.4223**	**0.9879**	**1.0000**
		0.0000	0.0000	0.0000	0.0004	0.0324
0.4		**0.0003**	**0.0307**	**0.4449**	**0.9445**	**0.9996**
		0.0000	0.0001	0.0014	0.0245	0.2049
0.6		**0.0033**	**0.0633**	**0.4549**	**0.9034**	**0.9967**
		0.0038	0.0073	0.0320	0.1467	0.4495
0.8		**0.0125**	**0.0945**	**0.4618**	**0.8703**	**0.9907**
		0.0500	0.0702	0.1549	0.3576	0.6577
1.0		0.0500	**0.1434**	**0.4800**	**0.8476**	**0.9828**
		0.1929	0.2318	0.3588	0.5719	0.7994
1.2		0.1620	**0.2522**	**0.5369**	**0.8443**	**0.9761**

TABUĽKA 5. Sily presných testov H_0 proti alternatíve H_1^* pre $n = 30$ a $\alpha = 0.05$.

mMood boli použité hodnoty $\alpha_1, \alpha_2, \delta$ rovnaké ako pre Mood*. Pre každú kombináciu (μ_1, σ_1^2) je zvýraznená najväčšia sila. Vzhľadom na výsledné sily testov odporúčame na testovanie H_0 proti H_1^* upravený tvar Moodovho testu.

Literatúra

[1] Arnold, B.C. - Shavelle, R. M. (1998) *Joint confidence sets for the mean and variance of a normal distribution.* The American Statistician **52**, 133 – 140.

[2] Casella, G. - Berger, R. L. (1990) Statistical Inference, *Duxbury Press, California*

[3] Choudhari, P. - Kundu, D. - Misra, N. (2001) *Likelihood ratio test for simultaneous testing of the mean and variance of a normal distribution.* Journal of Statistical Computation and Simulation **71**, 313 – 333.

[4] Fisher, R. A. (1928) *Statistical Methods for Research Workers*, 2nd Edition, 96 – 97.

[5] Kendall, M.G. - Stuart, A. (1961) The Advanced Theory of Statistics, Vol.2, *Charles Griffin.*

[6] Mood, A. M. (1950) Introduction to the Theory of Statistics, *New York, McGraw-Hill.*

[7] Wilks, S. S. (1962) Mathematical Statistics, *New York, John Wiley and Sons.*

Poďakovanie: Práca vznikla vďaka podpore grantov APVV-0096-10, VEGA 2/0019/10 a VEGA 2/0038/12.

Adresa: Ústav merania SAV, Dúbravská cesta 9, 841 04 Bratislava 4

E-mail: chvosta@gmail.com

ANALÝZA ČASOVÝCH ŘAD FORMÁLNÍ KOMUNIKACE OBCÍ

Radka Lechnerová[1], **Tomáš Lechner**[2]

Klíčová slova: e-Government, časové řady, komunikace

Abstract: Implementation of information and communication technologies in Public Administration called e-Government aims at increasing effectiveness of Public Administration and reducing public budgetary expenditure. One of the implemented e-Government tools is the record management system, which has been used by some municipalities since 2004. These systems provide statistical data, which describe formal communication between municipalities and other subjects. The paper deals with the basic analysis of three time series samples for incoming and outgoing communication separately. The data come from municipalities with extended powers that are authorities with the widest range of delegated power at the level of local government and thus with the widest spectrum of providing agendas through both local and delegated power. In the paper we analyze the time evolution of the overall communications and by using the chosen model we isolate trend and seasonal components.

Abstrakt: Implementace informačních a komunikačních technologií ve veřejné správě označovaná jako e-Government má za cíl zvyšování výkonu veřejné správy a snižování výdajů veřejných rozpočtů. Jedním ze zaváděných nástrojů jsou elektronické systémy spisové služby, které některé obce vyžívají již od roku 2004. Tyto systémy poskytují statistická data informující o formální komunikaci obcí s okolím. V rámci příspěvku se zabýváme základní analýzou tří vzorků časových řad příchozí a odchozí komunikace obcí s rozšířenou působností, které na úrovni územní samosprávy představují orgány s nejširším spektrem výkonu agend v přenesené i místní působnosti. Primárně analyzujeme časový vývoj celkové komunikace a pomocí zvoleného modelu izolujeme trend a sezónní složku.

1. Úvod

Veřejná správa v evropských zemích funguje na základě zákona a v jeho mezích. Je to zakotveno v ústavách jednotlivých evropských států. I veřejná správa, jakkoliv je vystavena na formalizovaných a víceméně neměnných pravidlech, musí reflektovat rozvoj informační společnosti, který probíhá již několik desetiletí. Využití výpočetní techniky ve veřejné správě, označované jako e-Government, má mnoho cílů, zejména směřovaných ke zvyšování efektivity výkonu veřejné správy a snižování výdajů veřejných rozpočtů na správu jako takovou. S ohledem na to mnoho evropských států, včetně České republiky, zařadilo e-Government do svého balíčku reakcí na ekonomickou krizi [7]. Odhlédneme-li od zmíněných cílů, má zavádění informačních a komunikačních

technologií ještě jeden výrazný důsledek. Veřejná správa začíná produkovat dále statisticky zpracovatelná data o své vlastní činnosti.

Přestože orgány veřejné moci smí vykonávat pouze to, co jim ukládá zákon, neexistuje v České republice celkový přehled o všech agendách a činnostech, které jsou na jednotlivých úrovních veřejné správy a jednotlivými orgány vykonávány. Tzv. mapa veřejné správy se buduje teprve nyní, a to v rámci zavádění základních registrů veřejné správy, jako jednoho ze stěžejních pilířů celého e-Governmentu [6].

Data, která jsou již zmíněným vedlejším produktem e-Governmentu, vycházejí z elektronizace jednotlivých procesů a jsou proto poměrně přesnými podklady o skutečném fungování veřejné správy. Samozřejmě, že před zveřejněním nebo poskytnutím ke zpracování musí být tato data s ohledem na ochranu osobnosti vycházející z občanského zákoníku a ochranou osobních, obzvláště pak citlivých, údajů podloženou zákonem o ochraně osobních údajů patřičným způsobem anonymizována. Na druhou stranu jsou-li produktem veřejné správy, jsou veřejně dostupná. Je např. veřejně známo, že datovou schránku mělo k 1. září 2012 zřízeno 24 534 fyzických osob[1], ale již nemůže být automaticky zveřejněno jméno každého občana, který si o zřízení datové schránky požádal. Uvedená data jsou tedy vhodná ke statistickému zpracování.

Velmi významným zdrojem statistických dat vypovídajících o činnosti jednotlivých institucí veřejné správy jsou elektronické systémy spisové služby. Povinnost jejich plného provozu mají sice jednotlivé úřady až od 1. července 2012, nicméně z hlediska příslušného zákona o archivnictví a spisové službě je elektronický způsob vedení spisové služby preferovaným způsobem od 1. července 2009 a dalšími právními předpisy umožněný (v rámci diskreční pravomoci) minimálně od roku 2000. Jako jedny z prvních implementovaly tento nástroj orgány územních samosprávných celků. Díky tomu máme v současné době k dispozici již relativně dlouhé časové řady dat, které vypovídají zejména o komunikaci jednotlivých orgánů s okolím. Domníváme se dále, že zmapování tohoto vývoje lze využít jednak pro porozumění jednotlivým vlivům a trendům, a jednak v rámci zpětné vazby ovlivňující další proces reformy veřejné správy, a to nejen prostředky informačních a komunikačních technologií.

Zkoumáním statistických dat poskytovaných elektronickými systémy spisových služeb implementovaných orgány územních samosprávných celků jsme se již zabývali v roce 2009 [3], avšak nešlo o časové vývojové řady, ale o statistické stanovení rozložení způsobů komunikace v období těsně před zavedením informačního systému datových schránek, které dalo poměrně jasné predikce následných úspor a umožnilo též zhodnocení hospodárnosti zavedení tohoto významného nástroje českého e-Governmentu [4]. Dále jsme již publikovali případovou studii vývoje komunikace jedné obce ukazující vztah způsobů komunikace a modelů financování, přičemž jsme v uvedené případové

[1]Zdroj: <http://www.datoveschranky.info>. Citace 1. září 2012.

studii došli k závěru, že finanční motivace není pro občany rozhodujícím kritériem volby komunikačního kanálu [5]. V tomto příspěvku analyzujeme vývoj celkové komunikace u tří vybraných obcí.

2. Data

Elektronické spisové služby evidují veškerou formální komunikaci orgánů veřejné moci s okolím. Ta se skládá ze dvou směrů; komunikace příchozí či vstupní a komunikace odchozí či výstupní. Vstupní komunikace je po právní stránce určena podmínkami pro podání vůči orgánům veřejné moci. Výstupní komunikace je určována pravidly pro doručování orgány veřejné moci. Pro oba směry je využíváno stejných komunikačních kanálů (způsobů komunikace), ale příslušná legislativní pravidla jsou rozdílná. Způsoby podání volí ten, kdo podání činí zcela svobodně z určené nabídky možností, zatím co při doručování je volba příslušného kanálu určena buď předchozí svobodnou volbou soukromoprávního subjektu, nebo poměrně přesně danými pravidly. Proto je vhodné oba směry komunikace zkoumat odděleně.

O každém podání a o každém doručení se zaznamenávají právními předpisy dané údaje, a proto jsou výsledná data poskytovaná jednotlivými orgány veřejné moci v obdobné struktuře, přestože jednotlivé orgány si mohou v rámci svých pravomocí vybrat dodavatele konkrétního softwarového řešení[2]. Data z těchto systémů nejsou poskytována automaticky, ale pouze na vyžádání v rámci žádostí podle zákona o svobodném přístupu k informacím. Z tohoto důvodu bylo v úvodní části výzkumu zvoleno pouze několik orgánů, konkrétně obcí s rozšířenou působností, které na úrovni územní samosprávy představují orgány s nejširším spektrem výkonu agend v přenesené i místní působnosti. Zvolili jsme tři zástupce jednotlivých velikostních kategorií v této oblasti[3], konkrétně *obec A*, která má přibližně 5 tis. obyvatel a získaná časová řada má rozsah od 1. 12. 2004 do 29. 4. 2011. *Obec B* má přibližně 14 tis. obyvatel a její časová řada je nejdelší a zachycuje data od 1. 1. 2004 do 28. 2. 2012. *Obec C* má necelých 24 tis. obyvatel a časová řada má rozsah od 1. 2. 2007 do 31. 3. 2011.

Výsledná data jsou k dispozici v následující struktuře údajů: *datum, způsob komunikace*; zvlášť pro příchozí a odchozí směr komunikace. Způsob komunikace není z hlediska zapisovaných údajů plošně standardizovaně kódovaným údajem, a proto je třeba jej před statistickým zpracováním upravit příslušným procesem překódování.

3. Metody

Z charakteru zkoumaných dat lze předpoklad jejich nezávislost, a to v každé řadě, tj. zvlášť pro příchozí, zvlášť pro odchozí komunikaci. Mezi vstupní

[2]U větších zakázek je tento výběr samozřejmě realizován prostřednictvím výběrových řízení.

[3]Počty obyvatel vycházejí ze zdroje [2] a jsou platné k 31.12.2010.

a výstupní komunikací již lze očekávat vzájemnou závislost, avšak ta není předmětem této analýzy.

Při zkoumání časových řad komunikace, v nichž jsou časové okamžiky událostí dány datem, se objevuje problém diskontinuit daných víkendy, státními svátky a jinými dny, kdy úřady ani nepřijímají zásilky, ani je neposílají. Zavedení elektronické komunikace sice otevřelo elektronické podatelny pro příjem podání v režimu 7×24, avšak u orgánů územních samosprávných celků se možnosti učinit takto podání v dny, kdy úřad není otevřen, využívají jen velmi vzácně, tj. jde o jednotky podání ročně ve srovnání s desítkami až stovkami podání během pracovních dní. Proto jsme souhrnná data za jednotlivé měsíce znormovali na 20 pracovních dní vztahem

$$y_t = \frac{20}{m_t} \sum_d q_d, \tag{1}$$

kde q_d je počet podání ve dni d, m_t je skutečný počet pracovních dní měsíce t.

Pro analýzu takto získaných časových řad Y_t^D resp. Y_t^O představující celkový objem příchozí resp. odchozí komunikace jsme použili aditivní dekompozici (viz např. [1]), na základě které můžeme získat jednotlivé složky časové řady. Konkrétně

$$Y_t = T_t + Sz_t + E_t, \tag{2}$$

kde T_t je trend, Sz_t je sezónní složka a E_t je reziduální složka. Reziduální složka je představována bílým šumem.

Pro vyjádření trendu jsme použili metodu centrovaných klouzavých průměrů se stejnými váhami (s ročním vyhlazením), tedy

$$T_t = \frac{1}{2m} \left(y_{t-p} + 2 \sum_{i=t-p+1}^{t+p-1} y_i + y_{t+p} \right), \text{ pro } m = 2p.$$

Následně, po trendovém očištění, tj. $a_t = y_t - T_t$, jsme sezónní složku (s roční periodou) spočetli průměrováním pro každý časový okamžik přes celou časovou řadu, tj.

$$I_j^* = \frac{1}{r} \sum_{i=1} r a_{(i-1)m+j}, \quad j = 1, ..., m,$$

kde r je počet let. Následně jsme získané hodnoty centrovali odečtením jejich aritmetického průměru, tj.

$$S_t = I_j^* - \frac{1}{m} \sum_{i=1}^{m} I_i^*,$$

kde kde t odpovídá j-tému měsíci v roce.

Reziduální složku jsme testovali, zda skutečně představuje bílý šum, tj. zda E_t jsou nezávislé, stejně rozdělené náhodné veličiny, a to pomocí testu založeného na znaménkách diferencí, viz [1] str. 322, a dále na nulovou střední

	rok	2004	2005	2006	2007	2008	2009	2010	2011
A	medián		1245	1960	1850	1689	1594	1561	
	průměr		1331	1991	2181	1730	1644	1579	
	směr. odchylka		214	197	1376	227	159	184	
B	medián	1736	1965	2251	2478	2523	2369	2444	2523
	průměr	1659	1980	2218	2422	2542	2403	2398	2618
	směr. odchylka	312	210	249	318	295	272	253	260
C	medián				5682	4990	4658	4659	
	průměr				5578	5126	4670	4637	
	směr. odchylka				443	566	363	520	

TABULKA 1. Základní charakteristiky měsíčních souhrnů pro příchozí komunikaci.

	rok	2004	2005	2006	2007	2008	2009	2010	2011
A	medián		1979	2086	2293	2067	1886	2015	
	průměr		2077	2134	2587	2146	1845	1944	
	směr. odchylka		347	257	1050	379	224	226	
B	medián	3225	3788	3502	2999	2892	2817	2951	3077
	průměr	3269	3815	3797	3058	2832	2875	3007	3043
	směr. odchylka	385	660	649	681	489	290	360	358
C	medián				4840	4488	3410	4507	
	průměr				4868	4509	3439	4552	
	směr. odchylka				828	664	274	504	

TABULKA 2. Základní charakteristiky měsíčních souhrnů pro odchozí komunikaci.

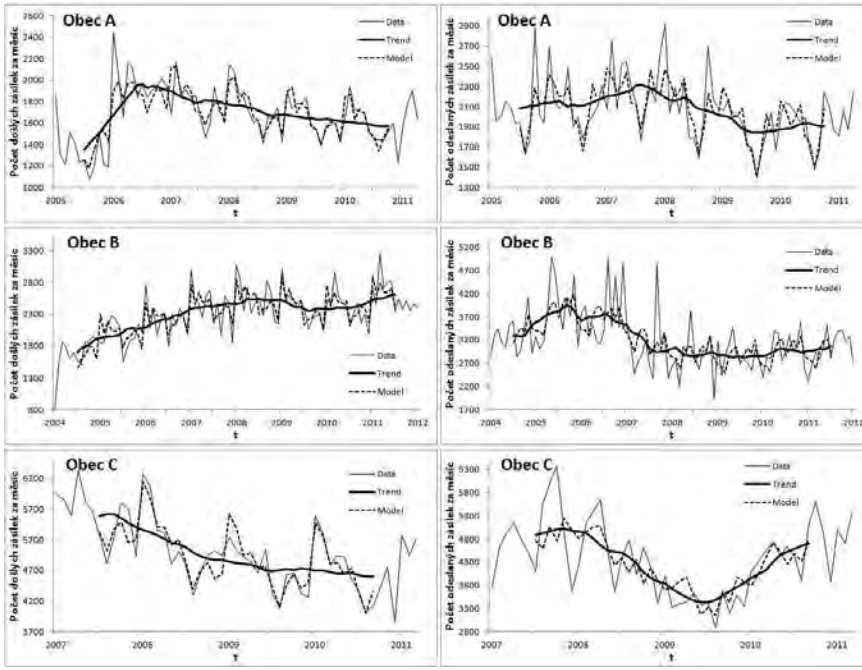
hodnotu. Případnou normalitu bílého šumu jsme testovali Shapiro-Wilkoxonovým testem.

4. **Numerické výsledky**

Základní charakteristiky měsíčních souhrnů přepočtených dle vzorce (1) jsou pro příchozí komunikaci v Tab. 1, pro odchozí v Tab. 2. Ve všech případech je směrodatná odchylka poměrně velká. Odlehlá pozorování v případě obce A v roce 2007 pro oba směry komunikace jsme pro další analýzu nahradili průměrnými hodnotami.

Výsledky trendové a sezónní složky modelu (2) jsou pro obce A, B a C zvlášť pro příchozí a odchozí komunikaci znázorněny v grafech na Obr. 1.

Podle provedených testů se ukazuje, že uvedený model dobře popisuje naměřená data. Reziduální složka pro příchozí komunikaci obce A a odchozí obce B je bílý šum, ve všech ostatních případech se jedná dokonce o Gasusovský bílý šum. Znamená to, že se v datech již nevyskytuje další cyklická složka, která by měla periodu (výrazně) kratší, než délka jednotlivých časových řad. Z grafického znázornění se zdá, že v některých případech zde



Obrázek 1. V grafech je zachycen průběh příchozí (levý sloupec) resp. odchozí (pravý sloupec) korespondence pro jednotlivé obce. Data jsou znázorněna souvislou křivkou, trend časové řady tučnou křivkou a přerušovaná křivka znázorňuje trend spolu se sezónní složkou.

mohou být cyklické složky delší (zejména v odchozí korespondenci) s periodou odpovídající délce uvedených časových řad. Tyto složky náš model nepostihuje a s ohledem na výraznou proměnlivost českého právního prostředí by byly i poměrně obtížně interpretovatelné.

Celkově se v trendech neprojevují v uvedeném období nějaké výrazné poklesy či nárůsty celkového objemu komunikace. Pro obec A můžeme sledovat mezi lety 2005 a 2006 nárůst příchozí pošty, který může být způsoben procesem učení se využití elektronického nástroje, tedy nikoliv skutečným nárůstem objemu komunikace, a poté následuje pozvolný pokles. V případě odchozí komunikace je trend téměř konstantní. Pro obec B je patrný pozvolný nárůst příchozí komunikace, v případě odchozí komunikace je trend proměnlivý, nicméně začátek a konec sledované časové řady je na stejné úrovni. Příchozí komunikace do obce C má klesající tendenci, kdežto u množství odchozí komunikace pozorujeme zřetelnou vlnu s minimem v polovině roku 2009.

Nyní se budeme věnovat sezónní složce modelu (2). Subjekty v rámci našeho malého vzorku tří obcí III. typu, které tvoří 1,5 % obcí III. typu, činí nejvíce podání vůči orgánům veřejné moci na začátku roku, nejméně pak v průběhu měsíců spadajících do období letních prázdnin a také na konci roku. Pro prosinec není tento propad ovlivněn obecně malým počtem pracovních dní tohoto měsíce, protože vliv konkrétního počtu pracovních dní v daném měsíci byl odstraněn přepočtem pomocí vztahu (1). U odchozí pošty nelze nějaké zobecnění na určitá roční období či určité měsíce napříč zkoumaným vzorkem určit. I to je ale také zajímavý výsledek, který ukazuje, že procesy probíhající na obcích mají různorodé doby trvání, díky čemuž sezónní složka odchozí komunikace nekopíruje sezónní složku příchozí komunikace.

5. Závěr

Provedená analýza časových řad příchozí a odchozí komunikace vybraných obcí III. typu ukazuje výrazné sezónní vlivy, které u příchozí komunikace mají navíc obdobný průběh pro všechny tři zkoumané obce, a poměrně nevýrazný nebo proměnlivý trend. Provedená analýza je prvním krokem ve výzkumu statistických dat poskytovaných elektronickými systémy spisových služeb implementovaných obcemi v ČR. Na tuto analýzu hodláme navázat výzkumem jednotlivých strukturních složek komunikace, které se s časem mění velmi výrazným způsobem.

Literatura

[1] Cipra, T. (2008) *Finanční ekonometrie*. Praha: Ekopress, s. r. o., 2008.

[2] ČESKÝ STATISTICKÝ ÚŘAD (2011) *Malý lexikon obcí ČR 2011*. Citace 05.06.2012. Dostupné na WWW: http://www.czso.cz/csu/2011edicniplan.nsf/p/1302-11.

[3] Lechner, T. (2009) *Způsoby formální komunikace veřejné správy v České republice v letech 2007–2008*. Veřejná ekonomika a správa 2009, Ostrava 08.09.2009 – 10.09.2009, P. Tománek, I. Vaňková (eds.). Ostrava: VŠB-TU 2009, 27 – 28.

[4] Lechner, T. (2010) *Hodnocení prvního roku existence datových schránek*. Veřejná správa 2010, Seč u Chrudimi 20.09.2010 – 21.09.2010. Pardubice: Univerzita Pardubice 2010, 131 – 141.

[5] Lechner, T. (2012) *Changes in Communication Thanks to eGovernment: Case Study of a Single Municipality in the Czech Republic*. European Journal of ePractice [online], 2012, **18**, 95 – 105.
Dostupné na WWW: http://www.epractice.eu/files/Journal_Volume 18.pdf.

[6] Mates, P., Smejkal, V. (2012) *E-Government v České republice: Právní a technologické aspekty. 2. podstatně přepracované a rozšířené vydání*. Praha: Leges, 2012.

[7] Ubaldi, B. Ch. (2011) *The impact of the Economic and Financial crisis on e-Government in OECD Member Countries*. European Journal of ePractice 2011, **5**(11), 5 – 18.

Poděkování: Tato práce byla podporována granty GAČR P201/10/0472 a VŠE IGS F5/2/2012.

Adresa: [1]Soukromá vysoká škola ekonomických studií, Katedra matematiky a IT, Lindnerova 575/1, 180 00 Praha 8 – Libeň, [2]Vysoká škola ekonomická

v Praze, Národohospodářská fakulta, Katedra práva, nám. W. Churchilla 4, 130 67 Praha 3 – Žižkov

E-mail: radka.lechnerova@svses.cz, tomas.lechner@gmail.com

THE FACTORS OF GROWTH OF SMALL FAMILY BUSINESSES. A ROBUST ESTIMATION OF THE BEHAVIORAL CONSISTENCY IN PANEL DATA MODELS

Eva Michalíková, Vladimír Benáček

Keywords: Family business, robust estimator, LTS, fixed effects

Abstract: The paper quantifies the role of factors associated with the growth (or decline) of micro and small businesses in European economies. The growth is related to employment and value added in enterprises as well as to ten institutional variables. We test the data for consistency of behavioral patterns in various countries and gradually remove outlying observations, quite a unique approach in panel data analysis that can lead to erroneous conclusions when using the classical estimators. In the first part of this paper we outline a highly robust method of estimation based on fixed effects and least trimmed squares (LTS). In its second part we apply this method on the panel data of 28 countries in 2002–2008 testing for the hypothesis that micro and small businesses in Europe use different strategies for their growth. We run a series of econometric tests where we regress employment and total net production in micro and small businesses on three economic factors: gross capital returns, labor cost gaps in small relative to large enterprises and GDP per capita. In addition, we test the role of 10 institutional factors in the growth of family businesses.

Abstrakt: Článek hodnotí faktory které ovlivňují růst malých rodinných podniků v evropské ekonomice. Tento růst souvisí se zaměstnaností a přidanou hodnotou v podnicích, stejně jako s institucionálními proměnnými. Testujeme chování různých evropských zemí a následně detekujeme odlehlá pozorování. Přítomnost těchto pozorování může vést k chybným závěrům, pokud jsou pro odhad dat použity klasické přístupy. V první části článku popisujeme robustní metodu odhadu založenou na odhadu pomoci fixních efektů a nejmenších useknutých čtvercích (LTS). V druhé části článku tuto metodu aplikujeme na data popisující 28 zemí v letech 2002–2008 a testujeme hypotézu, že růst mikro a malých rodinných podniků je založen na různých strategiích. V sérii ekonometrických testů vyjadřujeme zaměstnanost a přidanou hodnotu v mikro a malých podnicích jako funkci tří ekonomických veličin: hrubá kapitálová návratnost, relativní mzdové náklady a HDP na osobu. Dále do modelu vstupují některé z 10 různých institucionálních proměnných.

1. Introduction

As a consequenc of worldwide financial and economic crisis, there is a universally rising renewed interest in the performance of small and family businesses. Unique data on micro and small businesses in different countries do not represent a homogeneous pattern of behavior in firms that differ not only in sizes but also in institutional setups that also change in time. Thus mixing together of firms subject to different incentives could potentially lead to behavioral pattenrs that are not compatible. In this research we have tested the potential for such a heterogeneity in the behavior of small family businesses in various countries that could be even reflected in separating the original panel data into two subpopulations that are not compatible in their reaction to entrepreneurial stimuli. Hence, we have concentrated in our analysis on the techniques of robust estimation.

One of our innovation is the use ff robust estimations of the parameters in panel data models. In this paper in section 2 we describe and apply a robust version of the classical within-group estimators on data of two groups of family businesses. We transform the data by the subtraction of country-specific median. Then a robust Least Trimmed Squares estimator is applied on centered data. We decided to apply this method on economic data relating to family businesses grouped by company size. In section 3 we describe the role of family businesses in present economies. In section 4 we apply robust version of the within-group estimator on data for 28 European countries in 2002–2008 and we examine how employment and net production in family businesses depend on the measure of gross capital returns per value added and the relative gap between labor costs in small (or micro) and large enterprises. Additional explanatory variables include the GDP per capita and ten institutional variables. Section 5 concludes.

2. Robust Estimators and Robust Estimation of Panel Data Models

Outliers can be generated when the reporters mix up two or more subpopulations of data that represent agents whose behavior is mutually inconsistent. An example of this can be the case when the analysts presume that micro businesses (such as self-employed persons) and businesses up to 50 employees follow identical strategies for their growth in all studied countries. These kinds of inconsistency in carrying out observations are our main concern. Small family businesses are subject to specific circumstances that increase the uncertainty and inconsistency of their reported data. Firstly, their accountancy need not always be done by professionals and thus more open to errors and omissions. Secondly, their true production, employment and costs can be rigged due to much easier tax evasion. Thirdly, reporting to statistical offices is irregular. Thus a robust technique of setimation is a necessary and adequate approach in order to avoid the trap of data bias. Robust methods

date back to the history of statistics and the first basis for a theory of robust estimation was formed in the 1960's. Huber (1964) introduced a flexible class of M-estimators. However, robust estimation has not been the standard technique of analysis in this kind of panel data. Thus we will describe first our approach to data processing where the central issue rests in outliers. It is preferable to use such estimators of regression coefficients which are directly scale- and regression-equivariant as LMS or LTS estimator. Since LMS is only $\sqrt[3]{n}$-consistent, it is not asymptotically normal and not easy to evaluate, we will focus on the second applicable 50% breakdown point estimator – the least trimmed squares (LTS, Rousseeuw, 1983 and Rousseeuw and Leroy (1987).

Unfortunately, econometrics is limited to a scant amount of literature describing robust methods for panel data. This paper is an attempt at contributing to these techniques by focusing on the simple fixed effects panel data model of small businesses. We will try to find a robust alternative to the Within-Group estimator[1] which can be affected by the presence of outlying observations. Thus we will describe a high breakdown point estimator for the fixed effects panel data model based on LTS as an estimation procedure, which is less sensitive to the presence of aberrant observations.

We consider the following form of the fixed effects linear panel data model:

$$y_{it} = \alpha_i + x_{it}^{'}\beta + \varepsilon_{it} \qquad i = 1, \ldots, N \qquad t = 1, \ldots, T \tag{1}$$

where i denotes the cross-section dimension (number of countries) and t denotes the time-series dimension (number of years). x_{it} is a column vector of explanatory variables with dimension $K \times 1$ while β is a $K \times 1$ vector of regression parameters. α_i denotes the unobservable time-invariant individual fixed effects and ε_{it} denotes the error terms or disturbance terms, uncorrelated through time and through cross-sections.

The idea underlying Within-Group estimator is to center the series by mean when applying the within transformation and then transformed data are estimated by OLS. In order to get a *robust* version of this estimator we have to center the time series (in both the dependent and the explanatory variables) *robustly* and then a robust regression will be applied to the centered data. The difference between these two approaches is that the time-series must be centered by removing the median instead of mean because the mean is largely distorted by outliers since the median is known to be min-max robust (Huber, 1981). We will get:

$$\begin{aligned} \tilde{y}_{it} &= y_{it} - med_t(y_{it}) \\ \tilde{x}_{it}^{(j)} &= x_{it}^{(j)} - med_t(x_{it}^{(j)}) \end{aligned} \tag{2}$$

[1]Since our panel contain all countries of interenst, the fixed effects model is more appropriate than a random effects models for our dataset.

where $1 \leq i \leq N$, $1 \leq t \leq T$ and $1 \leq j \leq K$. $x_{it}^{(j)}$ denotes the j-th explanatory variable measured at time t in the i-th time-series. We can run a robust estimator (and regress $\tilde{y}_{it}$ on $\tilde{x}_{it}$ to identify the outliers). For this purpose we will apply the LTS estimator on centered data. LTS estimator is defined as $\hat{\beta}_{LTS}$ which minimizes the sum of the smallest h squared residuals:

$$\hat{\beta}_{LTS} = arg \min_{\beta} \sum_{k=1}^{h} [(\tilde{y}_k - \tilde{x}_k' \beta)^2]_i,$$

where

$$[(\tilde{y}_k - \tilde{x}_k' \beta)^2]_1 \leq [(\tilde{y}_k - \tilde{x}_k' \beta)^2]_2 \leq \ldots [(\tilde{y}_k - \tilde{x}_k' \beta)^2]_i \leq \ldots \leq [(\tilde{y}_k - \tilde{x}_k' \beta)^2]_{NT}$$

are ordered squared residuals (Rousseeuw, 1983). The value $1 \leq h \leq NT$ is a trimming value. As mentioned before, this estimator has a breakdown point attaining 50%. A default choice can be $h = [3NT/4]$ or $h = [4NT/5]$, making it possible to cope with up to 25% of outliers (or 20%, respectively). The LTS estimator in its basic version is regression, scale and affine equivariant. However, due to the nonlinearity of the centering transformation by the median β_{LTS} is only scale equivariant (Bramati and Croux, 2004). We can use this algorithm directly: centering the data by median, using least trimmed squares and discovering the outliers. However, it can also be employed in a different way by using outliers only as a diagnostic tool to recognize a "suspicious" behavior of an agent (Michalíková and Galeotti, 2010). In this paper we will identify the outliers in centered model, separate them and then use the LTS on the rest of data. This technique makes it possible to recognize outliers which are not able to be detected by eye or by means of traditional regression diagnostics. Once we have separated the observations (considered to be outliers), we can monitor if this subpopulation of data is subject to certain systemic regularity. Secondly, we may be watching if the removal of outliers brings some improvement in the estimated regression model. We may monitor the stability of estimated regression coefficients in the case of increasing h. Last but not least, we wonder if p-values of estimated regressors are improving as the outliers are dropped out from the model.

3. The Factors of Growth of Family Businesses

In the early 1990s, family-led enterprising was supposed to get a new boost as the pro-market forces triumphed. This was an error in judgement. Authentic small-scale family businesses were often squeezed out of the space for rapid development by surviving, former state-owned enterprises which were converted to corporations owned formally by thousands of petty stock-owners and a thin class of insiders with dominant stakes (Benáček, 2006). The parallel opening-up of globalisation offered new windows of opportunity

to large enterprises dominated by managers. In the late 1990s the floodgates to expansionary monetary policy opened up and government debt grew. These bubbles finally burst, which drove the economiesinto a lasting recession. Rising taxes, as a consequence of interventions, discriminated against small family business. The expectation is that the turnaround in the present recession should come from an increase in domestic aggregate spending and employment in SME dominated by family businesses, which in almost every country have been the main source of employment and job creation, but not the engine of spending dynamics. The main objective of this paper is to address the following question: which economic and institutional factors are associated with the development and growth of family businesses?

A firm is considered to be a family business if a member of one or more families is its controlling owner, implying a managerial commitment toward the business' overall performance. The main strength of a family business is the direct accountability and enforcement of property rights, without recourse to moral hazard and asset stripping. It also results in high wage flexibility. Micro and small businesses (i.e. MB and SB) cover 98.7% of all EU enterprises. In addition, approximately 50% of MB in the EU are formed by the self-employed. Therefore, in this paper we will use micro and small businesses as proxies of small family businesses. We will thus distinguish between two types of FB: those ranging in size from self-employed individuals to enterprises with 10 employees (i.e. MB) and enterprises with 10 to 50 employees (i.e. SB). We will measure the growth of MB and SB by their employment figures or, alternatively, by their net output. The following theoretical assumptions will be used as guidelines for hypotheses in our empirical tests:

a) The objective function of entrepreneurs is profit maximization. The maximization of gross capital returns per value added (KR/VA), where capital K is defined by reducing total labor compensation (W) from the net income of enterprises (VA)[2], is still a plausible criterion because it represents a social efficiency of capital allocated among businesses of various scales. We could set up a hypothesis that countries with higher KR/VA in any group of FB could also see the stronger development of FB. If the space for $K = VA - W$ increases (e.g. as a result of innovation or lower transaction costs), it will induce the entrepreneurs to expand their employment in order to bolster the sales and net output. This will result in an increase of labor income W and a raise in the wage rates per labor W/L. Nevertheless, is such a behavioral hypothesis valid for both employment and net production in reality? A very high KR/VA may also imply a shortage of capital (undercapitalization and/or too expensive capital). Then high capital returns could act as an impediment to FB growth, i.e. KR/VA could be negatively related to growth in employment.

[2]Net income (i.e. the value added) of enterprises is defined as difference between sales S and material inputs M.

b) FB development is not autonomous in isolation within their own SME categories because what also matters is an FB's relative performance vis-a-vis large businesses (LB). Small FB compete with LB for limited nationally available economic resources. The competition lies in costs and relative productivities. Thus the cost competition between FB and LB will depend on how well FB are able to depress wages, thus creating a wage gap relative to LB in order to gain a cost advantage once the prices of products are given. We will test whether (lower) wages per worker in FB related to (higher) wages per worker in LB are associated with higher growth in FB.

c) Another hypothesis about the determining factors of growth in FB that we will test concerns the degree of general economic development represented by GDP per capita.

d) Contemporary economics stresses the importance of institutions, as administrative bodiesdefining the 'rules of the game' or incentives whose purpose is to reduce uncertainties and transaction costs in business interaction. National institutions are important factors that may have both positive and negative impacts on businesses of different sizes.

Thus three economic indicators related to internal rates of gross capital returns (KR_{FB}/VA_{FB}), relative wages rates $(W_{FB}/L_{FB})/(W_{LB}/L_{LB})$, and GDP per capita, plus ten institutional indicators are selected as causal factors related to the growth of FB, i.e. the MB and SB.

4. **Results of Econometric Tests**

In this chapter we will test empirically the extent to which the growth in FB in 28 European countries was influenced during 2002–2008 by the three above-described economic factors and by political institutions. Sources of the data are: Small Business Act Factsheets (Eurostat and DG Enterprises and Industry); GDP statistics of the World Bank; Database on the Economic Freedoms (The Heritage Foundation). The robust version of the fixed effect panel data model will be used for the estimation of coefficients.

Dependet variables

L_{it}^{FB}: Employment in $FB = \{MB, SB\}$ quantified by the number of workers in country i and year t.

VA_{it}^{FB}: The value of net output (i.e. the value added) in MB or SB in country i and year t.

Economic explanatory variables:

$(KR/VA)_{it}^{FB}$: Gross capital returns in analyzed businesses per value added

$LC_{it}^{FB}/LC_{it}^{LB}$: Relative rates of full labour costs ($LC = W/L$), i.e. total labor compensation per worker in FB divided by similar compensation in LB

GDP_{it}/PC_{it}: GDP per capita in purchasing power parity.

Institutional explanatory variables:

$Regul_{it}$: Business freedom (regulation) index
$Trade_{it}$: Trade freedom (trade barriers) index
$Monet_{it}$: Monetary freedom (inflation and price control) index
$Govern_{it}$: Freedom from government (public spending) index
$Fiscal_{it}$: Fiscal freedom (taxation) index
$PropR_{it}$: Property rights index
$Invest_{it}$: Investment freedom (capital controls) index
$Financ_{it}$: Financial freedom (private banking security) index
$Corrupt_{it}$: Freedom from corruption (perception) index
$Labour_{it}$: Labor freedom index

N.B.: Institutional variables are the proxies of economic "freedoms" ranging in their values $\langle 0, 100\rangle$. The higher the percentage index, the more liberal and pro-market the local institutional arrangement.

The selection of 28 countries of Europe is highly representative, covering nearly all of the EU and potential accession countries (see Table 1). The first two explanatory variables are relevant for decision-making in enterprises. Reasons for having a high share of gross capital returns on the value added can be: a) Increasing labor productivity without compensating workers at a proportionally higher wage rate – that would imply high profits; b) Decreasing the marginal product of labor by overstaffing, which is reflected in disproportionally lower average wages in the enterprise. That would imply a high cost of capital that burdens the firm; c) Hiring and paying labor outside official contracts, which slashes total labor costs.

For different reasons that drive KR/VA upward, we cannot be sure whether this variable is related to FB growth negatively or positively.

The second variable LC^{FB}/LC^{LB} tests the relevance of low (reported) wages and of the gap in FB wage rates trailing behind LB. What matters is whether higher labor cost gap in FB is a driver or a retarder of FB growth. Once again we cannot be sure a priori about the nature of its sign. The third variable points to a general trend in development and is our only macroeconomic indicator. We should expect its sign to be positive.

ALL	Advanced Europe (14) + Emerging Europe (14)
Advanced Europe (14)	Austria, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, United Kingdom
Emerging Europe (14)	Albania, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Romania, Slovakia, Slovenia

TABLE 1. List of countries included in the analysis

The test of our robust regression analyses consist of four models related to micro and small enterprises, whose specifications are as follows:

$$
\begin{aligned}
L_{it}^{MB/SB} &= \alpha_1 (KR/VA)_{it}^{MB/SB} + \alpha_2 LC_{it}^{MB/SB}/LC_{it}^{large} + \\
&+ \alpha_3 GDP_{it}/PC_{it} + \alpha_x (INSTIT_{it}^{var} x) + \varepsilon_{it} \\
VA_{it}^{MB/SB} &= \beta_1 (KR/VA)_{it}^{MB/SB} + \beta_2 LC_{it}^{MB/SB}/LC_{it}^{large} + \\
&+ \beta_3 GDP_{it}/PC_{it} + \beta_x (INSTIT_{it}^{var} x) + \varepsilon_{it}
\end{aligned}
$$

where $i = 1, \ldots, 28$ are countries, $t = 2002, \ldots, 2008$ are the observed years, $x = \{4, 5, \ldots, 13\}$ indicates the respective number of institutional variable 4 through 13.

In Tables 2 and 3 we report the results of four regressions specified above[3]. In each regression we included three economic explanatory variables plus some relevant institutional explanatory variables. These variables were chosen according to the level of significance in individual models. The non-significant institutional variables were dropped from the model. In the first column for each regression we report results of fixed effects model, which was estimated by OLS from the data centered by median. In the following columns, we report the results of Least Trimmed Squares regression, applied on the data centered by median, with regard to different choice of h.

With regard to the results of our estimation in Tables 2 and 3, our first general observation is that parameters are mostly significant. In all four cases, the coefficient of determination (Adj.R^2) has been increasing and thus quality of model improved. If we focus on the signs of parameters, only in one case – in that of the coefficient of *KR/VA* in model 4 – the sign is unstable. Such a counter-intuitive reversal in sign could be, hypothetically, a result of multicolinearity, but variance inflation factor (VIF)[4] refuted that possibility.

In four models we use altogether 11 different variables. All three economic variables prove their clear dominance. The role of institutional factors seems to be only subsidiary, which is an unexpected finding of high importance. It signals that small family businesses are deeply dependent on market performance and policies are not so important in changing their strategic behavior. Variables *KR/VA* and *LC* have negative signs in models 1 and 2. This implies that job creation in small FB is conjoined with low pretentions to both capital returns and wage requirements. Thus saving on machines and prudent wage policy are traditional recipes for high employment in FB. There is also an important proviso to be added: a sustained or even widening gap in labor costs relative to large enterprises combined with lower capital endowments is a knife-edge enterprise strategy for gaining competitiveness in the short term that calls for low costs and prudence in expenditures on the one hand. A crucial piece of information is added by the third economic variable: rising GDP per capita enhances the employment in both types of FB. We can see that FB were the leading drivers of job creation throughout Europe during

[3]All estimates were obtained by Stata and Matlab

[4]Variance inflation factor (VIF) is common way for detecting multicollinearity. VIF is computed from the covariance matrix of parameter estimates (O'Brien, 2007).

Dep. variables	L_{it}^{MB} (Model 1)			
$h\%$	–	95%	85%	75%
$(KR/VA)_{it}^{MB}$	-0.080*	-0.329***	-0.210***	0.009
	(0.043)	(0.053)	(0.047)	(0.017)
$LC_{it}^{MB}/LC_{it}^{LB}$	-0.346***	-0.398***	-0.318***	-0.157***
	(0.062)	(0.051)	(0.039)	(0.025)
GDP_{it}/PC_{it}	0.509***	0.419***	0.405***	0.377***
	(0.039)	(0.029)	(0.021)	(0.018)
MONET	0.003**	0.0003	-0.001*	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)
FINANC	0.001**	0.002***	0.0006**	0.0004**
	(0.001)	(0.001)	(0.0003)	(0.002)
Number of obs.	196	187	167	147
Adj. R^2	0.525	0.603	0.700	0.772
Dep. variables	L_{it}^{SB} (Model 2)			
$h\%$	–	95%	85%	75%
$(KR/VA)_{it}^{SB}$	-0.164***	-0.157***	-0.166***	-0.005
	(0.021)	(0.015)	(0.010)	(0.051)
$LC_{it}^{SB}/LC_{it}^{LB}$	-0.330***	-0.167**	-0.016***	0.016
	(0.091)	(0.074)	(0.054)	(0.049)
GDP_{it}/PC_{it}	0.541***	0.496***	0.407***	0.423***
	(0.035)	(0.026)	(0.003)	(0.017)
FINANC	0.001**	0.001**	0.001**	0.001
	(0.001)	(0.0004)	(0.0003)	(0.001)
LABOR	-0.001	-0.001	-0.002**	-0.001***
	(0.001)	(0.001)	(0.001)	(0.0004)
Number of obs.	196	187	167	147
Adj. R^2	0.634	0.748	0.751	0.837

TABLE 2. Robust fixed effects regressions - models 1 and 2. Notes: The value for $h\%$ denotes how many observations were included into data set. * significant at 10%; ** significant at 5 %; *** significant at 1 %. Standard errors are in brackets. Dependent variables and GDP per capita are in logarithms.

the observed period. The conditions for job expansion in micro business are also in the prudent monetary policy (that sustains low inflation) and in the existence of efficient financial services.

The three most powerful findings occurred in models 3 and 4 (Table 3) explaining the mechanism of growth in net production in MB and SB. Firstly, our models point to the existence of a trade-off between employment and output expansion because the signs for the first two economic variables reversed

Dep. variables	VA_{it}^{MB} (Model 3)			
$h\%$	–	95%	85%	75%
$(KR/VA)_{it}^{MB}$	0.301***	0.299***	0.277***	0.503***
	(0.072)	(0.064)	(0.044)	(0.073)
$LC_{it}^{MB}/LC_{it}^{LB}$	0.448***	0.376***	0.388***	0.575***
	(0.103)	(0.087)	(0.061)	(0.063)
GDP_{it}/PC_{it}	1.736***	1.552***	1.404***	1.528***
	(0.067)	(0.060)	(0.045)	(0.036)
MONET	0.004**	-0.001	-0.002	-0.003**
	(0.002)	(0.002)	(0.002)	(0.001)
CORRUPT	0.005***	0.003**	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Number of obs.	196	187	167	147
Adj. R^2	0.823	0.825	0.880	0.938
Dep. variables	VA_{it}^{SB} (Model 4)			
$h\%$	–	95%	85%	75%
$(KR/VA)_{it}^{SB}$	-0.105***	0.052	0.456***	0.452***
	(0.032)	(0.148)	(0.128)	(0.098)
$LC_{it}^{SB}/LC_{it}^{LB}$	0.631***	0.408***	0.410***	0.478***
	(0.138)	(0.129)	(0.108)	(0.083)
GDP_{it}/PC_{it}	1.737***	1.576***	1.507***	1.386***
	(0.054)	(0.045)	(0.038)	(0.033)
GOVERNMENT	0.002*	0.002**	0.001	-0.0001
	(0.001)	(0.001)	(0.001)	(0.001)
INVEST	0.001	0.001	0.0003	0.0007**
	(0.001)	(0.005)	(0.0001)	(0.0003)
Number of obs.	196	187	167	147
Adj. R^2	0.866	0.877	0.909	0.934

TABLE 3. Robust fixed effects regressions - models 3 and 4. Notes: The value for $h\%$ denotes how many observations were included into data set. * significant at 10%; ** significant at 5 %; *** significant at 1 %. Standard errors are in brackets. Fixed effects are not reported. Dependent variables and GDP per capita are in logarithms.

from negative to positive. Secondly, the coefficients for GDP per capita increased approximately three-fold in their value, pointing to a high elasticity of FB output growth to aggregate demand. Thirdly, the results in Table 3 imply that value added VA is more sensitive to low labour costs LC (and with it to labour efficiency) than to high capital returns (capital efficiency). Therefore, by consolitating these results, we can draw an implication that increasing aggregate demand is driving production (and therefore probably also the profits) in FB more than its employment.

The growth in net output in FB is underpinned by high gross capital gains per value added. High GDP per capita is a crucial catalyst for such development accompanied by low corruption in the case of model 3.

Finally we will focus on observations which have been dropped out from the model. There are six countries that are generating the majority of outliers: Albania, Croatia, Greece, Latvia, Romania and Slovakia. With the exception of Greece they belong to countries of emerging post-Communist Europe that in the past had problems with macroeconomic stability and EU accession. These countries differ in their high growth of employment. Thus job creation in FB during 2002–2008 was faster in these emerging countries compared to other countries. In the case of value added this growth was even more significant.

5. **Conclusion**

In this paper we have analyzed the factors that were instrumental for growth in two types of small firms in 28 European countries. It has been revealed that growth related to employment and to net production was conditioned by very different internal incentives. As has been found, schemes (or incentives) targeting high employment can be in conflict with schemes concentrating on the growth in value added.

We have also tested the stability of behavioral patterns in FB throughout Europe. For that purpose we applied a robust methodology for fixed effect panel data models which allowed us to estimate a model where data were contaminated by outliers. Based on data for 28 European countries in 2002–2008 we ran a series of econometric tests in which we analyzed how two groups of businesses with up to 50 employees evolved over time by quantifying their growth in employment and net production. We regressed these two alternative indicators of development to a measure of gross capital returns per a unit of value added and to the relative gap between labor costs in small and large enterprises . In addition, we tested the role of GDP per capita and the significance of several institutional variables.

Our tests concluded with the finding that our three economic explanatory variables were highly statistically significant. With rising h(the number of deleted observations) results have been generally improving as the residual sum of squares was decreasing and the explanatory power of the model was gaining in strength. We can conclude from the results of four regressions that job creation in micro and small family businesses depends on a low pretention on capital returns. However, narrowing the gap in labor costs in family businesses relative to large corporations is negatively correlated with employment. In sharp contrast with this, both these economic variables are positively connected with the value added in micro and small business. The higher the gross capital gains per value added and the higher the relative labor costs in FB, the higher their growth in net production is.

Rising GDP per capita enhances both employment and value added in FB, even though the impact on the net output is markedly more intensive. We have also discovered that some less developed post-Communist countries (particularly Albania, Romania, Croatia, Latvia and Slovakia) were subject to highly different behavior of family businesses related to growth.

As a final point for discussion, our results imply that after all, hard economic fundamentals (factor costs, labour efficiency and the aggregate demand) are much more important for the development of small family businesses than soft institutional factors. This is in sharp contrast to the performance of large businesses, whose activities are found to be strongly influenced by policies and vertical transfers at the level of public administration, as was observed by Alfaro et al. (2008) or Benacek et al. (2011).

References

[1] Alfaro L., Kelemli-Ozcan S. and Volosovych V. (2008) *Why Doesn't Capital Flow from Rich to Poor Countries? An Empirical Investigation*. The Review of Economics and Statistics, 90(2), 347–368.

[2] Benáček V. (2006) *The Rise of Grand Entrepreneurs in the Czech Republic and Their Contest of Capitalism*. Czech Sociological Review, 42(6), 1151–1170.

[3] Benáček V., Lenihan H., Andreosso-O'Callaghan B. and Kan D. (2011) *Political Risk and Foreign Direct Investment: How Do they Get along in Various European Countries?* Working Papers, University of Limerick.

[4] Bramati M. C. and Croux C. (2004) *Robust Estimators for the Fixed Effects Panel Data Model*. Econometrics Journal, 10, 1–19.

[5] Huber P. J. (1964) *Robust Estimation of a Location Parameter*. Annals of Mathematical Statistics, 35, 73–101.

[6] Huber P. J. (1981) *Robust Statistics*. John Wiley, New York.

[7] Michalíková E. and Galeotti E. (2010) *Determinants of FDI in Czech Manucaturing Industries between 2000–2007*. South East European Journal of Economics and Business. 5(2), 21–32.

[8] O'Brien R. O. (2007) *A Caution Regarding Rules of Thumb for Variance Inflation Factors*. Quantity and Quality, 41, 673–690.

[9] Rousseeuw P. (1983) *Multivariate Estimation With High Breakdown Point*. Paper presented at Fourth Pannonian Symposium on Mathematical Statistics and Probability, Bad Tatzmannsdort, Austria.

[10] Rousseeuw P. and Leroy A. (1987) *Robust Regression and Outlier Detection*. Wiley.

Acknowledgement: Our research was supported by the Grant Agency of the Czech Republic, Grant No. P402/12/0982 Trade Flows in Times of Economic Boom and Slump and research project of VUT in Brno no. FP-S-12-1.

Address: Fakulta Podnikatelská VUT v Brně, Kolejní 2906/4, 612 00 Brno / Institut Ekonomických studií FSV UK, Opletalova 26, 110 00 Praha 1.

E-mail: michalikova@volny.cz

REGRESE V MODELECH OPRAV

Petr Novák

Klíčová slova: Analýza spolehlivosti, modely oprav.

Abstrakt: Při provozu systému, který podléhá opotřebení, je naší snahou odhadnout rozdělení doby do selhání pro optimalizaci plánování údržby. Pomocí vhodných modelů chceme popsat závislost tohoto rozdělení na případných regresorech. Běžně používané modely analýzy přežití, jako je Coxův model nebo model zrychleného času, je potřeba přizpůsobit systému s opravami. Například lze modelovat závislost na předchozích opravách a časově proměnných kovariátách odpovídajících stavu systému. V příspěvku takové modely popisujeme a předvádíme využití na datech z praxe.

1. Úvod - modelování životnosti jednoho zařízení

Studujeme data o historii zařízení, které podléhá opotřebení. Když se porouchá, je nutné provést opravu. Poruchám se snažíme předcházet preventivní údržbou. Označíme $T_1, ..., T_n$ časy zásahů (oprav nebo údržeb), $\Delta_1, ..., \Delta_n$ indikátor zda v j-tém čase byla provedena oprava, $X(t)$ vysvětlující proměnná. Chceme rozumný popis vlivu oprav, údržby a regresorů na životnost.

Zavedeme čítací procesy oprav a údržeb

$$N_\bullet(t) = \sum_{j=1}^{n} I(T_j \le t, \Delta_j = 1), \qquad M_\bullet(t) = \sum_{j=1}^{n} I(T_j \le t, \Delta_j = 0).$$

Označíme rizikovou funkci

$$\lambda(t) = \lim_{h \to 0} P(N_\bullet(t+h) - N_\bullet(t) \ge 1 | \mathcal{H}(t))/h$$

kde $\mathcal{H}(t)$ značí historii událostí do času t. Dále mějme kumulativní rizikovou funkci $\Lambda(t) = \int_0^t \lambda(s)ds$ a $f(t)$ a $S(t)$ příslušnou hustotu a funkci přežití. Předpokládejme, že oprava sice vrátí prvek jen do stavu těsně před poruchou, ale má vliv na rizikovou funkci. Rizikovou funkci vhodně parametrizujeme a parametry chceme odhadnout metodou maximální věrohodnosti. Věrohodnost lze přepsat jako

$$L = \prod_{j=1}^{n} \left(\frac{f(T_j^-)}{S(T_{j-1})} \right)^{\Delta_j} \left(\frac{S(T_j)}{S(T_{j-1})} \right)^{1-\Delta_j} = \prod_{j=1}^{n} \lambda(T_j^-)^{\Delta_j} \cdot S(T_n)$$

a log-věrohodnost má pak tvar

$$l = \sum_{j=1}^{n} \Delta_j \log \lambda(T_j^-) - \int_0^{T_n} \lambda(t)dt.$$

Coxův model

V Coxově modelu působí regresory multiplikativně na rizikovou funkci. Předpokládáme, že každá oprava či údržba multiplikativně sníží nebo zvýší riziko, stejně tak případné regresory. Uvažujeme rizikovou funkci ve tvaru (Percy & Alkali 2005):

$$\lambda(t) = \lambda_0(t)e^{M_\bullet(t)\rho+N_\bullet(t)\sigma+X^T(t)\beta} = \lambda_0(t)(e^\rho)^{M_\bullet(t)}(e^\sigma)^{N_\bullet(t)}(e^\beta)^{X(t)}.$$

Jako vysvětlující proměnnou $X(t)$ je možné použít např. náročnost poslední opravy. Pokud se hodnoty kovariáty mění jen v časech událostí, je možné snadno dosadit do logaritmické věrohodnosti a při parametrickém základním riziku maximalizovat.

Model zrychleného času

Můžeme také předpokládat, že každá oprava či údržba a regresory způsobí, že virtuální čas začne plynout pomaleji nebo rychleji (Accelerated Failure Time model, AFT). Využijeme transformaci času (Lin & Ying, 1995):

$$t \to \int_0^t e^{M_\bullet(s)\rho+N_\bullet(s)\sigma+X^T(s)\beta}ds =: h(t,\boldsymbol{\beta}),$$

kde jsme označili $\boldsymbol{\beta} = (\rho, \sigma, \beta)$. Riziková fukce pak má tvar

$$\lambda(t) = \lambda_0(h(t,\boldsymbol{\beta}))e^{M_\bullet(t)\rho+N_\bullet(t)\sigma+X^T(t)\beta}.$$

Pokud základní riziková funkce bude konstantní, oba modely splývají.

2. Inference při více pozorováních

Máme-li k dispozici data o n nezávislých zařízeních, pracujeme se sdruženou věrohodností. Můžeme buďto parametrizovat základní riziko a postupovat jako výše, nebo odhadnout základní riziko neparametricky. Mějme $\lambda_i(t)$, T_{ij}, Δ_{ij}, $j = 1, ...n_i$ a $X_i(t)$ rizikovou funkci, časy událostí, indikátory oprav a hodnoty regresorů i-tého prvku. Označíme

$$N_{ij}(t) = \Delta_{ij}I(T_{ij} \le t), \qquad M_{ij}(t) = (1-\Delta_{ij})I(T_{ij} \le t),$$

$$Y_{ij}(t) = I(T_{i,j-1} < t \le T_{ij}).$$

Pomocí $\bullet$ označíme součet přes příslušný index. Dostaneme log-věrohodnost

$$l = \sum_{ij}\int_0^\infty \Big(\log\lambda_i(t^-)dN_{ij}(t) - Y_{ij}(t)\lambda_i(t^-)dt\Big),$$

kde v rizikové funkci λ_i budou obsaženy počty oprav a údržeb $N_{i\bullet}$ a $M_{i\bullet}$.

Coxův model semiparametricky

Označíme $\boldsymbol{X}_i(t) = (N_{i\bullet}(t), M_{i\bullet}(t), X_i(t))$. Pak je pro Coxův model věrohodnost a skóre

$$l = \sum_{ij} \int_0^\infty \left((\log \lambda_0(t^-) + \boldsymbol{X}_i^T(t^-)\boldsymbol{\beta}) dN_{ij}(t) - Y_{ij}(t) e^{\boldsymbol{X}_i^T(t^-)\boldsymbol{\beta}} \lambda_0(t^-) dt \right),$$

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i^T(t^-) dN_{ij}(t) - Y_{ij}(t) \boldsymbol{X}_i^T(t^-) e^{\boldsymbol{X}_i^T(t^-)\boldsymbol{\beta}} d\Lambda_0(t) \right).$$

Skóre závisí na neznámé kumulované základní rizikové funkci $\Lambda_0(t)$. Tu můžeme nahradit Nelson-Aalenovým odhadem

$$\hat{\Lambda}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{dN_{\bullet\bullet}(s)}{\sum_{ij} e^{\boldsymbol{X}_i^T(s^-)\boldsymbol{\beta}} Y_{ij}(s)}.$$

Po dosazení získáme skóre ve tvaru

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i(t^-) - \frac{\sum_{ij} \boldsymbol{X}_i(t^-) e^{\boldsymbol{X}_i^T(t^-)\boldsymbol{\beta}} Y_{ij}(t)}{\sum_{ij} e^{\boldsymbol{X}_i^T(t^-)\boldsymbol{\beta}} Y_{ij}(t)} \right) dN_{ij}(t)$$

a pro nalezení odhadů parametrů řešíme rovnice $U(\boldsymbol{\beta}) = 0$.

AFT model semiparametricky

Pro každý prvek máme transformaci času $h_i(t, \boldsymbol{\beta})$. Zavedeme transformované procesy

$$N_{ij}^*(t, \boldsymbol{\beta}) = \Delta_{ij} I(h_i(T_{ij}, \boldsymbol{\beta}) \leq t), \quad M_{ij}^*(t, \boldsymbol{\beta}) = (1 - \Delta_{ij}) I(h_i(T_{ij}, \boldsymbol{\beta}) \leq t),$$

$$Y_{ij}^*(t, \boldsymbol{\beta}) = I(h_i(T_{i,j-1}, \boldsymbol{\beta}) < t \leq h_i(T_{ij}, \boldsymbol{\beta})), \quad X_i^*(t, \boldsymbol{\beta}) = X_i(h_i^{-1}(t, \boldsymbol{\beta})).$$

Přesné skóre má složitější tvar, je ale možné jej nahradit přibližným (Lin & Ying, 1995)

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \boldsymbol{X}_i^*(t^-, \boldsymbol{\beta}) \left(dN_{ij}^*(t, \boldsymbol{\beta}) - Y_{ij}^*(t, \boldsymbol{\beta}) d\Lambda_0(t) \right)$$

a dosadit odhad kumulované základní rizikové funkce

$$\hat{\Lambda}_0(t, \boldsymbol{\beta}) = \int_0^t \frac{dN_{\bullet\bullet}^*(s, \boldsymbol{\beta})}{\sum_{ij} Y_{ij}^*(t, \boldsymbol{\beta})}.$$

Získáme

$$U(\boldsymbol{\beta}) = \sum_{ij} \int_0^\infty \left(\boldsymbol{X}_i^*(t^-, \boldsymbol{\beta}) - \frac{\sum_{ij} \boldsymbol{X}_i^*(t^-, \boldsymbol{\beta}) Y_{ij}^*(t, \boldsymbol{\beta})}{\sum_{ij} Y_{ij}^*(t, \boldsymbol{\beta})} \right) dN_{ij}^*(t, \boldsymbol{\beta}).$$

Protože skóre není spojité v $\boldsymbol{\beta}$, odhadneme parametry minimalizací $\|U(\boldsymbol{\beta})\|$.

3. Modelování provozu čerpadla

Zkoumáme data o provozu ropných čerpadel za několik let (Kobbacy, 1997 a Percy, 2007). Pro jedno čerpadlo máme podrobnější údaje o časech údržeb, oprav a o náročnosti každého zásahu v člověkohodinách. Zde zkusíme parametricky modelovat životnosti při různých základních rizikových funkcích. U pěti pump jsou k dispozici jen časy oprav a údržeb, zde použijeme semiparametrické metody a porovnáváme s výsledky při použití parametrizovaného základního rizika.

Parametrické modelování provozu čerpadla

Máme k dispozici časy oprav, údržeb a náročnosti prací a podle metod z odstavce 1 odhadujeme parametry ρ, σ a β v Coxově i AFT modelu. Zkusíme maximalizovat věrohodnost pro exponenciální, Weibullovo $\lambda_0(t) = a\lambda^a t^{a-1}$, gamma $f(t) \propto t^{a-1}e^{-\lambda t}$, useknuté Gumbelovo $\lambda_0(t) = \lambda a^t$ a lognormální základní rozdělení pro oba popsané modely.

Model	λ_0	log - lik	$e^{\hat\rho}$	$e^{\hat\sigma}$	$e^{\hat\beta}$	$\hat\lambda$	$\hat a$
	Exp.	-213.8	1.407	0.980	1.0066	0.0015	–
Cox	Weibull	-213.5	1.266	0.924	1.0064	0.0017	1.672
	Gamma	-213.8	1.405	0.918	1.0066	0.0016	1.027
	Gumbel	**-210.2**	0.701	0.745	1.0063	0.0006	1.010
	LN	-214.8	1.541	0.913	1.0069	$\hat\mu$=6.3	$\hat\sigma$=1.66
AFT	Weibull	-212.7	1.278	0.918	1.0061	0.0014	1.639
	Gamma	-213.8	1.418	0.916	1.0066	0.0014	0.918
	Gumbel	**-210.2**	1.318	0.877	1.0050	0.0005	1.001
	LN	-218.1	1.300	1.050	1.0070	$\hat\mu$=5.25	$\hat\sigma$=0.89

Tabulka 1. Hodnota logaritmické věrohodnosti a odhadů parametrů při parametrickém modelování životnosti z dat o jednom čerpadlu

Porovnáním hodnoty věrohodnosti v tabulce 1 zjistíme, že je zde nejvyšší zároveň pro Coxův i AFT model s useknutým Gumbelovým rozdělením. Časová náročnost opravy zvyšuje riziko respektive zrychluje čas, protože $e^{\hat\beta} > 1$. Každá člověkohodina zásahu způsobí nárůst rizika či zrychlení času o zhruba $0.5 - 0.7\%$. Oprava samotná má vliv pozitivní ($e^{\hat\sigma} < 1$), jen u AFT modelu s lognormálním základním rozdělením je to naopak, ale tento případ má nejnižší věrohodnost. Zajímavé je, že všechny modely vyjma Gumbelova rozdělení v Coxově modelu vyhodnocují, že údržba má vliv negativní ($e^{\hat\rho} > 1$).

Semiparametrické modelování provozu čerpadla

Pro pět zařízení máme jen časy oprav a údržeb. Údaj o náročnosti opravy nebyl k dispozici u všech, takže budeme odhadovat jen regresní parametry ρ a σ. Zkusili jsme aplikovat Coxův i AFT model, jak parametricky se stejnými základními rozděleními jako výše, tak semiparametricky. V parametrickém postupu maximalizujeme věrohodnost, při semiparametrickém dosadíme odhad kumulované základní rizikové funkce a řešíme rovnice $U(\beta) = 0$.

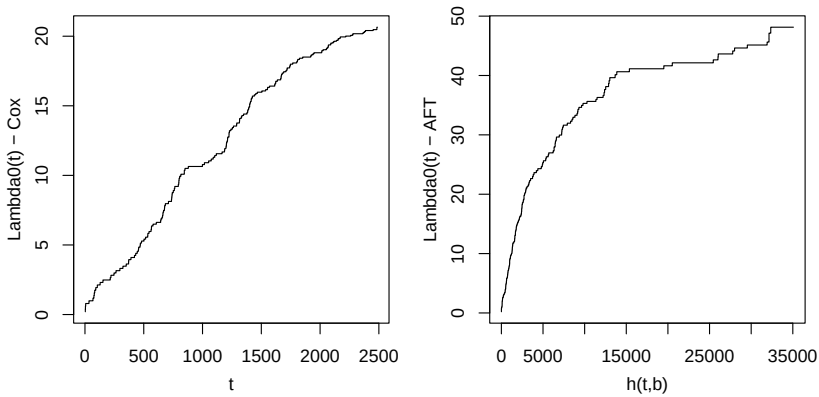
Model	λ_0	log - lik	$e^{\hat{\rho}}$	$e^{\hat{\sigma}}$	$\hat{\lambda}$	$\hat{a}$
	Exp.	-880.3	0.985	1.016	0.016	–
Cox	Weibull	-880.2	0.976	1.016	0.014	1.063
	Gamma	-880.1	0.988	1.016	0.015	0.811
	Gumbel	-880.3	0.994	1.016	0.016	0.999
	LN	-894.4	1.090	1.016	$\hat{\mu}$=3.22	$\hat{\sigma}$=0.89
AFT	Weibull	-880.2	0.980	1.015	0.014	1.038
	Gamma	-880.1	0.988	1.016	0.015	0.812
	Gumbel	**-875.1**	1.022	1.036	0.013	0.999
	LN	-879.5	1.284	1.158	$\hat{\mu}$=2.67	$\hat{\sigma}$=1.56
Cox	neparam.	–	1.043	1.020	–	–
AFT	neparam.	–	1.028	1.084	–	–

TABULKA 2. Hodnota logaritmické věrohodnosti a odhadů parametrů při modelování životnosti z dat o pěti různých čerpadlech

V tabulce 2 vidíme, že ve všech případech oprava zvýší riziko či zrychlí plynutí času ($e^{\hat{\sigma}} > 1$). Z parametrických modelů má nejvyšší logaritmickou věrohodnost Gumbelovo rozdělení v AFT modelu. U tohoto případu, stejně jako při lognormálním základním rozdělení a u neparametrických modelů, má údržba také negativní vliv, jinak má vliv pozitivní. Na obrázku 1 jsou znázorněny odhady kumulované základní rizikové funkce, čas pro AFT model je v transformované škále.

4. Závěr

Zkoumali jsme metody pro modelování vlivu údržby a oprav na životnost sledovaného zařízení. V Coxově modelu působí regresory vyjadřující počet a míru zásahů a příslušné parametry multiplikativně na rizikovou funkci, v AFT modelu přímo na rychlost plynutí vnitřního času. Při parametrizaci základního rizika nám pro získání odhadů stačí údaje o jednom zařízení. Pokud máme informace o více zařízeních, je možné základní riziko odhadnout neparametricky. Dalším předmětem zkoumání by mohlo být testování, zda je možné neparametrický odhad nahradit vhodnou parametrizovanou základní



Obrázek 1. Odhad kumulované základní rizikové funkce v semiparametrickém Coxově a AFT modelu

rizikovou funkcí. Také je možné prozkoumat jiné transformace pro model zrychleného času.

Literatura

[1] Kobbacy K.A.H., Fawzi B.B., Percy D.F., Ascher H.E. (1997) *A full history proportional hazards model for preventive maintenance scheduling.* Quality and Reliability Engineering Intl. **13**, 187 — 198.

[2] Lin D.Y., Ying Z. (1995) *Semiparametric inference for the accelerated life model with time-dependent covariates.* Journal od Statistical Planning and Inference **44**, 47 – 63.

[3] Percy D.F., Alkali B.M. (2005) *Generalized proportional intensities models for repairable systems.* IMA Journal of Management Mathematics **17**, 171 – 185.

[4] Percy D.F., Alkali B.M. (2007) *Scheduling preventive maintenance for oil pumps using generalized proportional intensities models.* International Transactions of Operational Research **14**, 547 – 563.

Poděkování: Tato práce byla podporována granty SVV 261315/2012 a MŠMT ČR 1M06047.

Adresa: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

E-mail: novakp@karlin.mff.cuni.cz

ESTIMATING DISTRIBUTION OF NEURONAL RESPONSE LATENCY

Zbyněk Pawlas

Keywords: asymptotic properties, deconvolution, empirical distribution function, nonparametric estimation, response latency

Abstract: A problem of inferring the distribution given a sample from the distribution of minimum with an independent random variable is considered. It is motivated by a neural coding model that describes neuronal response to an external stimulus. The aim is to estimate the unknown distribution of response latency. A natural nonparametric estimator of the complementary distribution function is given as a ratio of two empirical survival functions. The asymptotic properties of this estimator are investigated. The estimator need not to be a valid complementary distribution function. Therefore, its modification is proposed. The comparison of the finite sample performance of the estimators is carried out by a simulation study.

Abstrakt: Je uvažován problém odhadu rozdělení na základě výběru z rozdělení daného minimem s nezávislou náhodnou veličinu. Motivací pro tento problém je model neuronového kódování, který popisuje reakci neuronu na vnější stimulus. Cílem je odhadnout neznámé rozdělení latence odezvy neuronu. Přirozený neparametrický odhad funkce přežití je dán podílem dvou empirických funkcí přežití. Jsou zkoumány asymptotické vlastnosti tohoto odhadu. Odhad nemusí být platná doplňková distribuční funkce, proto je navržena vhodná modifikace. Pomocí simulační studie je provedeno srovnání kvality různých odhadů.

1. Introduction

Let X be a non-negative random variable with distribution function F and Y be a non-negative random variable with distribution function G. Assume that X and Y are independent, then $Z = \min(X, Y)$ has distribution function

$$H(t) = 1 - (1 - F(t))(1 - G(t)), \quad t \geq 0.$$

Our aim is to estimate the distribution function F based on independent random samples from the populations with distribution functions G and H.

This problem is related to the deconvolution problem that can be described as follows. Let $Z = X + Y$ be the sum of two independent random variables X and Y. If X has distribution function F and Y has distribution function G, then the distribution function H of Z is equal to the convolution of the functions F and G. The distribution of interest is that of X. The function G is usually assumed to be known or at least can be estimated. The aim is to estimate F based on a sample from H. A comprehensive overview of deconvolution problems can be found in [2].

The motivation for our study comes from neurophysiology. Neurons generate action potentials (more simply called spikes). Since both the shape and the amplitude of the spike are believed to carry minimal information, the main focus in neural coding is devoted to the timing of spikes (so called temporal coding). We consider a single neuron. The spikes that are fired at the presence of spontaneous activity are assumed to form a homogeneous sequence and will be refered to as spontaneous spikes. At a given moment, the stimulation period starts. At least a single evoked spike (usually a sequence) is generated as the reaction of a neuron to the stimulus. Response latency is defined as the time elapsed from stimulus onset to the occurence of a first evoked spike. In practice, we are able to measure first spike latency, that is the time to the occurence of a first spike after stimulus. This spike can be either spontaneous or evoked. However, we are not able to distinguish whether it was caused by spontaneous activity or by the response to the stimulus. It means that we observe the minimum Z (first spike latency) of time X (response latency) to the first evoked spike and time Y (spontaneous latency) to the first spontaneous spike after stimulus. We are interested in the estimation of response latency distribution. It will be based on the recordings obtained from repeated stimulations under identical conditions. For more details on this scenario, see [3] and [5].

If we ignored the effect of Y (presence of spontaneous activity) and approximated the distribution of X by the distribution of $Z = \min(X, Y)$ which can be directly estimated from data, then we would overestimate F (obviously, $H(t) \geq F(t)$). This procedure may be reasonable only if the probability that $Y \leq X$ is small. It is intuitively clear that larger values of $\mathbb{E}Y/\mathbb{E}X$ mean smaller chance that $Y \leq X$. A particular example of interest is when the random variables are exponentially distributed. If X has exponential distribution with intensity λ_F and Y has exponential distribution with intensity λ_G, then Z has distribution function $H(t) = 1 - \exp\{-(\lambda_F + \lambda_G)t\}$, i.e. it has exponential distribution with intensity $\lambda_H = \lambda_F + \lambda_G$. We will write shortly $X \sim \mathrm{Exp}(\lambda_F)$, $Y \sim \mathrm{Exp}(\lambda_G)$, and $Z \sim \mathrm{Exp}(\lambda_H)$. It is not difficult to see that $\mathbb{P}(Y \leq X)$ is a function $\mathbb{E}Y/EX = \lambda_F/\lambda_G$, namely

$$\mathbb{P}(Y \leq X) = \frac{\lambda_G}{\lambda_F + \lambda_G} = \frac{1}{1 + \mathbb{E}Y/\mathbb{E}X}.$$

Hence, the probability that the first spike after stimulus onset is spontaneous equals to

$$\mathbb{P}(Y \leq X) = \mathbb{P}(Z = Y) = \frac{\lambda_G}{\lambda_H} = \frac{\mathbb{E}Y}{\mathbb{E}Z}$$

and can be estimated from data. This may indicate whether the effect of spontaneous spikes can be neglected.

In Section 2 we define a nonparametric estimator of F that corrects the influence of Y on the observed minimum Z. We study the asymptotic properties as the number of observations increases. It is shown that the estimator is

asymptotically unbiased, asymptotically consistent, and asymptotically normal. The proposed estimator need not to be a valid distribution function. It is not necessarily either monotone or positive. Therefore, its modification is introduced in Section 3. In particular cases we determine the probabilities that the properties of distribution function are violated. Section 4 concludes with a simulation study comparing the performance of the considered estimators.

2. **Asymptotic properties**

Let us consider two independent random samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ such that the X_i have distribution function F and the Y_i have distribution function G. We suppose that the information contained in these random samples is available only through parallel minima $Z_i = \min(X_i, Y_i)$, $i = 1, \dots, n$. Therefore, the Z_i are iid random variables with distribution function H. Our aim is to estimate F. This resembles a classical problem of random censoring. The difference is that the indicators of censoring are unknown. Instead we have a random sample $\tilde{Y}_1, \dots, \tilde{Y}_m$ with distribution function G, that is independent of both $\{X_i\}$ and $\{Y_i\}$. In our motivating neurophysiology example, the Z_i are first spike latencies measured from n repeated stimulation trials and the $\tilde{Y}_i$ are obtained from the recordings during spontaneous activity of a neuron. If spontaneous spikes form a homogeneous Poisson process, then the times between two consecutive spikes (so called interspike intervals) have exponential distribution and can be used as the observations $\tilde{Y}_1, \dots, \tilde{Y}_m$ that are equal in distribution to Y.

We can estimate G and H by the corresponding empirical distribution functions

$$G_m(t) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{\tilde{Y}_i \le t\}, \quad H_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Z_i \le t\}, \quad t \ge 0,$$

respectively. Denote by $\tau_G = \sup\{t \in \mathbb{R} : G(t) < 1\}$ the right endpoint of G. Since

$$F(t) = 1 - \frac{1 - H(t)}{1 - G(t)}, \quad t < \tau_G,$$

a natural estimator of F is

$$F_{n,m}(t) = 1 - \frac{1 - H_n(t)}{1 - G_m(t)}.$$

The right-hand side is well defined for $t < \max_{i=1,\dots,m} \tilde{Y}_i$, otherwise we put $F_{n,m}(t) = 1$, i.e.

$$F_{n,m}(t) = 1 - \frac{1 - H_n(t)}{1 - G_m(t)} \mathbf{1}\{G_m(t) < 1\}, \quad t \ge 0. \tag{1}$$

The function $F_{n,m}(t)$ is piecewise constant and has jumps in the points $\tilde{Y}_1, \dots, \tilde{Y}_m$ and $Z_1, \dots, Z_n$. Note that the jumps may be negative, it means

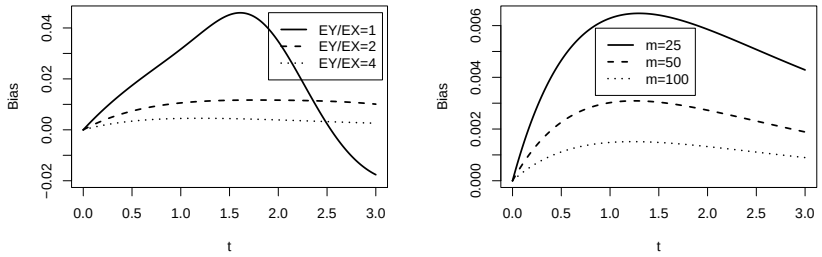


FIGURE 1. The bias of $\bar{F}_{n,m}(t)$ as the function of t for $X \sim \text{Exp}(1)$ and $Y \sim \text{Exp}(\lambda_G)$. Left: $m = 25$ and $\lambda_G \in \{1, 1/2, 1/4\}$. Right: $m \in \{25, 50, 100\}$ and $\lambda_G = 1/3$. The bias does not depend on n.

that $F_{n,m}$ does not have to be a distribution function. In the case $n = m$ we will use short notation $F_n = F_{n,n}$. If G is assumed to be known (corresponds to the limiting case $m \to \infty$), then we have

$$F_{n,\infty}(t) = 1 - \frac{1 - H_n(t)}{1 - G(t)}, \quad t < \tau_G, \tag{2}$$

which is an unbiased estimator of $F(t)$.

It will be useful to work with the survival functions. If $\bar{G} = 1 - G$ and $\bar{H} = 1 - H$, then the empirical survival functions

$$\bar{G}_m(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\tilde{Y}_i > t\}, \quad \bar{H}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_i > t\}, \quad t \geq 0,$$

are unbiased estimators of $\bar{G}(t)$ and $\bar{H}(t)$, respectively. Moreover, by rewriting (1),

$$\bar{F}_{n,m}(t) = 1 - F_{n,m}(t) = \frac{\bar{H}_n(t)}{\bar{G}_m(t)} \mathbf{1}\{\bar{G}_m(t) > 0\}, \quad t \geq 0,$$

is an estimator of $\bar{F}(t) = 1 - F(t)$. For $t < \max_{i=1,\ldots,m} \tilde{Y}_i$, it is a ratio of two empirical survival functions. Obviously, this is not an unbiased estimator, but it is asymptotically unbiased. Figure 1 shows the bias $\mathbb{E}\bar{F}_{n,m}(t) - \bar{F}(t)$ as the function of t for exponentially distributed generic random variables X and Y.

Theorem 1. *For any $t < \tau_G$ and any integer n, we have*

$$\mathbb{E}\bar{F}_{n,m}(t) \overset{m\to\infty}{\longrightarrow} \bar{F}(t).$$

Proof. Using independence of $\{Z_i\}$ and $\{\tilde{Y}_i\}$, we get

$$\mathbb{E}\bar{F}_{n,m}(t) = \mathbb{E}\bar{H}_n(t)\mathbb{E}\frac{\mathbf{1}\{\bar{G}_m(t) > 0\}}{\bar{G}_m(t)} = \bar{H}(t)\sum_{k=1}^{m}\frac{m}{k}\binom{m}{k}\bar{G}(t)^k G(t)^{m-k}.$$

Corollary 1 in [6] gives the asymptotic expansion of the sum. Consequently,

$$\mathbb{E}\bar{F}_{n,m}(t) = \bar{F}(t)\left(1 + \frac{G(t)}{m\bar{G}(t)} + O(\frac{1}{m^2})\right). \tag{3}$$

□

Next, we express the asymptotic variance.

Theorem 2. *Assume that there exists finite and positive λ such that $m/n \to \lambda$ as $m \to \infty$. Then*

$$m \operatorname{var} \bar{F}_{n,m}(t) \overset{m\to\infty}{\longrightarrow} \frac{\bar{F}(t)}{\bar{G}(t)}\left(\lambda H(t) + \bar{F}(t)G(t)\right)$$

for any $t < \tau_G$.

Proof. The independence of $\{Z_i\}$ and $\{\tilde{Y}_i\}$ yields

$$\begin{aligned}\mathbb{E}\bar{F}_{n,m}(t)^2 &= \mathbb{E}\bar{H}_n(t)^2\mathbb{E}\frac{\mathbf{1}\{\bar{G}_m(t) > 0\}}{\bar{G}_m(t)^2} \\ &= \left(\frac{1}{n}\bar{H}(t)H(t) + \bar{H}(t)^2\right)\sum_{k=1}^{m}\frac{m^2}{k^2}\binom{m}{k}\bar{G}(t)^k G(t)^{m-k}.\end{aligned}$$

The sum can be expressed by Corollary 1 in [6] as

$$\sum_{k=1}^{m}\frac{m^2}{k^2}\binom{m}{k}\bar{G}(t)^k G(t)^{m-k} = \frac{1}{\bar{G}(t)^2}\left(1 + \frac{3G(t)}{m\bar{G}(t)} + O(\frac{1}{m^2})\right).$$

Combining this with (3), we obtain

$$\operatorname{var}\bar{F}_{n,m}(t) = \frac{\bar{F}(t)H(t)}{n\bar{G}(t)} + \frac{\bar{F}(t)^2 G(t)}{m\bar{G}(t)} + O(\frac{1}{mn}) + O(\frac{1}{m^2}). \tag{4}$$

□

Since $H(t) \geq F(t)$ and $0 \leq G(t) \leq 1$, the asymptotic variance is always larger or equal to $\lambda F(t)\bar{F}(t)$. If $0 < F(t) < 1$, then this lower bound is attained if and only if $G(t) = 0$. In this case the Y_i do not influence the observations, all observed minima Z_i are equal to X_i.

As a consequence of (3) and (4), the estimator $\bar{F}_{n,m}$ is L^2-consistent. By the strong law of large numbers, it is also strongly consistent.

Theorem 3. *The estimator $\bar{F}_{n,m}(t)$ is both L^2-consistent and strongly consistent estimator of $F(t)$ for any $t < \tau_G$.*

Proof. From (3) and (4) we see that

$$\mathbb{E}\left(\bar{F}_{n,m}(t)-\bar{F}(t)\right)^2=\frac{\bar{F}(t)H(t)}{n\bar{G}(t)}+\frac{G(t)\bar{F}(t)^2}{m\bar{G}(t)}+O(\frac{1}{mn})+O(\frac{1}{m^2})$$

as both m and n go to ∞.

The strong law of large numbers implies

$$H_n(t)\overset{n\to\infty}{\longrightarrow}H(t)\quad\text{a.s.,}\quad\text{and}\quad G_m(t)\overset{m\to\infty}{\longrightarrow}G(t)\quad\text{a.s.}$$

Therefore,

$$F_{n,m}(t)\overset{n,m\to\infty}{\longrightarrow}F(t)\quad\text{a.s.,}$$

i.e. $F_{n,m}(t)$ is strongly consistent estimator of $F(t)$. □

In the remainder of this section we consider the case $n=m$ and study weak convergence of the empirical process given by

$$Y_n(t)=\sqrt{n}\left(F_n(t)-F(t)\right)=\sqrt{n}\left(\bar{F}(t)-\bar{F}_n(t)\right),\quad t\in[0,\delta], \tag{5}$$

where $0<\delta<\tau_G$ is a fixed constant.

Theorem 4. *The empirical process* (5) *converges weakly in the Skorohod space* $D[0,\delta]$ *to the zero mean Gaussian process* $\{Y(t),t\in[0,\delta]\}$ *with covariance function*

$$\mathbb{E}Y(s)Y(t)=\frac{\bar{F}(s\vee t)}{\bar{G}(s\wedge t)}H(s\wedge t)+\frac{G(s\wedge t)}{\bar{G}(s\wedge t)}\bar{F}(s)\bar{F}(t),\quad s,t\in[0,\delta],$$

where $s\wedge t=\min(s,t)$ *and* $s\vee t=\max(s,t)$.

Proof. We can rewrite $Y_n(t)$ as

$$\begin{aligned}
Y_n(t)&=\sqrt{n}\left(\bar{F}(t)-\bar{F}_n(t)\right)=\sqrt{n}\left(\frac{\bar{H}(t)}{\bar{G}(t)}-\frac{\bar{H}_n(t)}{\bar{G}_n(t)}\right)\\
&=\sqrt{n}\left(\frac{\bar{H}(t)}{\bar{G}(t)}-\frac{\bar{H}(t)}{\bar{G}_n(t)}+\frac{H_n(t)-H(t)}{\bar{G}_n(t)}\right)\\
&=\sqrt{n}\left(\frac{\bar{F}(t)(G(t)-G_n(t))}{\bar{G}_n(t)}+\frac{H_n(t)-H(t)}{\bar{G}_n(t)}\right)\\
&=\frac{\sqrt{n}}{\bar{G}_n(t)}\left[\bar{F}(t)(G(t)-G_n(t))+H_n(t)-H(t)\right],
\end{aligned}$$

which, by the Slutsky lemma, has the same weak limit in the Skorohod space $D[0,\delta]$ as

$$\frac{\sqrt{n}}{\bar{G}(t)}\left[\bar{F}(t)(G(t)-G_n(t))+H_n(t)-H(t)\right],\quad t\in[0,\delta]. \tag{6}$$

It is well-known that the process $\sqrt{n}\left(G_n(t)-G(t)\right)$ converges weakly in the space $D[0,\delta]$ to a Brownian bridge $W_G^0(t)$ which is given by the covariance function $\mathbb{E}W_G^0(s)W_G^0(t)=G(s\wedge t)-G(s)G(t)$, see Theorem 14.3 in [1]. Similarly, $\sqrt{n}\left(H_n(t)-H(t)\right)$ converges weakly in $D[0,\delta]$ to $W_H^0(t)$ with covariance function $\mathbb{E}W_H^0(s)W_H^0(t)=H(s\wedge t)-H(s)H(t)$. Therefore, using

independence of G_n and H_n, the weak limit of (6) (and consequently also of $\{Y_n(t), t \in [0,\delta]\}$) is

$$Y(t) = \frac{1}{\bar{G}(t)} W_H^0(t) - \frac{\bar{F}(t)}{\bar{G}(t)} W_G^0(t), \quad t \in [0,\delta],$$

where W_G^0 and W_H^0 are independent Brownian bridges. □

In particular, $F_n(t)$ is asymptotically normal for any $t < \tau_G$.

3. **Violation of distribution function properties**

The disadvantage of the estimator $F_{n,m}$ is that it is not necessarily monotone and may take negative values (see Figure 2 left). Therefore, the following modification

$$\tilde{F}_{n,m}(t) = 1 - \frac{1 - \inf_{s \geq t} F_{n,m}(s)}{1 - \inf_{s \geq 0} F_{n,m}(s)}, \quad t \geq 0,$$

was introduced in [3]. In the case $n = m$ we write shortly $\tilde{F}_n = \tilde{F}_{n,m}$. Figure 2 shows the estimates $F_n(t)$ and $\tilde{F}_n(t)$ computed from $n = 25$ observations of exponentially distributed Z_i and $m = 25$ observations of exponentially distributed $\tilde{Y}_i$. Both estimators have jumps at the same points.

In the case of known G, the estimator $F_{n,\infty}$, given by (2), still may be non-monotone and may attain negative values (see Figure 2 right). If G is increasing, then $F_{n,\infty}$ is decreasing on the intervals $(Z_{(i)}, Z_{(i+1)})$, where $Z_{(1)} \leq \cdots \leq Z_{(n)}$ is the order statistics of $Z_1, \ldots, Z_n$. We define the modified estimator

$$\tilde{F}_{n,\infty}(t) = 1 - \frac{1 - \inf_{s \geq t} F_{n,\infty}(s)}{1 - \inf_{s \geq 0} F_{n,\infty}(s)}, \quad t \geq 0,$$

which is already non-decreasing and non-negative function of t.

If $G_m(t) = k/m$, $k = 0, \ldots, m-1$, and $H_n(t) = l/n$, $l = 0, \ldots, n$, then $F_n(t) = (lm - kn)/m(n-k)$. Hence, the distribution of $F_{n,m}$ is given by the probabilities

$$\begin{aligned}\mathbb{P}(F_{n,m}(t) = j) &= \sum_{k,l: j = \frac{lm-kn}{m(n-k)}} \mathbb{P}(G_m(t) = k/m)\mathbb{P}(H_n(t) = l/n) \\ &= \sum_{k,l: j = \frac{lm-kn}{m(n-k)}} \binom{m}{k} G(t)^k \bar{G}(t)^{m-k} \binom{n}{l} H(t)^l \bar{H}(t)^{n-l}\end{aligned}$$

for $j \in \mathbb{Q}$.

We can ask for the probability that the estimator is negative. Let

$$P_{n,m}(t) = \mathbb{P}\left(F_{n,m}(t) < 0\right), \quad t \geq 0.$$

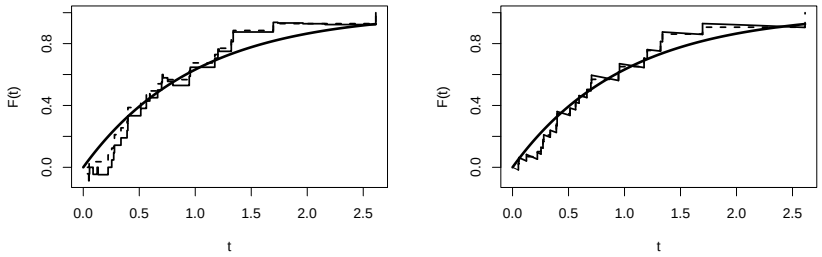


FIGURE 2. The estimators obtained from realizations of Z_i, $\tilde{Y}_i$, $i = 1, \ldots, 25$, where $Z_i \sim \mathrm{Exp}(4/3)$ and $\tilde{Y}_i \sim \mathrm{Exp}(1/3)$. Left: the estimates $F_n(t)$ (solid line) and $\tilde{F}_n(t)$ (dashed line). Right: the estimates $F_{n,\infty}$ (solid line) and $\tilde{F}_{n,\infty}$ (dashed line). For comparison, also theoretical distribution function F (bold) is shown in both cases.

Obviously,

$$\begin{aligned} P_{n,m}(t) &= \mathbb{P}\left(H_n(t) < G_m(t)\right) \\ &= \sum_{k,l:\, l/n<k/m} \binom{m}{k} G(t)^k \bar{G}(t)^{m-k} \binom{n}{l} H(t)^l \bar{H}(t)^{n-l}. \end{aligned}$$

Figure 3 shows $P_{n,n}(t)$ as the function of t for exponentially distributed random variables. Let t_n be such t at which the maximum of $P_{n,n}(t)$ is attained. For $n = 25$, $X \sim \mathrm{Exp}(1)$, and $Y \sim \mathrm{Exp}(\lambda_G)$ the values of t_n are becoming larger with increasing $\mathbb{E}Y/\mathbb{E}X$, see Figure 3 left. Specifically, we have $t_{25} \doteq 0.0289$ for $\lambda_G = 1$, $t_{25} \doteq 0.0328$ for $\lambda_G = 1/2$, and $t_{25} \doteq 0.0358$ for $\lambda_G = 1/4$. For $X \sim \mathrm{Exp}(1)$ and $Y \sim \mathrm{Exp}(1/3)$ with increasing n the values of t_n decrease while the maximal probabilities slowly increase, see Figure 3 right. In particular, we get $P_{25,25}(t_{25}) \doteq 0.09201$, $P_{50,50}(t_{50}) \doteq 0.09214$, and $P_{100,100}(t_{100}) \doteq 0.09220$.

In the case of known G, the probability

$$P_{n,\infty}(t) = \mathbb{P}\left(F_{n,\infty}(t) < 0\right) = \mathbb{P}\left(H_n(t) < G(t)\right) = \sum_{l<nG(t)} \binom{n}{l} H(t)^l \bar{H}(t)^{n-l}$$

is shown in Figure 4 as the function of t for exponential distribution of X and Y. For $0 < t < \min_{i=1,\ldots,n} Z_i$, $H_n(t) = 0$ while $G(t) > 0$, this causes $F_{n,\infty}(t) < 0$. Therefore, $P_{n,\infty}(t)$ is close to 1 for small t.

Further, we consider the probability that the monotonicity is violated. Let

$$Q_{n,m}(t_1, t_2) = \mathbb{P}\left(F_{n,m}(t_1) > F_{n,m}(t_2)\right), \quad t_1 < t_2.$$

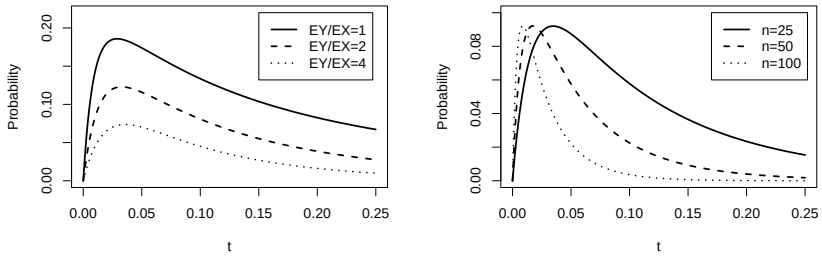


FIGURE 3. Probabilities $P_{n,m}(t)$ of negative estimator as the function of t for $X \sim \text{Exp}(1)$ and $Y \sim \text{Exp}(\lambda_G)$. Left: $n = m = 25$ and $\lambda_G \in \{1, 1/2, 1/4\}$. Right: $n = m \in \{25, 50, 100\}$ and $\lambda_G = 1/3$.

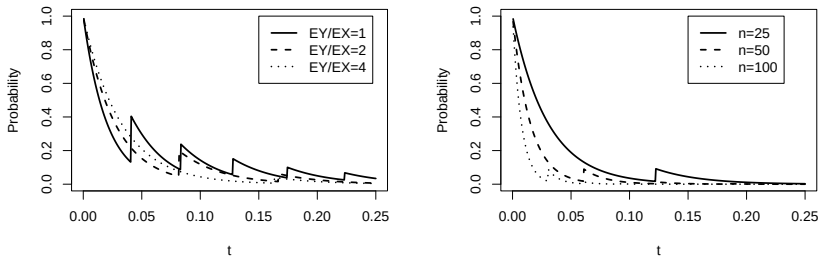


FIGURE 4. Probabilities $P_{n,\infty}(t)$ of negative estimator as the function of t for $X \sim \text{Exp}(1)$ and $Y \sim \text{Exp}(\lambda_G)$. Left: $n = 25$ and $\lambda_G \in \{1, 1/2, 1/4\}$. Right: $n \in \{25, 50, 100\}$ and $\lambda_G = 1/3$.

For $t_1 < t_2$ we see that $F_{n,m}(t_1) > F_{n,m}(t_2)$ if and only if

$$\frac{1 - H_n(t_1)}{1 - H_n(t_2)} < \frac{1 - G_m(t_1)}{1 - G_m(t_2)}. \tag{7}$$

Let us denote

$$N^H(s,t) = \sum_{k=1}^{n} \mathbf{1}\{Z_k \in (s,t]\} \quad \text{and} \quad N^G(s,t) = \sum_{k=1}^{m} \mathbf{1}\{\tilde{Y}_k \in (s,t]\}.$$

Then (7) becomes

$$\frac{N^H(t_1,\infty)}{N^H(t_2,\infty)} < \frac{N^G(t_1,\infty)}{N^G(t_2,\infty)}$$

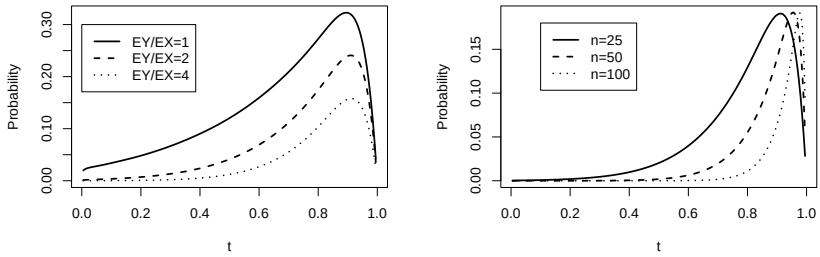


FIGURE 5. Probability that $F_n(t) > F_n(1)$ as the function of t for $X \sim \text{Exp}(1)$ and $Y \sim \text{Exp}(\lambda_G)$. Left: $n = 25$ and $\lambda_G \in \{1, 1/2, 1/4\}$. Right: $n \in \{25, 50, 100\}$ and $\lambda_G = 1/3$.

which is equivalent to

$$\frac{N^H(t_1, t_2)}{N^H(t_2, \infty)} < \frac{N^G(t_1, t_2)}{N^G(t_2, \infty)}.$$

The joint distribution of $(N^H(t_1, t_2), N^H(t_2, \infty))$ is multinomial with parameters n, $H(t_2) - H(t_1)$ and $\bar{H}(t_2)$. Similarly, the joint distribution of $(N^G(t_1, t_2), N^G(t_2, \infty))$ is multinomial with parameters m, $G(t_2) - G(t_1)$ and $\bar{G}(t_2)$. Hence, $Q_{n,m}(t_1, t_2)$ can be determined numerically.

We present the graphs of $Q_{n,n}(t, 1)$ in the case of exponential distributions with intensities $\lambda_F = 1$ and λ_G. In Figure 5 we take $n = 25$ and various values of λ_G (left) and $\lambda_G = 1/3$ and several values of n (right). Let t_n be the point at which the function $Q_{n,n}(t, 1)$ attains its maximum. It is becoming larger for larger values of $\mathbb{E}Y/\mathbb{E}X$, see Figure 5 left. Specifically, $t_{25} \doteq 0.895$ for $\lambda_G = 1$, $t_{25} \doteq 0.910$ for $\lambda_G = 1/2$, and $t_{25} \doteq 0.912$ for $\lambda_G = 1/4$. If $\lambda_G = 1/3$, then with larger n both t_n and $Q_{n,n}(t_n, 1)$ are becoming larger. The maximal probabilities are $Q_{25,25}(t_{25}) \doteq 0.1908$, $Q_{50,50}(t_{50}) \doteq 0.1922$, and $Q_{100,100}(t_{100}) \doteq 0.1924$.

4. **Simulation study**

We performed a comparative simulation study of the estimators using R [4]. We simulated exponentially distributed random variables $\tilde{Y}_1, \ldots, \tilde{Y}_m$ and $Z_1, \ldots, Z_n$, and calculated the following estimators of F: $F_{n,m}$, $\tilde{F}_{n,m}$, $F_{n,\infty}$, $\tilde{F}_{n,\infty}$, and H_n. The last three estimators require only $Z_1, \ldots, Z_n$. For $F_{n,\infty}$ and $\tilde{F}_{n,\infty}$ we assume knowledge of $G(t) = 1 - \mathrm{e}^{-\lambda_G t}$, $t \geq 0$. The quality of each estimator, say $\hat{F}_n$, was assessed by the Kolmogorov-Smirnov distance

$$d_{\text{KS}}(\hat{F}_n, F) = \sup_{t \geq 0} |\hat{F}_n(t) - F(t)|$$

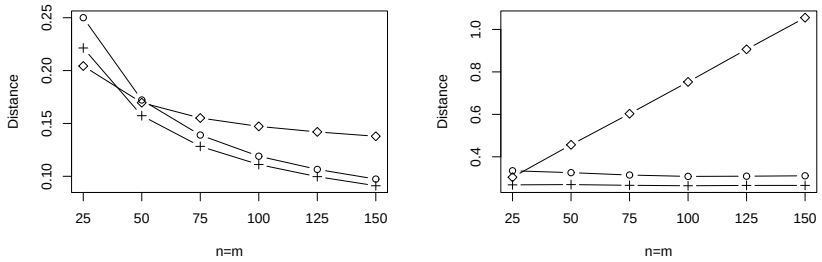


FIGURE 6. The sample means of Kolmogorov-Smirnov distances (left) and Cramér-von Mises statistics (right) computed from 10 000 simulation experiments with generic random variables $X \sim \mathrm{Exp}(1)$ and $Y \sim \mathrm{Exp}(1/3)$. The following estimators are considered: $F_{n,m}$ (circles), $F_{n,\infty}$ (crosses), and H_n (diamonds) for several choices of $n = m$.

and Cramér-von Mises statistics

$$d_{\mathrm{CvM}}(\hat{F}_n, F) = n \int_0^\infty \big(\hat{F}(t) - F(t)\big)^2 f(t)\,\mathrm{d}t,$$

where $f(t) = F'(t) = \lambda_F \mathrm{e}^{-\lambda_F t}$, $t \geq 0$, is the density function corresponding to the distribution function F. We always took $\lambda_F = 1$.

For each choice of m, n, and λ_G we generated 10 000 simulation experiments. The sample means (over 10 000 trials) of d_{KS} and d_{CvM} are shown in Figure 6 and Figure 7. Figure 6 presents the dependence on $n = m$ for $\lambda_G = 1/3$ (or equivalently, $\mathbb{E}Y/\mathbb{E}X = 3$) while in Figure 7 we have the dependence on $\mathbb{E}Y/\mathbb{E}X = \lambda_F/\lambda_G$ for $n = m = 50$.

Not surprisingly, simulations reveal that knowledge of G always improves the estimation. Of course, with misspecified G we would get worse results. We do not show the results for modified estimators $\tilde{F}_{n,m}$ and $\tilde{F}_{n,\infty}$ because they are quite similar to those for $F_{n,m}$ and $\tilde{F}_{n,\infty}$, respectively. In all cases modified estimators lead to slightly smaller d_{KS} and slightly larger d_{CvM}. We also investigated the estimator H_n which simply takes the observed minima Z_i and does not try to correct the influence of spontaneous latency (random variable with distribution function G). It performs particularly poorly for larger sample size $n = m$ and smaller $\mathbb{E}Y/\mathbb{E}X$.

References

[1] Billingsley P. (1999) *Convergence of Probability Measures*, 2nd edition, John Wiley & Sons, New York.

[2] Meister A. (2009) *Deconvolution Problems in Nonparametric Statistics*, Springer-Verlag, Berlin.

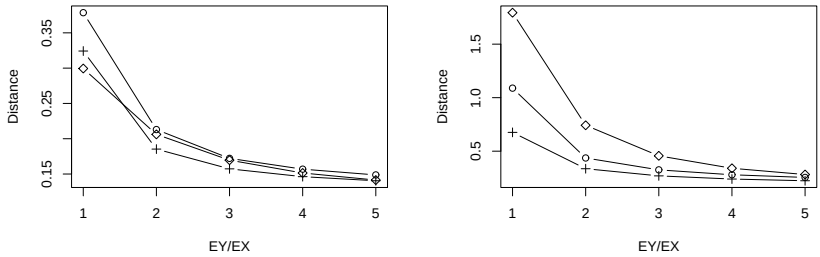


FIGURE 7. The sample means of Kolmogorov-Smirnov distances (left) and Cramér-von Mises statistics (right) computed from 10 000 simulation experiments with generic random variables $X \sim \mathrm{Exp}(1)$ and $Y \sim \mathrm{Exp}(\lambda_G)$. The estimators are based on samples of sizes $n = 50$ and $m = 50$. We consider $F_{n,m}$ (circles), $F_{n,\infty}$ (crosses), and H_n (diamonds) for several choices of λ_G.

[3] Pawlas Z., Klebanov L. B., Beneš V., Prokešová M., Popelář J. and Lánský P. (2010) *First-spike latency in the presence of spontaneous activity*, Neural Computation **22**, 1675 – 1697.

[4] R Development Core Team (2012) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org`.

[5] Tamborrino M., Ditlevsen S. and Lansky P. (2012) *Identification of noisy response latency*, Physical Review E **86**, 021128.

[6] Wuyungaowa and Wang T. (2008) *Asymptotic expansions for inverse moments of binomial and negative binomial*, Statistics and Probability Letters **78**, 3018 – 3022.

Address: Charles University, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Praha 8 – Karlín

E-mail: `pawlas@karlin.mff.cuni.cz`

METODY ODHADU EXTRÉMNÍCH SRÁŽEK A JEJICH APLIKACE V ČR

Jan Picek[1], **Jan Kyselý**[1,2], **Ladislav Gaál**[2,3]

Klíčová slova: Extrémní srážky; odhad návrhové hodnoty; pravděpodobná maximální srážka; regionální frekvenční analýza

Abstrakt:
Příspěvek shrnuje koncepce a metodické postupy, na jejichž základě byla řešena problematika odhadu pravděpodobností extrémních srážek a návrhových hodnot v ČR v nedávném období. Tyto práce byly inspirovány jak výskytem několika mimořádných srážkových událostí provázených povodněmi s rozsáhlými materiálními škodami, tak pokroky v oblasti statistického modelování extrémů. Metody, které představujeme, lze rozdělit do dvou hlavních celků: metody odhadu tzv. pravděpodobné maximální srážky a metody odhadu návrhových srážek na základě různých alternativ regionální frekvenční analýzy. Závěrem uvádíme některé otevřené otázky a možné směry výzkumu do budoucna.

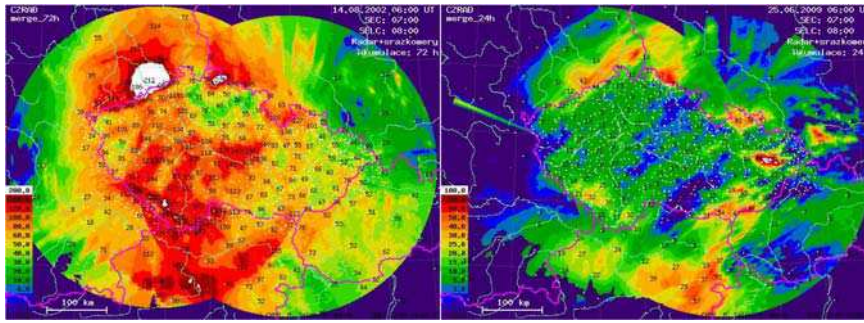
The paper summarizes concepts and methods applied in estimation of probabilities of extreme precipitation amounts and design precipitation in the Czech Republic in the recent past. This work was motivated by the occurrence of several extraordinary precipitation events that resulted in floods with enormous material damages, but also by advances in the field of statistical modelling of extremes. The methods that we introduce are split in two main parts: methods for estimation of probable maximum precipitation and methods for estimation of design precipitation using different variants of the regional frequency analysis. In the concluding section several open issues and possible directions of follow-up research and improvements are outlined.

1. Úvod

Extrémní srážkové jevy jsou provázeny velkými dopady na lidskou společnost i přírodní prostředí a proto je odhadování pravděpodobnostních rozdělení vysokých srážkových úhrnů na základě pozorovaných dat nebo jejich změn v simulacích klimatických modelů věnována značná pozornost. V podmínkách střední Evropy jsou hlavním negativním projevem srážkových extrémů (dále SE) povodně. Ty mohou postihovat rozsáhlá území a trvat mnoho dní (např. povodně v červenci 1997 na Moravě a ve Slezsku a v srpnu 2002 v Čechách – v obou případech se jednalo o největší plošně rozsáhlé povodně na daném území přinejmenším od počátku 20. století), nebo relativně menší povodí s rychlejším nástupem i následným poklesem průtoků (tzv. bleskové povodně, např. v červenci 1998 v podhůří Orlických hor nebo v červnu a červenci 2009 na více místech ČR). Pro ilustraci dopadů lze uvést, že např. během povodně v červenci 1997 zahynulo 49 osob, bylo zničeno přes 2000 domů (dalších více

než 5000 se stalo dlouhodobě neobyvatelnými) a celkové škody byly odhadnuty na 63 mld. Kč, což je více než dvojnásobek stávajícího státního rozpočtu na VaV.

Z výše uvedeného je rovněž zřejmé, že SE mohou mít více příčin a projevů. Může se jednat o prostorově rozsáhlé intenzivní srážky trvající několik dnů (Obr. 1 vlevo), nebo prostorově i časově mnohem omezenější srážkové události (Obr. 1 vpravo). Z meteorologického hlediska souvisí první typ s atmosférickými frontami a oblačnými pásy spojenými s tlakovými níženi, zatímco druhý s oblačností konvekčního původu, která s frontami nemusí (ale může) souviset. Tato typologie je navíc velmi schematická a oba základní „typy“ SE se mohou prolínat. Prostorová proměnlivost srážkových úhrnů je zejména v případě velkoprostorových událostí určena jak konkrétním charakterem polí zejména tlaku a vlhkosti v atmosféře, tak i orografií a konfigurací terénu vůči převládajícímu proudění.



Obr. 1: *Příklady prostorového rozložení srážkových úhrnů souvisejících s povodňovými situacemi – kombinovaná srážkoměrná a radarová data. Vlevo: 3-denní srážkové úhrny 11.–13. 8. 2002 (velkoměřítkové povodně v Čechách). Vpravo: 1-denní srážkové úhrny 24. 6. 2009 (blesková povodeň na Novojičínsku). Zdroj dat: ČHMÚ (P. Novák, M. Šálek).*

V souvislosti se změnou klimatu (která už je často označována za „probíhající“, nikoli „budoucí“) lze navíc na základě teoretických úvah (Trenberth a kol. [1]) i modelových simulací (např. Hanel a Buishand [2], Kyselý a kol. [3]) předpokládat častější a intenzivnější SE v teplejším klimatu, v důsledku obecně zesíleného hydrologického cyklu. Je proto přirozené, že odhady pravděpodobnosti pozorovaných SE (jaká je perioda opakování daného srážkového úhrnu), odhady návrhových hodnot SE (odpovídajících zvolené n-letosti) a jejich změn do možné budoucnosti (reprezentované různými odhady vývoje společnosti) jsou tématy, která mají význam i mimo akademickou rovinu.

Pod pojmem „návrhová srážka“ se rozumí srážkový úhrn za daný časový interval (od minut přes hodiny po dny) odpovídající určité periodě opakování (n-letosti), např. 50 let (častá v „teoretických“ pracích) nebo 150 let (využívaná pro návrhy vodohospodářských staveb [4]). Návrhové srážky se používají

pro účely plánování, výstavby, provozu apod. ve vodním hospodářství a stavebnictví [5]. Ke stanovení návrhových srážek se aplikuje široká paleta nástrojů matematické statistiky s různou komplexností, od jednoduchých modelů (např. lokální odhad návrhové hodnoty na základě proložení zvolené distribuční funkce – nejčastěji Gumbelova rozdělení nebo zobecněného rozdělení extrémních hodnot GEV – výběrovým souborem tvořeným např. ročními maximy) až po pokročilé matematické modely (různé modely regionální frekvenční analýzy zohledňující prostorovou strukturu výběrových souborů nebo trendy v důsledku klimatické změny, Khaliq a kol. [6]; kaskádové modely škálování intenzit srážek, Veneziano a kol. [7]; prostorové mnohorozměrné modely extrémů, Davison a kol. [8] atd.). Prezentace takto komplexního metodického přehledu přesahuje rámec tohoto příspěvku, jehož záběr omezíme na nedávné aplikace uvedených metod na odhady extrémních srážek na území ČR. Tyto práce byly limitovány rovněž dostupnými výběrovými soubory reprezentujícími pozorovaná data, ve všech případech se jednalo o analýzy denních nebo vícedenních SE. Analýza krátkodobých SE (tj. úhrnů za dobu kratší než 24 hodin) pomocí jiných než „jednoduchých“ modelů (viz výše) se začíná řešit teprve v současné době.

Příspěvek vznikl se záměrem shrnout koncepce a metodické postupy, na jejichž základě byla řešena problematika odhadu extrémních srážek v ČR v nedávném období, a uvést některé možnosti budoucího vývoje v této oblasti. Metody, které budou dále představeny, lze rozdělit do dvou hlavních tematických celků: první tvoří metody odhadu tzv. pravděpodobné maximální srážky (část 2.1), druhým okruhem je odhad návrhových srážek na základě různých alternativ regionální frekvenční analýzy (část 2.2). V části 3 uvádíme některé otevřené otázky a možné směry výzkumu do budoucna.

2. **Přehled metod a vybraných výsledků**

2.1. **Pravděpodobná maximální srážka**

Návrhové hodnoty hydrologických proměnných pro dimenzování vodohospodářských staveb v ČR byly až do konce 20. století odhadovány na základě návrhové srážky odpovídající době opakování 150 let. Mimořádné povodňové události v roce 1997 (povodí Odry a Moravy) a 2002 (povodí Labe) jasně ukázaly potřebu revidovat běžně používané metodické postupy, mj. proto, že 150-leté návrhové hodnoty byly na mnoha místech výrazně překročeny, především u 2- až 5-denních úhrnů (Řezáčová a kol. [9]). Jednou z otázek, která byla po povodních 1997 v souvislosti s problematikou extrémních povodňových situací řešena, byl odhad tzv. pravděpodobné maximální srážky (*probable maximum precipitation*, PMP) na území ČR.

Podle manuálu Světové meteorologické organizace (WMO [10]) je PMP definována jako maximální fyzikálně možný úhrn atmosférických srážek pro oblast dané velikosti v dané geografické poloze, v daném období roku a za

daný časový interval (např. 1 hodina, 1 den, 5 dní). Pro odhad PMP existuje více metod, které jsou vesměs zatíženy určitými subjektivními volbami a nelze definovat univerzálně platný „standardní“ způsob výpočtu PMP (WMO [10], Casas a kol. [11]).

Pro bodový odhad PMP na území ČR se aplikovaly dvě metody v závislosti na časovém intervalu srážek. PMP pro dobu kratší než 1 den byly určovány na základě tzv. modelu srážek (*storm model*; např. Collier a Hardaker [12]). Tento postup je založený na fyzikální parametrizaci srážkových procesů (vliv energie dostupné v přízemní vrstvě, vliv orografie na vznik vzestupných pohybů, vliv konvergence) a následné objektivní maximalizaci jejích složek (Řezáčová a kol. [9]). Odhady PMP pro intervaly 1 den a delší byly naproti tomu získány pomocí statistických postupů, které doporučuje WMO [10]. Na základě tohoto přístupu je PMP pro danou lokalitu a trvání t vyjádřena vztahem

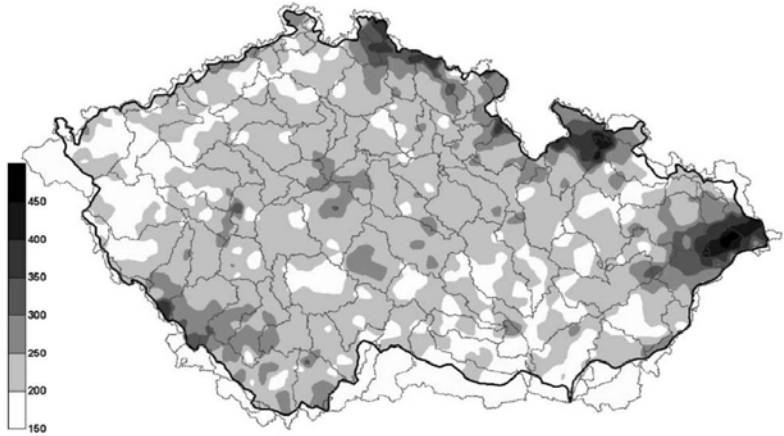
$$PMP(t) = X(t) + K(t)\sigma(t)$$

kde X (σ) je střední hodnota (směrodatná odchylka) maximálních ročních (příp. sezónních) úhrnů v dané lokalitě a K je tzv. parametr extrémnosti, pro jehož odhad obsahuje manuál WMO [10] doporučené postupy. Původní metodika WMO byla modifikována zohledněním klimatických podmínek ČR pro určení parametru K a pomocí korekcí pro odlehlé hodnoty ve výběrových souborech (Řezáčová a kol. [4]), což nutně vnáší do odhadů PMP určitou míru subjektivity.

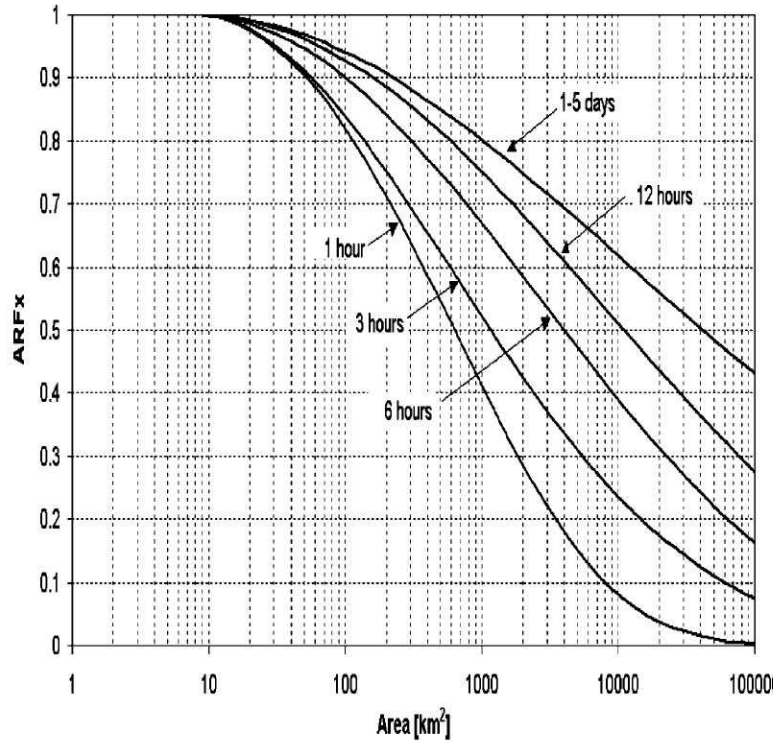
Aplikací výše uvedených postupů byly získány bodové odhady PMP v síti s horizontálním krokem 1 km, která pokrývá území ČR (Řezáčová a kol. [9]). Bodové odhady PMP (Obr. 2) však není možné přímo využít např. pro dimenzování vodohospodářských staveb, je nezbytné je transformovat na plošné hodnoty. Bodové úhrny srážek se obecně konvertují na plošné úhrny prostřednictvím tzv. plošného redukčního faktoru (*area reduction factor*, ARF), který kromě trvání t závisí i na ploše uvažovaného území S, resp. pravděpodobnosti překročení (Sivapalan a Blöschl [13]). V případě extrémů jako je PMP se uvažuje maximální hodnota plošného redukčního faktoru ARFx, která je funkcí plochy a trvání:

$$PMP(S,t) = PMP(t) \times ARFx(S,t)$$

Pro určení hodnot ARFx byla v práci Řezáčové a kol. [4] využita radarová měření odrazivosti pole srážek (snímky v kroku 10-minut, rozlišení 2x2 km, 256x256 pixelů, období 1996–2001). Jednotlivé kroky analýzy, od transformace maximální hodnoty odrazivosti na srážkové intenzity, přes analýzu 2D pole pixelů s cílem určit hodnoty ARF(S,t) až po konstrukci obalové křivky pro získané hodnoty ARF(S,t) jsou podrobně popsané v [4]. Výsledkem analýzy radarových měření bylo analytické a grafické vyjádření závislosti maximální hodnoty plošného redukčního faktoru ARFx na ploše území a trvání srážek (Obr. 3).



Obr. 2: *Prostorové rozložení bodových odhadů denní pravděpodobné maximální srážky (PMP, v mm) získané interpolací hodnot PMP na srážkoměrných stanicích v ČR (převzato z Řezáčová a kol. [4]).*



Obr. 3: *Závislost maximální hodnoty plošného redukčního faktoru (ARFx) na ploše území. Křivky odpovídají různým trváním srážek (převzato z Řezáčová a kol. [4]).*

Výsledkem výše uvedeného postupu bylo kvantitativní vyhodnocení 1- až 5-denních srážkových úhrnů, které vedly k povodním v letech 1997 a 2002, vyjádřené v procentech plošných hodnot PMP na povodích 3. stupně pokrývajících území ČR (Řezáčová a kol. [4]).

2.2. **Regionální frekvenční analýza**

Cílem frekvenční analýzy srážkových úhrnů je odhadnout četnosti (pravděpodobnosti) srážkových úhrnů za daný časový interval (zde se zabýváme 1-denními a vícedenními úhrny) na určité stanici nebo území. V případě metody blokových maxim, na kterou se v tomto přehledu omezujeme, jde u frekvenční analýzy o proložení zvolené distribuční funkce (DF) souborem ročních nebo sezónních maxim (u metody peaks-over-threshold souborem nadprahových hodnot) a odhad parametrů této DF. Z plně určené DF lze stanovit kvantily rozdělení (tj. návrhové srážky odpovídající danému kvantilu, resp. n-letosti) a pravděpodobnosti srážkových úhrnů (obvykle nás zajímá oblast pravého chvostu rozdělení, tj. vysoké úhrny, malé pravděpodobnosti překročení, velké doby opakování). Univerzální „standardní“ realizace výše uvedeného postupu však neexistuje a v literatuře lze najít mnoho alternativ frekvenční analýzy, lišících se např. v konstrukci výběrového souboru (lokální/regionální data), volbě DF, metodě odhadu parametrů DF, metodě stanovení neurčitosti odhadnutých kvantilů a návrhových hodnot atd. (např. Gaál [14]).

Tradiční variantou frekvenční analýzy je lokální přístup, kdy výběrový soubor tvoří pouze údaje z místa pozorování. V případě srážkových dat, pro která je délka dostupných řad na konkrétních stanicích typicky jen několik desetiletí (zatímco úkolem frekvenční analýzy je odhadnout kvantily rozdělení odpovídající i výrazně vyšší n-letosti), je tato metoda ve velké míře zatížena výběrovou proměnlivostí pramenící z velké prostorové proměnlivosti SE. Například srážková událost, která vyvolala bleskovou povodeň na Odře na Novojičínsku 24. 6. 2009 (10 obětí na životech, velké materiální škody), byla charakteristická velmi malým prostorovým rozsahem příčinné srážkové události (Obr. 1 vpravo) – stanice Bělotín zaznamenala denní úhrn 123,8 mm (z toho 114 mm během 3 hodin), čímž byl překonán předchozí maximální denní úhrn z r. 1967 téměř o 50 mm, zatímco na některých stanicích ve vzdálenosti pouhých 50 km byly srážkové úhrny do 1 mm (Kyselý a kol. [15]). Jedná se přitom o oblast, která není výrazně exponována z hlediska orografie a kde je prostorová proměnlivost statistických charakteristik SE dána především výběrovými efekty.

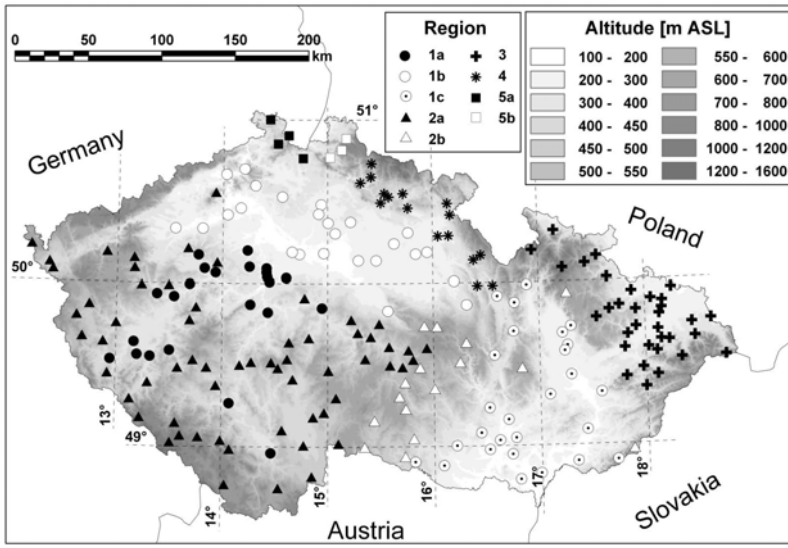
Na rozdíl od lokálních metod je regionální přístup založen na současné analýze výběrových souborů z více stanic. Základním požadavkem regionální frekvenční analýzy je to, aby stanice tvořily homogenní region, tj. normovaná rozdělení četnosti byla pro celý region (všech N stanic) totožná. V praxi se vyžaduje dostatečná podobnost empirických DF, která se posuzuje pomocí

testů regionální homogenity založených na L-momentech (ekvivalenty konvenčních modelů založené na pořádkových statistikách a široce používané v regionální frekvenční analýze), např. tzv. H-testu (Hosking a Wallis [16]) nebo $X10$-testu (Lu a Stedinger [17]). Při splnění předpokladu regionální homogenity platí vztah

$$Q_i(n) = \mu_i q(n),\ i = 1, \ldots, N,$$

kde $Q_i(n)$ je návrhová hodnota odpovídající době opakování n na stanici i, $q(n)$ je bezrozměrná regionální kvantilová funkce (růstová křivka), která je společná pro všechny stanice v regionu, a μ_i je indexová hodnota pro stanici i, kterou je nejčastěji průměr výběrového souboru. „Klasickou“ koncepci regionální frekvenční analýzy podle Hoskinga a Wallise [16], která je založená na pevně vymezených regionech, se pokusili pro území ČR adaptovat Kyselý a Picek [18]. Na základě tehdy dostupných dat ze 78 srážkoměrných stanic (1961–2000) byly vymezeny 2 větší a 2 menší homogenní regiony pro roční maxima 1- až 7-denních srážkových úhrnů, byly odhadnuty regionální růstové křivky a návrhové hodnoty pro vybrané n-letosti spolu s příslušnými intervaly spolehlivosti. Frekvenční analýza poukázala na zvláštnost regionu severní Moravy a Slezska (oblast Jeseníků a Moravskoslezských Beskyd), kde jednak z testu dobré shody vycházelo jako vhodnější zobecněné logistické rozdělení místo „standardního“ GEV rozdělení, které bylo plně vyhovující v ostatních regionech, jednak testy na hodnotu indexu chvostu rozdělení (Jurečková a Picek [19]) naznačily, že vícedenní úhrny srážek zde mohou mít rozdělení s natolik těžkými chvosty, že pro ně L-momenty neexistují a celá frekvenční analýza potom může být nekorektní a odhady vychýlené.

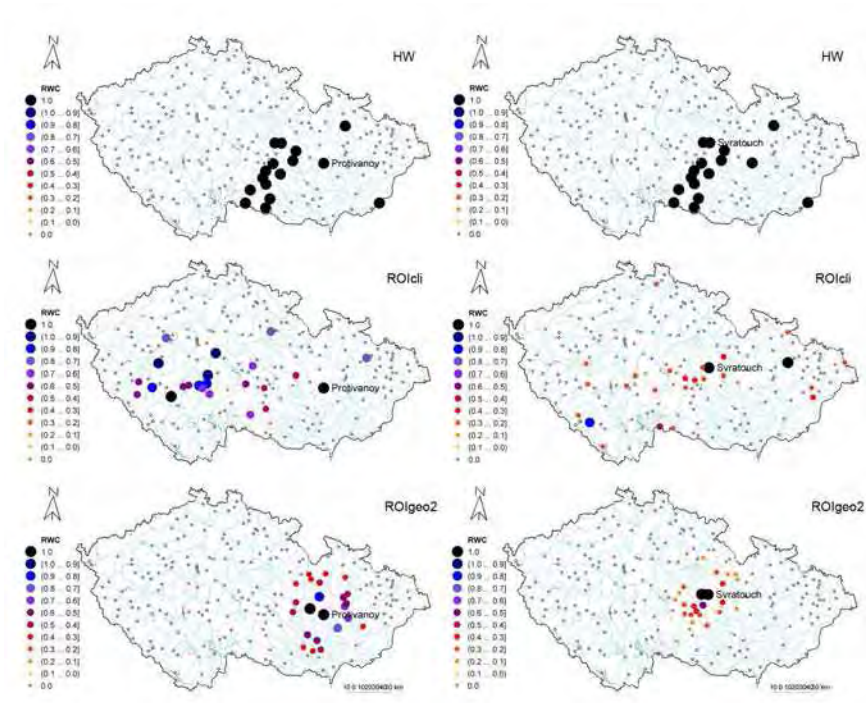
Po získání údajů z dalších srážkoměrných stanic v ČR (celkem 209) a doplnění dat do r. 2007 (celkem 9573 staničních let ve srovnání s 3120 v původní analýze) byla výše uvedená regionalizace revidována v Kyselý a kol. [15]. Ukázalo se, že oba větší regiony z původní regionalizace se staly heterogenními a bylo třeba je dále rozdělit na 3, resp. 2 menší (1a–c, 2a–b; Obr. 4). Redefinice byla nutná i pro oblast severních Čech. Jedním ze závěrů práce [15] proto bylo, že na orograficky členitém území, jakým je ČR, je obtížné vymezit homogenní regiony, které jsou robustní např. vůči doplnění dat o nová pozorování (jediným regionem z původní analýzy, který si zachoval homogenitu, byl výše diskutovaný region severní Moravy a Slezska).



Obr. 4: *Homogenní regiony pro regionální frekvenční analýzu v ČR – revidované vymezení pro 209 stanic (převzato z Kyselý a kol. [15]).*

Pokus o redukci subjektivních rozhodnutí při vytváření homogenních regionů se přitom stal již dříve motivací pro rozvoj alternativního způsobu regionální frekvenční analýzy, v němž regiony nejsou fixní, ale definují se flexibilním způsobem. Metoda oblasti vlivu (region of influence, ROI; Burn [20]; Castellarin a kol. [21]) se vyznačuje tím, že pro každou analyzovanou stanici je identifikován vlastní homogenní region, který se skládá z určitého počtu dostatečně podobných stanic. Metoda byla navržena pro hydrologická data (průtoky), práce Gaál a kol. [22] a Gaál a Kyselý [23] patřily k prvním pokusům ji modifikovat pro srážková data. Podobnost stanic se posuzuje na základě vhodně zvoleného souboru staničních atributů, u kterých se předpokládá, že mohou mít vliv na pravděpodobnostní rozdělení SE; mohou jimi být klimatologické, hydrologické, geografické nebo jiné charakteristiky, resp. jejich různé kombinace (Castellarin a kol. [21]; Gaál a kol. [22]; Gaál a Kyselý [23]; Zrinji a Burn [24]). Proces odhadu kvantilů na základě metody I se skládá z následujících kroků: (i) na základě vybraných staničních atributů se definuje míra podobnosti mezi stanicemi; (ii) pomocí regionálního testu homogenity se hledá oblast vlivu pro analyzovanou stanici; (iii) pro každou stanici v oblasti vlivu se definuje regionální váhový koeficient (RWC) a (iv) parametry DF pro analyzovanou stanici se určí jako vážený průměr odpovídajících parametrů z jednotlivých stanic v oblasti vlivu, přičemž váhami jsou hodnoty RWC (Gaál a Kyselý [23]). Postup se opakuje pro všechny studované stanice.

Obr. 5 znázorňuje základní rozdíl mezi regionální frekvenční analýzou na základě pevně vymezených a flexibilních regionů. Nezávisle na tom, která stanice je předmětem zájmu v rámci daného regionu, zůstává v Hosking-Wallisově metodě složení regionu, resp. příslušné hodnoty RWC neměnné (Obr. 5 nahoře). Na rozdíl od toho má v metodě ROI každá stanice odlišné složení homogenního regiónu a váhové koeficienty RWC nejsou jednotkové, ale proměnné (závisí na míře podobnosti s analyzovanou stanicí; Obr. 5 uprostřed a dole). (Poznámka: RWC na Obr. 5 jsou znázorněny za předpokladu, že velikost výběrového souboru, tj. délka pozorování, je na všech stanicích stejná. Tento faktor se rovněž bere do úvahy při regionálním vážení, a tak v případě rozdílných hodnot n na různých stanicích regionu by ani na Obr. 5 nahoře nebyly všechny hodnoty RWC jednotkové.)



Obr. 5: *Složení homogenních regionů a odpovídající regionální váhové koeficienty (RWC) pro 2 vybrané stanice (Protivanov, Svratouch) pro regionální frekvenční analýzu ročních maxim 1-denních srážkových úhrnů, založenou na pevně vymezených regionech – metoda Hoskinga a Wallise (HW, nahoře), resp. flexibilních regionech – metoda oblasti vlivu (ROIcli na základě klimatologických charakteristik, uprostřed; ROIgeo2 na základě geografické vzdálenosti stanic, dole).*

Gaál a Kyselý [23] se pomocí Monte Carlo simulací pokusili identifikovat kombinace klimatologických a geografických atributů v definici míry podobnosti stanic (u metody ROI), které vedou k nejlepším výsledkům (podle

střední kvadratické chyby růstových křivek pro vysoké kvantily) v relativně husté staniční síti v ČR (jedna stanice připadala na území přibližně 20 × 20 km). Podle jejich výsledků hraje pro roční maxima 1-denních úhrnů nejvýznamnější roli v určování podobnosti rozdělení pravděpodobnosti geografická vzdálenost mezi stanicemi, pro roční maxima 5-denních úhrnů (které jsou často považovány za zástupné ukazatele velkoměřítkových povodní, zejména v simulacích z klimatických modelů) vychází jako mírně lepší míra podobnosti, v níž kromě vzdálenosti vystupují (s menší vahou) vybrané klimatologické charakteristiky (např. průměrný roční úhrn srážek, podíl srážek v letním a zimním období). Tento závěr je v souladu se skutečností, že vícedenní extrémy mají silnější vazbu na typické „vzory“ režimu srážek v oblasti střední Evropy než jednodenní.

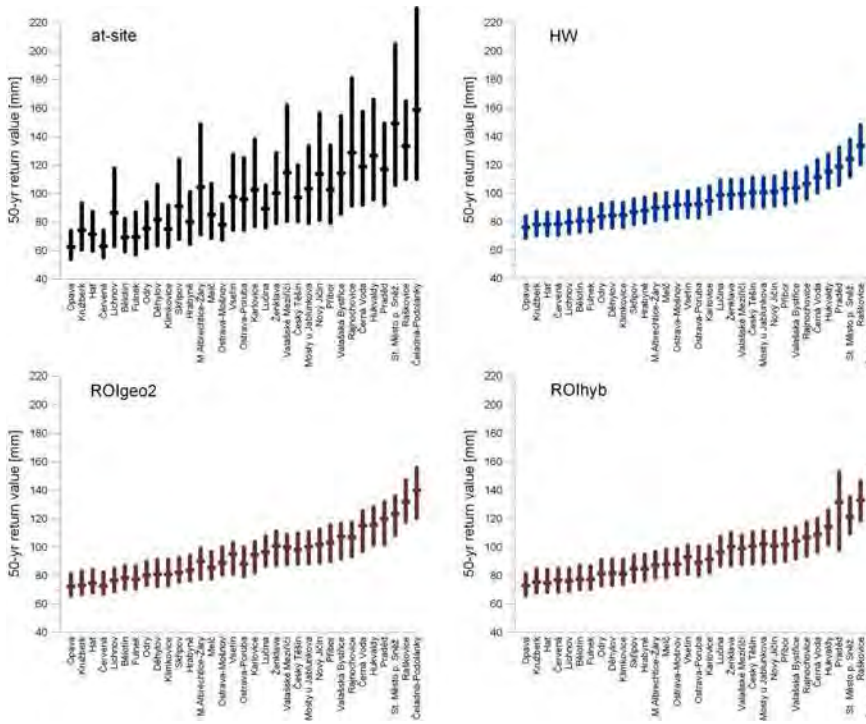
Výše zmiňovaná práce [15] se zaměřila na komplexní srovnání různých variant frekvenční analýzy: dva regionální přístupy (metoda oblasti vlivu s flexibilními regiony a Hosking-Wallisova metoda s pevně vymezenými regiony) byly pomocí Monte Carlo simulací porovnány mezi sebou a s lokálním odhadem. Jako nejvýhodnější regionální model byla vyhodnocena varianta metody I, v níž se podobnost stanic definuje na základě geografické vzdálenosti. Rozdíly mezi jednotlivými regionálními metodami jsou přitom velmi malé ve srovnání s rozdílem jakékoli regionální metody vůči lokálnímu odhadu vysokých kvantilů, který je zatížen mnohem větší střední kvadratickou chybou. Regionální metody výrazně redukují rozptyl odhadů parametrů DF a výsledných kvantilů, který pramení z náhodné výběrové proměnlivosti (Obr. 6).

Přednosti regionální frekvenční analýzy ilustrovali Kyselý a kol. [15] na příkladu odhadu doby opakování výše zmíněné srážkové události z 24. 6. 2009, která vyvolala bleskovou povodeň na Novojičínsku. Odhad doby opakování denního srážkového úhrnu 123,8 mm na stanici Bělotín na základě lokálního přístupu je nereálný, zatížený extrémní neurčitostí a zásadním způsobem závisí na tom, zda uvažovaná událost je nebo není zahrnuta do lokálního výběrového souboru (Tab. 1; odpovídá odhadům „před“ a „po“ události, která představuje výrazně odlehlou hodnotu vůči historickým datům). Naproti tomu regionální metody se shodují na řádovém odhadu doby opakování (několik set let) a její neurčitosti. Robustnost regionálních metod dokládá i skutečnost, že odhady se jen málo změní při zahrnutí uvažované mimořádné události do analýzy (Tab. 1).

Datový soubor	At-site	HW	Igeo2	Ihyb
1961–2007	45110	483	657	695
	(605 – $2{,}12\times10^9$)	(359 – 920)	(458 – 1340)	(392 – 1568)
1961–2007 +	283	353	419	415
úhrn z 24. 6. 09	(85 – 69730)	(275 – 641)	(333 – 864)	(279 – 944)

Tab. 1: *Odhady periody opakování a 90% intervalu spolehlivosti (v závorce) mimořádné srážkové události (denní úhrn 123,8 mm) z 24. 6. 2009 na stanici Bělotín, na základě 4 frekvenčních modelů a 2 různých datových souborů (bez/s rokem 2009):*

at-site = lokální analýza, HW = regionální analýza podle Hosking-Wallisovy metody (pevně vymezené regiony), ROIgeo2 = metoda oblasti vlivu, kde mírou podobnosti je skutečná geografická vzdálenost mezi stanicemi, a ROIhyb = metoda oblasti vlivu, kde mírou podobnosti je kombinace geografických a klimatologických staničních atributů.



Obr. 6: *Návrhové hodnoty ročních maxim 1-denních úhrnů srážek s dobou opakování 50 let a příslušné 90% intervaly spolehlivosti pro stanice z oblasti Jeseníků a Moravskoslezských Beskyd (severovýchodní Morava) na základě 4 frekvenčních modelů: at-site = lokální analýza, HW = regionální analýza na základě Hosking-Wallisovy metody (pevně vymezené regiony), ROIgeo2 = metoda oblasti vlivu, kde mírou podobnosti je geografická vzdálenost mezi stanicemi, a ROIhyb = metoda oblasti vlivu, kde mírou podobnosti je kombinace geografických a klimatologických staničních atributů (podrobnosti v [15]).*

3. Závěr

Z výše uvedeného je zřejmé, že metody odhadu pravděpodobné maximální srážky (PMP) a odhadu návrhových srážek (frekvenční analýza) vycházejí z odlišných a do určité míry nekompatibilních východisek a přístupů. Zatímco v případě PMP je cílem odhad „nepřekročitelné“ hodnoty pro srážkový úhrn daného trvání, frekvenční analýza „pouze“ připisuje libovolnému srážkovému

úhrnu pravděpodobnost překročení (a její neurčitost). Protože jsou pro rozdělení srážkových extrémů (SE) typické těžké chvosty (např. Katz a kol. [25]), dochází přitom k paradoxní situaci, že libovolně vysokému úhrnu (i teoreticky zjevně nerealistickému) je připsána nenulová pravděpodobnost. Jedná se však spíše o akademický problém, protože nenulové pravděpodobnosti nerealisticky vysokých úhrnů nemají praktický význam a jsou zanedbatelné, z hlediska modelu však užitečné. Určitou nevýhodou metod odhadu PMP je větší ovlivnění subjektivními volbami při jejím výpočtu ve srovnání s frekvenční analýzou, a dále skutečnost, že výsledkem je „pouze" jedna hodnota (pro daný časový interval a lokalitu nebo území), zatímco frekvenční analýza poskytuje informaci o celém pravděpodobnostním rozdělení. Oba výše popsané přístupy jsou však v praxi užitečné a mohou se doplňovat.

Jak ilustrují výše uvedené příklady, u frekvenční analýzy SE je regionální přístup třeba jednoznačně upřednostňovat na úkor lokálního (který však dosud naopak jednoznačně převládá v klimatologické i hydrologické praxi). Platí to bez ohledu na to, že volba mezi dvěma základními alternativami regionální frekvenční analýzy, tj. mezi koncepcí pevně vymezených regionů a flexibilních seskupení stanic, nemusí být triviální. Výsledky simulačních experimentů každopádně ukazují jen malé rozdíly ve statistických charakteristikách simulovaných kvantilů mezi těmito dvěma přístupy, srovnáme-li je s (velkým) zlepšením vůči lokálnímu přístupu (Gaál a Kyselý [23]; Kyselý a kol. [15]). Výběr nejvhodnějšího modelu regionální analýzy může být ovlivněn dostupnou sítí srážkoměrných stanic nebo „připraveností" pevně vymezených regionů – podmínku homogenity je pro nová data třeba testovat a zkušenost ukazuje, že homogenita regionů má tendenci se po doplnění dat o nové stanice nebo roky měření změnit, což vyžaduje předefinování regionů nebo alespoň přeřazení stanic do jiného regionu a nové testování homogenity.

Domníváme se, že v případě srovnatelné kvality těchto dvou regionálních přístupů hovoří tři zásadní argumenty ve prospěch metody oblasti vlivu (ROI): (1) vytváření flexibilních regionů vyžaduje mnohem menší subjektivní zásahy do analýzy než bývá obvyklé v procesu formování pevně vymezených regionů, (2) metodu ROI lze poměrně snadno implementovat v případě nutnosti opakování stejné analýzy pro velký počet datových souborů, např. u simulací klimatických modelů (viz Kyselý a kol. [3]), a to na rozdíl od metody pevně vymezených regionů, kde by homogenitu regionů bylo třeba testovat (s poměrně velkou pravděpodobností negativních výsledků a potřeby redefinice regionů) pro každou simulaci zvlášť, příp. i pro různá časová období, a (3) v metodě ROI se nevyskytuje problém se skokovými změnami odhadnutých kvantilů rozdělení (návrhových hodnot), který se objevuje na hranicích pevně definovaných regionů.

Vzhledem ke své flexibilitě a možnosti implementace ve velkých databázích bez nutnosti subjektivních zásahů je metoda ROI v současné době aplikována i na odhad jedno- a vícedenních návrhových srážek v ČR v databázi Českého

hydrometeorologického ústavu (všechny srážkoměrné stanice s alespoň 20 lety měření) a byla použita i k odhadům rozdělení SE ve výstupech regionálních klimatických modelů nad Evropou (v tomto případě se jedná o výpočetně poměrně náročnou úlohu, protože regionální frekvenční analýza je aplikována v síti řádově několika desítek tisíc uzlových bodů a ve velkém počtu takových simulací, zvlášť pro jednotlivá roční období, různé časové horizonty a doby trvání SE).

Navzdory uvedeným pokrokům v oblasti vývoje a aplikací pokročilejších metod odhadu pravděpodobností SE zůstává mnoho otázek nedořešených a otevřených dalšímu výzkumu. Potřebu a možnosti další práce vidíme zejména v těchto směrech:

- aplikace výše uvedených metod na krátkodobé SE (pro agregace $<$ 24 hod) – limitována kvalitou a délkou dostupných dat, databáze kombinující data z ombrografů (většinou do r. 2000) a automatických stanic je v současné době připravována ke zpracování
- vývoj metod regionální frekvenční analýzy zahrnujících závislost studované proměnné na čase nebo obecně kovariátách (viz příspěvek Hanel a Buishand v tomto čísle) – nevýhodou metod založených na L-momentech je obtížnost takového zobecnění
- adaptace metod regionální frekvenční analýzy pro přístup peaks-over-threshold (POT), který využívá všechna nezávislá překročení zvolené prahové hodnoty, nikoli roční nebo sezónní maxima
- vývoj metod založený na jiném přístupu než jsou metody POT nebo bloková maxima, tj. např. bayesovský přístup
- vývoj a aplikace metod pro odhady SE v datech z klimatických modelů – tyto metody by měly umožňovat jak prostorové „shlazení“ (redukci šumu), tak zahrnutí závislosti SE na čase, příp. dalších proměnných (vzhledem k možné přítomnosti trendu nebo jiných závislostí v datech)
- vývoj metod vícesložkových rozdělení extrémů, s využitím pro srážky rozdělené podle původu na převážně velkoprostorové (vrstevnaté) a konvekční; toto rozdělení může alespoň částečně reprezentovat různé fyzikální mechanismy vedoucí ke SE (např. Willems [26])
- modely extrémů využívající kombinace staničních měření (bodové charakteristiky, relativně dlouhé řady) a radarových, příp. satelitních dat (plošné charakteristiky, relativně krátké řady)
- vývoj regionálních metod pro statistické modelování závislosti mezi různými charakteristikami srážkových událostí (trvání srážkové události, celkový úhrn, průměrná a/nebo maximální intenzita), především pomocí kopulí (např. Zhang [27])
- vývoj prostorových mnohorozměrných modelů SE, na nichž v posledních letech pracuje zejména skupina kolem prof. Davisona z EPFL Lausanne (např. [8])

- vývoj metod pro odhad chyby (neurčitosti) odhadů, kombinující chyby vycházející z výběrové proměnlivosti, použitého statistického modelu a neurčitosti v důsledku scénářů změny klimatu
- redefinice pojmu n-letosti pro data obsahující trend nebo v případech, kdy lze trend předpokládat do budoucnosti, a to zejména s ohledem na praktické aplikace.

Z výše uvedeného je zřejmé, že navzdory pokroku v posledních letech zůstává mnoho otázek nezodpovězených a k jejich řešení je interakce klimatologů, hydrologů a statistiků nezbytným předpokladem. Stejně důležité je pokoušet se o přenos těchto poznatků do klimatologické a hydrologické praxe, aby mohli odborníci i veřejnost (a v první řadě ti, kdo se zabývají např. návrhem protipovodňových opatření) aspoň o málo víc důvěřovat tvrzením, že „byla překročena n-letá srážka“ (nebo průtok). Maximalistickým přáním autorů tohoto příspěvku pak je, aby byl spolu s odhadem n-letosti automaticky spojen i pokus o odhad chyby – rozpětí intervalu spolehlivosti obvykle spolehlivě prozradí, zda byla k odhadu použita lokální nebo regionální metoda, a může odradit od publikace některých podle našeho názoru nedostatečně podložených výsledků vycházejících z lokální metody odhadu, s nimiž se lze stále poměrně běžně setkat (např. při hodnocení nedávných povodní).

Literatura

[1] TRENBERTH, K.E. – DAI, A. – RASMUSSEN, R.M. – PARSONS, D.B. (2003) *The changing character of precipitation.* Bulletin of American Meteorological Society, 84, 1205 – 1217.

[2] HANEL, M. – BUISHAND, T.A. (2011) *Analysis of precipitation extremes in an ensemble of transient regional climate model simulations for the Rhine basin.* Climate Dynamics, 36, 1135 – 1153.

[3] KYSELÝ, J. – GAÁL, L. – BERANOVÁ, R. – PLAVCOVÁ, E. (2011) *Climate change scenarios of precipitation extremes in Central Europe from ENSEMBLES regional climate models.* Theoretical and Applied Climatology, 104, 529 – 542.

[4] ŘEZÁČOVÁ, D. – PEŠICE, P. – SOKOL, Z. (2005) *An estimation of the probable maximum precipitation for river basins in the Czech Republic.* Atmospheric Research, 77, 407 – 421.

[5] KOLEKTIV (2002) *Hydrológia – Terminologický výkladový slovník.* Vědecký redaktor O. Majerčáková. 1. vyd. Ministerstvo životného prostredia SR, 158 s.

[6] KHALIQ, M.N. – OUARDA, T.B.M.J. – ONDO, J.-C. – GACHON, P. – BOBÉE, B. (2006) *Frequency analysis of a sequence of dependent and/or non-stationary hydrometeorological observations: A review.* Journal of Hydrology, 329, 534 – 552.

[7] VENEZIANO, D. – LEPORE, CH. – LANGOUSIS, A. – FURCOLO, P. (2007) *Marginal methods of IDF estimation in scaling and non-scaling rainfall.* Water Resources Research, 43, W10418.

[8] DAVISON, A.C. – PADOAN, S.A. – RIBATET, M. (2012) *Statistical modeling of spatial extremes.* Statistical Science, 27, 161 – 186.

[9] ŘEZÁČOVÁ, D. – SOKOL, Z. – PEŠICE, P. (2001) *Deterministické a statistické odhady pravděpodobné maximální srážky pro území České republiky.* In „Vývoj metod pro odhad extrémních povodní“ – Sborník přednášek ze semináře k výsledkům

grantového projektu VaV/510/97, 23.4. 2001, Praha. Česká vědeckotechnická vodohospodářská společnost, Praha, s. 24 – 35.
[10] WMO (WORLD METEOROLOGICAL ORGANISATION) (1986) *Manual for estimation of probable maximum precipitation*. Operational hydrology report Nr. 1, WMO-No. 332. Geneva, 269 pp.
[11] CASAS, M.C. – RODRÍGUEZ, R. – PROHOM, M. – GÁZQUEZ, A. – REDANO, A. (2011) *Estimation of the probable maximum precipitation in Barcelona (Spain)*. International Journal of Climatology, 31, 1322 – 1327.
[12] COLLIER, C.G. – HARDAKER, P.J. (1997) *Estimating probable maximum precipitation using a storm model approach*. Journal of Hydrology, 183, 277 – 306.
[13] SIVAPALAN, M. – BLÖSCHL, G. (1998) *Transformation of point rainfall to areal rainfall: intensity–duration–frequency curves*. Journal of Hydrology, 204, 150 – 167.
[14] GAÁL, L. (2009) *Metódy výpočtu štatistických charakteristík návrhových hodnôt úhrnov zrážok na Slovensku*. Key Publishing, Ostrava, 224 s.
[15] KYSELÝ, J. – GAÁL, L. – PICEK, J. (2011) *Comparison of regional and at-site approaches to modelling probabilities of heavy precipitation*. International Journal of Climatology, 31, 1457 – 1472.
[16] HOSKING, J.R.M. – WALLIS, J.R. (1997) *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, Cambridge; New York; Oakleigh.
[17] LU, L.-H. – STEDINGER, J.R. (1992) *Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test*. Journal of Hydrology, 138, 223 – 245.
[18] KYSELÝ, J. – PICEK, J. (2007) *Regional growth curves and improved design value estimates of extreme precipitation events in the Czech Republic*. Climate Research, 33, 243 – 255.
[19] JUREČKOVÁ, J. – PICEK, J. (2001) *A class of tests on the tail index*. Extremes, 4, 165 – 183.
[20] BURN, D.H. (1990) *Evaluation of regional flood frequency analysis with a region of influence approach*. Water Resources Research, 26, 2257 – 2265.
[21] CASTELLARIN, A. – BURN, D.H. – BRATH, A. (2001) *Assessing the effectiveness of hydrological similarity measures for flood frequency analysis*. Journal of Hydrology, 241, 270 – 287.
[22] GAÁL, L. – KYSELÝ, J. – SZOLGAY, J. (2008) *Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia*. Hydrology and Earth System Sciences, 12, 825 – 839.
[23] GAÁL, L. – KYSELÝ, J. (2009) *Comparison of region-of-influence methods for estimating high quantiles of precipitation in a dense dataset in the Czech Republic*. Hydrology and Earth System Sciences, 13, 2203 – 2219.
[24] ZRINJI, Z. – BURN, D.H. (1994) *Flood frequency analysis for ungauged sites using a region of influence approach*. Journal of Hydrology, 153, 1 – 21.
[25] KATZ, R.W. – PARLANGE, M.B. – NAVEAU, P. (2002) *Statistics of extremes in hydrology*. Advances in Water Resources, 25, 1287 – 1304.
[26] WILLEMS, P. (2000) *Compound intensity/duration/frequency-relationships of extreme precipitation for two seasons and two storm types*. Journal of Hydrology, 233, 189 – 205.
[27] ZHANG, Q. – SINGH, V.P. – LI, J. – JIANG, F. – BAI, Y. (2012) *Spatio-temporal variations of precipitation extremes in Xinjiang, China*. Journal of Hydrology, 434–435, 7 – 18.

Poděkování:
Vývoj metod regionální frekvenční analýzy byl podpořený granty GA AV ČR (B300420601, B300420801) a GA ČR (P209/10/2045, P209/10/2265), v

rámci kterých byla také získána použitá data. Za pomoc při jejich přípravě děkujeme především P. Skalákovi a P. Štěpánkovi (ČHMÚ), za pomoc při přípravě rukopisu M. Kašparovi (ÚFA AV ČR) a M. Šálkovi (ČHMÚ).

Příspěvek vznikl s podporou projektu ESF „Zapojení týmu KLIMATEXT do mezinárodní spolupráce“ (reg. č. CZ.1.07./2.3.00/20.0086).

Adresa:
(1) Katedra aplikované matematiky, Technická Univerzita v Liberci
(2) Ústav fyziky atmosféry AV ČR, Praha
(3) Stavební fakulta, Slovenská technická univerzita v Bratislavě
E-mail: kysely@ufa.cas.cz, ladislav.gaal@stuba.sk, jan.picek@tul.cz

A LOWER BOUND FOR THE MIXTURE PARAMETER IN THE BINARY MIXTURE MODEL AND ITS ESTIMATOR

Bobosharif K. Shokirov

Keywords: mixture model, mixture parameter, estimator, lower bound, empirical stochastic process

Abstract: With a sample $X_1, \ldots, X_n$ of size n drawn from a distribution function $H(x;\theta)$ represented as a mixture of two distribution functions $F(x)$ and $G(x)$, where one of them is unknown, the paper discusses an approach to estimate the mixture parameter θ. By utilizing the behavior of the family of random variables $\{\theta_n^*(x),\ x \in [0,\ 1],\ n = 1, 2, \ldots\}$, a number of properties of the estimator of the parameter θ are derived. In particular, it is shown that this family of random variables contains an unbiased estimator of the mixture parameter. Based on approach and results of [3], an inequality which bounds the mixture parameter from below is obtained. The lower bound of the inequality is estimated, which serves as an estimator of the mixture parameter.

Abstrakt: Pomoci vyběru $X_1, \ldots, X_n$ z distribuční funkce $H(x, \theta)$, který je reprezentován jako směs dvou distribučních funkcí $F(x)$ a $G(x)$, kde jeden z nich je neznámý, tento článek studuje postup k odhadu mixujicího parametru θ. Jsou odvozeny nektere vlastností parametru θ na zaklade chování rodiny náhodných proměnných $\{\theta_n^*(x),\ x \in [0,\ 1],\ n = 1, 2, \ldots\}$. Zejména, jeli prokázáno, že tato rodina náhodných veličin obsahuje nestranný odhad mixujiciho parametru. Na základě přístupu a výsledcích [3], se získá nerovnost, která omezuje mixujicího parametru. Dostano odhad dolní mez tato nerovnosti, který slouží jako odhad mixujicího parametru.

1. The problem

Let $X_1, \ldots, X_n$ be a sample of size n drawn from a distribution function (d.f.) $H(x;\theta)$ of the form

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad (\theta \in (0,1)). \tag{1}$$

In representation (1) $F(x)$ is a known d.f., while d.f. $G(x)$ and a parameter $\theta \in [0,1]$ are unknown. This is a binary (or two-component) mixture model with one unknown component and our aim is to estimate θ, which we call a mixture parameter. Throughout of the paper we use notations $H(x;\theta)$, $H(x)$ and $H_\theta(x)$ interchangeably just to emphasize that d.f. $H(x)$ depends on the proportions of the components in the mixture; θ should not be interpreted as an unknown parameter from a parametric family of d.f.'s $H(x;\theta)$.

This model appears in many contexts. In the problem of multiple hypotheses testing, such as testing differentially expressed genes the p-values of

the test statistics, assuming that they are independent, under the null hypotheses are distributed uniformly on the interval [0, 1], while under the alternative their distribution is unknown ([2, 5]). In terms of the model (1) $F(x)$ is a uniform distribution. In such settings the aim would be estimating the proportion of false null hypotheses θ (and d.f. $G(x)$). An efficient estimator of θ is important once we want to control the multiple error rates, such as the false discovery rate (FDR) (see [1]).

In contrast to the usual classical mixture problem, where the mixture consists of a combination of two or more, mainly specified or partially specified distributions, the mixture in the right-hand side of (1) contains an unknown component and hence suggested classical methods cannot be applied here. Instead, the approach proposed in [3] to a binary survival model seems to be more promising to drive an inequality for the mixture parameter and estimate its lower bound.

In the classical mixture problem (partially) specified components could be already considered as a certain type of restrictions imposed on the family of distributions that together with other restrictions and assumptions makes the problem well-defined, in particular, identifiable. Basically, it is the identifiability of the mixture models which allows one to estimate the mixture parameter. Proceeding from this principle, basically, means that it is impossible to estimate the mixture parameter in the model (1) without the model being indentifiable. Indentifiability of this model previously was considered in [7]. To be able to derive certain properties of the parameter θ, model (1) was studied under certain assumptions, which could be summarized as following: estimate parameter θ in the model

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad x \in [0,\,1], \quad (\theta \in (0,1)), \tag{2}$$

where $F(x)$ is a known d.f., while d.f. $G(x)$ is unknown, both $F(x)$ and $G(x)$ are continuous and satisfies the following conditions

$$G(x) > F(x), \quad \forall x \in (0,\,1) \tag{3}$$

and

$$S_G \subset [0,\,1-\delta], \quad \text{for some} \quad \delta > 0, \tag{4}$$

where S_G denotes the support of d.f. $G(x)$.

Model (2) is, basically, transformed into interval [0, 1] model (1), meaning that S_F, the support of d.f. $F(x)$ tansformed into the compact set [0, 1] and condition (3) guarantees that S_G to be a proper subset of S_F: $S_G \subseteq S_F$. Indeed, let $x_0 \in [0,\,1]$ be such that $G(x_0) = 1$. Due to monotonicity and condition (3) it follows that $x_0 < 1$, that is, d.f. $G(x)$ reaches its maximum before the end of the interval [0, 1]. This shows that the support of d.f. $G(x)$ is a proper subset of the interval [0, 1].

Although in such setting model (1) becomes well-defined, conditions (3-4) still cannot guarantee its identifiability. However, it seems that even without

ensuring identifiability, one can deal with the problem of estimating the parameter θ in the model (1), in particular, one can derive certain bounds for the mixture paprameter. Based on the idea, approach and results of [3], we derive an inequality, which bounds the mixture parameter from below, and estimate its lower bound which, basically, is the estimator of parameter θ. It should be noted that among others, issues that bring up similar problems were considered in [4, 8].

2. Estimator of mixture parameter as an empirical stochastic process

In this section we give some properties of the mixture parameter in (2) previously proved in [7].

Consider the family of random variables $\{\theta_n^*(x),\ x \in [0,\ 1]\}$, where

$$\theta_n^*(x) = \frac{1 - H_n(x)}{1 - F(x)}. \tag{5}$$

Expression (5) is obtained from (2) under assumptions (3-4) with replacing d.f. $H(x)$ to $H_n(x)$, the empirical d.f., constructed by the sample $X_1, \ldots, X_n$. The right-hand side of (5) represents, basically, an empirical stochastic process. Based on knowledge on the behavior of this process, some properties of the mixture parameter and its estimator was studied in [7]. Treated empirical process $\theta_n^*(x)$ as a function of the random variable x in the interval $[0, 1]$, it was shown that the expected value of the process $\theta_n^*(x)$ is a nondecreasing function in $S_F \cap S_G$, the intersection of S_F and S_G, and is constant (or "almost" constant) in $S_F \backslash S_G$, the complement of S_G to S_F.

The following statement is true.

Theorem 1. *Assume condition (3) holds. Let d.f.'s $F(x)$ and $G(x)$ are continuously differentiable and satisfy the relation*

$$\frac{F'(x)}{1 - F(x)} \leq \frac{G'(x)}{1 - G(x)}. \tag{6}$$

Then the expected value of the process $\{\theta_n^*(x),\ x \in [0,1]\}$ is monotonic nonincreasing on the interval $[0,\ 1]$ with x and $\theta \leq \mathbb{E}[\theta_n^*(x)] \leq 1,\ x \in [0,1]$.

Proof. Proof is given in [7].

Corollary 1. *If condition (4) holds, then $\forall x \in [1 - \delta, 1]$ the expected value of the family of random variables $\{\theta_n^*(x),\ x \in [0,1]\}$ is an unbiased estimator for θ: $\mathbb{E}[\theta_n^*(x)] = \theta$.*

Theorem 2. *If conditions (3) and (4) hold, then the variance $\sigma^2(x; \theta, n)$ of the family of random variables $\{\theta_n^*(x),\ x \in [0,\ 1]\}$ has the from*

$$\sigma^2(x; \theta, n) = \frac{H(x)(1 - H(x))}{n(1 - F(x))^2}. \tag{7}$$

Corollary 2. (a) *If (4) holds, then*

$$\sigma^2(x;\theta,n) = \frac{\theta}{n}\left(\frac{1}{1-F(x)} - \theta\right), \quad \textit{for} \quad 1-\delta < x \le 1. \tag{8}$$

(b) *If (3) holds, then*

$$\begin{aligned}\sigma^2(x;\theta,n) &= \frac{1}{n(1-F(x))}\left[\theta\left(1-\frac{1-G(x)}{1-F(x)}\right)+\frac{1-G(x)}{1-F(x)}\right] - \\ &- \frac{1}{n}\left[\theta\left(1-\frac{1-G(x)}{1-F(x)}\right)+\frac{1-G(x)}{1-F(x)}\right]^2, \quad \textit{for} \quad 0 \le x \le 1-\delta.\end{aligned}$$

Theorem 3. *Let conditions (3) and (4) are satisfied. Then if in addition to (6), the condition*

$$2\frac{F'(x)}{1-F(x)} \ge \frac{G'(x)}{1-G(x)} \tag{9}$$

also holds, then the variance $\sigma^2(x;\theta,n)$, defined by (7) is a monotonic non-decreasing function of x for all $x \in [0,1]$.

Proof. Calculate derivative of

$$\sigma^2(x;\theta,n) = \frac{H(x)(1-H(x))}{n(1-F(x))^2}$$

with respect to x. We have

$$\frac{d}{dx}\left(\sigma^2(x;\theta,n)\right) = \sigma^2(x;\theta,n)\left(\frac{H'(x)}{H(x)} - \frac{H'(x)}{1-H(x)} + 2\frac{F'(x)}{1-F(x)}\right).$$

For the expression in the brackets we get

$$\frac{H'(x)}{H(x)} - \frac{H'(x)}{1-H(x)} + 2\frac{F'(x)}{1-F(x)} \ge \left(1+\frac{1}{H(x)}\right)\frac{F'(x)}{1-F(x)} - \frac{H'(x)}{1-H(x)}.$$

From (6) it follows that

$$\frac{H'(x)}{1-H(x)} \le \frac{G'(x)}{1-G(x)}.$$

Since

$$1+\frac{1}{H(x)} \ge H(x)+\frac{1}{H(x)} \ge 2,$$

by virtue of (9) we have $\frac{d}{dx}\left(\sigma^2(x;\theta,n)\right) \ge 0$. □

Due to Corollary 2 the variance of $\theta_n^*(x)$ is increasing (non-decreasing) for $x \in S_F \backslash S_G$ regardless of the condition (9) but not in $S_F \cap S_G$. Therefore, condition (9) guarantees $\sigma^2(x;\theta,n)$ to be increasing (or at least non-decreasing) on the whole interval [0, 1]. We suggest that for certain (fully specified) distributions this condition could be removed or at least replaced by a weaker one.

2.1. **The right border of the support of d.f.** $G(x)$

Theorem 1 states that the expected value of the family of random variables $\{\theta_n^*(x), n = 1, \ldots, x \in [0,\ 1]\}$ is non-increasing on the interval $[0,\ 1]$ and Corollary 1 shows that beyond the right border of S_G, which is assumed to be less than 1, it is an unbiased estimator for the mixture parameter. But the Corollary does not specify the right border of S_G. If $1 - \delta$, the right border of the support S_G of the d.f. $G(x)$ was known, then, intuitively, any values of $\theta_n^*(x)$ for $x \geq 1-\delta$ could be taken as an (unbiased) estimator of θ. Therefore, it seems that finding (or estimating) the right border of S_G should be an essential step towards estimating of the mixture parameter. But how to find or estimate $1 - \delta$?

Here we propose an approach to find a point $x_0 \in (0,\ 1)$ that can serve as estimator of the right border of S_G and such that $\theta_n^*(x_0) = \hat{\theta}$. The algorithm is as following.

1. Divide interval $[0,\ 1]$ into m equal parts such that $0 < x_0 \leq x_1 \leq \ldots \leq x_m < 1$.
2. Calculate the values of $\theta_n^*(x)$ at the last points: x_m, x_{m-1}, x_{m-2}.
   (i) If $\theta_n^*(x_m) = \theta_n^*(x_{m-1})$ and $\theta_n^*(x_{m-2}) > \theta_n^*(x_{m-1})$, then take the point x_{m-1} as the right border of S_G;
   (ii) If $\theta_n^*(x_m) = \theta_n^*(x_{m-1}) = \theta_n^*(x_{m-2})$ then calculate $\theta_n^*(x_{m-3})$;
   (iii) If $\theta_n^*(x_{m-1}) = \theta_n^*(x_{m-2})$ and $\theta_n^*(x_{m-3}) > \theta_n^*(x_{m-2})$, then take the point x_{m-2} as the right border of S_G.
3. If $\theta_n^*(x_m) > \theta_n^*(x_{m-1})$ or $\theta_n^*(x_{m-2}) > \theta_n^*(x_{m-3})$ then we divide interval $[0,\ 1]$ into smaller parts and repeat steps 1-3. Repeat this procedure until $\theta_n^*(x_m) = \theta_n^*(x_{m-1})$ and $\theta_n^*(x_{m-2}) > \theta_n^*(x_{m-1})$ or $\theta_n^*(x_{m-1}) = \theta_n^*(x_{m-2})$ and $\theta_n^*(x_{m-3}) > \theta_n^*(x_{m-2})$.

By this algorithm we obtain a point that can be accepted as the right border of S_G, that is, we derive $\hat{\delta}$, the estimator of δ. Since for $x \geq 1 - \hat{\delta}$ we have $\mathbb{E}[\theta_n^*(x)] = \theta$, then as a naive estimator of θ could be taken any values of $\theta_n^*(x)$ for $x \geq 1 - \delta$. But Theorem 2 shows that the variance of $\theta_n^*(x)$ increases drastically as soon as x reaches the right border of S_G. Since we would like the estimator of θ to be as close as possible to the right border of S_F, then as its estimator we can take a value $\theta_n^*(x^*),\ x^* > 1 - \hat{\delta}$ with maximum admissible variance.

3. **The lower bound of the mixture parameter estimator**

In this section we derive an inequality for the mixture parameter that bounds the mixture parameter from below and is expressed via the components of the mixture model and the sample size. We obtain an estimator of the lower bound of that inequality, which serves as an estimator of the mixture parameter in the model (2). The idea of this section takes its origin from [3].

For the further discussion, we need the following lemma.

Lemma 1. *Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be a sample of size n drawn from d.f. $H(x)$. Then sample $\mathbb{Y}_n = \{Y_1, \ldots, Y_n\}$ of size n drawn from the complementary cumulative distribution function (c.c.d.f.) $(1 - H(x))/(1 - F(x))$ could be obtained from $\mathbb{X}_n$ by*

$$y = \overline{H}^{-1}\left(\frac{1 - H(x)}{1 - F(x)}\right), \quad \overline{H(x)} = 1 - H(x).$$

Proof. We have

$$\overline{H(x)} = \mathbb{P}\{X > x\}.$$

Therefore for every $x \in [0,\ 1]$

$$\frac{1 - H(x)}{1 - F(x)} = \frac{\mathbb{P}\{X > x\}}{1 - F(x)} = \mathbb{P}\{Y > y\} = \overline{H(y)},$$

from which follows the statement of lemma.

Let us call $\mathbb{X}_n$ the original sample and $\mathbb{Y}_n$ its transformed sample.

Theorem 4. *Let $\mathbb{X}_n$ be the original sample and $\mathbb{Y}_n$ be its transformed sample and $1 \le k \le n$. Assume the following conditions hold:*

$$G(x) > F(x), \quad \forall x \in (0,\ ), \tag{10}$$

$$S_G \subset [0,\ 1 - \delta], \quad \textit{for some} \quad \delta > 0, \tag{11}$$

and

$$\frac{F'(x)}{1 - F(x)} \le \frac{G'(x)}{1 - G(x)}. \tag{12}$$

Assume that $\varphi(x)$ is a strictly decreasing function on the interval $[0,\ 1]$ such that $\varphi(0) = -\varphi'(0) = 1$ and satisfies the relation

$$\frac{d^2}{dx^2}\left[\varphi^{-1}\left(\frac{1 - H(x)}{1 - F(x)}\right)\right] \ge 0. \tag{13}$$

Then for the mixture parameter in the model (2) the inequality

$$\theta \ge 1 - \frac{H(X) - F(X)}{\overline{F(X)}(1 - \varphi(YR_H(y_0)))} \tag{14}$$

holds and the estimator of its lower bound, which serves as an estimator of the mixture parameter θ in the model (2), can be defined as

$$\theta_n^* = \max\left\{1 - \frac{k}{n[1 - \varphi(YR_n(y_0))]}, 0\right\}, \tag{15}$$

where Y is defined as

$$\max\{Y_1, \ldots, Y_k\} \le Y \le \min\{Y_{k+1}, \ldots, Y_n\}, \quad k \le n, \tag{16}$$

$y_0 \in (0,\ Y)$, x_0 is such that $\overline{H(y_0)} \cdot \overline{F(x_0)} = \overline{H(x_0)}$ and

$$R_n(y_0) = \frac{1}{y_0}\varphi^{-1}\left(\frac{1 - H_n(x_0)}{1 - F(x_0)}\right),$$

$H_n(x)$ is the empirical d.f., constructed by the sample $\{X_1, \ldots, X_n\}$.

Proof. Rewrite (2) it in the form

$$1 - H(x) = \theta(1 - F(x)) + (1 - \theta)(1 - G(x)), \tag{17}$$

where d.f.'s $F(x)$ and $G(x)$ are defined on the interval $[0,\ 1]$. Assuming $F(x) \neq 1$ for $x \in [0,\ 1]$, divide both sides of (17) by $1 - F(x)$. We obtain

$$\frac{1 - H(x)}{1 - F(x)} = \theta + (1 - \theta)\frac{1 - G(x)}{1 - F(x)}. \tag{18}$$

In fact the last assumption is satisfied by virtue of monotonicity of d.f. $F(x)$ and conditions (10 - 11). Since $S_G = [0, 1 - \delta] \subseteq [0,\ 1]$, therefore $1 - G(x)$ vanishes earlier than $1 - F(x)$ does, hence both ratios $(1 - H(x))/(1 - F(x))$ and $(1 - G(x))/(1 - F(x))$ have meaning $\forall x \in [0,\ 1]$ and $0 \leq (1 - H(x))/(1 - F(x)) \leq 1$ and $0 \leq (1 - G(x))/(1 - F(x)) \leq 1$. By virtue of conditions (10) and (12) both of these ratios are monotonically decreasing in the interval $[0,\ 1]$ and represent complementary cumulative distribution functions (or tail distribution or survival function): $(1 - H(x))/(1 - F(x)) = \mathbb{P}_H\{Y > x\}$ and $(1 - G(x))/(1 - F(x)) = \mathbb{P}_G\{Y > x\}$.

From (13) it follows that $f(x) = \varphi^{-1}\left(\frac{1-H(x)}{1-F(x)}\right)$ is a convex function. Further denote

$$R(x, y) = \frac{f(x) - f(y)}{x - y}$$

$R(x, y)$ is a symmetric with respect to x and y function. If $f(x)$ is convex then $R(x, y)$ is nondecreasing with x for fixed y and vise versa if $R(x, y)$ is nondecreasing then $f(x)$ is a convex function. Simply speaking, convexity of $f(x)$ is equivalent to nondecreasing property of the function $R(x, y)$. We have a convex function $f(x)$ and we need to show that the first part of our claim is true, i.e. $R(x, y)$ is nondecreasing with x for every fixed y. Take the derivative of $R(x, y)$ with respect to x and show that it is nonnegative. We have

$$\frac{dR(x, y)}{dx} = \frac{f'(x)(x - y) - (f(x) - f(y))}{(x - y)^2} \geq 0$$

if only $f'(x)(x - y) - (f(x) - f(y)) \geq 0$. Therefore, we need to show that if $f(x)$ is convex then $f(x) \geq f(y) + f'(y)(x - y)$ holds. (In fact the last confirmation holds in both directions but in this particular case we need only the first part of it.) From the convexity of $f(x)$ it follows that for $0 \leq \alpha \leq 1$

$$f(x + \alpha(y - x)) \leq \alpha f(y) + (1 - \alpha)f(x) = \alpha f(y) - \alpha f(x) + f(x).$$

Rewrite the last expression in the form

$$f(y) \geq f(x) + \frac{f(x + \alpha(y - x)) - f(x)}{\alpha}.$$

But

$$\lim_{\alpha \to 0} \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}(y - x) = f'(x)(y - x).$$

Therefore,

$$f(y) \geq f(x) + f'(x)(y - x).$$

This shows that $R(x,y)$ is nondecreasing with x for fixed y, that is, for any real $x \geq z > 0$ we have

$$R(z,y) \leq R(x,y)$$

or

$$\frac{f(z)-f(y)}{z-y} \leq \frac{f(x)-f(y)}{x-y}.$$

Since $f(0) = \lim_{y\to 0} f(y) = \varphi^{-1}(1) = 0$, we have

$$\frac{1}{z}\varphi^{-1}\left(\frac{1-H(z)}{1-F(z)}\right) \leq \frac{1}{x}\varphi^{-1}\left(\frac{1-H(x)}{1-F(x)}\right)$$

or

$$\varphi^{-1}\left(\frac{1-H(x)}{1-F(x)}\right) \geq \frac{x}{z}\varphi^{-1}\left(\frac{1-H(z)}{1-F(z)}\right).$$

Applying $\varphi(x)$ to both sides of the last inequality we get

$$\frac{1-H(x)}{1-F(x)} \leq \varphi\left(\frac{x}{z}\varphi^{-1}\left(\frac{1-H(z)}{1-F(z)}\right)\right). \tag{19}$$

Next denote

$$R_H(y_0) = \frac{1}{y_0}\varphi^{-1}\left(\frac{1-H(x_0)}{1-F(x_0)}\right)$$

and

$$R_G(y_0) = \frac{1}{y_0}\varphi^{-1}\left(\frac{1-G(x_0)}{1-F(x_0)}\right),$$

where y_0 corresponds to x_0. By lemma 1 we can define such X that the corresponding Y be from (16). Then from (19) we obtain

$$\frac{1-H(X)}{1-F(X)} \leq \varphi\left(\frac{Y}{y_0}\varphi^{-1}\left(\frac{1-H(x_0)}{1-F(x_0)}\right)\right) = \varphi(YR_H(y_0)). \tag{20}$$

and

$$\frac{1-G(X)}{1-F(X)} \leq \varphi\left(\frac{Y}{y_0}\varphi^{-1}\left(\frac{1-G(x_0)}{1-F(x_0)}\right)\right) = \varphi(YR_G(y_0)). \tag{21}$$

Due to (10) $\forall x \in (0,\ 1)$ we have

$$\frac{1-H(x)}{1-F(x)} \geq \frac{1-G(x)}{1-F(x)}. \tag{22}$$

From (20) and (21) it follows that

$$\varphi(YR_G(y_0)) \leq \varphi(YR_H(y_0)). \tag{23}$$

Therefore, by virtue of the inequalities (23) and (22) from equation (18) we obtain

$$\frac{1-H(X)}{1-F(X)} = \theta + (1-\theta)\left(\frac{1-G(X)}{1-F(X)}\right) \leq \theta + (1-\theta)\varphi(YR_G(y_0)) \leq$$
$$\leq \theta + (1-\theta)\varphi(YR_H(y_0)) = \varphi(YR_H(y_0)) + \theta(1-\varphi(YR_H(y_0))).$$

From the last relation it follows that

$$\theta \geq \frac{\frac{1-H(X)}{1-F(X)} - \varphi(YR_H(y_0))}{1 - \varphi(YR_H(y_0))}, \tag{24}$$

which is equivalent to to (14).

Since $(1 - H(x))/(1 - F(x)) = \mathbb{P}_H\{Y \geq y\}$, the ratio $(1 - H(X))/(1 - F(X))$ could be consistently estimated by $(n - k)/n$, (see [3]). Hence the estimator of the lower bound for θ becomes

$$\theta_n^* = \max\left\{1 - \frac{k}{n[1 - \varphi(YR_n(y_0))]}, 0\right\}, \tag{25}$$

where $R_n(y_0)$ is the empirical counterpart of $R_H(y_0)$. □

Literature

[1] Benjamini, Y., Hochberg, Y., (1995) *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. J. Roy. Statist. Soc. Ser. B **57** 289 – 300.

[2] Efron, B. (2010) *Large-scale inference*, Institute of Mathematical Statistics Monographs. **1**, Cambridge University Press, Cambridge.

[3] Klebanov, L. B., Yakovlev, A. A., (2007) *A New Approach to Testing for Sufficient Follow-up in Cure Rate Analysis*. Journal of Statistical Planning and Inference, **137**, 3557 – 3569.

[4] Meinhausen, N., Rice, J. P. (2006) *Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses*, The Annals of Statistics, **34**, 373 – 393.

[5] Robin, S., Bar-Hen, A., Daudin, J.-J., Pierre, L. (2007) *A semi-parametric approach for mixture models: application to local false discovery rate estimation*. Comput. Statist. Data Anal. **51**, 5483 – 5493.

[6] Patra,R., Sen, B. (2012) *Estimation two component mixture model with application to multiple testing*. Comput. Statist. Data Anal. **51**, 5483 – 5493.

[7] Shokirov, B. K. (2010) *On a problem connected with the mixture paramter estimation*. In Antoch, J., Dohnal, G. (Eds.), Informacni Bulletin České statistické společnosti. **3**, 95 – 102.

[8] Wu, W. B. (2008) *On false discovery control under dependence*. The Annals of Statistics, **36**, 364 – 380.

Acknowledgement: The author thanks professor Lev B. Klebanov for statement of the problem and his continuous help to complete the paper.

Address: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

E-mail: bobosari@karlin.mff.cuni.cz

FEKM ALGORITHM: A MODIFICATION

Marta Žambochová[1]

Abstract: This paper describes the modification of one of the designing algorithms for clustering in large data sets – the algorithm FEKM (Fast and Exact K-Means). This is one of the variants of the k-means algorithm to process very large data sets that do not fit into memory. The effectiveness of the algorithm is based on minimizing the number of passes through the entire data set.

Just one pass through the entire data set suffices to become an exceptionally good case. It is necessary to have the same number of passes through the entire data set as the classical k-means algorithm in the worst case scenario. The number of passages depends strongly on the initial selection of the sample data. Therefore, ways to create an appropriate selection so that the number of passes is minimized and preferably stable, are proposed in the article.

Abstrakt: Příspěvek se zabývá popisem modifikace algoritmu FEKM. Algoritmus FEKM (Fast and Exact K-Means) je jednou ze známých variant algoritmu k-průměrů umožňující zpracování velmi rozsáhlých datových souborů.

Hlavní myšlenkou algoritmu je prvotní vytvoření přiměřeně velkého výběrového souboru z původního souboru dat. V rámci tohoto souboru jsou vytvořeny shluky pomocí klasického algoritmu k-průměrů. V jednotlivých iteracích v průběhu shlukování výběrového souboru jsou zaznamenávána všechna centra a k nim předem definované popisné statistiky. Pomocí všech těchto center pak dochází k vytváření cílového shlukování celého souboru při minimálním počtu průchodů celým datovým souborem.

Ve výjimečně příznivém případu stačí pouze jeden průchod celým datovým souborem. V nejhorším možném případu je nutný stejný počet průchodů jako u klasického algoritmu k-průměrů. Malý počet průchodů celým souborem potřebný k provedení celého algoritmu tak, jak deklarují jeho autoři, byl však potvrzen pouze ve výjimečných případech. Počet průchodů silně závisí na prvotním výběrovém vzorku dat. Algoritmus FEKM tvoří vzorek dat náhodným výběrem. Hlavní myšlenkou navrhované modifikace je proto vytvoření vzorku dat nikoliv náhodně, ale za pomoci jistých datových struktur (stromů). Výsledný vzorek pak lépe vypovídá o rozložení dat a následně se sníží potřebný počet průchodů celým datovým souborem.

Keywords: Clustering, large data set, sampling, algorithm fekm, trees.

Klíčová slova: Shlukování, velký soubor dat, vzorkování, algoritmus FEKM, stromy.

[1], University of J.E. Purkyne, Faculty of Social and Economic Studies, Department of Mathematics and Informatics, Moskevská 54, Ústí nad Labem; marta.zambochova@ujep.cz

1 Introduction

It is often necessary to process even very large data sets in a data analysis. Processing these files is very difficult – first of due to the time demanded by the processing, but also due to the fact that the data file does not fit into the main memory. This problem is dealt with very differently by various authors in the area of cluster analysis – by algorithmically small, almost cosmetic, modifications to the fundamental changes. One possibility may be a sampling, during which a representative sample set is selected from the whole data set, which contains only such number of objects that can be clustered within a reasonable time limit. A set of clusters is first created in this selected subset of the objects. Then the remaining objects are assigned to already established clusters. Another approach is to attempt to minimize the passage of the entire dataset.

The FEKM algorithm (Fast and Exact K-Means) will first be examined in detail in the article. After this, the test results of this algorithm from the test data and the subsequent comparison of such with some selected algorithms will be described. Finally a proposal for the pre-processing of the algorithm FEKM and its influence on the behaviour of the algorithm will be set out.

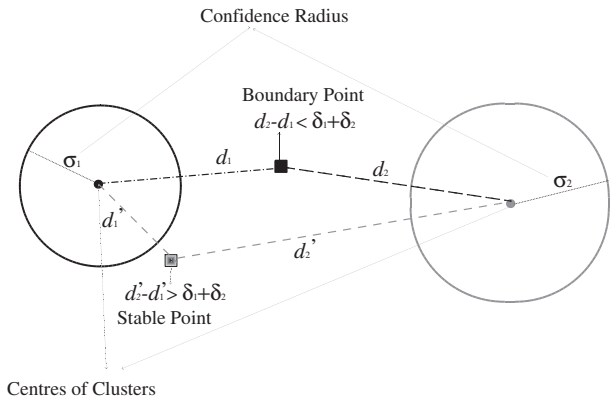
2 Algorithm FEKM

The FEKM algorithm (Fast and Exact K-Means) was described by the authors of the article listed in [1]. This is one of variants of the k-means algorithm to process very large data sets that do not fit into memory. The main idea of the algorithm is the initial creation of a reasonably large sample of the original data set. The clusters are created using the traditional k-means algorithm in this sample set. All the centres of individual clusters and their descriptive statistics are logged into each iteration of the algorithm during the clustering in the sample file. Then, the target clustering of the whole set with a minimum number of passes through the entire data set is created by means of the application of all of these centres. The data file, which is essential for the clustering, is the input of the algorithm. The initial centres and stopping criterion are input too as in the classical k-means algorithm. The target distribution of the all objects in the clusters is the output of the algorithm.

Firstly, a random sample data set is created by the application of random sampling. The algorithm clusters the sample set by means of using classical k-means methods in the first stage of this algorithm. In the second phase, the algorithm passes through the entire data set. The centre that is closest to the given object is found for each data object and each iteration which is recorded in the first phase. Each data object is assigned to a particular cluster by using its usage.

The second phase is to detect all objects that are suspected of changes resulting in a clustering comparison with the clustering of sample. This pro-

blem applies particularly to objects located at the edge of the clusters. It is evident that the objects deep inside the clusters are unencumbered from this perspective. Identified suspicious points are stored. If the object is not identified as a suspected boundary point, we can assume that this object by means of the clustering of the entire file has been assigned to the same cluster to which it is assigned by an additional assignment. We thereby calculate once again the characteristics of this cluster in this case. The situation about the stable and the boundary points is graphically illustrated in Figure 1.



Obrázek 1: Stable and boundary points

In the third phase, the algorithm deals with suspect border objects that were revealed and subsequently saved in the previous phase. At the beginning of this phase, the algorithm recalculates the clusters for each of the iterations from the first phase. It does this by virtue of the assumption that we actually assigned all suspicious boundary objects to their nearest centre. Recalculation is performed using the suspect boundary objects together with the preserved statistics describing each cluster. Should a recalculated centre that is distant from the original possess more than the pre-specified critical value, the algorithm returns to the second phase and once again passes through the entire data set. The authors of the algorithm call this above mentioned critical value as the confidence radius. It is evident that only one passing through the entire data set suffices in the case of being exceptionally and sufficiently good enough. It is necessary to possess the same number of passes as the classical k-means algorithm in the worst case scenario.

The number of suspect boundary points should not exceed 20% of the size of the entire data set. In addition, these points should fit into the RAM. If any violation of these conditions has occurred during the algorithm, it is necessary to reduce the individual confidence radiuses ratting by setting a special parameter for the algorithm. This will have the effect of reducing

the number of suspect boundary points. The authors in their paper [1] confirmed the correctness and effectiveness of the proposed algorithm. The authors further develop the algorithm and thus provide its further acceleration in the article [2].

3 Description of data files

All programmed algorithms were tested on three data sets. The first two files are available at [8]. The last file is a generated file.

IRIS file

- 150 objects
- 4 numeric variables (individual dimensions of the calyx and of the petals of flowers)
- 3 clusters – three different kinds of genus iris
- 50 representatives of each kind
- The exploratory analysis showed that one of the kinds differs considerably in the described attributes.
- The remaining two kinds differ only slightly, even values of their attributes penetrate.

VOWEL file

- 528 objects
- 10 numeric variables
- 11 clusters – eleven monosyllabic words
- The contents of the file refer to the pronunciation of vowels in British English. Eight speakers (four men and four women) read 11 monosyllabic words six times, thereby producing different vowel pronunciations (heed, hid, head, had, hard, hud, hod, hoard, hood, who'd, heard). Thus, every word was spoken 48 times. It was recorded and converted into sound with ten numeric values in each of these cases. Each record is one object of the file.

GENER file

- 1,000,000 objects

- 2 numerical variables
- 20 clusters
- The coordinates of objects in the individual clusters were generated as random values belonging to a Gaussian distribution within the given parameters. The parameters of the distribution in individual clusters were again randomly generated as the values of the continuous uniform distribution (the mean value μ randomly from the interval $(0, 10)$ and the variance σ^2 from the interval $(0, 3)$).

4 The testing of algorithms

The algorithm FEKM was programmed in the MATLAB development environment. In the same environment the other algorithms were also programmed. These were selected algorithms, which either use different types of sampling, or limit the number of passes through the entire data set. Both of these principles are the basis of the algorithm FEKM. These selected algorithms and their implementations are comprehensively described in [7].

Dependence on the input parameters is the big disadvantage for most of the tested algorithms (including FEKM). In addition, there is no universally valid recommendation for setting the values of individual parameters.

The other disadvantage of the majority of tested algorithms is that sampling provides poor-quality for the clustering. In particular, the quality of the resulting clustering by means of using the algorithm FEKM depends very much on the fabricated sample, ranging from very poor to excellent quality. This behaviour bore similarity to the fact that the quality of clustering by using k-means algorithm is very strongly dependent on the choice of initialization centres. This is described in detail in [6].

The algorithm FEKM had one special disadvantage according to the experiments carried out. The authors declare a small number of passes by a set needed to perform the algorithm. This small number was confirmed only in exceptional cases. The number of passes depended strongly on the initial selection of the sample data.

The algorithm FEKM is the only algorithm which uses a random sample of data. The other above mentioned algorithms represent samples using certain data structures (trees). The resulting sample therefore provides a better reflection of the distribution of the data. This fact gave rise to the idea that it would actually be possible to combine the algorithm FEKM with any of these other algorithms to achieve better results.

The algorithm BIRCH k-means was chosen from all of the examined algorithms. The reason was the main principle of the both algorithms BIRCH - amounting to one fundamental pass through the entire data file. The second reason was the fact that the primary output of both BIRCH algorithms is the set of the centres from the clusters. These centres can therefore also be used directly as input for the algorithm FEKM. These two properties meet

both algorithms BIRCH. The last reason for our choice of algorithm was its relative simplicity in comparison with the classical BIRCH algorithm.

5 BIRCH algorithm *k*-means

The paper denoted in [3] describes the clustering method, which is a variant of algorithm k-means, and which is influenced by the algorithm BIRCH. The algorithm BIRCH is described in detail in [4], [5]. This method gives the special procedure prior to processing by using Lloyd's algorithm. We thus receive the distribution of objects into groups, which is useful for further work, by means of applying this procedure.

The algorithm BIRCH k-means has two parameters – the allowed number of objects in the cluster and the maximum allowable value of the "radius" of the cluster (i.e. critical value of variability). This algorithm does not work with CF-trees as is the case in the classical algorithm BIRCH.

Also the CF-characteristic of the classical procedure BIRCH is replaced by another equivalent characteristic. The information contained in both characteristics is sufficient to calculate the centres of clusters, i.e. the distances between clusters and the measure of the compactness of clusters. This information is necessary and suffices for the implementation of the algorithm. It is true that the characteristic of the cluster that was given rise to by the merger of two disjoint clusters is equal to the sum of the characteristics of the original clusters.

The characteristic used in the algorithm BIRCH k-means is a triplet (m, q, c), where m is the number of objects in the cluster, q is the quality of the cluster (calculated as the sum of squared distances of all objects of a given cluster to the centre of the cluster) and c is the centre of the cluster. This characteristic is maintained for each cluster.

The algorithm works in three phases. The beginning of the first phase contains a set of all the tracked objects and the set of clusters is empty. Then in the cycle after this, the algorithm always selects an object from the set of objects and tries to find the "nearest" cluster of a set of clusters, in which the addition of an object does not exceed either the limit of the number of elements in a cluster or the critical value of variability of the cluster. If no such cluster is already in existence, it creates a new cluster that includes only this object. The object is deleted from the set of all objects after the successful inclusion of the object into the cluster. The cycle is performed until contains empty sets of objects.

The centres of all clusters that have arisen in the first phase are clustered in the second phase of the algorithm. This clustering can be performed using any clustering method. In [3] a special variant of the method of the quadratic k-means algorithm is used.

In the third phase, the original objects are classified into clusters. Each object is assigned to the closest of the centres created in the second phase. This phase is given in [3] as the final phase.

In my experiments, however, I added a fourth phase, in which the division into clusters arising from the first three phases is taken as the initial clustering and running of the classical Lloyd's algorithm. The results of clustering were significantly improved by means of this. The disadvantage of this treatment, however, is the increase in computational time.

6 Results of testing the modified FEKM algorithm

The insertion of the first phase of the algorithm BIRCH k-means prior to the proper processing of the algorithm FEKM establishes a method, which combines the advantages of both algorithms. The centres of the clusters generated in the first phase of algorithm BIRCH k-means were taken as a sample data file, which is required as input to the algorithm FEKM. The number of necessary passes through the entire data set was stabilized in comparison to the algorithm FEKM. The core of the modification shows that the minimum number of passes is two. There are cases where only a single passage through the entire dataset sufficed in the original algorithm. On the other hand, there have been cases where the sample was generated data "inappropriately" and there was a very high number of passage data file. It has resulted in the improvement of the quality of clustering compared to algorithm BIRCH k-means. Unfortunately, this also gave rise to an extension of the processing time (on average 2 – 3 times) compared to the BIRCH algorithm k-means.

References

[1] Goswami, A., Ruoming, J., Agrawal, G. (2004): Fast and exact out-of-core k-means clustering Data Mining. *ICDM apos;04. Fourth IEEE International Conference on Volume.*

[2] Goswami, A., Ruoming, J., Agrawal, G. (2006): Fast and exact out-of-core and distributed k-means clustering. *Knowledge and Information Systems* **10**, 17–40.

[3] Kogan, J. (2007): *Introduction to Clustering Large and High-Dimensional data.* Cambridge University Press, New York.

[4] Zhang, T., Ramakrishnan, R., Livny, M. (1996): An efficient data clustering method for very large databases. *ACM SIGMOD Record* **25**, 103–114.

[5] Zhang, T., Ramakrishnan, R., Livny, M. (1997): BIRCH: A new data clustering algorithms and its applications. *Journal of Data Mining and Knowledge Discovery* **1**, 141–182.

[6] Žambochová, M. (2009): Inicializační rozdělení do shluků a jeho vliv na konečné shlukování v metodách k-průměrů. *Sborník prací účastníků vědeckého semináře doktorského studia FIS VŠE, Praha*, 243–250.

[7] Žambochová, M. (2012): Clustering Algorithms Based on Sampling. *Informační bulletin České statistické společnosti, Praha* **23**, 10–19.

[8] http://archive.ics.uci.edu/ml/datasets/

VYUŽITÍ REGULAČNÍCH DIAGRAMŮ PŘI TVORBĚ SYSTÉMU INTELIGENTNÍCH ALARMŮ V ENERGETICKÉM PROVOZU JADERNÝCH ELEKTRÁREN

USAGE OF CONTROL CHARTS FOR CREATION OF SYSTEM OF INTELLIGENT ALARMS IN ENERGETIC FACILITY OF NUCLEAR POWER PLANTS

Josef Bednář, Radomil Matoušek

Adresa: Vysoké učení technické v Brně, Fakulta strojního inženýrství, Technická 2, 616 69 Brno; bednar@fme.vutbr.cz, matousek@fme.vutbr.cz

Abstrakt: Při tvorbě systému inteligentních alarmů v energetickém provozu jaderných elektráren byl řešen problém jak rychle a efektivně (on-line) odhalit začínající nestabilitu procesu popsaného cca 140 procesními charakteristikami a varovat velín, že bez zásahu do systému dojde velice pravděpodobně k mimořádné události. Ze statistických přístupů klasické vícerozměrné statistické metody selhaly (především proto, že jsme v historických datech nebyli schopni rozeznat, zda jsou změny vyvolány zásahem operátora, nebo zda k nim dochází samovolně). Proto bylo navrženo využití klasických regulačních diagramů pro sledování stability procesů. Díky tomu, že mnoho charakteristik je autokorelovaných nebo nestacionárních, byly vyhodnoceny jako nejlepší regulační diagramy diagramy rozpětí a klouzavého rozpětí, které budou v budoucnu implementovány do softwarové aplikace.

Abstract: In creation of systems of intelligent alarms in energetic facility of nuclear plants was solved a problem how fast and effectively (on-line) reveal beginning instability of process described by about 140 process characteristics and warn control centre that without intervention is highly probable to occur an emergency event. From statistical approaches the multidimensional statistic methods have failed (mainly because we could not recognize from historical data if changes are result of operator's actions or they emerge spontaneously). Therefore it has been proposed to use classic control chart for monitoring of process' stability. Thank to fact that many of the characteristics are autocorrelated and non-stationary came out as best control chart of range and moving range which will be in the future implemented into software application.

Klíčová slova: Regulační diagram, SPC, x-bar R diagram, I-MR diagram

Keywords: Control Chart, Stability, SPC, x-bar R chart, I-MR chart

1. Úvod

Při tvorbě systému inteligentních alarmů v energetickém provozu jaderných elektráren byl řešen problém jak rychle a efektivně (real-time) odhalit začínající nestabilitu procesu popsaného cca 140 procesními charakteristikami a varovat velín, že bez zásahu do systému dojde velice pravděpodobně k události. Ze statistických přístupů klasické vícerozměrné statistické metody selhaly (především proto, že jsme v historických datech nebyli schopni rozeznat, zda jsou změny vyvolány zásahem operátora nebo zda k nim dochází samovolně). Proto bylo navrženo využití klasických regulačních diagramů pro sledování stability procesů. Procesní charakteristiky byly charakteristiky teploty, tlaku, výšky hladiny, průtoku a výkonu apod. Protože jde o důvěrné informace, jsou data transformována a veličiny označeny pouze typově.

2. Regulační diagramy

Regulační diagram má obecně sloužit jako diagnostický nástroj k posouzení, zda se sledovaný proces (představovaný nějakou měřenou veličinou nebo veličinami, které jej charakterizují) chová tak, jak očekáváme, zvláště pak, nedošlo-li k nečekané změně procesu. Došlo-li k takové změně, je třeba ji interpretovat – vysvětlit a případně přistoupit k nějakému zásahu. Proces, ve kterém není třeba přistupovat k zásahům, nazýváme stabilní a poznáme ho tak, že se v něm vyskytují pouze (přirozené) náhodné příčiny kolísání. Těchto příčin je široká škála a každá přispívá ke změně procesu jen nepatrně. Stabilní proces se chová v každém okamžiku stejně, tudíž je predikovatelný. Predikovatelné procesy jsou z hlediska nákladu na jakost levnější než procesy, které se chovají chaoticky. Kromě náhodných příčin kolísání proces ovlivňují i vymezitelné příčiny kolísání, působením těchto příčin již dochází k zásadním změnám procesu (odlehlé hodnoty, posunutí procesu, unášení procesu).

2.1. Typy regulačních diagramů

Pro spojitá data většinou používáme jeden ze tři základních diagramů:

- I-MR diagram - individuální hodnoty a klouzavá rozpětí,
- Xbar–R diagram - aritmetický průměr a rozpětí,
- Xbar–S diagram - aritmetický průměr a směrodatná odchylka.

Pro atributivní data používáme dle typu dat diagramy:

- np diagram počet nestandardních výrobků v sériích stejného rozsahu,
- p diagram podíl nestandardních výrobků v sériích různého rozsahu,
- c diagram počet neshod na stejně velkých jednotkách,
- u diagram počet neshod na různě velkých jednotkách.

2.2. Testy vymezitelných příčin

Abychom určili, zda je proces statisticky stabilní a tedy není vhodné do něj zasahovat, jsou zavedeny tzv. testy vymezitelných příčin. Tyto testy hledají taková seskupení bodů (hodnot) v regulačním digramu, která jsou málo pravděpodobná. V dalším byly využity především následující testy:

- Test 1: 1 bod dále než 3 směrodatné odchylky od střední hodnoty,
- Test 2: 9 bodů v řadě na stejné straně od střední hodnoty,
- Test 3: 6 bodů v řadě rostoucích resp. klesajících,
- Test 4: 14 bodů v řadě pravidelně kolísá nahoru dolu.

3. Aplikace regulačních diagramů na sledované procesní charakteristiky

V různých časech, kdy lze dle operátorů proces považovat za stabilní, jsme vybrali interval o rozsahu 1000 hodnot a zaznamenávali jsme všechny sledované charakteristiky. Každou charakteristiku lze vykreslit v čase pomocí průběhového diagramu (ilustrujeme na charakteristice „vodní hladina“, viz Obr. 1), toto zobrazení není příliš přehledné, proto použijeme regulační diagram Xbar-S (Obr. 2), kde velikost skupiny volíme 10. Z tohoto regulačního diagramu vidíme, že ač jsou data v globálním pohledu nestabilní, směrodatné odchylky počítané z deseti následujících hodnot jsou stabilní.

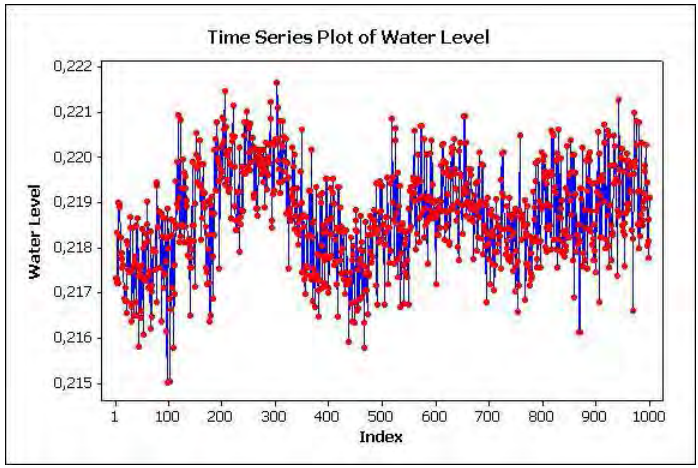
Podobné výsledky dostaneme i v okamžiku, kdy je v datech patrný trend, viz Obr. 3. V tomto případě Xbar-S diagram sice vykazuje odlehlé hodnoty, ale riziko poplašného signálu lze omezit rozšířením regulačních mezí pro danou charakteristiku (např. 4,5 násobek směrodatné odchylky).
Samozřejmě jsou zde i charakteristiky, pro které jsou regulační diagramy, založené na sledování stability krátkodobé variability, nepoužitelné viz Obr. 4.

4. Mimořádná událost (porucha)

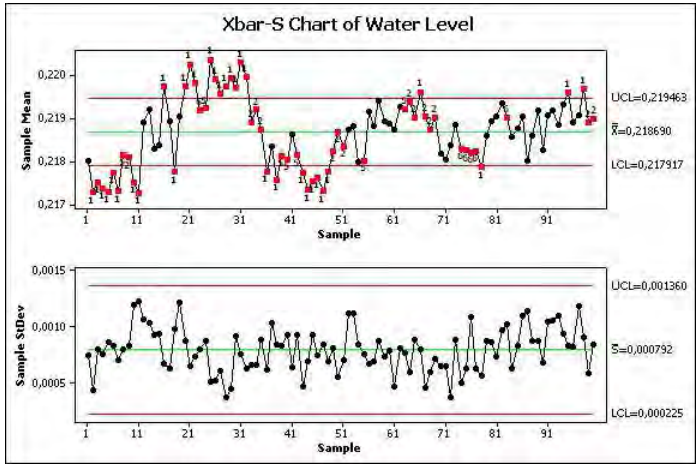
Pro další vyhodnocení byly využity časové průběhy sledovaných charakteristik, u kterých došlo k mimořádné události. Předpokládejme, že máme 1801 sledovaných hodnot daných procesních charakteristik ze dne, kdy došlo k poruše. Dále bylo vyhodnoceno, že po cca 307 hodnotách došlo k poruše (časový rámec pro daný příklad není podstatný). V tomto případě je i pouhým okem vidět, že většina charakteristik se změnila, viz Obr. 5. Neřešíme co je příčina a co důsledek. Většina charakteristik se začala výrazně měnit během cca 5 hodnot (*Water Surface, Output*), některé charakteristiky reagovaly se značným zpožděním (*Value*), jinde není porucha z trendu patrná (*Level*).

5. Vytvoření alarmu

Nyní se pokusíme vytvořit alarm, který by automaticky reagoval na změnu v procesu. Použijeme data z předchozí kapitoly a vykopírujeme 201 až 315 hodnotu, tedy v datech by mělo dojít k poruše ve 107 hodnotě. Protože



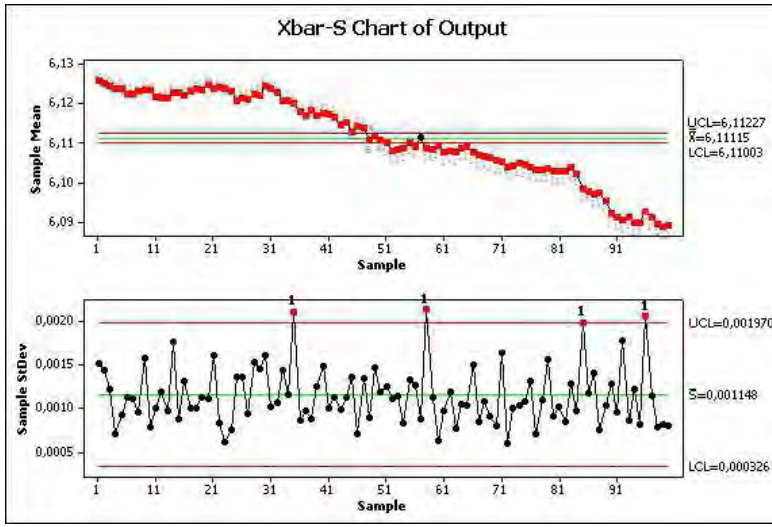
Obrázek 1. Příklad posuzované časové řady dpecifického procesu.



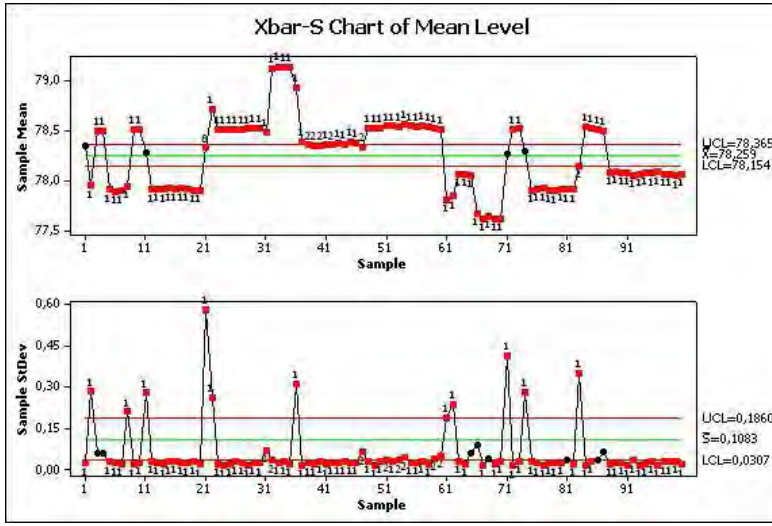
Obrázek 2. Xbar-S regulační diagram procesu dle Obr. 1.

chceme reagovat na poruchu co nejdříve, nebudeme pracovat s regulačními diagramy pro podskupiny, ale použijeme individuální hodnoty. Abychom omezili riziko falešného alarmu, aktivován je pouze test 1.

Jako nejvhodnější nástroj pro detekci poruch se jeví diagram klouzavého rozpětí (*Moving Range*). Tento diagram velice dobře reagoval ve většině případů a nedocházelo u něj k falešným alarmům, jako při využití diagramu individuálních hodnot, viz Obr. 6. Tento diagram je rovněž netečný k datům



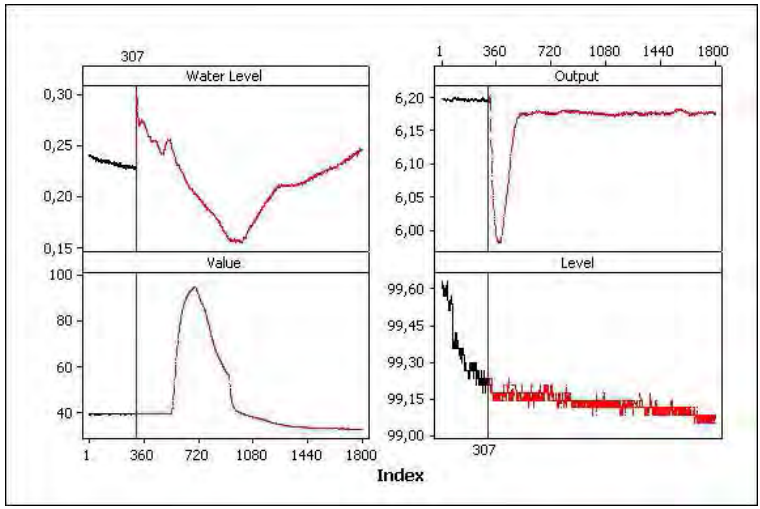
Obrázek 3. Xbar-S regulační diagram procesu vykazujícího trend.



Obrázek 4. Příklad procesu kde Xbar-S regulační diagram selže.

s trendovou složkou, viz Obr. 7. Samozřejmě pokud se charakteristika pravidelně přepíná mezi několika diskrétními stavy, je tento regulační diagram nevhodný, viz Obr. 8.

Test 1, který je standardně formulován: „*1 bod dále než 3 směrodatné odchylky od střední hodnoty*“, zobecníme na „*1 bod dále než K směrodatných*



Obrázek 5. Příklady sledovaných procesních charakteristik před a po události. Svislá čára v grafech odděluje data před a po detekované změně procesu.

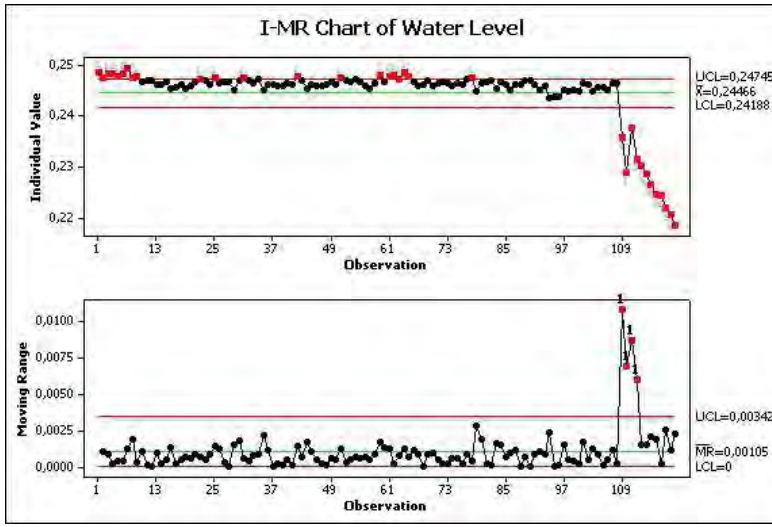
odchylek od střední hodnoty“ a budeme hledat vhodné K pro různé charakteristiky popisující proces tak, abychom minimalizovali pravděpodobnost chyby falešného alarmu i chybějícího alarmu.

6. Závěr

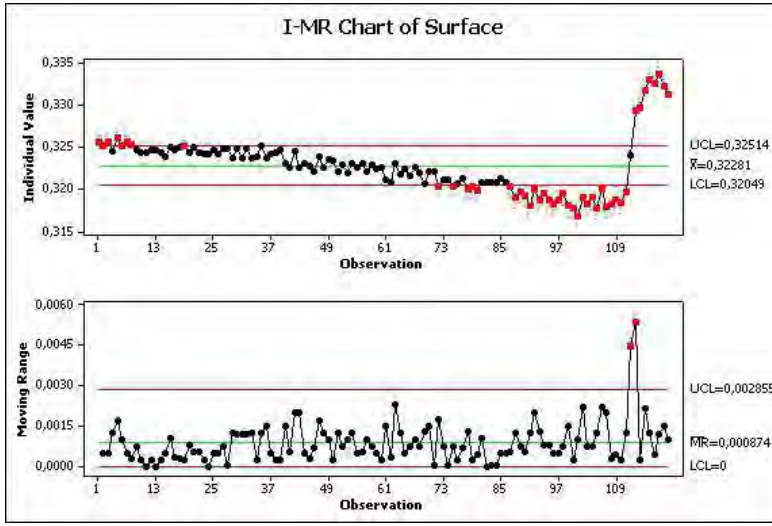
Klasické regulační diagramy s modifikovanými testy vymezitelných příčin jsou velice rychlým a efektivním nástrojem vhodným pro tvorbu systému real-time inteligentních alarmů v energetickém provozu jaderných elektráren. Pro většinu sledovaných veličin jsou nejvhodnější regulační diagramy založené na sledování stability krátkodobé variability, protože nereagují na dlouhodobé trendy, které se v datech vyskytují a mohou být například způsobeny cíleným zásahem obsluhy nebo regulačního mechanismu. Alarmy založené na uvedených regulačních diagramech budou implementovány do softwarové aplikace, která již nyní pomocí jiných mechanismů vyhodnocuje stabilitu sledovaných procesů.

Literatura

[1] Cézová E. Ekonomicko-statistický návrh regulačního diagramu, *sborník konference Request'08*, CQR, VUT Brno, 2008.
[2] ČSN ISO 8258 Shewhartovy regulační diagramy. Praha: ČNI, 1993.
[3] Kupka, K. *Statistické řízení jakosti.* 1. vyd. Pardubice: TriloByte, 2001. 191 s. ISBN 80-238-1818-X.



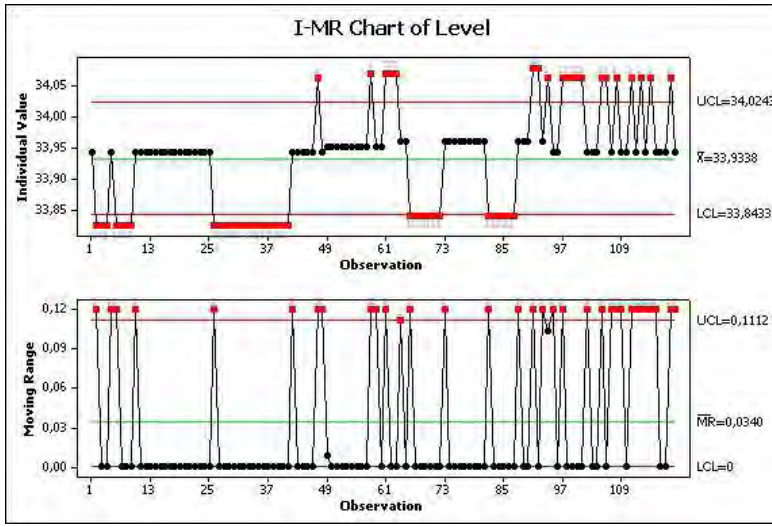
Obrázek 6. I-MR diagram detekující změnu procesu. *Moving Range* (dole).



Obrázek 7. I-MR diagram detekující změnu procesu při trendu hodnot.

[4] Meloun, M.; Militký, J. *Kompendium statistického zpracování dat.* 2. vyd. Praha: Academia, nakladatelství Akademie věd České republiky, 2006. 982 s. ISBN 80-200-1396-2.

[5] Tošenovský, J.; Noskievičová, D. *Statistické metody pro zlepšování jakosti.* 1. vyd. Ostrava: Montanex, a. s., 2000. 362 s. ISBN 80-7225-040-X.



Obrázek 8. Příklad procesu kde I-MR regulační diagram selže.

[6] Runger, G. C. Multivariate statistical process control for autocorrelated processes. *Int. Journal of production Research*, 1996, Vol. 34, pp. 1715-1724.

[7] Nasimi, E.; Gabbar, H.A., Development of support tool for control design of nuclear power plant using hierarchical control chart (HCC), In *Int. J. of Process Systems Engineering*, 2011 Vol.1, No.2, pp.150 - 168, DOI: 10.1504/IJPSE.2011.038943

[8] Steady-State Trend Data Analysis for Applications in Predictive Maintenance and Operations. *Technical Report (corporate document)* EPRI, Palo Alto, CA: 2006. 1010200.

[9] Rossetti, M.D.; Zhe Li; Qu, P., Exploring exponentially weighted moving average control charts to determine the warm-up period, In *Simulation Conference, 2005 Proceedings of the Winter* pp.10,4-4, Dec. 2005 DOI: 10.1109/WSC.2005.1574321

Poděkování: Práce vznikla jako součást řešení projektu TAČR TA02021449: Systém inteligentních alarmů v energetickém provozu jaderných elektráren.

EKONOMICKO-STATISTICKÁ OPTIMALIZACE REGULAČNÍHO DIAGRAMU

ECONOMICAL AND STATISTICAL OPTIMIZATION OF CONTROL CHART

Eliška Cézová, Gejza Dohnal

Adresa: ČVUT v Praze, Fakulta strojní, Ústav technické matematiky, Karlovo nám. 13, 121 35 Praha 2

Abstrakt: Hlavním cílem této práce je studium postupů a algoritmů vhodných pro ekonomicko-statistickou optimalizaci regulačních diagramů. Je zde navržen a podrobně rozebrán návrh nového typu zónového regulačního diagramu. Vedle toho se práce soustřeďuje na otázku výběru vhodného regulačního diagramu v praxi nejenom na základě jeho matematicko-statistických vlastností, ale též s přihlédnutím k ekonomické stránce jeho nasazení v praxi. Pozornost je též věnována metodám numerického řešení problémů ekonomicko-statistické optimalizace regulačních diagramů.

Abstract: The main objective of this thesis is the study of methods and algorithms suitable for optimizing economic and statistical control chart. New type of zone control chart is designed and its properties analyzed. Moreover, the work focuses on the issue of selecting the appropriate control chart in practice, which is based not only on its mathematical and statistical characteristics, but also take into accout the economics of deployment. Attention is also paid to the numerical methods of solving the problemes of economical and statistical optimization of control charts.

Klíčová slova: regulační diagram, optimalizace, ekonomicko-statistiká optimalizace, SPC, ARL, CUSUM, EWMA, Shewhartův regulační diagram, zónový regulační diagram, ekonomicko-statistický model s údržbou, ekonomicko-statistický model bez údržby, scénář

Keywords: control chart, optimization, optimization of economics statistics, SPC, ARL, CUSUM, EWMA, Shewhart of control chart, control chart zone, economics statistics model with maintenance, economics statistics model without maintenance, scenario

1. Úvod

Ekonomicko-statistická optimalizace spočívá v minimalizování předpokládaných ztrát za jednotku času. Za určitých podmínek a z určitého hlediska lze považovat výrobní proces za proces obnovy, tedy proces pracující v určitých cyklech zvaných cykly obnovy. O těchto cyklech předpokládáme, že jejich délky jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti. Označme očekávané náklady za celý cyklus v procesu jako C a očekávanou

délku cyklu v procesu jako T, potom můžeme vyjádřit očekávanou střední ztrátu $E(L)$ za jednotku času jako poměr $E(C)$ ku $E(T)$:

$$E(L) = \frac{E(C)}{E(T)}. \tag{1}$$

Výrobní proces v praktickém provozu je spojen s řadou nákladových položek, které určují jeho efektivitu. Nízká efektivita použití regulačních diagramů bohužel často vede k jejich odmítání ze strany vedení u firem. Cílem ekonomicko-statistické optimalizace je nalézt takové parametry regulačních diagramů, které ztráty minimalizují a přinesou maximální zisk, [6], [11], [17].

První ekonomicko-statistický model, který vycházel z Shewhartova regulačního diagramu, navrhl v roce 1956 A. J. Duncan [6]. V daném modelu uvažoval náklady na inspekce, náklady na vadné produkty, náklady na vyhledání falešného signálu, náklady na vyhledání a odstranění zjistitelné příčiny. Předpokládáme přitom, že známe parametry, to je střední hodnotu procesu δ_0, posun v procesu δ a směrodatnou odchylku měření σ. Mezi optimalizované proměnné parametry patří rozsah výběru n, délka intervalu mezi inspekcemi h a šířka regulačních mezí k, viz [5], [6], [7], [8], [18], [20].

V tomto příspěvku uvedeme nejčastěji používaný rozšířený ekonomicko-statistický model T. J. Lorenzena a L. C. Vance. Tento model předpokládá standardní průběh statistické regulace jakožto procesu obnovy. Jednotlivé cykly obnovy začínají počátkem pozorování procesu který je pod statistickou kontrolou, detekcí změny způsobené zjistitelnou příčinou, její opravou a opětovným uvedením procesu do stavu pod statistickou kontrolou. Model nezahrnuje předpoklad provádění preventivní údržby. Údržba, která je v praxi běžně prováděna, může změnit parametry celého regulačního procesu. Zároveň má vliv na ekonomické dopady regulace – zpravidla přináší úspory plynoucí z menší pravděpodobnosti běhu procesu mimo statistickou kontrolu a tedy menší pravděpodobnosti výskytu nekvalitních produktů.

2. **Model Lorenzen-Vance**

V roce 1986 T. J. Lorenzen a L. C. Vance navrhli rozšířený ekonomicko-statistický model, který v sobě zahrnuje náklady na vyhledání falešného signálu, náklady na vyhledání a odstranění zjistitelné příčiny, náklady na vadné produkty a náklady na inspekce.

Délka cyklu T v tomto modelu je složena z doby ve stavu pod statistickou kontrolou T_{In} a doby mimo statistickou kontrolu T_{Out}. Pro střední hodnoty platí

$$E(T_{In}) = \frac{1}{\lambda} + \frac{(1-\gamma_z)s\ T_f}{ARL_0}, \tag{2}$$

kde první člen na pravé straně rovnice (2) vyjadřuje očekávanou dobu do poruchy běžícího procesu, λ je intenzita výskytu zjistitelné příčiny (poruchy). Druhý člen prodlužuje dobu, kdy je proces pod statistickou kontrolou o intervaly, ve kterých proces stojí a my detekujeme falešný signál po dobu T_f.

Parametr γ_z je indikátorem běhu procesu: je roven jedné, pokud proces po dobu detekce běží a je roven nule v případě, že proces v průběhu detekce neběží. Symbol ARL_0 označuje průměrný počet inspekcí které proběhnou než dojde k falešnému signálu, je-li proces pod statistickou kontrolou. Je $ARL_0 = \frac{1}{\alpha}$, kde $\alpha = P(\text{nastane signál} \,|\, \text{proces pod kontrolou})$.

Celková doba, kdy je proces mimo statistickou kontrolu zahrnuje pět částí:

$$E(T_{Out}) = h - \nu + h(ARL_\delta - 1) + T_{gn} + T_z + T_r. \tag{3}$$

První je doba, která uběhne od poslední inspekce před tím, než nastane zjistitelná příčina do okamžiku vzniku této příčiny. Tato doba je rovna $(h-\nu)$. ν je doba od vzniku zjistitelné příčiny do první následující inspekce. Druhou část tvoří doba, která signalizuje stav mimo statistickou kontrolu je $h(ARL_\delta - 1)$. Symbol ARL_δ označuje průměrný počet inspekcí které proběhnou než dojde k signálu v případě, že proces je mimo statistickou kontrolu s posunem cílové hodnoty o $\delta\sigma$. Třetí část obsahuje dobu potřebnou k zakreslení a výpočtu standardních testů jednoho výsledku při inspekci (obsahující n měření), je-li proces mimo statistickou kontrolu, kterou označíme jako T_{gn}. Čtvrtá a pátá část jsou reprezentovány dobou na vyhledání zjistitelné příčiny T_z a dobou k odstranění zjistitelné příčiny (odstranění poruchy) v procesu, kterou označíme jako T_r.

Střední náklady na cyklus v tomto modelu jsou složeny ze tří částí:

– střední náklady na na odběr vzorků za cyklus

$$E(C_S) = \frac{C_s\left(\frac{1}{\lambda} - \nu + T_{gn} + h(ARL_\delta) + \gamma_z T_z + \gamma_r T_r\right)}{h},$$

kde C_S jsou jednotkové náklady na jednu inspekci,

– očekávané náklady na detekci a opravu zjistitelné příčiny

$$E(C_D) = \frac{sC_f}{ARL_0} + C_{zr},$$

kde C_f jsou náklady na detekci falešného poplachu a C_{zr} náklady na detekci a opravu zjistitelné příčiny,

– očekávané náklady plynoucí z nekvalitní výroby za cyklus

$$E(C_Q) = \frac{C_I}{\lambda} + C_O\left(- \nu + T_{gn} + h(ARL_\delta) + \gamma_z T_z + \gamma_r T_r\right),$$

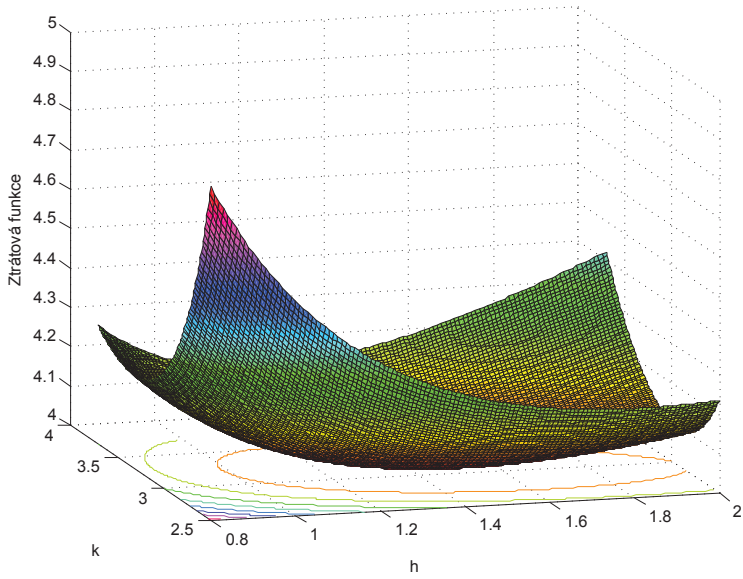
kde C_I jsou jednotkové náklady (ztráta) z nekvalitní výroby v době, kdy je proces pod statistickou kontrolou, C_O jsou jednotkové náklady (ztráta) z nekvalitní výroby v době, kdy je proces mimo statistickou kontrolu.

Celkovou ztrátovou funkci lze potom vyjádřit jako:

$$E(L) = \frac{E(C_Q) + E(C_D) + E(C_S)}{E(T)}.$$

Podrobnosti o tomto modelu lze nalézt například v [5], [1], [11], [12], [13].

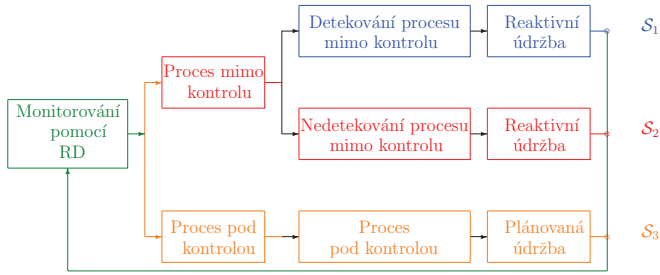
Optimalizace spočívá v nalezení parametrů n, h, k, při nichž ztrátová funkce nabývá svého minima. V tomto případě je je třeba použít některou z numerických metod pro hledání extrému funkce více proměnných. K nalezení optimálních parametrů rozsahu výběru m, dobu mezi inspekcemi h a šířkou regulačních mezí k byly v literatuře publikovány počítačové programy napsané v jazyce FORTRAN (viz [12], [13]). Nabízí se zde také Nelderova-Meadova simplexová metoda (viz [15]), snadno aplikovatelná například v prostředí Matlab.



Obrázek 1. Průběh ztrátové funkce v modelu Lorenzena-Vance při hodnotě rozsahu výběru $n = 5$.

3. **Ekonomicko-statistické modely s údržbou**

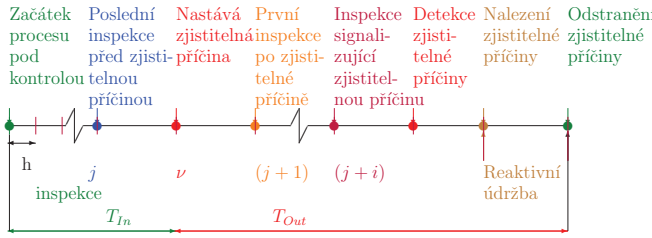
Základní ekonomicko-statistický model s údržbou se skládá ze tří scénářů, jak je vidět na 2. Průběh procesu zaznamenáváme do regulačních diagramů, z kterých zjišťujeme, zda je proces pod statistickou kontrolou nebo není. Pokud je ve stavu pod statistickou kontrolou, provedeme v plánovaném čase údržbu procesu, která předchází poruše v procesu a reaktivní údržbě. Reaktivní údržbu provádíme tehdy, kdy regulační diagram detekuje proces mimo statistickou kontrolu. Po provedení reaktivní nebo plánované údržby se proces vrací do stavu pod statistickou kontrolu, viz [5], [10]. Ve všech třech scénářích předpokládáme, že začínáme ve stavu pod statistickou kontrolou,



Obrázek 2. Základní ekonomicko-statistický model s údržbou.

inspekce provádíme po h hodinách a po údržbě se proces vrací do stavu „jako nový“.

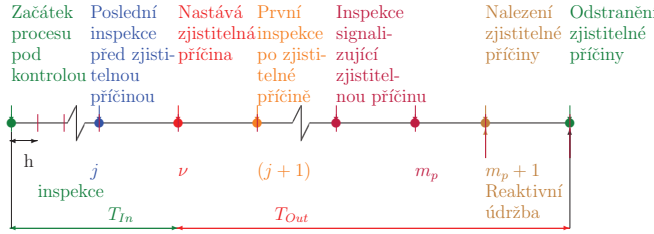
Ve scénáři $(\mathcal{S}_1)$ podle 3 předpokládáme, že v intervalu mezi j-tou a $(j+1)$-ní inspekcí dojde ke zjistitelné příčině, která posune proces mimo statistickou kontrolu. Po zakreslení výsledků do regulačního diagramu, detekujeme zjistitelnou příčinu. Regulační diagram signalizuje stav mimo statistickou kontrolu mezi $(j+i)$-tou inspekcí. Zjistíme, zda signál je oprávněný. Pokud ano, hledáme zjistitelnou příčinu, která způsobila posun v procesu. Po identifikování zjistitelné příčiny provedeme reaktivní údržbu, která vrátí proces do stavu pod statistickou kontrolu.



Obrázek 3. Scénář $(\mathcal{S}_1)$ (detekování procesu mimo kontrolu).

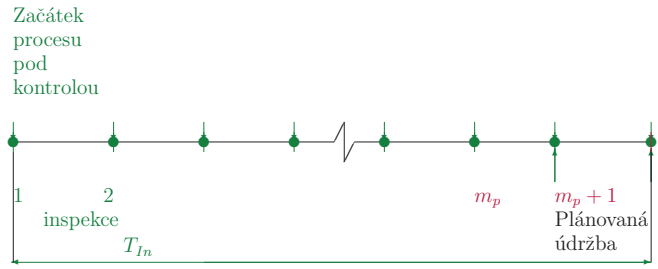
Ve scénáři $(\mathcal{S}_2)$ podle 4 dojde k posunu procesu v intervalu mezi j-tou a $(j+1)$-ní inspekcí, nicméně proces pokračuje, protože regulační diagram nedetekoval posun v procesu, a tudíž nevyslal signál, že je proces mimo statistickou kontrolu před plánovanou údržbou. V (m_p+1)-ním intervalu by měla začít plánovaná preventivní údržba nicméně, protože je indikován stav mimo statistickou kontrolu bude provedena reaktivní údržba, která odstraní problém na zařízení a uvede jej do stejného stavu jako plánová preventivní údržba.

Ve scénáři $(\mathcal{S}_3)$ dle 5 začínáme ve stavu pod statistickou kontrolou a jsme ve stavu pod statistickou kontrolou i před provedením plánované údržby. Plánovaná údržba se uskutečňuje v (m_p+1)-ním vzorkovacím intervalu, tato



Obrázek 4. Scénář ($\mathcal{S}_2$) (nedetekování procesu mimo kontrolu).

údržba zabrání poruše v procesu. Plánovaná údržba je méně nákladná než reaktivní údržba, protože přípravné práce mohou být provedeny za běhu procesu před zahájením této údržby.



Obrázek 5. Scénář ($\mathcal{S}_3$) (proces pod kontrolou před plánovanou údržbou).

Ekonomicko-statistická analýza

Vycházíme ze vzorce (1) pomocí něhož spočítáme ztrátovou funkci za hodinu v procesu.

Očekávanou délku cyklu vyjádříme jako součet všech dob cyklu vynásobených pravděpodobnostmi jednotlivých scénářů, která je dána vztahem:

$$E(T) = E[T|\mathcal{S}_1]P(\mathcal{S}_1) + E[T|\mathcal{S}_2]P(\mathcal{S}_2) + E[T|\mathcal{S}_3]P(\mathcal{S}_3).$$

Podobně, očekávané náklady za cyklus jsou součtem očekávaných nákladů z jednotlivých scénářů, které násobíme pravděpodobnostmi jednotlivých scénářů. Můžeme je vyjádřit následovně:

$$E(C) = E[C|\mathcal{S}_1]P(\mathcal{S}_1) + E[C|\mathcal{S}_2]P(\mathcal{S}_2) + E[C|\mathcal{S}_3]P(\mathcal{S}_3).$$

Proces začíná ve stavu pod statistickou kontrolou s mechanismem poruchy, které má Weibullovo rozdělení s hustotou $f(t) = \lambda^\rho \rho t^{\rho-1} e^{-(\lambda t)^\rho}$, kde $\lambda, \rho, t \geq 0$ a distribuční funkcí, kterou označíme $F(t)$. Pravděpodobnosti jednotlivých

scénářů vyjádříme těmito vztahy:

$$\begin{aligned} P[\mathcal{S}_1] &= F(m_p h)P(\text{signál}\,|\,\text{stav mimo kontrolu}), \\ P[\mathcal{S}_2] &= F((m_p+1)h) - F(m_p h)P(\text{signál}\,|\,\text{stav mimo kontrolu}), \\ P[\mathcal{S}_3] &= 1 - F((m_p+1)h). \end{aligned}$$

Ve scénáři $(\mathcal{S}_1)$ předpokládáme detekci zjistitelné příčiny, která posune proces do stavu mimo statistickou kontrolu, po které následuje provedení reaktivní údržby. V době pod statistickou kontrolou počítáme se střední dobou do poruchy a počtem zjištěných falešných signálů, které vyjádříme pomocí useknutého Weibullova rozdělení na intervalu $\langle 0, (m_p+1)h\rangle$ s hustotou:

$$f^{Tr}(t; m_p+1) = \frac{\lambda^\rho \rho t^{\rho-1} e^{-(\lambda t)^\rho}}{1 - e^{-(\lambda(m_p+1)h)^\rho}}, \quad 0 \le t \le (m_p+1)h.$$

Označme ještě $E(T_{m_p}) = \int_0^{m_p h} t f^{Tr}(t; m_p+1)dt$.

Celková očekávaná doba ze scénáře $(\mathcal{S}_1)$, která je dána součtem doby pod statistickou kontrolou a doby mimo statistickou kontrolu ze scénáře $(\mathcal{S}_1)$, je:

$$E[T|\mathcal{S}_1] = E[T_{in}|\mathcal{S}_1] + E[T_{out}|\mathcal{S}_1].$$

Vezmeme-li do úvahy možnost zastavení procesu v dobách identifikace falešných signálů (viz (2)), dostáváme:

$$E[T_{in}|\mathcal{S}_1] = E(T_{m_p}) + (1-\gamma_z)\frac{sT_f}{ARL_0},$$

kde s je počet inspekcí, které proběhnou, když je proces pod statistickou kontrolou ve scénáři $(\mathcal{S}_1)$.

$$E[T_{out}|\mathcal{S}_1] = hARL_\delta - \nu + T_{gn} + T_z + T_R,$$

kde $\nu = \sum_{i=0}^{m_p} \int_{ih}^{(i+1)h} (t-ih) f^{Tr}(t; m_p+1)dt$.

Celkové očekávané náklady ze scénáře $(\mathcal{S}_1)$ jsou součtem očekávaných nákladů ze ztráty kvality, očekávaných nákladů na vzorkování a očekávaných nákladů na vyhledání falešného signálu a údržbu:

$$E[C|\mathcal{S}_1] = E[C_Q|\mathcal{S}_1] + E[C_S|\mathcal{S}_1] + E[C_D|\mathcal{S}_1].$$

Očekávané náklady ze ztráty kvality ve scénáři $(\mathcal{S}_1)$ jsou:

$$E[C_Q|\mathcal{S}_1] = C_I E(T_{m_p}) + C_O\left[hARL_\delta - \nu + T_{gn} + \gamma_z T_z + \gamma_R T_R\right].$$

Očekávané náklady na vzorkování ve scénáři $(\mathcal{S}_1)$ jsou:

$$E[C_S|\mathcal{S}_1] = C_s \frac{E(T_{m_p}) + h(ARL_\delta) - \nu + T_{gn} + \gamma_z T_z + \gamma_R T_R}{h}.$$

Očekávané náklady na vyhledání falešného signálu a údržbu jsou:

$$E[C_D|\mathcal{S}_1] = \frac{sC_f}{ARL_0} + C_R.$$

Ve scénáři ($\mathcal{S}_2$) předpokládáme, že zjistitelná příčina není detekována a proces je mimo statistickou kontrolu, provádíme reaktivní údržbu.

Celková očekávaná doba ze scénáře ($\mathcal{S}_2$), která je dána součtem doby pod statistickou kontrolou a doby mimo statistickou kontrolu ze scénáře ($\mathcal{S}_2$), je:

$$E[T|\mathcal{S}_2] = E[T_{in}|\mathcal{S}_2] + E[T_{out}|\mathcal{S}_2],$$

$$E[T_{in}|\mathcal{S}_2] = E(T_{m_p+1}) + (1-\gamma_z)\frac{sT_f}{ARL_0},$$

$$E[T_{out}|\mathcal{S}_2] = (m_p+1)h - E(T_{m_p+1}) + T_R.$$

kde $E(T_{m_p+1}) = \int_0^{(m_p+1)h} tf^{Tr}(t; m_p+1)dt$.

Celkové očekávané náklady ze scénáře ($\mathcal{S}_2$) jsou součtem očekávaných nákladů ze ztráty kvality, očekávaných nákladů na vzorkování a očekávaných nákladů na vyhledání falešného signálu a údržbu:

$$E[C|\mathcal{S}_2] = E[C_Q|\mathcal{S}_2] + E[C_S|\mathcal{S}_2] + E[C_D|\mathcal{S}_2].$$

Očekávané náklady ze ztráty kvality ve scénáři ($\mathcal{S}_2$) jsou:

$$E[C_Q|\mathcal{S}_2] = C_I E(T_{m_p+1}) + C_O\left[(m_p+1)h - \int_0^{(m_p+1)h} tf(t)dt + \gamma_R T_R\right].$$

Očekávané náklady na vzorkování ve scénáři ($\mathcal{S}_2$) jsou:

$$E[C_S|\mathcal{S}_2] = m_p C_s.$$

Očekávané náklady na vyhledání falešného signálu a provedení reaktivní údržby ve scénáři ($\mathcal{S}_2$) jsou:

$$E[C_D|\mathcal{S}_2] = \frac{sC_f}{ARL_0} + C_R.$$

Ve scénáři ($\mathcal{S}_3$) předpokládáme stav pod statistickou kontrolou, provedeme pouze plánovanou údržbu, která předchází zjistitelné příčině. Dobu potřebnou pro realizaci plánované údržby budeme dále označovat symbolem T_P. Celková očekávaná doba ve scénáři ($\mathcal{S}_3$) je:

$$E[T|\mathcal{S}_3] = (m_p+1)h + (1-\gamma_z)\frac{m_p T_f}{ARL_0} + T_P.$$

Celkové očekávané náklady ze scénáře ($\mathcal{S}_3$) jsou součtem očekávaných nákladů ze ztráty kvality, očekávaných nákladů na vzorkování a očekávaných nákladů na vyhledání falešného signálu a údržbu:

$$E[C|\mathcal{S}_3] = E[C_Q|\mathcal{S}_3] + E[C_S|\mathcal{S}_3] + E[C_D|S_3].$$

Očekávané náklady ze ztráty kvality ve scénáři ($\mathcal{S}_3$) jsou (γ_P je indikátor běhu procesu v průběhu plánované údržby):

$$E[C_Q|\mathcal{S}_3] = C_I[(m_p+1)h + \gamma_P T_P].$$

Očekávané náklady na vzorkování ve scénáři ($\mathcal{S}_3$) jsou:

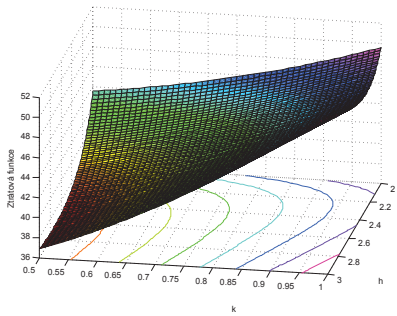
$$E[C_S|\mathcal{S}_3] = m_p C_s.$$

Očekávané náklady na plánovanou údržbu ve scénáři ($\mathcal{S}_3$) jsou:

$$E[C_D|\mathcal{S}_3] = \frac{m_p C_f}{ARL_0} + C_P.$$

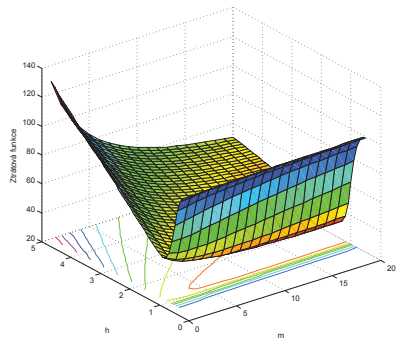
Příklad: Uvažujme proces řízený Shewhartovým regulačním diagramem pro střední hodnotu. Předpokládáme posunutí ve střední hodnotě $\delta = 2$, pravděpodobnost doby vzniku do poruchy má Weibullovo rozdělení s parametrem měřítka $\lambda = 0.05$ a tvaru $\rho = 1$. Mezi známé parametry, které se týkají nákladů jsou: náklady za hodinu procesu ve stavu mimo statistickou kontrolu $C_O = 100$ Kč, náklady za hodinu procesu ve stavu pod statistickou kontrolou $C_I = 0$ Kč, náklady na falešný signál $C_f = 5$ Kč, náklady na vyhledání a odstranění zjistitelné příčiny $C_{zr} = 25$ Kč, náklady na provedení reaktivní údržby $C_R = 50$ Kč, náklady na provedení plánované údržby $C_P = 75$ Kč, náklady závislé na odebírání počtu vzorků a zakreslení výběrového bodu do regulačního digramu $C_V = 1.0$ Kč a náklady, které nezávisí na počtu odběru vzorků $C_F = 5$ Kč. Předpokládáme, že známe dobu potřebnou k zakreslení a výpočtu standardních testů jednoho výsledku při inspekci, je-li proces mimo statistickou kontrolu $T_g = 0.05$ h, dobu k nalezení zjistitelné příčiny $T_r = 2$ h, dobu na vyhledání falešného signálu $T_f = 1$ h, doba na vyhledání zjistitelné příčiny $T_z = 1$ h, dobu k provedení reaktivní údržby $T_R = 3$ h a dobu k provedení plánované údržby $T_P = 8$ h. Pokračuje-li proces během vyhledávání zjistitelné příčiny, reaktivní údržby a plánované údržby stanovíme parametr $\gamma_z = \gamma_R = \gamma_P = 1$. Dále předpokládáme počet vzorků před plánovanou údržbou $m_p = 300$.

Průběh ztrátové funkce pro hodnotu rozsahu výběru $m = 5$ je na obrázku 6. Minimální hodnoty ztrátové funkce L = 36.6120 je dosaženo při době mezi inspekcemi $h = 2$ a šířkou regulačních mezí $k = 1.24$. Optimalizace zde byla provedena pomocí Nelderovy-Meadovy simplexové metody v prostředí Matlab.



Obrázek 6. Průběh ztrátové funkce pro základní ekonomicko-statistický model s údržbou pro hodnotu rozsahu výběru $m = 5$.

Tento výsledek je ovšem příkladem toho, že samotná ekonomická optimalizace nevede vždy k nejlepším výsledkům. V daném případě nízká hodnota šířky regulačních mezí k vedla k nepřijatelně nízké hodnotě $ARL_0 = 4.6517$, tedy k velmi četným falešným signálům. V tomto případě je potřeba provést statisticko-ekonomickou optimalizaci. Nejprve určíme optimální hodnotu šířky regulačních mezí k_0 pro požadovanou minimální hodnotu ARL_0. Potom opět provedeme ekonomickou optimalizaci pro tuto optimální hodnotu šířky regulačních mezí k. V našem příkladu jsme dostali šířku regulačních mezí $k_0 = 3.024$ pro požadovanou hodnotu $ARL_0 = 400$.



Obrázek 7. Průběh ztrátové funkce pro základní ekonomicko-statistický model s údržbou pro hodnotu šířky regulačních mezí $k = 3.024$.

Následnou ekonomicko-statistickou optimalizací jsme získali hodnotu ztrátové funkce $L = 31.3277$, která je zobrazena na 7, při rozsahu výběru $m = 11$ a dobou mezi inspekcemi $h = 2.4$ při hodnotě $ARL_0 = 400.8716$.

4. Závěr

V příspěvku byl popsán model pro ekonomicko-statistickou optimalizaci statistické regulace procesu s údržbou. Tento model je rozšířením obvykle používaného modelu Lorenzena-Vance. Optimalizační algoritmus je založen na simplexové Nelderově-Meadově metodě a byl naprogramován v prostředí Matlab.

Literatura

[1] Baud-Lavigne B., Bassetto S., Penz B. (2009) *A broader view of the economic design of the X-bar chart in semiconductor industry.* International Journal of Production Research, pp. 1 – 12

[2] Cézová E. (2008) *Ekonomicko-statistický návrh regulačního diagramu*, Request 08, CQR VUT Brno, ISBN 978-80-214-3774-6

[3] Cézová E. (2009) *Ekonomické aspekty statistické regulace.* Dostupné online z `http://www.isq.cz/npj/2009/`

[4] Cézová E. (2011) *Optimální regulační diagramy.* Dostupné online z http://www.isq.cz/npj/2011/06_2011.eps/

[5] Cézová E. (2012) *Ekonomická analýza statistické regulace výrobního procesu s údržbou.* dostupné online z http://www.isq.cz/npj/2012/09_CezovaNL2012.eps

[6] Duncan A. J. (1956) *The economic design of X-charts used to maintain current control of a process.* Journal of the American Statistical Association, Vol. 51, No. 274, pp. 228 – 242

[7] Duncan A. J. (1971) *The economic design of X-charts when there is a multiplicity of assignable causes.* Journal of the American Statistical Association, Vol. 66, No. 333, pp. 107 – 121

[8] Engin A. B. (2008) *Determination of optimum economic inspection by economic control chart design and by machine efficiency estimation: An application in weaving industry.* Simulation modelling practice and theory 16, pp. 147 – 170

[9] James R. Evans, Willliam M. Lindsay (1993) *The Management and Control of Quality* ., Second Edition, West publishing Company, USA

[10] Linderman K., Anderson J. C., McKone-Sweet K. E. (2005) *An integrated systems approach to process control and maintenance.* Europen Journal of Operational Research 164, pp. 324 – 340

[11] Lorenzen T. J. a Vance, L. C. (1986) *The economic design of control charts: a unified approach.* Technometrics 28, pp. 3 – 10

[12] McWilliams P. T. (1994) *Economic, statistical and economic-statistical X chart designs.* Journal of Quality Technology, vol. 26, No.3, pp. 227 – 238

[13] McWilliams P. T., Saniga E.M., Davis D.J. (1995) *Economic, statistical and economic-statistical design of attribute charts.* Journal of Quality Technology, vol. 27, No.1, pp. 56 – 73

[14] Douglas C. Montgomery (2001) *Introduction to Statistical Quality Control.*, Four Edition

[15] Nelder J. A., Mead R. (1965) *A simplex method for function minimization.* The computer journal 7(4), pp. 308 – 313

[16] Nelson L. (1984) *The Shewhart control chart tests for special causes.* J. Quality Technology 16, pp. 237-239.

[17] Nenes G., Tagaras G. (2007) *The economically designed two-sided Bayesian X control chart.* European Journal of Operational Research 183, pp. 263 – 277

[18] Prabhu S. S., Montgomery D. C., Runger G. C. (1997) *Economic-statistical design of an adaptive X chart.* Int. J. Production Economics 49, pp. 1 – 15

[19] Shewhart W. A. (1931) *Economic control of quality of manufactured product.* New York: Van Nostrand.

[20] Vommi V. B., Seetala Murty, S. N. (2007) *A simple approach for robust economic design of control charts.* Computers & Operations Research 34, pp. 2001 – 2009

MATICE VZTAHŮ A VÝZNAMNOSTI PARAMETRŮ PROCESŮ V MRM, JEJÍ TVORBA, STRUKTURA, ANALÝZA, VÝZNAM Z HLEDISKA KONKRETIZACE VAZEB A VZTAHŮ

RELATION AND PARAMETERS SIGNIFICANCE MATRIX IN MRM METHOD, ITS CREATION, STRUCTURE,ANALYSIS, IMPORTANCE FROM RELATIONS CONCRETIZATION POINT OF VIEW

Radim Flegl

Adresa: ISQ PRAHA, s.r.o., Pechlátova 19, 150 00 Praha 5; flegl.radim@isq.cz

Abstrakt: Příspěvek pojednává o podstatě a základních principech metody relačních matic, jakožto vhodné metody pro hodnocení parametrů procesů a objektů v oblasti výroby, služeb, veřejné a státní správy. Autor v příspěvku vysvětluje doporučenou strukturu relačních matic, uvádí jejich význam a možnosti využití pro zkoumání vztahů mezi parametry a požadavky procesů a objektů. Autor se dále zabývá i doporučeným postupem pro tvorbu relačních matic, způsoby jejich výpočtu a možnostmi analýz.

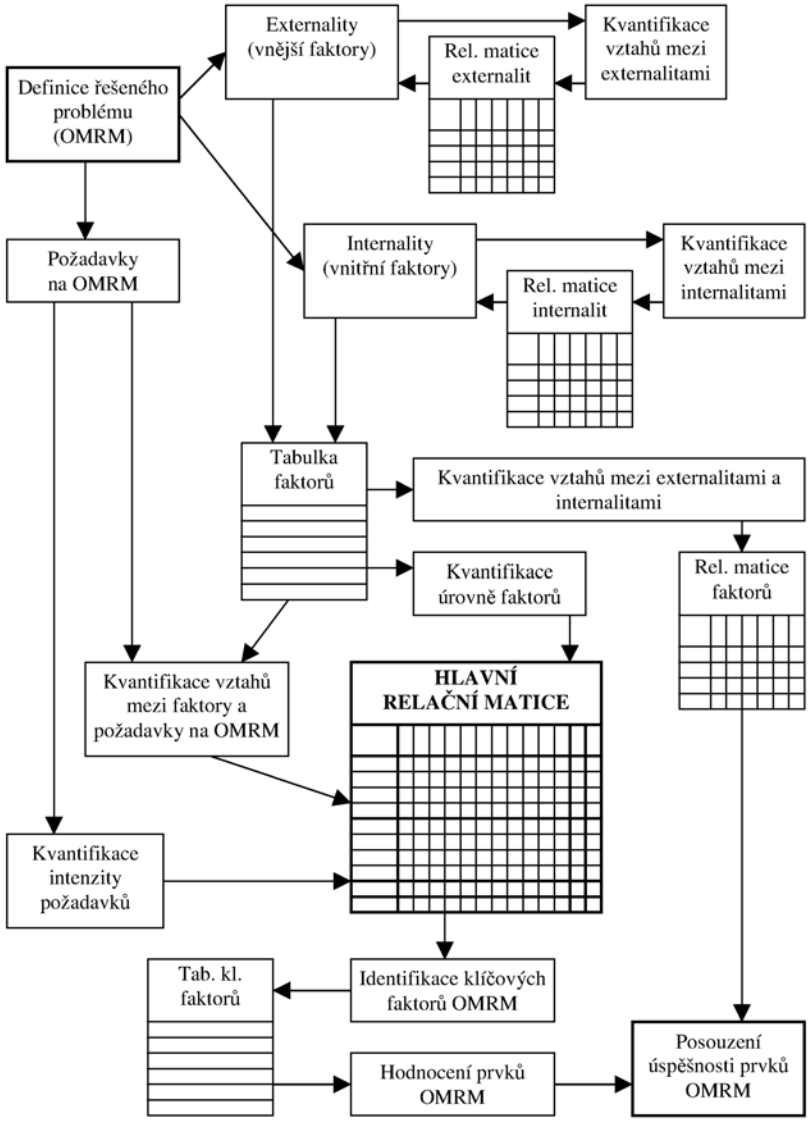
Abstract: The contribution deals with basis and basic principles of Relation Matrix Method as an appropriate method for process and objects parameters evaluation in the field of production providing services and public and state administration. Author explains recommended structure of relation matrices, their importance and possibilities of utilization in analysis of relations between parameters and requirements of processes and objects. Author follows up recommended procedure for relation matrices creation, ways of enumeration and analysis possibilities.

Klíčová slova: Kvantifikace, management procesů, management projektů

Keywords: Quantification, process management, project management

1. Úvod

Pro vlastní implementaci metody MRM je využíván metodický postup uvedený v příspěvku „Význam a možnosti využití MRM ve výrobě, službách, veřejné a státní správě“ Ing. O. Krále, PhD., CSc., zveřejněném v tomto buletinu. Cílem tohoto příspěvku je ukázat, jakým způsobem je tento postup v praxi naplňován a jaké skýtá metoda MRM možnosti pro analýzu zkoumaných objektů zájmu.



Obrázek 1. Schema MRM

2. Postup pro sestavení a výpočet relačních matic

2.1. Krok 1 - Identifikace internalit, externalit a požadavků

Cílem této etapy je identifikovat a kvantifikovat hlavní vlivy (internality a externality) na plnění požadavků kladených na MRM (např. zadání projektu, očekávané či již dosažené výsledky projektu atp.).

Přípravná fáze této etapy zahrnuje:

- identifikaci hlavních požadavků na OMRM (zadání projektu),
- identifikaci faktorů, které ovlivňují výstupy OMRM ve smyslu plnění požadavků na něj kladených,
- kategorizaci na internality a externality.

V této fázi je výhodné využití metody brainstormingu. Výstupem fáze je:

(1) seznam požadavků na OMRM,
(2) seznam internalit,
(3) seznam externalit.

2.2. Krok 2 - Kvantifikace vztahu internalit a externalit uvnitř skupin mezi sebou

Cílem této etapy je identifikovat a kvantifikovat vztahy jednotlivých internalit, resp. externalit uvnitř skupiny mezi sebou. To nám umožňuje internality (resp. externality) sdružit do skupin a omezit tak jejich počet.

V další fázi je sestrojen maticový diagram typu „L“, do kterého jsou zanášeny vazby mezi jednotlivými internalitami, resp. externalitami. Hodnocení provádí expertní skupina, experti provádějí hodnocení opět samostatně jako v předchozí etapě a výsledek je agregován průměrováním.

Experti hodnotí vztah mezi jednotlivými internalitami, resp. externalitami. Pro své hodnocení používají 2 proměnné:

- sílu vztahu internalit (SI), resp. externalit (SE) a
- kompozitní korekční koeficient vztahu internalit (KI), resp. externalit (KE).

Výsledné bodové hodnocení intenzity vztahu (II), resp. (IE) vznikne součinem síly a vztahu mezi internalitami, resp. externalitami a příslušným korekčním koeficientem:

$$II_{ij} = SI_{ij} \cdot KI_{ij}, \qquad resp. \qquad IE_{ij} = SE_{ij} \cdot KE_{ij}$$

kde:

i, j ... čísla internalit, resp. externalit,
II_{ij} ... intenzita vztahu i-té a j-té internality,
IE_{ij} ... intenzita vztahu i-té a j-té externality,
SI_{ij} ... síla vztahu i-té a j-té internality,
SE_{ij} ... síla vztahu i-té a j-té externality,
KI_{ij} ... kompozitní korekční koeficient vztahu i-té a j-té internality,
KE_{ij} ... kompozitní korekční koeficient vztahu i-té a j-té externality,

Kompozitní korekční koeficienty vztahu internalit, resp. externalit mohou být konstruovány různými způsoby dle charakteru řešeného problému. Nejjednodušší je využití 2 základních proměnných, obdobně jako u ostatních korekčních koeficientů:

(1) Korekční koeficient vztahu internalit resp. externalit dle charakteru OMRM (KIC, resp.KEC),
(2) Korekční koeficient stavu OMRM vztahu internalit resp. externalit dle stavu OMRM (KIS, resp. KES).

Kompozitní koeficient je zkonstruován jako aritmetický průměr obou koeficientů:

$$KI_{ij} = \frac{KIC_{ij} + KIS_{ij}}{2}; \qquad KE_{ij} = \frac{KEC_{ij} + KES_{ij}}{2}$$

		Internality									
		I_1	I_2	I_3							I_m
Internality	I_1										
	I_2										
	I_3										
	I_m										

Obrázek 2. Relační matice vztahu internalit mezi sebou

Následně se provádí vyhodnocení tabulek s cílem, pokud se vyskytnou internality se silnou vazbou na jiné internality (resp. pokud se vyskytnou

		Externality									
		E_1	E_2	E_3							E_n
Externality	E_1										
	E_2										
	E_3										
	E_n										

Obrázek 3. Relační matice vztahu externalit mezi sebou

externality se silnou vazbou na jiné externality), provést sloučení těchto internalit, resp. externalit, což sníží celkový počet internalit, resp. externalit a hlavní analýza MRM se stane přehlednější.

Určení, které faktory a jakým způsobem lze sloučit není jen matematická záležitost, ale vyžaduje i řízenou diskusi expertů. Při slučování faktorů do skupin je k přehlednému znázornění vhodné použít afinitní diagram.

2.3. **Krok 3 - Kvantifikace vztahu internalit a externalit**

Cílem této etapy je identifikovat a kvantifikovat vztahy jednotlivých internalit a externalit.

Fáze navazuje etapu kvantifikace vlivů na plnění požadavků na OMRM, ve které byly:

(1) identifikovány internality,
(2) identifikovány externality,
(3) stanoveno hodnocení anticipované úrovně externalit a internalit (faktorů) na základě:

- výpočtů z tzv. tvrdých dat,
- expertního hodnocení.

V další fázi je sestrojen maticový diagram typu „L“, do kterého jsou zanášeny vazby mezi jednotlivými internalitami a externalitami. Hodnocení provádí expertní skupina, experti provádějí hodnocení opět samostatně jako v předchozí etapě a výsledek je agregován průměrováním. Hodnocení vychází jak z výpočtů z tzv. tvrdých dat, tak i vlastních úsudků jednotlivých expertů.

Experti hodnotí vztah mezi jednotlivými internalitami a externalitami. Pro své hodnocení používají 2 proměnné:

- sílu vztahu (SF) a
- kompozitní korekční koeficient vztahu internalit a externalit (KF).

Výsledné bodové hodnocení intenzity vztahu k-té internality a l-té externality (IF_{kl}) vznikne součinem síly vztahu mezi internalitou a externalitou a kompozitním korekčním koeficientem k-té internality a l-té externality:

$$IF_{kl} = SF_{kl} \cdot KF_{kl}$$

kde:

k ... číslo internality,
l ... číslo externality,
IF_{kl} ... intenzita vztahu k-té internality a l-té externality,
SF_{kl} ... síla vztahu k-té internality a l-té externality,
KF_{kl} ... kompozitní korekční koeficient k-té internality a l-té externality.

Kompozitní korekční koeficient vztahu internalit a externalit může být konstruován různými způsoby dle charakteru řešeného problému. Nejjednodušší je využití 2 základních proměnných, obdobně jako u ostatních korekčních koeficientů:

(1) Korekční koeficient vztahu internalit a externalit dle charakteru OMRM (KFC),
(2) Korekční koeficient stavu OMRM vztahu internalit a externalit dle stavu OMRM (KFS).

Kompozitní koeficient je potom zkonstruován jako aritmetický průměr obou koeficientů:

$$KF_{ij} = \frac{KFC_{ij} + KFS_{ij}}{2}$$

Sumace intenzit vztahů jednotlivých internalit pronásobené anticipovanými úrovněmi internalit nám po znormování dávají výsledné priority jednotlivých internalit:

		Externality												
		E_1	E_2	E_3							E_n	**Úroveň Internalit**	**Σ**	**Výsledné hodnocení internalit**
Internality	I_1													
	I_2													
	I_3													
	I_m													
Úroveň externalit														
Σ														
Výsledné hodnocení externalit														

Obrázek 4. Relační matice vztahu externalit mezi sebou

$$PIF_k = \frac{\sum_{l=1}^{n}(IF_{kl} \cdot UI_k)}{\sum_{k=1}^{m}\sum_{l=1}^{n}(IF_{kl} \cdot UI_k)} \cdot 100[\%]$$

kde:

k ... číslo internality,
l ... číslo externality,
m ... počet internalit (vnitřních faktorů),
n ... počet externalit (vnějších faktorů),
IF_{kl} ... intenzita vztahu k-té internality a l-té externality,
UI_k ... anticipovaná úroveň k-té internality,
PIF_l ... priorita l-té internality vzhledem k externalitám.

Sumace intenzit vztahů jednotlivých externalit pronásobené anticipovanými úrovněmi externalit nám po znormování dávají výsledné priority jednotlivých externalit:

$$PEF_l = \frac{\sum_{k=1}^{m}(IF_{kl} \cdot UE_l)}{\sum_{l=1}^{n}\sum_{k=1}^{m}(IF_{kl} \cdot UE_l)} \cdot 100[\%]$$

kde:

k	... číslo internality,
l	... číslo externality,
m	... počet internalit (vnitřních faktorů),
n	... počet externalit (vnějších faktorů),
IF_{kl}	... intenzita vztahu k-té internality a l-té externality,
UE_l	... anticipovaná úroveň l-té externality,
PEF_l	... priorita l-té externality vzhledem k internalitám.

Stanovení priorit nám z hlediska praktického řízení projektu umožňuje:

- Soustředit se na zlepšování internalit, které mají silné vazby na externality (a jsou proto důležité).
- Identifikovat externality, které mají silné vazby na internality (a potřeba je proto sledovat).

2.4. **Krok 4 – Hodnocení úrovně faktorů a požadavků**

Součástí expertního hodnocení je hodnocení anticipované úrovně faktorů (internalit a externalit) a intenzity požadavků na OMRM. Logika tohoto přístupu vychází z toho, že:

- Faktory (internality či externality), které očekáváme, že budou problematické (např. máme zkušenosti z předchozích projektů, existují určité indicie, že zde hrozí problém apod.), je potřeba v hodnocení zdůraznit.
- Je třeba zohlednit i intenzitu jednotlivých požadavků na OMRM –některé požadavky může zadavatel chápat jako zásadní, jiné spíše doplňkové. Zásadní požadavky je potřeba v hodnocení zdůraznit.

2.5. **Krok 5 – Hodnocení vztahů mezi faktory a požadavky**

Experti hodnotí vztah mezi jednotlivými faktory a požadavky na MRM. Pro své hodnocení používají:

- sílu vztahu (SP), která reprezentuje funkční vazbu mezi faktory a požadavky a
- kompozitní korekční koeficient vztahu faktorů a požadavků (KP). Výsledné bodové hodnocení intenzity vztahu i-tého faktoru a j-tého požadavku (IP_{ij}) vznikne součinem síly vztahu mezi požadavkem a faktorem a korekčním koeficientem:

$$IP_{ij} = SP_{ij} \cdot KP_{ij},$$

kde:

i ... číslo faktoru,
j ... číslo požadavku,
IP_{ij} ... intenzita vztahu i-tého faktoru a j-tého požadavku,
SP_{ij} ... síla vztahu i-tého faktoru a j-tého požadavku,
KP_{ij} ... kompozitní korekční koeficient i-tého faktoru a j-tého požadavku.

Kompozitní korekční koeficient vztahu faktorů a požadavků může být konstruován různými způsoby dle charakteru řešeného problému. Nejjednodušší je využití 2 základních proměnných:

(1) Korekční koeficient vztahu faktorů a požadavků dle charakteru OMRM (KPC),
(2) Korekční koeficient vztahu faktorů a požadavků dle stavu OMRM (KPS).

Oba koeficienty nabývají hodnoty 1-5. Kompozitní koeficient je potom zkonstruován jako aritmetický průměr obou koeficientů:

$$KP_{ij} = \frac{KPC_{ij} + KPS_{ij}}{2}$$

Charakterem OMRM může být např. projekt, investice, produkt, podnik, instituce, hodnocení prvků systému atp. Hodnota korekčního koeficientu charakteru OMRM 5 se přidělí, pokud je vztah významný s ohledem na charakter OMRM, hodnota 1 se přidělí, pokud vztah vzhledem k charakteru OMRM je nevýznamný. Zkoumaný vztah může být kupř. velmi významný u rozsáhlých projektů, méně významný u malých projektů a nevýznamný u hodnocení produktů.

Stavem OMRM může být např. fáze projektu (návrh, příprava, realizace, vyhodnocování). Hodnota korekčního koeficientu stavu OMRM 5 se přidělí, pokud je vztah významný s ohledem na stav OMRM, hodnota 1 se přidělí, pokud vztah vzhledem ke stavu OMRM je nevýznamný. Zkoumaný vztah může být kupř. velmi významný v definiční fázi projektu, méně ve fázi realizace a nevýznamný ve fázi vyhodnocení projektu.

2.6. **Krok 6 - Sestavení a výpočet hlavní relační matice**

V další fázi je sestrojen maticový diagram typu „L“, do kterého jsou zanášeny vazby mezi jednotlivými faktory (internality a externality) a požadavky na OMRM. Hodnocení provádí expertní skupina, experti provádějí hodnocení samostatně (na rozdíl od předchozí týmové fáze) a výsledek je agregován průměrováním.

Sumace intenzit vztahů jednotlivých faktorů (internalit a externalit) pronásobené anticipovanými úrovněmi faktorů nám po znormování dávají výsledné priority jednotlivých faktorů:

$$PFP_i = \frac{\sum_{j=1}^{p}(IP_{ij} \cdot UF_i)}{\sum_{i=1}^{n+m}\sum_{j=1}^{p}(IP_{ij} \cdot UF_i)} \cdot 100[\%]$$

kde:

i ... číslo faktoru (externality či internality),
j ... číslo požadavku,
m ... počet internalit (vnitřních faktorů),
n ... počet externalit (vnějších faktorů),
p ... počet požadavků,
IP_{ij} ... intenzita vztahu i-tého faktoru a j-tého požadavku,
UF_i ... anticipovaná úroveň i-tého faktoru,
PFP_i ... priorita i-tého faktoru vzhledem k požadavkům.

Sumace intenzit vztahů jednotlivých požadavků pronásobené anticipovanými úrovněmi intenzit požadavků nám po znormování dávají výsledné priority jednotlivých požadavků:

$$PPP_j = \frac{\sum_{i=1}^{n+m}(IP_{ij} \cdot UP_j)}{\sum_{j=1}^{p}\sum_{i=1}^{n+m}(IP_{ij} \cdot UP_j)} \cdot 100[\%]$$

kde:

i ... číslo faktoru (externality či internality),
j ... číslo požadavku,
m ... počet internalit (vnitřních faktorů),
n ... počet externalit (vnějších faktorů),
p ... počet požadavků,
IP_{ij} ... intenzita vztahu i-tého faktoru a j-tého požadavku,
UP_j ... anticipovaná úroveň j-tého požadavku,
PPP_j ... priorita j-tého požadavku vzhledem k faktorům.

		Požadavky na OMRM										Úroveň faktorů	Σ	Výsledné hodnocení faktrorů
		P_1	P_2	P_3							P_p			
Internality	I_1													
	I_2													
	I_3													
	I_m													
Extrernality	E_1													
	E_2													
	E_3													
	E_n													
Intenzita požadavků														
Σ														
Výsledné hodnocení požadavků														

Obrázek 5. Relační matice vztahu faktorů (internalit a externalit) a požadavků na OMRM

2.7. **Krok 7 – Vyhodnocení hlavní relační matice**

Smyslem hlavní relační matice je stanovení významu jednotlivých faktorů (internalit a externalit) a požadavků z hlediska síly jejich vazeb, očekávané úrovně plnění faktorů a intenzity požadavků.

Stanovení priorit nám z hlediska praktického řízení projektu umožňuje:

- Soustředit se na zlepšován internalit, které mají silné vazby na požadavky na OMRM (a jsou proto významné), dle stanovené úrovně závažnosti.
- Soustředit se na zlepšování internalit, které mají špatnou úroveň a mohou se tedy stát pro projekt ohrožujícími, dle stanovené úrovně závažnosti.
- Identifikovat, analyzovat a kvantifikovat požadavky na OMRM, které mají silné vazby na faktory (a jsou proto významné).
- Identifikovat, analyzovat a kvantifikovat požadavky na OMRM, které budou obtížně splnitelné z důvodu jejich vysoké intenzity.

Vhodné grafické znázornění výstupů z kvantifikace v OMRM následně umožňuje rychlou orientaci v řešené problematice a rozhodování o dalším postupu prací či přijetí rozhodnutí.

2.8. **Krok 8 - Hodnocení prvků OMRM dle klíčových faktorů**

Dalším nezbytným krokem metody MRM je hodnocení kvality prvků OMRM dle klíčových faktorů. Klíčovými faktory chápeme internality a externality, které byly v předchozích etapách identifikovány a vyhodnoceny jako pro činnost OMRM významné.

Tato etapa umožňuje podrobnější pohled na strukturu OMRM z hlediska výše uvedených faktorů. Z praktického hlediska nám tato analýza může pomoci hodnotit úspěšnost jednotlivých prvků systému, např. organizačních jednotek, tj. útvarů, pracovišť, pracovních týmů atp. a porovnávat je mezi sebou.

Hodnocení prvků OMRM podle stanovených faktorů se provádí podle následující stupnice.

Tab. č. 1 - Klasifikační stupnice hodnocení neshod

Hodnotící stupnice prvků OMRM	
1	Shoda s požadavky – splnění požadavků na OMRM.
2	Méně závažná neshoda – neshoda komplikující realizaci činností, neohrožující funkci OMRM.
3	Významnější neshoda – neshoda způsobující snížení schopnosti funkce OMRM.
4	Významná (kritická) neshoda – neshoda přímo ohrožující funkci OMRM.

Pro účely kvantitativního hodnocení lze u každého z realizovaných prvků OMRM vypočítat ukazatel hodnocení HA podle následujícího vztahu:

$$HA = w_e \cdot HA_e + w_i \cdot HA_i[\%]$$

kde:

w_e ... váha neshod typu externality, např. $w_e = 0,666$,
w_i ... váha neshod typu internality, např. $w_i = 0,333$,
HA_e ... dílčí ukazatel hodnocení externalit,
HA_i ... dílčí ukazatel hodnocení internalit.

$$HA_e = \left(1 - \frac{P_{e1} \cdot 0 + P_{e2} \cdot 1 + P_{e3} \cdot 2 + P_{e4} \cdot 3}{P_{ec} \cdot 3}\right) \cdot 100[\%]$$

kde:

P_{ec} ... je celkový počet externalit, hodnocených v rámci hodnocení prvků OMRM,
P_{e1} ... počet externalit, hodnocených v rámci daného prvku stupněm 1,
P_{e2} ... počet externalit, hodnocených v rámci daného prvku stupněm 2,
P_{e3} ... počet externalit, hodnocených v rámci daného prvku stupněm 3,
P_{e4} ... počet externalit, hodnocených v rámci daného prvku stupněm 4.

$$HA_i = \left(1 - \frac{P_{i1} \cdot 0 + P_{i2} \cdot 1 + P_{i3} \cdot 2 + P_{i4} \cdot 3}{P_{ic} \cdot 3}\right) \cdot 100[\%]$$

kde:

P_{ic} ... je celkový počet internalit, hodnocených v rámci hodnocení prvků OMRM,

P_{i1} ... počet internalit, hodnocených v rámci daného prvku stupněm 1,

P_{i2} ... počet internalit, hodnocených v rámci daného prvku stupněm 2,

P_{i3} ... počet internalit, hodnocených v rámci daného prvku stupněm 3,

P_{i4} ... počet internalit, hodnocených v rámci daného prvku stupněm 4.

Hodnocení zjištění daného prvku OMRM ukazatelem HA charakterizuje míru způsobilosti (připravenosti) posuzovaného prvku k plnění požadavků na OMRM.

Prvek OMRM	Internality							Externality							Popis zjištění	HA_i	HA_e	HA
	1	2					N	1	2					M				
Počty zjištěných internalit/ externalit																		
. . .																		

Obrázek 6. Rámec vyhodnocovací tabulky pro hodnocení prvků OMRM

POROVNÁNÍ OPERAČNÍCH TECHNIK POMOCÍ NEPARAMETRICKÉ PREDIKTIVNÍ INFERENCE

COMPARISON OF SURGERY TECHNIQUES BASED ON NONPARAMETRIC PREDICTIVE INFERENCE

Kateřina Janurová

Adresa: Kateřina Janurová, VŠB-TU Ostrava, Fakulta elektrotechniky a informatiky, Katedra aplikované matematiky; katerina.janurova@vsb.cz

Abstrakt: Cílem tohoto článku je na jedné straně porovnání dvou různých operačních technik a na straně druhé představení relativně nové neparametrické indukční metody. Neparametrická prediktivní inference (NPI) je vhodnou alternativou ke standardním metodám využívaných v analýze přežití, jako jsou Kaplan-Meierův odhad funkce přežití a log-rank test, protože umí zacházet s daty, která obsahují cenzorovaná pozorování. Data tohoto typu mohou vzniknout z různých situací, např. měřením času přežití pacientů v medicínských studiích, životností mechanických zařízení v industriální spolehlivosti nebo délky časového úseku nezaměstnanosti v modelování zaměstnanosti. Možné pole působnosti je proto velmi širové, NPI přístup může být využit např. v medicíně, inženýrství, ekonomice, sociologii a pojišťovnictví.

Abstract: The goal of this article is to compare two different surgery techniques on one hand and present relatively new nonparametric predictive inferential method on the other hand. Nonparametric predictive inference (NPI) is a proper alternative to the standard nonparametric approach in survival analysis represented for example by the Kaplan-Meier estimator of survival function or the log-rank test, because it can deal with data consisting of event times and right-censoring times. This type of data may arise from various types of situations: the survival times of patients in medical trials, the lifetimes of machine components in industrial reliability or the duration of periods of unemployment in duration modeling. The possible range of applications is therefore quite wide, NPI could be used in medicine, engineering, economics, sociology and insurance industry.

Klíčová slova: neparametrická prediktivní inference (NPI), dolní a horní pravděpodobnost, analýza přežití, zprava cenzorovaná medicínská data, porovnání chirurgických technik

Keywords: nonparametric predictive inference (NPI), lower and upper probability, survival analysis, right-censored medical survival data, comparison of surgery techniques

1. Úvod

Jednou z nejzásadnějších otázek, které vyvstávají obecně u chirurgických zákroků s několika možnými operačními technikami je, kterou z nich zvolit, abychom pacientům po zákroku garantovali celkově delší dobu dožití. Během devadesátých let minulého století stoupl v chirurgii výrazně poměr minimálně invazivních metod a laparoskopické techniky tak v některých případech částečně, nebo dokonce zcela nahradily klasické otevřené operační techniky. Kolorektální operace karcinomu, které je věnována analýza v tomto článku, není výjimkou. Mezi všeobecně známé výhody minimálně invazivních metod patří obvykle menší operační stres pacienta, příznivější pooperační průběh a kratší doba hospitalizace po výkonu. Po takovém výčtu pozitiv se zdá být zjevné, že laparoskopické techniky generují celkově delší čas přežití, na druhou stranu u nich ale existuje mnoho méně známých negativních faktorů, které se mohou velkou měrou podílet na úmrtnosti pacientů (riziko protržení střev u kapnoperitonea, delší operační čas a extrémní pozice pacientů při zákroku). Porovnání morbidity a mortality u obou typů operačních technik je často publikovaným výsledkem mnoha medicínských studií. Například konsensus Evropské asociace endoskopických chirurgů tvrdí, že rozdíl v morbiditě u laparoskopické a otevřené operace kolonu neexistuje [1].

V tomto článku jsou analyzována medicínská, zprava cenzorovaná data 844 pacientů, kteří podstoupili kolektomii ve Fakultní nemocnici v Ostravě v letech 2001 - 2009, za účelem srovnání obou chirurgických technik a zodpovězení otázky která operační technika je méně riskantní, jestli laparoskopická nebo otevřená. Porovnání je provedeno pomocí nově používané metody neparametrické prediktivní inference (NPI), jejíž výsledky jsou konfrontovány s klasickým přístupem reprezentovaným Kaplan-Meierovým odhadem funkce přežití a log-rank testem.

2. Neparametrická prediktivní inference

Neparametrická prediktivní inference (NPI) pro zprava cenzorovaná data, představena Coolenem a Yanem [2], vychází z Berliner-Hillovy metody pro neparametrickou analýzu přežití [3]. Berliner a Hill využili předpokladu $A_{(n)}$ představeného Hillem [4] pro predikci pravděpodobností výskytu budoucího jednoho (nebo více) pozorování na základě n předchozích pozorování. Předpoklad $A_{(n)}$ je definován na principu de Finettiho zaměnitelnosti posloupnosti nezávislých náhodných kvantit a tvrdí že pokud máme n uspořádaných kvantit $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, pak pravděpodobnost pořadí následující náhodné kvantity X_{n+1} vstupující do studie je rovnoměrně rozdělená mezi hodnoty od 1 do $n+1$ ($P(X_{n+1}(x_{(i)}, x_{(i+1)})) = 1/(n+1)$, pro $i = 0, \ldots, n$, kde $x_{(0)} = 0$ a $x_{(n+1)} = \infty$). Jinými slovy je stejně pravděpodobné, že co se týká pořadí sestaveného podle velikosti, bude další pozorovaná kvantita stejně pravděpodobně první největší, druhá největší, třetí atd.. Předpoklad $A_{(n)}$

tak poskytuje částečně specifikované prediktivní rozdělení pravděpodobnosti pro budoucí pozorování, které je popsáno pomocí pravděpodobnostních hodnot přiřazených jednotlivým otevřeným intervalům mezi pozorovanými časy událostí. Tyto pravděpodobnostní hodnoty jsou omezeny na daný interval, ale jejich rozložení v intervale není blíže specifikováno ani omezováno. Modifikovaný předpoklad $A_{(n)}$ pro zprava cenzorovaná data určí v dalším textu prediktivní pravděpodobnosti obdobným způsobem, proto je zavedeno označení pro tyto pravděpodobnostní hodnoty jako M-funkce.

Definice (M-funkce). Částečná specifikace rozdělení pravděpodobnosti reálné náhodné kvantity T může být určena pomocí pravděpodobnostních hodnot přiřazených jednotlivým intervalům, bez dalších omezení kladených na rozdělení pravděpodobnosti uvnitř těchto intervalů. Pravděpodobnostní hodnota přiřazená tímto způsobem intervalu (a, b) je označena jako hodnota M-funkce pro kvantitu T na (a, b), ozn. $M_{T(a,b)}$.

Předpoklad $A_{(n)}$ bohužel není dostatečný k odvození klasických přesných pravděpodobností splňujících Kolmogorovy axiomy, což je nutné u mnoha zajímavých problémů, na druhou stranu ale poskytuje dostatečné hranice pro pravděpodobnosti pro všechny problémy zahrnující X_{n+1}. Těmito hranicemi jsou dolní a horní pravděpodobnosti z teorie intervalové pravděpodobnosti [5] a jako takové mají silné konsistenční vlastnosti [6]. Dolní a horní pravděpodobnost jsou označeny jako $\underline{P}$ a $\overline{P}$ a platí $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$.

NPI přístup je rámcem pro statistickou teorii a metody, které využívají dolní a horní pravděpodobnosti, založené na předpokladu $A_{(n)}$ a obsahuje také několik modifikací $A_{(n)}$, které jsou vhodné pro různé aplikace, např. pro Bernoulliho data, mulninomiální data, nebo pro zprava cenzorovaná data.

2.1. Předpoklad zprava cenzorovaného $A_{(n)}$

Aby se mohly u zprava cenzorovaných dat vzít v úvahu i informace obsažené v cenzorování, zevšeobecnili Coolen a Yan [2] Hillův předpoklad pro zprava cenzorovaná data na zprava cenzorované $A_{(n)}$ (ozn. rc-$A_{(n)}$).

Definice (rc-$A_{(n)}$). Předpokládejme náhodně cenzorovaná data, která mohou být výsledkem experimentu, kde je mezi n nezávislými pozorováními $m \leq n$ pozorovaných úmrtí $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$; p $(= n - m)$ zprava cenzorovaných pozorování $c_{(1)} < c_{(2)} < \ldots < c_{(p)}$ a že ve výběru neexistují shody (nakládání se shodami je diskutováno Coolenem a Yanem [2]). Nechť $t_{(0)} = 0$ a $t_{(m+1)} = \infty$. Potom rc-$A_{(n)}$ částečně specifikuje rozdělení pravděpodobnosti pro budoucí pozorování T_{n+1} pomocí M-funkce jako

$$M_i^T = M_{T_{n+1}}(t_{(i)}, t_{(i+1)}) = \frac{1}{n+1} \prod_{\{r:c_{(r)}<t_{(i)}\}} \frac{\tilde{n}_{c_{(r)}} + 1}{\tilde{n}_{c_{(r)}}}, \tag{1}$$

$$(2) \qquad M_{i,k}^T = M_{T_{n+1}}(c_{(k)}^i, t_{(i+1)}) = \frac{1}{(n+1)\tilde{n}_{c_{(k)}^i}} \prod_{\left\{r:c_{(r)}<\tilde{n}_{c_{(k)}^i}\right\}} \frac{\tilde{n}_{c_{(r)}}+1}{\tilde{n}_{c_{(r)}}},$$

kde $i = 0, 1, \ldots, m; k = 1, 2, \ldots, p$ a $\tilde{n}_{c_{(r)}}$ je počet pacientů v riziku úmrtí (tedy těch, kteří jsou stále naživu) těsně před časem $c_{(r)}$.

Jestliže je součin prováděn přes prázdnou množinu, je definován jako roven jedné. Jedna z $n+1$ hodnot M-funkce, které částečně specifikují rozdělení pravděpodobnosti pro T_{n+1} podle rc-$A_{(n)}$, je přiřazena ke každému z n pozorování na otevřeném intervale od tohoto pozorování do dalšího pozorovaného času úmrtí pacienta (nebo nekonečna) a pro otevřený interval $(0, t_{(1)})$. Součet hodnot M-funkce pro T_{n+1} na všech intervalech je roven jedné a každá hodnota M-funkce je z intervalu [0, 1]. Jestliže ve výběru nejsou žádná cenzorovaná pozorování, je rc-$A_{(n)}$ rovno $A_{(n)}$. Implicitním předpokladem pro rc-$A_{(n)}$ je neinformativní cenzorování [2].

Částečně specifikované pravděpodobnostní rozdělení pro T_{n+1}, založeno na předpokladu rc-$A_{(n)}$, potom umožňuje odvození dolní a horní pravděpodobnosti pro události týkající se T_{n+1}. Dolní pravděpodobnost pro $\underline{P}(T_{n+1} \in B)$ je odvozena jako největší dolní hranice pro $P(T_{n+1} \in B)$, kterou získáme sečtením všech hodnot M-funkce pro T_{n+1} na intervalech náležících zcela do intervalu B. Horní pravděpodobnost $\overline{P}(T_{n+1} \in B))$ je odvozena jako nejmenší horní hranice $P(T_{n+1} \in B)$ vzniklá sečtením všech hodnot M-funkce z intervalů, které mají s intervalem B neprázdný průnik.

Přesné pravděpodobnosti pro $T_{n+1} \in (t_{(i)}, t_{(i+1)})$ mohou být odvozeny díky tomu, že intervaly na kterých jsou specifikovány hodnoty M-funkce (založené na předpokladu rc-$A_{(n)}$), jsou každý plně obsažený v jediném intervale $(t_{(i)}, t_{(i+1)})$, což vede k

$$(3) \qquad P_i^T = P(T_{n+1}) \in (t_{(i)}, t_{(i+1)}) = \frac{1}{n+1} \prod_{\left\{r:c_{(r)}<t_{(i+1)}\right\}} \frac{\tilde{n}_{c_{(r)}}+1}{\tilde{n}_{c_{(r)}}},$$

kde $i = 0, 1, \ldots, m$, $\tilde{n}_{c_{(r)}}$ je počet pacientů v riziku úmrtí těsně před časem $c_{(r)}$ a součin přes prázdnou množinu je definován jako roven jedné.

2.2. **Horní a dolní funkce přežití**

V této sekci se zaměříme na funkci přežití, která udává v každém časovém okamžiku t pravděpodobnost, že čas přežití pacienta bude delší než t ($S(t) = P(T_{n+1} \in (t, \infty))$). Horní($\overline{S}$) a dolní ($\underline{S}$) funkce přežití, založeny na rc-$A_{(n)}$ jsou odvozeny [2], pro všechna $i = 0, 1, \ldots, m$, následovně

$$(4) \qquad \overline{S}_{T_{n+1}}(t) = \sum_{j=i}^{m} \left[M_{T_{n+1}}(t_{(j)}, t_{(j+1)}) + \sum_{k=1}^{l_j} M_{T_{n+1}}(c_{(k)}^j, t_{(j+1)}) \right]$$

pro všechna $t \in (t_{(i)}, t_{(i+1)})$, kde $c^i_{(k)} \in (t_{(i)}, t_{(i+1)}), k = 1, \ldots, l_i$,

$$\underline{S}_{T_{n+1}}(t) = \overline{S}(t_{(i+1)}) + \sum_{\left\{k: c^i_{(k)} \geq t\right\}} M_{T_{n+1}}(c^i_{(k)}, t_{i+1}), \tag{5}$$

pro všechna $t \in (t_{(i)}, t_{(i+1)})$.

Obě, dolní i horní funkce přežití jsou, stejně jako Kaplan-Meierův odhad funkce přežití, schodovité funkce. Jsou konstantní mezi jednotlivými časy pozorování, zatímco horní funkce přežití klesá jen v každém pozorovaném čase úmrtí (cenzorování má však vliv na velikost poklesu), dolní funkce přežití klesá v každém pozorovaném čase události (jak v čase úmrtí, tak v čase cenzorování). Důsledkem je zvýšení rozdílu mezi oběma funkcemi v každém pozorovaném cenzorovaném čase c_r, které trvá až do dalšího pozorovaného času úmrtí, což odpovídá v teorii intervalové pravděpodobnosti ztrátě informací. Horní funkce pravděpodobnosti je na intervalu $[0, t_{(1)})$ rovna jedné a na $[t_{(m)}, \infty)$ kladné konstantě, zatímco dolní funkce přežití je za posledním pozorováním nulová.

2.3. **Neparametrické prediktivní porovnání dvou skupin pacientů**

NPI pro porovnání dvou nezávislých skupin obsahujících data o přežití pacientů, včetně zprava cenzorovaných pozorování, byla představena Coolenem a Yanem [7]. Označme dvě skupiny dat X a Y, porovnání je pak provedeno vypočítáním dolní a horní pravděpodobnosti pro událost, že budoucí pozorování X_{n+1} skupiny X je menší než budoucí pozorování Y_{n+1} skupiny Y. Výpočet je založen na n_x pozorováních ze skupiny X a n_y pozorováních ze skupiny Y a předpokladech rc-$A_{(n_x)}$ a rc-$A_{(n_y)}$. Předpokládejme, že máme ve skupině X m_x pozorovaných časů úmrtí, označených $x_{(1)} < x_{(2)} < \ldots < x_{(m_x)}$ a $p_x (= n_x - m_x)$ zprava cenzorovaných pozorování $c_{(x,1)} < c_{(x,2)} < \ldots < c_{(x,p_x)}$. Nechť je dále $x_{(0)} = 0, x_{(m_x+1)} = \infty$ a označme $l_{x,i}$ počet zprava cenzorovaných pozorování v intervale $(x_{(i)}, x_{(i+1)}), x_{(i)} < c^i_{(x,1)} < c^i_{(x,2)} < \ldots < c^i_{(x,l_{x,i})} < x_{(i+1)}$, tedy že součet $l_{x,i}$ na všech intervalech je roven p_x. Analogickým způsobem předpokládejme, že ve skupině Y máme m_y pozorovaných časů úmrtí, označených $y_{(1)} < y_{(2)} < \ldots < y_{(m_y)}$ a $p_y (= n_y - m_y)$ zprava cenzorovaných pozorování $c_{(y,1)} < c_{(y,2)} < \ldots < c_{(y,p_y)}$. Nechť je dále $y_{(0)} = 0, y_{(m_y+1)} = \infty$ a $l_{y,j}$ počet zprava cenzorovaných pozorování v intervale $(y_{(j)}, y_{(j+1)}), y_{(j)} < c^j_{(y,1)} < c^j_{(y,2)} < \ldots < c^j_{(y,l_{y,j})} < y_{(j+1)}$, takže je součet $l_{y,j}$ na všech intervalech roven p_y. Potom jsou NPI dolní a horní pravděpodobnosti pro událost $X_{n+1} < Y_{n+1}$ definovány jako

Tabulka 1. porovnání skupin

Skupina	**Celkem**	**Úmrtí**	**Cenzorování**	**Poměr cenz.** [%]
Laparoskopická	457	160	295	64,99
Otevřená	387	205	182	47,03
Celkem	**844**	**365**	**477**	**56,75**

$$\overline{P}(X_{n_x+1} < Y_{n_y+1}) = \sum_{i=0}^{m_x}\sum_{j=0}^{m_y} P_j^X \left\{ \mathbf{1}\left\{x_i < y_{j+1}\right\} M_i^X + \sum_{k=1}^{l_{x,i}} \mathbf{1}\left\{c_{x,k}^i < y_{i+1}\right\} M_{i,k}^X \right\}, \tag{6}$$

$$\overline{P}(X_{n_x+1} < Y_{n_y+1}) = \sum_{i=0}^{m_x}\sum_{j=0}^{m_y} P_i^X \left\{ \mathbf{1}\left\{x_{i+1} < y_j\right\} M_j^Y + \sum_{k=1}^{l_{y,j}} \mathbf{1}\left\{x_{i+1} < c_{y,k}^j\right\} M_{j,k}^Y \right\}, \tag{7}$$

kde $M_i^X(M_j^Y), M_{i,k}^X(M_{j,k}^Y), P_i^X(P_j^Y)$ jsou dány popořadě vzorci (1), (2), (3), a kde $\mathbf{1}\{E\}$ je charakteristická funkce definovaná jako jedna jestliže podmínka E nastane, nula jinak.

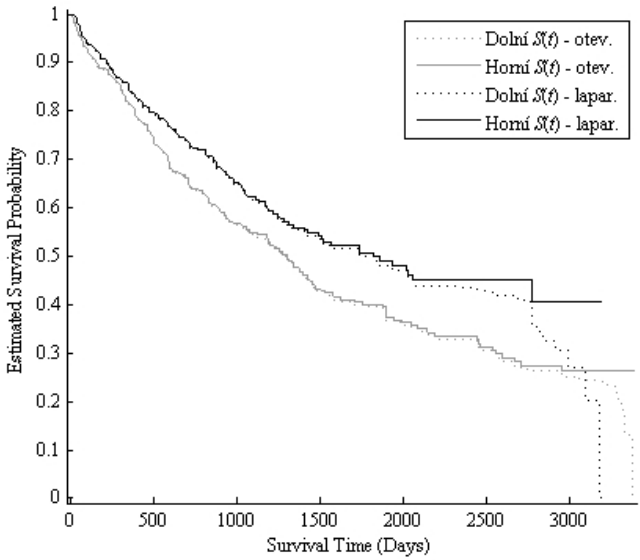
Výsledek $\underline{P}(X_{n_x+1} < Y_{n_y+1}) > 0,5$ je poté interpretován jako silný důkaz pro závěr $X_{n_x+1} < Y_{n_y+1}$.

3. Výsledky

Z dat 844 pacientů byly zkonstruovány, podle rovnic (4) a (5), modely dolní a horní funkce pravděpodobnosti pro budoucího pacienta vstupujícího do studie a to jak pro skupinu pacientů operovaných klasickou otevřenou technikou, tak pro skupinu pacientů operovaných technikou laparoskopickou. Výsledky jsou uvedeny na Obrázku 1, kde jsou různé křivky přiřazeny různým operačním technikám.

Tabulka 1 obsahuje dodatečné informace o obou skupinách pacientů. Vidíme zde celkový počet pacientů, pozorovaný počet úmrtí pacientů, počet zprava cenzorovaných pozorování a celkový poměr těchto pozorování ve studii v jednotlivých skupinách.

Porovnání obou skupin bylo provedeno nejdříve pomocí NPI přístupu a poté konfrontováno s výsledky získanými klasickým log-rank testem. Jestliže O_{388} a L_{458} jsou dvě náhodné kvantity reprezentující čas úmrtí budoucího



Obrázek 1. Dolní a horní funkce přežití pro obě skupiny pacientů

pacienta vstupujícího do studie (označeno popořadě do otevřené a laparoskopické skupiny pacientů), porovnání NPI přístupem vede k dolním a horním pravděpodobnostem:

$$\underline{P}(O_{388} < L_{458}) = 0,511 \qquad \overline{P}(O_{388} < L_{458}) = 0,626, \tag{8}$$

$$\underline{P}(L_{458} < O_{388}) = 0,374 \qquad \overline{P}(L_{458} < O_{388}) = 0,489, \tag{9}$$

což můžeme v terminologii NPI přístupu interpretovat jako silný důkaz pro $O_{388} < L_{458}$, tedy jako důkaz pro závěr, že celkový čas přežití pacientů operovaných laparoskopickou technikou je výrazně delší než celkový čas přežití pacientů operovaných klasickou otevřenou technikou. Výsledkem klasického log-rank testu, za předpokladu nulové hypotézy, že mezi danými skupinami neexistuje statisticky významný rozdíl v délce přežití pacientů, jsou hodnoty

$$\chi^2 = 8,024, p - value = 0,005,$$

což nás opravňuje k závěru, že mezi danými skupinami pacientů existuje statisticky významný rozdíl v délce přežití na hladině spolehlivosti 99 %.

4. Závěr

Hlavní výhodou využití NPI přístupu v analýze přežití je způsob, jakým zachází s cenzorovanými informacemi.

Na rozdíl od tradičního Kaplan-Meierova odhadu funkce přežití, NPI dolní funkce přežití klesá také v každém čase cenzorovaného pozorování a poskytuje tak podstatně více informací o pozorovaných událostech, zejména v pozdním čase studie, jestliže je poslední čas pozorování cenzorovaný.

Na základě výsledků obou neparametrických přístupů můžeme tvrdit, že laparoskopická operační technika přináší ve srovnání s klasickou otevřenou technikou výrazně delší dobu přežití pacientů. Tento závěr je v kontrastu s konsensem Evropské asociace endoskopických chirurgů [1], ale výsledky (8) a (9) naprosto korespondují s výsledky získanými klasickým log-rank testem.

5. Literatura

Literatura

[1] Veldkamp R, Gholghesaei M, Bonjer HJ, Meijer DW, Buunen M, Jeekel J, et al.. Laparoscopic resection of colon cancer: Consensus of the European Association of Endoscopic Suregry. Surg. Endosc. 2004, 18:1163-85.

[2] Coolen F.P.A., Yan K.J. Nonparametric predictive inference with right-censored data. Journal of Statistical Planning and Inference, No. 126, 2004, pp. 25-54.

[3] Berliner L.M., Hill B.M. Bayesian nonparametric survival analysis (with discussion). Journal of the American Statistical Association, No.83, 1988, pp.772-784.

[4] Hill B.M. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of the American Statistical Association, No. 63, 1968, pp. 677-691.

[5] Weichselberger, K.. Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervalwahrscheinlichkeit als umfassendes Konzept (in German). Physika, Hei- delberg, 2001.

[6] Augustin, T. and Coolen, F.P.A. Nonparametric predictive inference and interval probability. Journal of Statistical Planning and Inference 124, 2004, 251-272.

[7] Coolen F.P.A., Yan K.J. Nonparametric Predictive Comparison of Two Groups of Lifetime Data. ISIPTA'03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications, 2003, pp. 148-161.

Poděkování: Tato práce vznikla s podporou grantu SGS SP2012/108 VŠB-TU Ostrava.

INDEXOVÁ ANALÝZA S BOOTSTRAPEM
INDEX ANALYSIS WITH BOOTSTRAP

Zdeněk Karpíšek, Veronika Lacinová, Alena Kocmanová, Zdeněk Sadovský

Adresa: AKADEMIE STING, Stromovka 1, 637 00 Brno; karpisek@sting.cz

Abstrakt: Příspěvek je zaměřen na možnosti využití metody bootstrap při výpočtu intervalových odhadů středních hodnot cen a množství, individuálních jednoduchých indexů, individuálních složených indexů a souhrnných indexů z pozorovaných hodnot množstevních a cenových znaků.

Abstract: The article is focused on the possibilities to use bootstrap method for calculation of the interval estimations of mean prices and quantities, individual simple idexes, individual composite indexes, and general indexes from the observed values of quantity and price variables.

Klíčová slova: Bootstrap, bootstrapový odhad, bootstrapový odhad indexu

Keywords: Bootstrap, bootstrap estimate, bootstrap estimate of index

1. Úvod

Metoda bootstrapových intervalových odhadů je užitečná, potřebujeme-li určit intervalové odhady parametrů pozorované náhodné veličiny, resp. náhodného vektoru, nebo testovat statistické hypotézy o těchto parametrech, ale:

- neznáme nebo neodhadneme rozdělení pravděpodobnosti dané veličiny, resp. vektoru,
- rozsah výběru není dostatečně velký, abychom mohli aplikovat asymptotické odhady.

Od vydání prvního článku [1] se metoda bootstrap velmi rozvinula [2], [3] a našla uplatnění v mnoha oblastech aplikací matematické statistiky. Její základní postupy byly proto implementovány do některých profesionálních statistických softwarových produktů.

2. Princip metody bootstrap

Ze statistického souboru $(x_1, \dots x_n)$ pozorovaných hodnot náhodné veličiny X vytvoříme nový statistický soubor $(x_1^*, \dots, x_n^*)$ náhodným výběrem hodnot x_i s opakováním (s vracením). Takto získaný náhodný výběr se nazývá ***bootstrapový výběr***, resp. ***bootstrapový soubor***. Bootstrapový výběr pak B-krát opakujeme [2], [3]. Počet všech různých bootstrapových výběrů je

$$\binom{n+B-1}{B}.$$

Obecný postup aplikace metody bootstrap:

(1) Získání původního statistického souboru.
(2) Výpočet statistik pro původní statistický soubor.
(3) Vytvoření bootstrapových výběrů.
(4) Výpočet bootstrapových statistik.
(5) Výpočet bootstrapových intervalových odhadů, resp. testování hypotéz.

3. Bootstrapové odhady

Předpokládejme, že chceme odhadnout parametr θ pozorované náhodné veličiny (vektoru) X. Bootstrapový odhad parametru θ je založen na těchto krocích [2], [3]:

(1) Z pozorovaných hodnot $(x_1, \dots x_n)$ náhodného výběru $(X_1, \dots, X_n)$ vypočítá-me odhad $\hat{\theta}$ parametru θ.
(2) Realizujeme B náhodných bootstrapových výběrů $(x_1^*, \dots, x_n^*)$o rozsahu n z pozorovaných hodnot $(x_1, \dots x_n)$. Obvykle přitom volíme $B \gg n$.
(3) Z každého bootstrapového výběru vypočítáme odhad $\hat{\theta}_{b,j}$ parametru θ, $j = 1, \dots, B$. Odtud získáme ***bootstrapový odhad rozptylu*** $D\left(\hat{\theta}\right)$

$$\hat{D}\left(\hat{\theta}\right)_b = \frac{1}{B-1} \sum_{j=1}^{B} \left(\hat{\theta}_{b,j} - \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_{b,i} \right)^2$$

a ***bootstrapový odhad směrodatné odchylky*** $\sigma\left(\hat{\theta}\right)$

$$\hat{\sigma}\left(\hat{\theta}\right)_b = \sqrt{\hat{D}\left(\hat{\theta}\right)_b}.$$

Odhad $\hat{\theta}$ vypočtený z původního statistického souboru $(x_1, \dots x_n)$ je bodovým odhadem parametru θ, ale můžeme jej dle potřeby také nahradit aritmetickým průměrem

$$\frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_{b,j}.$$

Pomocí bootstrapových výběrů získáme ***bootstrapové intervalové odhady*** se spolehlivostí $1-\alpha$ střední hodnoty, rozptylu a směrodatné odchylky náhodné veličiny X. NechÂť $\bar{x}$ je aritmetický průměr a s^2 je rozptyl původního statis-tického souboru $(x_1, \dots x_n)$, $\bar{x}_{b,i}$ je aritmetický průměr a $s^2_{b,i}$ je rozptyl statis-tického souboru z j-tého bootstrapového výběru, $j = 1, \dots, B$. Potom:

(1) ***Bootstrapový intervalový odhad střední hodnoty*** $E(X)$ ***se spolehlivostí*** $1-\alpha$ je

$$\left\langle \bar{x} - t_{b,1-\alpha/2}\frac{s}{\sqrt{n-1}}; \bar{x} - t_{b,\alpha/2}\frac{s}{\sqrt{n-1}} \right\rangle,$$

kde $t_{b,P}$ je P-kvantil statistického souboru $(t_{b,1}, ..., t_{b,B})$ a $t_{b,j} = \frac{\bar{x}_{b,j}-\bar{x}}{s_{b,j}}\sqrt{n-1}$, $j = 1, \ldots, B$.

(2) ***Bootstrapový intervalový odhad rozptylu*** $D(X)$ ***se spolehlivostí*** $1-\alpha$ je

$$\left\langle \frac{ns^2}{\chi^2_{b,1-\alpha/2}}; \frac{ns^2}{\chi^2_{b,\alpha/2}} \right\rangle,$$

kde $\chi^2_{b,P}$ je P-kvantil statistického souboru $\left(\chi^2_{b,1}, ..., \chi^2_{b,P}\right)$ a $\chi^2_{b,j} = \frac{ns^2_{b,j}}{s^2}$, $j = 1, \ldots, B$.

(3) ***Bootstrapový intervalový odhad směrodatné odchylky*** $\sigma(X)$ se spolehlivostí $1-\alpha$ obdržíme z bootstrapového odhadu $D(X)$ pomocí odmocniny.

Existuje řada dalších způsobů stanovení intervalových odhadů uvedených i jiných parametrů založených na bootstrapových výběrech [2], [3]. Mimo popsaných oboustranných intervalových odhadů se používají dle potřeby také jednostranné bootstrapové intervalové odhady.

4. Bootstrapové odhady indexů

Indexy patří mezi poměrové kvantitativní statistické znaky a vyjadřují změnu sledovaného kvantitativního znaku nebo souboru znaků u jedné nebo více statistických jednotek během nějakého časového intervalu nebo vlivem nějakého faktoru [4], [5]. Indexy se obvykle konstruují ve tvaru zlomku, kde v čitateli je hodnota znaku ve srovnávaném, tzv. ***běžném období***, a ve jmenovateli hodnota tohoto znaku v tzv. ***základním období***.

Porovnávané znaky (veličiny) dělíme na ***intenzitní***, vyjadřující cenu, intenzitu apod., které značíme písmenem p a ***extenzitní***, vyjadřující množství, objem, produkci apod., které značíme písmenem q. V ekonomických aplikacích se nejčastěji užívají [4], [5]:

- ***cenové indexy*** pro intenzitní znaky,
- ***množstevní indexy*** pro extenzitní znaky,
- ***hodnotové indexy*** pro spojení extenzitních a intenzitních znaků.

Individuální jednoduché indexy:

(1) Individuální jednoduchý cenový index $I_p = \frac{p_1}{p_0}$.

(2) Individuální jednoduchý množstevní index $I_q = \frac{q_1}{q_0}$.

(3) Individuální jednoduchý hodnotový index $I_h = \frac{q_1 p_1}{q_0 p_0} = I_q I_p$.

Individuální složené indexy:

(1) Individuální složený cenový index (index proměnlivého složení, index průměrných cen)

$$I_{prom.sloz.} = \frac{\bar{p}_1}{\bar{p}_0} = \frac{\frac{\sum\limits_i p_1^{(i)} q_1^{(i)}}{\sum\limits_i q_1^{(i)}}}{\frac{\sum\limits_i p_0^{(i)} q_0^{(i)}}{\sum\limits_i q_0^{(i)}}} = \frac{I_h}{I_q} \quad .$$

(2) Individuální složený množstevní index $I_q = \frac{\sum\limits_i q_1^{(i)}}{\sum\limits_i q_0^{(i)}}$.

(3) Individuální složený hodnotový index $I_h = \frac{\sum\limits_i q_1^{(i)} p_1^{(i)}}{\sum\limits_i q_0^{(i)} p_0^{(i)}}$.

Souhrnné indexy:

(1) Laspeyresův souhrnný index pro intenzitní veličinu $I_p^L = \frac{\sum\limits_i q_0^{(i)} p_1^{(i)}}{\sum\limits_i q_0^{(i)} p_0^{(i)}}$.

(2) Laspeyresův souhrnný index pro extenzitní veličinu $I_q^L = \frac{\sum\limits_i q_1^{(i)} p_0^{(i)}}{\sum\limits_i q_0^{(i)} p_0^{(i)}}$.

(3) Paascheho souhrnný index pro intenzitní veličinu $I_p^P = \frac{\sum\limits_i q_1^{(i)} p_1^{(i)}}{\sum\limits_i q_1^{(i)} p_0^{(i)}}$.

(4) Paascheho souhrnný index pro extenzitní veličinu $I_q^P = \frac{\sum\limits_i q_1^{(i)} p_1^{(i)}}{\sum\limits_i q_0^{(i)} p_1^{(i)}}$.

(5) Souhrnný hodnotový index (index obratu, index nákladů) $I_h = \frac{\sum\limits_i q_1^{(i)} p_1^{(i)}}{\sum\limits_i q_0^{(i)} p_0^{(i)}}$.

(6) Fisherův ideální index pro intenzitní veličinu $I_p^F = \sqrt{I_p^L \; I_p^P}$.

(7) Fisherův ideální index pro extenzitní veličinu $I_q^F = \sqrt{I_q^L \; I_q^P}$.

Bootstrapové intervalové odhady se spolehlivostí $1 - \alpha$ středních hodnot, rozptylů a směrodatných odchylek cen, množství, individuálních jednoduchých indexů, individuálních složených indexů a souhrnných indexů určíme tímto postupem:

(1) Pozorováním dvojic cen a množství u n statistických jednotek ve dvou různých časových obdobích nebo regionech získáme čtyřrozměrný statistický soubor

$$\begin{pmatrix} p_{01}, \ldots, p_{0n} \\ q_{01}, \ldots, q_{0n} \\ p_{11}, \ldots, p_{1n} \\ q_{11}, \ldots, q_{1n} \end{pmatrix}$$

(2) Řádky čtyřrozměrného souboru z kroku 1 jsou jednorozměrné statistické soubory cen a množství

$$(p_{0_1}, \ldots, p_{0_n}), (p_{1_1}, \ldots, p_{1_n}), (q_{0_1}, \ldots, q_{0_n}), (q_{1_1}, \ldots, q_{1_n})$$

a výpočtem individuálních jednoduchých indexů obdržíme jednorozměrné statistické soubory individuálních jednoduchých indexů:

$$(I_{p_1}, \ldots, I_{p_n}), (I_{q_1}, \ldots, I_{q_n}), (I_{h_1}, \ldots, I_{h_n}).$$

(3) Ze čtyřrozměrného souboru z kroku 1 provedeme B bootstrapových výběrů podle sloupců.
(4) Ze všech bootstrapových výběrů dostaneme analogickým způsobem jako v kroku 2 jednorozměrné bootstrapové výběry cen, množství a individuálních jednoduchých indexů. Z těchto bootstrapových výběrů a čtyřrozměrného souboru z kroku 1 vypočteme jednotlivé hodnoty individuálních složených indexů a souhrnných indexů.
(5) Bootstrapové výběry pak zpracujeme způsobem popsaným v oddílu 3.

4.1. **Příklad**

Statistickým šetřením u náhodně vybraných 20 řidičů z České republiky byly zjištěny měsíční nákupy a ceny benzínu Natural 95 v květnu a září 2012. Získané hodnoty jsou v levé části tabulky 1 (jde o expertně simulovaný soubor) a v pravé části této tabulky jsou vypočtené individuální jednoduché indexy.

Základní číselné charakteristiky statistických souborů z tabulky 1 jsou v tabulce 2. Tabulka 3 obsahuje konfidenční a bootstrapové intervalové odhady cen, množství a indexů se spolehlivostí 0,95. Počet bootstrapových výběrů byl $B = 500$. Hodnoty uvedené v tabulkách 2 a 3 byly získány pomocí statistického softwaru STATGRAPHICS Centurion XV.

Z cen a množství z tabulky 1 byly v EXCELU vypočteny průměrné ceny, individuální složené indexy a souhrnné indexy, které jsou uvedeny v tabulce 4. Pro souhrnné indexy byla použita stejná data, což odpovídá např. zohlednění vlivu typu (značky) automobilu na spotřebu benzínu.

Tabulka 1. Ceny (Kč/litr), množství (litr/měsíc) a individuální jednoduché indexy

i	p_0	q_0	p_1	q_1	I_p	I_q	I_h
1	35,2	400	35,9	380	1,0199	0,9500	0,9689
2	35,2	350	36,2	350	1,0284	1,0000	1,0284
3	35,6	350	35,6	310	1,0000	0,8857	0,8857
4	35,8	320	36,5	290	1,0196	0,9063	0,9240
5	35,9	410	36,4	350	1,0139	0,8537	0,8655
6	35,9	420	36,9	380	1,0279	0,9048	0,9300
7	36,5	450	36,9	450	1,0110	1,0000	1,0110
8	36,6	280	37,1	310	1,0137	1,1071	1,1223
9	36,9	290	37,8	250	1,0244	0,8621	0,8831
10	37,1	240	37,5	190	1,0108	0,7917	0,8002
11	37,2	190	37,9	200	1,0188	1,0526	1,0724
12	37,5	200	38,3	180	1,0213	0,9000	0,9192
13	37,9	280	37,5	250	0,9894	0,8929	0,8834
14	38,3	150	38,6	200	1,0078	1,3333	1,3438
15	38,3	180	38,9	220	1,0157	1,2222	1,2414
16	38,5	180	39,2	190	1,0182	1,0556	1,0747
17	38,9	210	39,4	250	1,0129	1,1905	1,2058
18	39,1	160	39,9	170	1,0205	1,0625	1,0842
19	39,2	130	39,5	120	1,0077	0,9231	0,9301
20	39,5	210	40,2	200	1,0177	0,9524	0,9693

Tabulka 2: Základní číselné charakteristiky

	p_0	q_0	p_1	q_1	I_p	I_q	I_h
Average	37,255	270,0	37,81	262,0	1,015	0,992	1,007
Median	37,15	260,0	37,65	250,0	1,017	0,951	0,969
Stnd. deviation	1,401	99,737	1,389	86,912	0,009	0,138	0,140
Var. coeff. (%)	3,759	36,94	3,673	33,17	0,906	13,94	13,93
Minimum	35,2	130,0	35,6	120,0	0,99	0,792	0,800
Maximum	39,5	450,0	40,2	450,0	1,028	1,333	1,344
Range	4,3	320,0	4,6	330,0	0,039	0,542	0,544
Stnd. skewness	0,124	0,725	0,291	0,944	-2,035	1,782	1,617
Stnd. kurtosis	-1,200	-1,048	-1,027	-0,465	1,892	0,513	0,305

Tabulka 5 obsahuje bodové odhady průměrných cen, individuálních a souhrnných indexů vypočtených z bootstrapových výběrů vytvořených z původního souboru hodnot v tabulce 1. Jde o tzv. ***primární bootstrapové výběry***.

Tabulka 6 obsahuje konfidenční a bootstrapové intervalové odhady se spolehlivostí 0,95, vypočtené v softwaru STATGRAPHICS Centurion XV pomocí tzv. ***sekundárních bootstrapových výběrů*** ze souboru cen a indexů získaných primárním bootstrapem.

Tabulka 3: Konfidenční a bootstrapové intervalové odhady průměrů

		p_0		q_0	
Conf. Intervals	Mean	36,5995	37,9105	223,322	316,678
Bootstrap Intervals	Mean	36,67	37,87	228,0	315,0
		p_1		q_1	
Conf. Intervals	Mean	37,1601	38,4599	221,324	302,676
Bootstrap Intervals	Mean	37,215	38,45	231,5	298,0
		I_p		I_q	
Conf. Intervals	Mean	1,01067	1,01929	0,92758	1,05708
Bootstrap Intervals	Mean	1,0105	1,0186	0,9382	1,0513
		I_h			
Conf. Intervals	Mean	0,94151	1,07283		
Bootstrap Intervals	Mean	0,95181	1,07761		

Tabulka 4: Průměrné ceny, individuální složené indexy a souhrnné indexy

$\bar{p}_0$	$\bar{p}_1$	$I_{prom.sloz}$	I_q	I_h	
36,844	37,474	1,01711	0,97037	0,98697	
I_p^L	I_q^L	I_p^P	I_q^P	I_p^F	I_q^F
1,01525	0,97211	1,01528	0,97214	1,01527	0,97213

Tabulka 5: Bodové odhady průměrných cen, individuálních složených indexů a souhrnných indexů

	$\bar{p}_0$	$\bar{p}_1$	$I_{prom.sloz}$	I_q	I_h
Average	36,962	37,594	1,0171	0,9690	0,9856
Median	36,906	37,581	1,0173	0,9704	0,9849
Stnd. deviation	0,3350	0,3297	0,0022	0,0225	0,0228
Var. coeff. (%)	0,906	0,877	0,219	2,321	2,313
Minimum	36,415	37,055	1,0120	0,9320	0,9467
Maximum	37,601	38,184	1,0211	1,0149	1,0330
Range	1,184	1,129	0,0090	0,0828	0,0863
Stnd. skewness	0,4876	0,6167	-0,6928	0,5822	0,6018
Stnd. kurtosis	-0,3233	-0,3658	0,0339	-0,4780	-0,3341

	I_p^L	I_q^L	I_p^P	I_q^P	I_p^F	I_q^F
Average	1,0154	0,9705	1,0155	0,9706	1,0155	0,9706
Median	1,0156	0,9716	1,0154	0,9719	1,0155	0,9717
Stnd. deviation	0,0022	0,0228	0,0021	0,0228	0,0022	0,0228
Var. coeff. (%)	0,2165	2,3514	0,2075	2,3468	0,2119	2,3491
Minimum	1,0106	0,9332	1,0110	0,9333	1,0108	0,9333
Maximum	1,0189	1,0173	1,0190	1,0171	1,0110	1,0172
Range	0,0084	0,0841	0,0080	0,0838	0,0082	0,0840
Stnd. skewness	-0,4372	0,5621	-0,2761	0,5517	-0,3590	0,5569
Stnd. kurtosis	-0,2387	-0,4453	-0,2483	-0,4596	-0,2410	-0,4525

Tabulka 6: Konfidenční a bootstrapové intervalové odhady průměrných cen, individuálních složených indexů a souhrnných indexů

		$\bar{p}_0$		$\bar{p}_1$	
Conf. Intervals	Mean	36,809	37,114	37,444	37,744
Bootstrap Intervals	Mean	36,823	37,092	37,461	37,717
		$I_{prom.sloz}$		I_q	
Conf. Intervals	Mean	1,01610	1,01813	0,95873	0,97921
Bootstrap Intervals	Mean	1,01614	1,01811	0,95781	0,97833
		I_h			
Conf. Intervals	Mean	0,97517	0,99593		
Bootstrap Intervals	Mean	0,97616	0,99536		
		I_p^L		I_q^L	
Conf. Intervals	Mean	1,01443	1,01643	0,96014	0,98091
Bootstrap Intervals	Mean	1,01450	1,01628	0,96100	0,98021
		I_p^P		I_q^P	
Conf. Intervals	Mean	1,01453	1,01645	0,96022	0,98095
Bootstrap Intervals	Mean	1,01460	1,01633	0,96108	0,97994
		I_p^F		I_q^F	
Conf. Intervals	Mean	1,01448	1,01644	0,96018	0,98093
Bootstrap Intervals	Mean	1,01449	1,01635	0,96149	0,97949

5. **Závěr**

Z tabulky 3 je vidět, že intervalové odhady (ostatně jako vždy) oproti bodovým odhadům realizovaným číselnými charakteristikami z tabulky 2 vypovídají daleko více o středních hodnotách, směrodatných odchylkách a mediánech cen množství i indexů. Navíc vypočtené bootstrapové intervaly středních hodnot a směrodatných odchylek jsou povětšinou menší než vypočtené konfidenční intervaly týchž parametrů, takže jsou při stejné spolehlivosti 0,95 přesnější. Nezanedbatelná je také skutečnost, že bootstrapové intervaly nezávisí na rozděleních pravděpodobnosti pozorovaných cen a množství, takže není nutno verifikovat obvyklý požadavek, že jde o normální rozdělení. Navíc předpoklad normálního rozdělení zejména cen asi neobstojí, protože ceny stanovují prodejci.

Bootstrapové intervalové odhady v tabulce 6 byly získány dvoustupňovým bootstrapem. Oprávněnost výpočtu „klasických“ intervalových odhadů (konfidenčních intervalů), uvedených také v tabulce 6, závisí na předpokladu, že soubory jednotlivých indexů získané ve druhém stupni bootstrapu (tj. bootstrapu z primárního bootstrapu) pochází z normálních rozdělení pravděpodobnosti. Hypotézy o shodě těchto rozdělení s rozdělením normálním byly úspěšně testovány více metodami na hladině významnosti 0,05.

Aplikace bootstrapových odhadů v indexové analýze je dosti netradiční. Bohužel ani konfidenční intervalové odhady se v ekonomickém průzkumu moc nepoužívají nebo jejich použití nebývá seriózní a nejčastěji se počítají pouze

bodové odhady středních cen a indexů. Naopak bootstrapové intervalové odhady umožňují vyrovnat se s náhodností pozorovaných cen a množství, s problémem jejich rozdělení pravděpodobnosti, s extrémně odchýlenými hodnotami a také s nepříliš velkými rozsahy statistických souborů.

Metoda bootstrap nachází uplatnění i v jiných oblastech matematické statistiky, např. při fitování rozdělení pravděpodobnosti kategoriální veličiny [6]. O jiném přístupu k určení intervalových odhadů indexů ekonomických ukazatelů, a to z předem zadaných intervalových odhadů cen a množství, je pojednáno v [7].

Poděkování: Článek je součástí řešení projektu GAČR P403/11/2 085 *Konstrukce metod pro vícefaktorové měření komplexní podnikové výkonnosti ve vybraném odvětví*, výzkumného úkolu AKADEMIE STING *Podpora řízení firem s využitím kvantitativních metod* a grantového projektu TAČR TA02021449 *Sytém inteligentních alarmů v energetickém provozu jaderných elektráren.*

Literatura

[1] EFRON, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7(1), 1979, pp. 1 - 26.

[2] EFRON, B., TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. New York: Chapmann & Hall, 1993. ISBN 0-412-04231-2.

[3] DAVISON, A. C., HINKLEY, D. V. *Bootstrap Methods and their Applications*. Cambridge: Cambridge Univ. Press, 2006. ISBN 0-521-57471-4.

[4] SEGER, J., HINDLS, R. *Statistické metody v tržním hospodářství*. Praha: Victoria Publishing, 1995. ISBN 80-7187-058-7.

[5] KARPÍŠEK, Z., SADOVSKÝ, Z. *Matematické metody 2*. Elektronický učební text. Brno: AKADEMIE STING, 2005.

[6] KARPÍŠEK, Z., LACINOVÁ, V. Odhady diskrétního rozdělení pravděpodobnosti s použitím kvazinorem a bootstrapu. *Analýza dat 2010/II - Statistické metody pro technologii a výzkum*. Pardubice: TriloByte, CQR, 2010, pp. 131-145. ISBN 978-80-904053-3-2.

[7] KARPÍŠEK, Z., KOCMANOVÁ, A., KRÁL, D., LACINOVÁ, V. Aplikace intervalové aritmetiky v indexové analýze. *ACTA STING* 1 (1), 2012, pp. 13-24. ISSN 1805-1391 (Print), ISSN 1805-6873 (Online).

METODIKA KOMPLEXNÍHO NÁVRHU REGULAČNÍHO DIAGRAMU

METHODOLOGY FOR A COMPLEX DESIGN OF CONTROL CHART

Jan Král

Adresa: ISQ PRAHA, s.r.o., Pechlátova 19, 150 00 Praha 5; kral.jan@isq.cz

Abstrakt: Při řízení procesu je naším cílem zvládnutí variability procesu do té míry, že na něj působí pouze inherentní složky variability. K tomuto účelu se v praxi používají nejčastěji regulační diagramy. Tento přístup lze využít v oblasti výroby, služeb i veřejné a státní správy. Uvedený příspěvek prezentuje přístup využívající nově navrženou metodiku pro Komplexní návrh regulačního diagramu. To vše s ohledem na konkrétní podmínky s využitím soudobých moderních přístupů a znalostí z oblasti aplikované statistiky spolu s praktickými zkušenostmi z řady průmyslových provozů.

Abstract: Our goal in process control is mastering process variability to the extent that it influences only inherent variability component. Most frequently control charts are used for this purpose in practice. This approach can be used in manufacturing, services and public and state administration. This contribution presents an approach that uses a newly designed methodology for complex design of Control chart.

Klíčová slova: Řízení procesu, metodika, SPC, regulační diagram

Keywords: Process control, methodology, SPC, control chart

1. Úvod

Od doby vzniku prvních regulačních diagramů do současnosti bylo vytvořeno velké množství různých přístupů, modifikací a rozšíření. Řada teoretiků se zabývá použitím regulačních diagramů v nestandardních podmínkách zaměřených na konkrétní výrobní procesy a využívajících moderních statistických metod. Tyto výsledky, bezesporu přínosné, však nejsou v naší podnikové praxi využívány, nebo pouze v omezené míře.

Podmínky a předpoklady pro tyto nové metody jsou často pro praxi příliš akademické a není zcela zřejmé, jak je v praxi využít. V literatuře lze nalézt návody a postupy, jak tyto nové metody implementovat, avšak neexistuje systematická analýza, která by vyústila v ucelenou metodologii návrhu regulačního diagramu na potřebné teoretické úrovni tak, aby byla relativně snadno aplikovatelná v praktickém provozu. V literatuře jsou dosud obvykle dána dílčí řešení některých částí, jako je optimalizace parametrů regulačních diagramů, návrhy různých typů regulačních diagramů pro specifické situace, začlenění strategií údržby do regulačního procesu a mnohé další.

Prezentovaná metodika byla navržena a řešena jako reakce na konkrétní aktuální požadavky strojírenských podniků, které požadují průkazné, jednoznačné a rychlé řešení vznikajících problémů za pomoci SW podpor a jen nezbytně nutných finančních nákladů. Dále byla požadována garance dostatečné vědecko-teoretické úrovně a metodické podpory při praktickém zavádění nově navržené metodiky i pro pracovníky, kteří budou pro využívání metod SPC zaškoleni.

Nově formulovaná metodika pro komplexní návrh regulačního diagramu poskytuje cílovému uživateli návod řešící v jednotlivých krocích etapy implementace SPC. Nedílnou součástí je taktéž ověření předpokladů pro užití konkrétních metod. Tato etapa je v praxi běžně opomíjena, což vede ke zkreslení výsledku, demotivaci pracovníků reagujících na základě planého signálu o změně nastavení procesu a ekonomické ztrátě. Zmíněná ztráta vyplývá z vícepráce na procesu, který v statisticky nezvládnutém stavu produkuje vyšší podíl neshodných jednotek.

Pokud zvolíme metodu SPC s využitím prezentované metodiky, bude návrh regulačního diagramu teoreticky správný, s optimalizovanými náklady na provádění statistické regulace a s problémově zaměřenou SW podporou.

2. Popis navržené metodiky

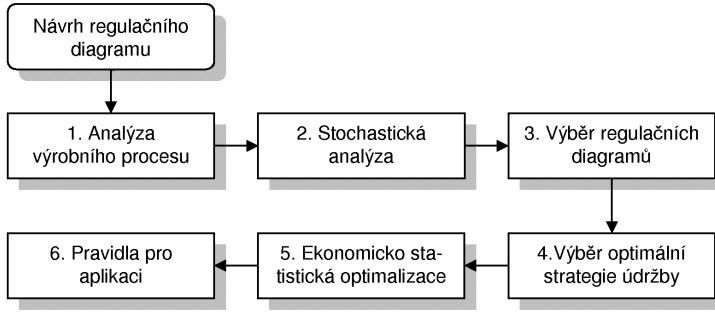
Návrh regulačního diagramu byl realizován v následujících krocích, jejichž posloupnost je vnitřní náplní prezentované metodiky. Tato metodika byla formulována na základě kritické analýzy dosavadních poznatků a názorů zjištěných při studiu literatury a dále na základě vlastních zkušeností.

- **Analýza výrobního procesu:** úkolem je vymezení cílové kvalitativní veličiny, kvalitativních znaků, jejich závislostí, identifikovat jednotlivé vazby mezi proměnnými a z toho plynoucí proměnné použitelné pro řízení. Pro tuto etapu je nutno znát podrobně výrobní proces, strojní zařízení, materiál. Tento krok vyžaduje součinnost s výrobním expertem.
- **Stochastická analýza:** parametry a proměnné ve výrobním procesu mají charakter náhodných veličin, a pokud je sledujeme v čase, dostáváme stochastické procesy. Pro jejich řízení je potřeba co nejpodrobněji poznat pravděpodobnostní charakteristiky a stochastické závislosti. Za tímto účelem je třeba provést důkladnou stochastickou analýzu, kterou by v tomto případě měl provádět statistik, nebo alespoň pracovník znalý základních statistických metod.
- **Výběr regulačních diagramů:** od 30. let minulého století vznikla celá řada různých typů regulačních diagramů aplikovatelných na různé typy proměnných za různých předpokladů chování sledovaných veličin. Proto je důležité na základě předchozí stochastické analýzy a analýzy výrobního procesu vybrat ten nejlepší typ regulačního diagramu. Přitom je třeba respektovat podmínky, pro které byl tento

regulační diagram konstruován a za kterých je schopen poskytovat správné výsledky.

- **Výběr optimální strategie údržby:** regulační diagramy mají za cíl ochránit výrobní proces před důsledky poruch, které se mohou vyskytnout náhodně v čase. Těmto poruchám se dá často předejít účinnou preventivní údržbou a pravidelnými opravami. Na druhou stranu údržba bývá často nákladná a nelze ji provádět příliš často. Použití regulačních diagramů ve spojitosti s optimální strategií preventivní údržby může výrazně zvýšit efektivitu SPC.
- **Ekonomicko statistická optimalizace:** použití regulačních diagramů závisí na řadě parametrů, jako je například doba mezi inspekcemi, rozsah výběru při inspekčním měření, stanovení reakčních mezí a dalších. Tyto parametry jsou v současné době určovány převážně expertními odhady, nebo tradičními doporučeními. Vhodná volba regulačních diagramů ještě nezaručuje jeho optimální funkci z hlediska nákladů a statistických vlastností (četnost planých poplachů, včasnost detekce poruch). Proto je třeba výběr regulačních diagramů doplnit o optimalizaci jeho parametrů pro daný účel.
- **Pravidla pro aplikaci:** všechny předchozí kroky nebudou dostatečně účinné, pokud nebudou přesně určena závazná pravidla a zajištěny vhodné podmínky pro jejich realizaci. Zde přichází opět ke slovu výrobní expert, který zná podrobně možnosti výrobního procesu, technické podmínky, vybavení a organizaci práce.

Nově navržená metodika komplexního návrhu regulačního diagramu spočívá v postupné důsledné aplikaci těchto šesti kroků při návrhu SPC, které jsou zobrazeny na následujícím obrázku 1, s důrazem na předběžnou analýzu výrobního procesu a na statisticko-ekonomickou optimalizaci zvoleného detekčního algoritmu. Zvláštní důraz je kladen na zajištění technologicko-organizačních předpokladů pro úspěšnou aplikaci ve výrobě.



Obrázek 1. Postup při návrhu RD

Budou-li tyto kroky dodrženy, bude zajištěno, že sledovaná veličina je vhodně zvolena a popisuje charakteristiku mající hlavní vliv na požadovanou

vlastnost produktu. Pro tuto veličinu je dále nezbytné navrhnout vhodný systém měření. Tím bude zajištěno, že získaná data z procesu budou dostatečně přesná (počet platných desetinných míst u kvantitativního znaku kvality) a chyba měření nebude významně ovlivňovat variabilitu sledovaného procesu.

Na základě takto získaných dat je možné stanovit teoreticky správný stochastický model, vyčíslit jeho pravděpodobnostní charakteristiky, které dále využíváme k řízení procesu a provést výběr optimálního typu regulačního diagramu.

Za optimální návrh regulačního diagramu považujeme takový, který je teoreticky správný (generuje pouze očekávané procento falešných signálů) a ekonomicky únosný (přiměřená frekvence inspekcí s dostatečnou technicky nebo ekonomicky zdůvodnitelnou velikostí vzorku-podskupiny).

Nedílnou součástí péče o proces je taktéž údržba, která má synergický efekt na stabilitu procesu. Zde je nezbytné stanovit přiměřenou míru preventivních a plánovaných opatření tak, aby byl proces ochráněn před důsledky fatálních poruch s přihlédnutím k ekonomickým hlediskům (náklady na nekvalitu, náklady na údržbu, potenciální ztráta dobrého jména).

Pro úspěšnost produkce je důležitá výše nákladů a technická úroveň, která je podmíněna konstrukcí, technologií, systémem měření, použitými materiály, prostředím, lidským činitelem atd.

V těsné návaznosti na statistické řízení a využívání poznatků z analýz výrobního procesu je proces minimalizace ztrát z nekvality. K tomuto cíli je proto nezbytné využívat současných moderních systémů údržby TQM (Total Quality Maintenance), které jsou provozně propracované a dostupné včetně SW podpory. Cílem těchto systémů je prioritně zabezpečit spolehlivost strojů a zařízení, což je deklarováno jako schopnost plnit trvale stanovené požadavky.

V následující kapitole se budu blíže věnovat bodu 3 citované metodiky, ve kterém je volba metody regulace rozpracována až na úroveň rozhodovacích stromů.

3. Metodická schémata pro návrh regulačního diagramu

Cílem těchto schémat je poskytnout výkonným pracovníkům z praxe ověřené a teoreticky podložené postupy, vedoucí ke korektní implementaci statistických nástrojů pro řízení a sledování znaků kvality výrobku či procesu, tj. aby se grafické a numerické výsledky statistické analýzy skutečně vztahovaly k reálné situaci, která panuje ve výrobním procesu.

Rozhodovací schéma zobrazené na obrázku 1 je věnováno postupu při volbě vhodného regulačního diagramu podle typu znaku kvality. Tento znak kvality může být jednorozměrný, nebo vícerozměrný, u kterého dále vyšetřujeme, zda nelze snížit jeho dimenzi a transformovat ho na jednorozměrný znak.

V další etapě provádíme autokorelační analýzu, která ověří, zdali nejsou jednotlivá měření jako časová řada korelována. V případě pozitivně zjištěné korelace se pokoušíme nalézt její model a pomocí něho data „očistit“. S takto upravenými daty (rezidui) lze již pracovat jako s nekorelovanými. V případě, že se nepodaří pro autokorelovaná data nalézt model, zahrnuje se tato složka do inherentní variability procesu a zvolíme některý z regulačních diagramů, který je robustní vůči autokorelaci dat.

Metody SPC se dále dělí podle typu znaku kvality na regulační diagramy pro spojitá či atributivní data. Dalším rozhodujícím faktorem je i velikost podskupiny, která má být u spojitých dat konstantní. U atributivních dat se může velikost podskupiny měnit.

Typ regulačního diagramu lze volit i s ohledem na požadovanou rychlost průměrné doby odezvy na změnu neboli zmenšení chyby 2. druhu. Pokud se na regulační diagram díváme jako na sekvenční test, je vhodné zvláště u procesů citlivých na změny použít modernější typy regulačních diagramů jako jsou diagramy EWMA či CUSUM.

Volba typu regulačního diagramu pro spojitý znak je rozpracována v samostatném schématu na obrázku 5. Na obrázku 6 je rozpracována volba typu regulačního diagramu pro diskrétní znak.

4. **Statistická optimalizace**

Součástí prezentované metodiky jsou také nově navržené a řešené postupy statistické optimalizace vedoucí k teoreticky správnému a účinnému návrhu rozšířených regulačních mezí v případě, že monitorovaný proces je statisticky zvládnut v „širším slova smyslu“, kdy se připouští určitá, procesu vlastní a neodstranitelná variabilita střední hodnoty a případně i směrodatné odchylky. Tato situace je demonstrována na následujícím obrázku 1.

Dle míry statistického zvládnutí procesu je navrženo užití následujícího typu regulačních mezí pro výběrové průměry:

1) Shewhartův proces – statisticky zvládnutý v „užším slova smyslu“

$$UCL = \bar{\bar{X}} + A_3(n)\bar{s}, LCL = \bar{\bar{X}} - A_3(n)\bar{s},$$

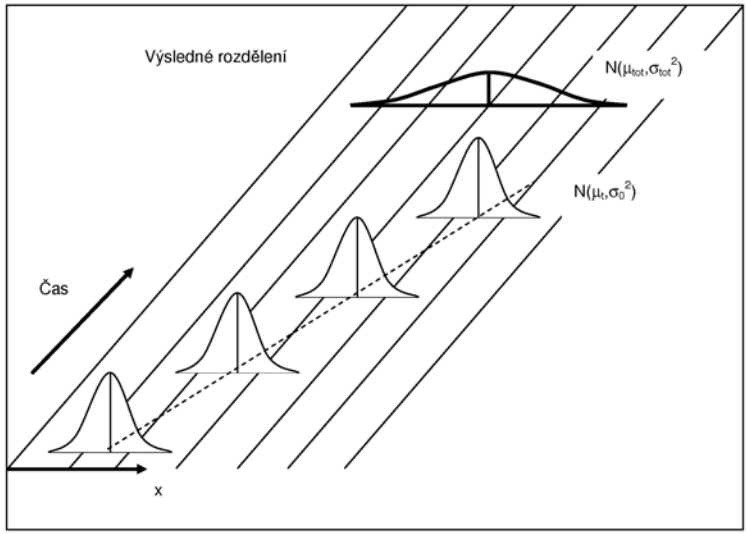
pro výpočet regulačních mezí založený na výběrové směrodatné odchylce.

$$UCL = \bar{\bar{X}} + A_2(n)\bar{R}, LCL = \bar{\bar{X}} - A_2(n)\bar{R}.$$

pro výpočet regulačních mezí založený na výběrovém rozpětí. (Pozn.: Příslušné koeficienty jsou tabelovány v normě ČSN ISO 8258) [3].

2) Proces se střední hodnotou měnící se v čase, nebo proces se střední hodnotou měnící se v čase s inherentním trendem

2.1 Rozšířené regulační meze s využitím celkové směrodatné odchylky σ_{tot}



Obrázek 2. Proces se střední hodnotou měnící se v čase s inherentním trendem

Celková směrodatná odchylka σ_{tot} se odhaduje ze všech napozorovaných hodnot v k podskupinách stejného rozsahu n na základě vztahu:

$$\sigma_{tot} = \sqrt{\frac{1}{kn-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-\bar{\bar{X}})^2}$$

Při znalosti střední hodnoty μ a σ_{tot} vypočteme rozšířené regulační meze podle předpisu:

$$UCL = \bar{\bar{X}} + A_3(n)\sigma_{tot}C_4(n),$$
$$LCL = \bar{\bar{X}} - A_3(n)\sigma_{tot}C_4(n).$$

2.2 Rozšířené regulační meze s využitím směrodatné odchylky výběrových průměrů

Směrodatnou odchylku výběrových průměrů $s_{\bar{x}}^2$ vypočteme podle vztahu:

$$\hat{\sigma}_{\bar{\bar{x}}}^2 = s_{\bar{x}}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2.$$

Takto konstruované regulační meze vypočteme podle předpisu:

$$UCL = \bar{\bar{x}} + u_{1-\alpha}s_{\bar{x}};$$
$$LCL = \bar{\bar{x}} - u_{1-\alpha}s_{\bar{x}}.$$

Za $u_{1-\alpha}$, kvantil normovaného normálního rozdělení je možné dosadit hodnotu

$u_{1-\alpha} = 3$, což odpovídá riziku planého poplachu $\alpha = 0,00135$ pro kterou je konstruován původní Shewhartův regulační diagram.

2.3 Rozšířené regulační meze s využitím rozšiřující konstanty $\triangle$

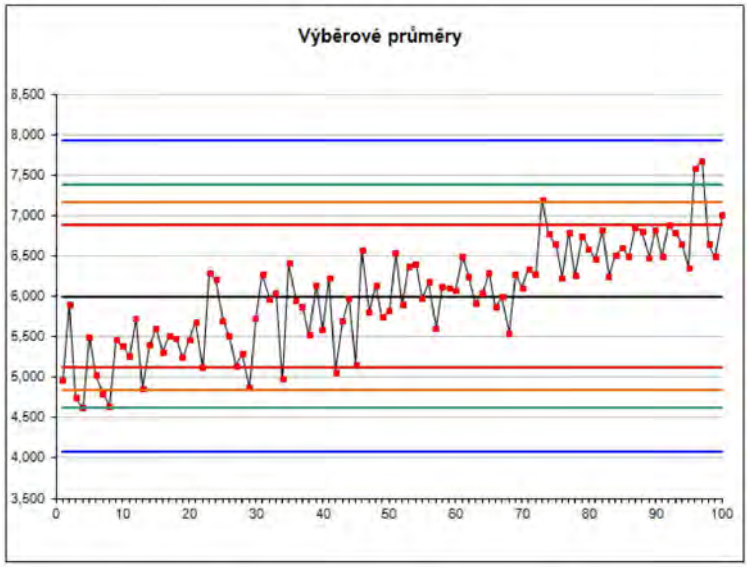
V případě, že aritmetické průměry podskupin $\bar{X}_i$ nelze popsat jako realizace náhodné veličiny, např. jsou svázány lineárním trendem, či sezónními systematickými vlivy, či vlivy plynoucími ze změny nástrojů, je možné použít postup, který takovéto neodstranitelné vlivy bude respektovat. Původní Shewhartovy meze rozšíříme o vhodně zvolenou konstantu $\triangle > 0$.

$$UCL = \bar{\bar{X}} + A_3(n)\bar{s} + \triangle,$$

$$LCL = \bar{\bar{X}} - A_3(n)\bar{s} - \triangle.$$

Rozšíření ve formě pásu šířky $2\triangle$ je voleno tak, aby tento pás popsal a zohlednil možné chování nastavení jednotlivých logických podskupin v maximální míře. Volba parametru $\triangle$ silně závisí na povaze a znalostech výrobního procesu.

Na obrázku 5 je zobrazen regulační diagram pro výběrové průměry se zakreslenými regulačními mezemi. Od středové čáry to jsou postupně meze Shewhartovy; rozšířené s_{TOT}; rozšířené $\triangle$; rozšířené $s_{\bar{x}}$.



Obrázek 3. Regulační diagram pro výběrové průměry

Výpočet a zakreslení těchto rozšířených regulačních mezí není dosud standardně implementován ani v renomovaných statistických SW, např. v ČR rozšířeném programu Minitab.

4.1. Navrhovaný postup pro případ zamítnutí hypotézy o normalitě znaku kvality

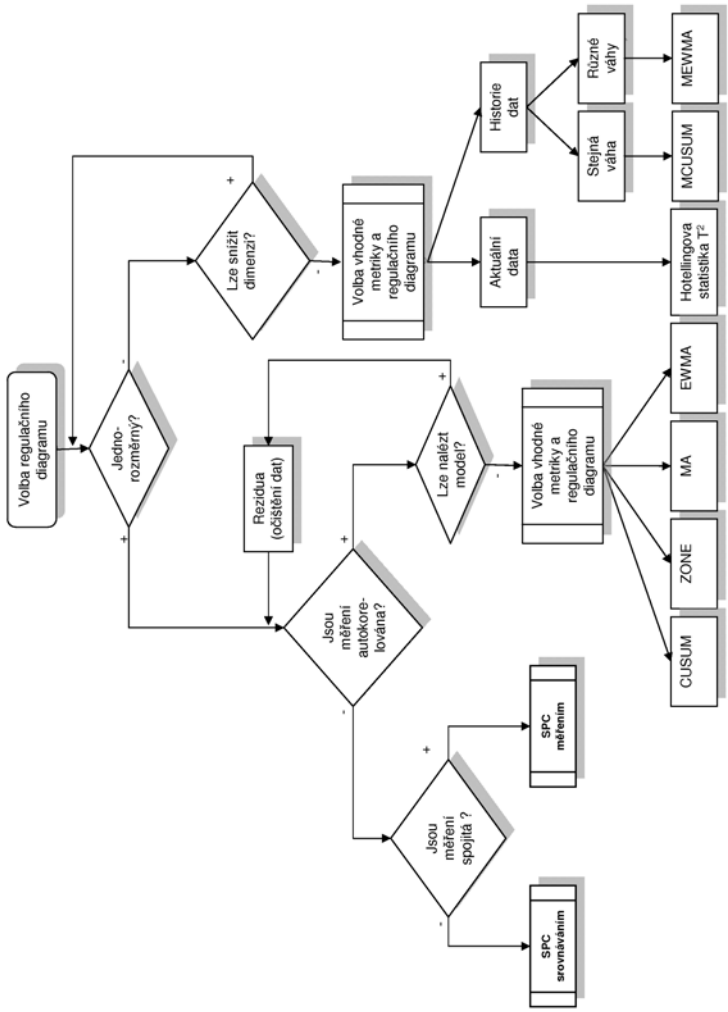
V případě, že dojde k zamítnutí hypotézy o normalitě znaku kvality, avšak rozdělení znaku kvality je blízké normálnímu rozdělení, je v praxi běžným přístupem použít regulační diagramy pro výběrové průměry sledovaného znaku kvality s odkazem na Centrální limitní větu. Tento přístup má však své omezení spočívající ve velikosti realizovatelné logické podskupiny nezbytné pro dosažení normality výběrového souboru. Pro některé procesy vychází velikost korektní logické podskupiny větší než 30 kusů!

V praxi se běžně využívá k regulaci výběrových průměrů ze tří až pěti hodnot. Takovýto přístup však může vést k signifikantnímu nárůstu počtu falešných poplachů, a proto jej nelze doporučit.

Autorem byl navržen postup, kterým lze zachovat přiměřeně velkou, ekonomickou (realizovatelnou) logickou podskupinu s využitím následujících teoreticky správných přístupů (tato problematika je podrobněji analyzována na řešených příkladech v autorem publikované literatuře [4]:

- identifikovat typ rozdělení a pracovat s kvantily identifikovaného rozdělení;
- provést Box Coxovu či Johnsonovu transformaci nenormálně rozdělených dat na normální a vyhodnotit požadované kvantily;
- pomocí zpětné transformace (například: Excel, nástroj „Hledání řešení“) stanovit odpovídající kvantily v původně rozdělených datech a využít jich pro stanovení regulačních mezí s riziky 0,00135 a 0,99865, odpovídajících rizikům planého poplachu, se kterými pracují Shewhartovy regulační diagramy;
- V krajním případě, když není možné aplikovat výše uvedené postupy, je přípustné odhadnout příslušné percentily z většího množství napozorovaných dat, případně s využitím metody „Bootstrap“.

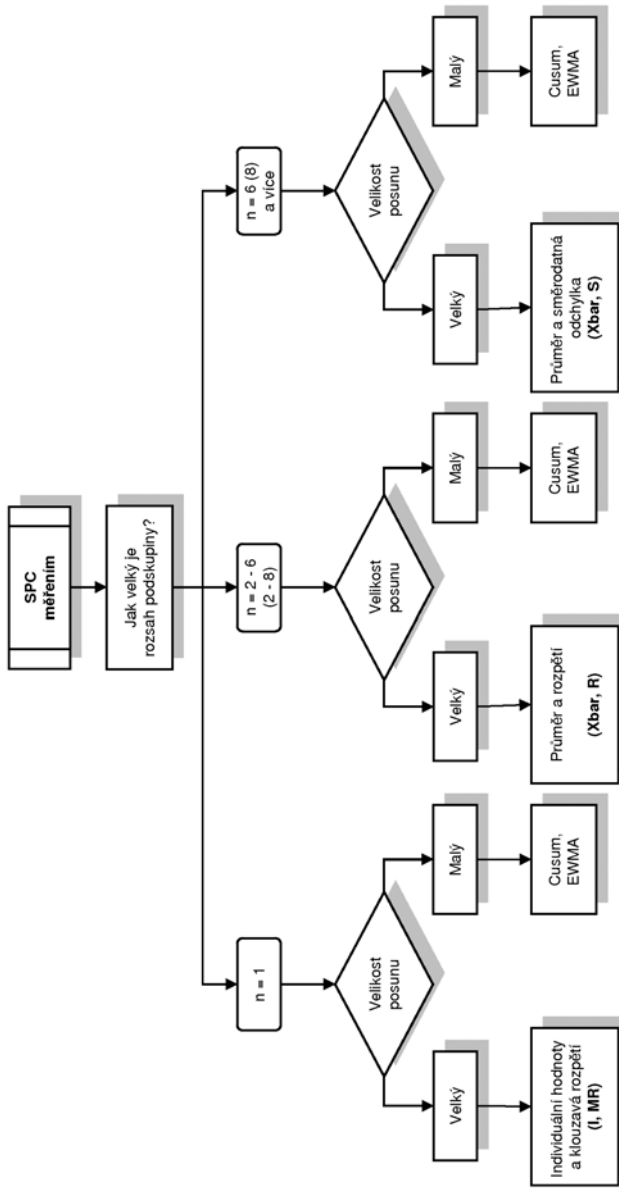
Poznámka: Uvedené metody jsou seřazeny sestupně dle vhodnosti implementace.



Obrázek 4. Schéma pro volbu regulačního diagramu

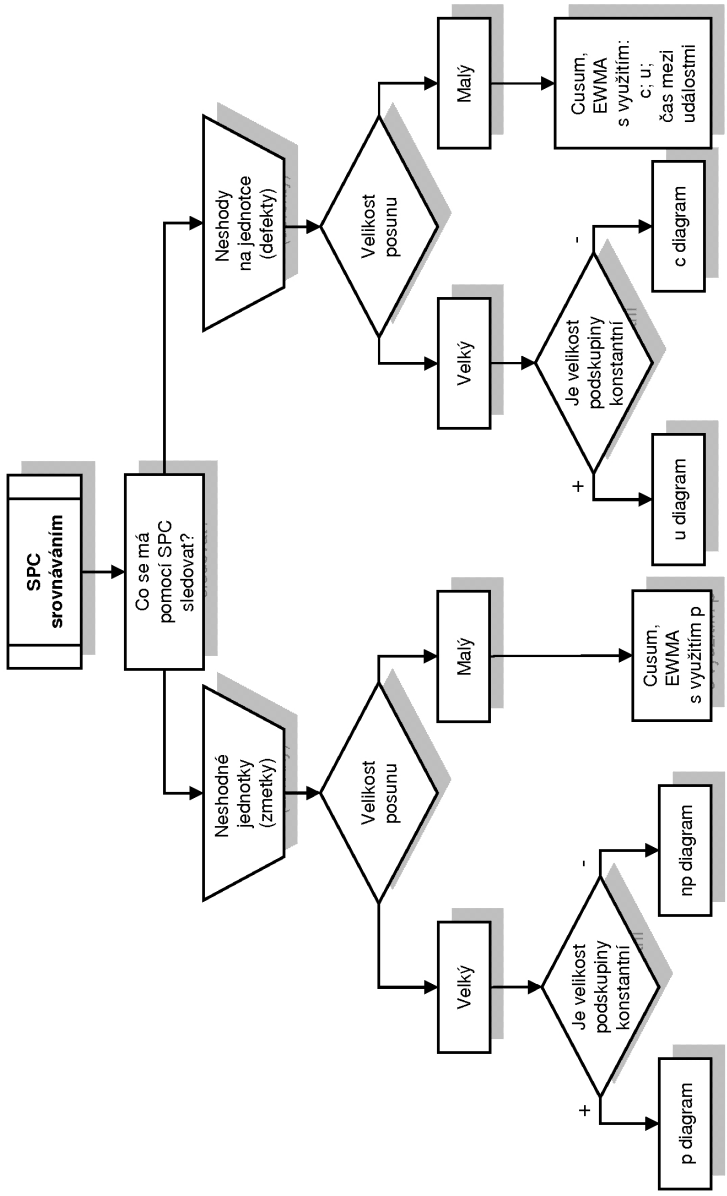
5. **Závěr**

Na řízení výrobního procesu jsou v současné době kladeny kriteriální požadavky směřující k minimalizaci nákladů a maximalizaci realizovaného výnosu.



Obrázek 5. Volba SPC pro spojitá neautokorelovaná data

K naplnění těchto požadavků je nezbytné tyto procesy trvale monitorovat a řídit. Za optimální situaci se dá považovat stav, kdy na sledovaný proces



Obrázek 6. Volba SPC pro diskrétní neautokorelovaná data

působí pouze náhodné složky variability, proces je centrován a není přerušován operativními korekčními zásahy technické obsluhy procesu. Pro zajištění tohoto stavu je velice důležité monitorovaný proces analyzovat, identifikovat a definovat dílčí komponenty variability a v maximální možné míře eliminovat vymezitelné příčiny. K tomu je nezbytné mít k dispozici stochastický model procesu a na jeho základě je možné rozhodnout o volbě vhodného prostředku k řízení procesu - regulačního diagramu. Volba regulačního diagramu má zásadní vliv na naši schopnost proces řídit. V tomto kontextu hovoříme o stochastickém řízení procesu ve smyslu realizování řídících a korekčních aktů na základě objektivních informací o chodu procesu. Tyto informace nám poskytuje právě regulační diagram tím, že vyšle signál v okamžiku, kdy na proces začne působit jiná než náhodná složka variability. Touto nežádoucí variabilitou může být například zhoršení průměrné výstupní kvality od dodavatele, opotřebení nástroje, či náhlá změna technického stavu výrobního zařízení.

Při nesprávné volbě typu regulačního diagramu zde však hrozí nebezpečí, že bude docházet k vyslání takzvaného „falešného signálu“ za situace, kdy v procesu nedošlo k reálné změně. I na tento falešný signál musí obsluha reagovat. Vzhledem k tomu, že vyslání signálu není založeno na skutečné změně v nastavení procesu, obsluha tuto příčinu nemůže identifikovat a adekvátně na ni reagovat. Toto se negativně projeví na:

- důvěře v metodu regulace a měření,
- nezájmu o implementace SPC,
- zvýšených nákladech na vlastní analýzu falešného signálu,
- celkových ztrátách kapacit i financí v důsledku nekvality v celém reprodukčním procesu,
- zvýšené, nežádoucí variabilitě procesu vyvolané neopodstatněným zasahováním do procesu.

Metodika spočívá v důsledné aplikaci šesti kroků popsaných v předchozí stati. Takováto komplexní metodologie v současné době není v našich výrobních podnicích k dispozici, což často vede k opomíjení důležitých zásad při aplikaci regulačních diagramů a v důsledku toho zapříčiňuje nedosažení potřebné efektivity, spolu s nedůvěrou v účinnost těchto metod.

Literatura

[1] DOHNAL G.: *Design of Control Charts.* ROBUST 2008. ČStS, JČMF a SŠDS 2008, ISBN 80-7015-004-07

[2] KRÁL J.: *Metodika komplexního návrhu regulačního diagramu.* Doktorská disertační práce, ČVUT v Praze, 2012

[3] ČSN ISO 8258: *Shewhartovy regulační diagramy*, 1994.

[4] KRÁL, J., MICHÁLEK, J., KŘEPELA, J.: *Shewarts Control Charts of Sample Means for Nonnormal Distribution of Quality Variables (Shewhartovy RD výběrových průměrů v případě nenormálního rozdělení znaku jakosti)* In: 5th Annual International Travelling Conference for Young Researchers and PhD. Students ERIN 2011: Proceedings: 13th - 16th April 2011. Prešov: Harmony Apeiron Non-profit Association, 2011. s. 285-294. ISBN: 978-80-89347-04-9.

VÝZNAM A MOŽNOSTI VYUŽITÍ MRM VE VÝROBĚ, SLUŽBÁCH A VEŘEJNÉ A STÁTNÍ SPRÁVĚ

IMPORTANCE AND POSSIBILITIES OF MRM METHOD UTILIZATION IN THE SPHERE OF PRODUCTION, SERVICES, PUBLIC AND STATE ADMINISTRATION

Otakar Král, Gejza Dohnal

Adresa: ISQ PRAHA, s.r.o., Pechlátova 19, 150 00 Praha 5; kral.otakar@isq.cz
FS ČVUT, ÚTM, Karlovo náměstí 13, 121 35-Praha 2; dohnal@nipax.cz

Abstrakt: V přednášené zprávě je uveden vývoj, současné využití, další postup prací a oblasti možného využívání Metody relačních matic - MRM. Původní záměr v CQR deklarovat MRM jako nástroj pro kvantifikaci parametrů procesů v oblasti výroby a služeb byl překonán využitím pro veřejnou a státní správu, jmenovitě v ČSÚ při SLDB 2011. Význam MRM se dále rozšiřuje jako průkaz a dokumentace dosažených výstupů na straně dodavatelů i objednatelů. Nejnověji se MRM implementuje do oblasti inovací, kde významnou měrou svým novým způsobem kvantifikace parametrů procesů identifikuje a prokazuje dosažený stupeň inovace a predikuje, respektive prokazuje objektivně dosažené technické i ekonomické výsledky.

Abstract: The contribution deals with development, contemporary utilization, future course o action and spheres of possible utilization of MRM method. The original intention to declare MRM method in CQR as instrument of process parameters quantification in the sphere of production and services was exceeded by its utilization in the sphere of public and state administration, particularly in Czech Statistical Office in Population and Housing Census 2011. Importance of MRM method is increasing as a proof and documentation of achieved outputs of suppliers and customers. MRM method is newly implemented in the sphere of innovations, where significantly by new way of process parameters quantification identifies and proves achieved level of innovation and predicts, or more precisely proves, really achieved technical and economical results.

Klíčová slova: kvantifikace, procesní management, projektový management

Keywords: quantification, process management, project management

1. Úvod

Metoda relačních matic (MRM) je nově vyvinutým a v praxi ověřeným nástrojem pro hodnocení a **optimalizaci rozhodování vedoucích pracovníků** v oblasti výroby, služeb, veřejné a státní správy.

Tato metoda je rozsáhlou modifikací a inovací jednoho z realizačních výstupů CQR, projektu MŠMT 1M06047, řešeného v letech 2006-2011.

V období 08/2010 až 03/2012 byla implementována metoda MRM na ČSÚ, v rozsáhlém projektu SLDB 2011, s výsledkem viz vyjádření ČSÚ na obrázku. č.1.

V současné situaci roste důraz na **zodpovědnost za důsledky a výsledky řešení.** V období recese je zcela oprávněný nárůst požadavku na průkaz účelného vynaložení kapacit, finančních prostředků, dodržení termínů a stanovených technických a ekonomických parametrů zadání. V této situaci je nezbytné mít k dispozici **metodiku hodnocení požadovaných parametrů pro oblast výroby, služeb, veřejné a státní správy** pro komplexní hodnocení objektů zájmu (t.j. podniků, institucí, projektů, procesů atd.).

MRM je univerzálním interdisciplinárním nástrojem manažerů, který vychází z již v praxi ověřených a úspěšně uplatněných metod, soudobých trendů a forem řízení v celé struktuře podnikání a veřejné a státní správě. Nově jsou formulovány a řešeny požadavky na **řízení celého reprodukčního cyklu procesů organizací a institucí (vývoj, realizace, využívání, zlepšování až po ekologickou likvidaci)** procesů dle jejich požadovaných parametrů.

Metodika MRM zahrnuje a **důsledně vyžaduje kvantifikaci všech podstatných vlivů** na chod objektu zájmu (OZ) a umožňuje získat kvantifikované a objektivizované podklady pro finální rozhodování, či dokumentování dosažených /nedosažených parametrů. Metodika MRM je navržena a ověřena na základě pěti zásad, které jsou podstatné pro zabezpečení akceptovatelných realizačních výstupů a doporučení finálních závěrů, rozhodnutí. Nová a inovovaná pojetí v MRM představují následující zásady:

(1) **Kategorizace vlivů** na OZ MRM do vlivů **interních a externích**, jejich nositelů a důsledků působností.
(2) Důsledná kvantifikace metodami exaktními u průkazných (tvrdých dat) a expertními u dat měkkých, již ověřenými i nově stanovenými metodami, **která umožní porovnání působností shodných i analogických objektů** mezi sebou i objektů zcela samostatných. Kvantifikace působností umožňuje rovněž **vyhodnocení nositelů** z hledisek dodržování stanovených **kvalitativních ukazatelů**.
(3) Implementace **matematicko statistických metod** pro stabilitu sledovaných probíhajících procesů, vč. specifických nových procedur.
(4) Maximalizace **nasazení SW podpor**, komerčních i autorsky zpracovaných.

(5) **Důsledná, průkazná dokumentace** pro realizovaná rozhodnutí vedoucích pracovníků. Toto garantuje kvalitu řízení i účelné vynaložení kapacit a finančních prostředků na základě optimalizovaného, dokumentovaného, kvalifikovaně regulovaného realizačního výstupu vycházejícího z provedení kvantifikace zadaných, respektive v průběhu řešení vzniklých významných procesů a činností OZ MRM zjištěných požadavků.

Tím je vytvořena **objektivní multifunkční baze** pro rozhodování ve všech typech manažerského řízení. Je nutno zdůraznit, že především oblast ekonomická, uplatnění vyšších MSM a informační je oproti jiným metodám významně zdůrazněna a objektivizována.

Výše uvedený postup při zabezpečování nových projektů i významných racionalizačních akcí z oblasti výroby, služeb a veřejné i státní správy garantuje, že při řešení za podmínek dodržení stanoveného postupu dle MRM **bude řešení dokumentovat a splňovat veškeré podstatné požadavky na zadání a jeho optimalizaci** v celém rozsahu reprodukčního cyklu procesů řešeného objektu, s minimalizací možností opomenutí významnějších náležitostí v řešení. Metoda MRM byla již úspěšně ověřena v velkém strojírenském podniku a realizována v oblasti veřejných a státních služeb v projektu Sčítání lidu, domů a bytů 2011.

Pro implementaci metody MRM je využíván následující základní metodický postup.

(1) Identifikace faktorů a požadavků: V této fázi se provede identifikace:
- internalit (vnitřní pracovníky ovlivnitelné působnosti na prvky OMRM),
- externalit (vnější pracovníky neovlivnitelné působnostina prvky OMRM),
- požadavků na OMRM.

(2) Hodnocení úrovně vztahů mezi internalitami a externalitami uvnitř skupin mezi sebou navzájem: V této fázi jsou na základě objektivních informací, resp. na základě expertního hodnocení stanoveny úrovně vztahů mezi jednotlivými internalitami (resp. externalitami). Tato analýza umožňuje identifikovat závislé internality, resp. externality a snížit tak počet internalit, resp. externalit (seskupením resp. sloučením). V této fázi se pro přehledné znázornění používá afinitní diagram.

(3) Hodnocení úrovně vztahů mezi internalitami a externalitami: V této fázi jsou na základě objektivních informací resp. na základě expertního hodnocení stanoveny úrovně vztahů mezi faktory - tj. internalitami, externalitami. Smyslem této analýzy je soustředit se na zlepšování internalit, které mají silné vazby na externality (a jsou proto důležité) a identifikovat externality, které mají silné vazby na interní faktory (a je proto potřeba je sledovat a řešit).

(4) Hodnocení úrovně faktorů a intenzity požadavků: V této fázi jsou na základě objektivních zdrojů informací resp. na základě v MRM stanoveného expertního hodnocení kvantifikovány úrovně jednotlivých faktorů, tj. internalit, externalit i úrovně požadavků na OMRM jako celek.
(5) Hodnocení úrovně vztahů mezi faktory a požadavky: V této fázi jsou na základě objektivních informací resp. na základě expertního hodnocení stanoveny ohodnocené, kvantifikované úrovně vztahů mezi jednotlivými faktory (tj. internalitami, externalitami) a požadavky na OMRM.
(6) Sestavení hlavní relační matice: V této fázi je již možné sestavit hlavní relační matici a provést v MRM stanovené příslušné výpočty.
(7) Vyhodnocení hlavní relační matice: Cílem vyhodnocení je:
- identifikovat prioritně internality, které mají silné vazby na požadavky na OMRM (a je potřeba je proto analyzovat, hodnotit a zlepšovat).
- identifikovat prioritně internality, které mají špatnou úroveň a mohou se tedy stát pro projekt ohrožujícími a optimalizovat je.
- identifikovat, formulovat a utřídit požadavky na OMRM, které mají silné vazby na faktory (a jsou proto důležité).
- identifikovat požadavky na OMRM, které budou obtížně splnitelné z důvodu jejich vysoké intenzity a působení alternativně je substituovat z hledisek jejich naplnění.
- oponentní řízení se zadavatelem a externími odborníky o dosaženém stupni řešení MRM, případně korekce a alternativní řešení v oblasti věcné, termínové, finanční, kapacitní.

(8) Hodnocení prvků OMRM podle klíčových faktorů na základě stanoveného rozsahu, obsahu, formy, termínů a postupu řešení a doporučení oponentury.
(9) Závěry: Na základě vyhodnocení metodou relačních matic jsou vyvozeny závěry, které jsou projednány a odsouhlaseny na řešitelském týmu a jsou formulována doporučení pro zadavatele.

Uvedeným postupem lze řešit zadané problematiky pro MRM v etapách návrhu, vývoje, realizace, využívání a zlepšování. Metodika MRM, již ověřená v praxi, umožnila svými výstupy kvantifikovaně hodnotit dosažení stanovených cílů a podílů zúčastněných na přínosech a ztrátách.

V neposlední řadě MRM svou dokumentací a kvantifikací činností a procesů dokládá včasnost, úplnost, stanovené parametry pro obsahy, rozsahy a formy výstupů vč. účelného využití finančních prostředků.

2. **Projekt SLDB 2011**

V tomto příspěvku je uveden z hlediska ČR i EU významný projekt **Sčítání lidu, domů a bytů 2011 (SLDB 2011)**, řešený v letech 2004-2014. Z hlediska náročnosti kapacitní, odborné i nákladové bylo období 2010 až 2011,

kdy sčítání bylo realizováno ve své podstatné věcné i nákladové fázi, je za toto období provedeno i zásadní vyhodnocení kvality celého projektu SLDB 2011. Představu o náročnosti podává objem plánovaných finančních prostředků, celkem cca 2,5 miliardy Kč (vč. přípravy a následného zpracování) za účasti zahraničních i tuzemských dodavatelů.

Bylo nezbytné vypracovat a zabezpečit realizaci pro zcela nové dokumenty v SLDB 2011 – **Projekt systému kvality** v souladu s ČSN EN ISO 9001, **Projekt auditů**, školení interních auditorů, spolupracovat na **Pracovních postupech** a **Realizačním projektu**. Velmi významné bylo rovněž ustavení Řešitelského týmu řízení kvality (ŘTK), kde kromě specialistů na kvalitu působili i externí odborníci dodavatelů a pracovníci ČSÚ, kde byl navržen a vyřešen **Projekt indikátorů kvality SLDB 2011** dle požadavků EUROSTATu.

Zde již nastal uvedený zásadní problém, tj. jak, čím, kdo a kdy stanoví měřítka, standardy neboli metriky pro posuzování úrovně splnění v jednotlivých krocích návrhu celého projektu SLDB 2011, jeho realizace, využívání a zlepšování. Tento problém se promítá do všech výše uvedených navržených a přijatých dokumentací systému managementu.

Tam, kde lze získávat ověřená, tvrdá data, není problém stanovovat ukazatele (metriky) v jednotkách finančních, fyzických, časových atd. a provádět jejich ověřování. Problém nastává u parametrů procesů, kde je nezbytné provádět jejich kvantifikace expertně a následně veškerá hodnocení kompletovat do objektivních, technicky, ekonomicky i časově akceptovatelných hodnot, doporučení a závěrů.

3. Kvantitativní vyhodnocování zjištění z auditů v rámci SLDB 2011

Kvantitativní vyhodnocení zjištění z ověřování a následných vyhodnocování v rámci Projektu SLDB 2011 vede k výpočtu ukazatele hodnocení ověřování na základě provedených auditů (HA) pro jednotlivá SM.

V tomto příspěvku se soustředíme na hodnocení sběrných míst (SM), center pro projekt sčítání. V rámci Projektu SLDB byly hodnocena i jiná pracoviště, například call-centrum, pracoviště příjmu a třídění formulářů, režimové pracoviště IT, krajská pracoviště ČSÚ atd.

Podmínkou kvantitativního vyhodnocení bylo to, že auditoři ve svých zprávách z ověřování funkčnosti uvedli a dokumentovali základní typy neshod, jak v oblasti internalit tak i externalit, vzhledem ke stanoveným pracovním postupům a nově vznikajícím problémům.

Zkoumání vazeb mezi jednotlivými internalitami a externalitami a definovanými požadavky na projekt pomocí relačních matic vedlo k určení 7 internalit a 7 externalit - klíčových faktorů, které ovlivňují požadované výstupy cca z 80%. Ostatní faktory nebyly v dalším hodnocení uvažovány.

Klíčové internality:

(1) Nepřítomnost plánovaných pracovníků na auditovaném místě.
(2) Nedostatky ve znalosti v rozsahu PP, RP, Plánu kvality.
(3) Nedostatky v záznamech o provedeném školení.
(4) Nedostupnost dokumentace pro SLDB 2011 na pracovišti.
(5) Neshody sčítacích komisařů (SK) v rámci distribuce a sběru sčítacích formulářů (SF).
(6) Nevyužití asistentů v problémových lokalitách.
(7) Nedostatky v ostatních záznamech SM.

Klíčové externality:

(1) Nedostatečné kapacity personálu SM (VSM, PSM, SK).
(2) Nedostatky ve využití, způsobilosti a disponibilitě informačního systému (IS).
(3) Nedostatky v přípravě a distribuci podkladů k SLDB 2011.
(4) Nedostatky v organizaci, provedení a celkovém zabezpečení školení pro SLDB 2011.
(5) Nedostatky v činnosti CC, v komunikaci CC a SM navzájem a vzhledem k PO.
(6) Nedostatky v kapacitě CC vzhledem k interakci s PSM, VSM a SK.
(7) Nedostatky infrastruktury.

Expertní tým pro hodnocení pracovišť definoval čtyři stupně:

(1) Shoda s požadavky PP, RP, PK, resp. dalšími požadavky SLBD 2011, hodnocení 1.
(2) Neshoda jen komplikující realizaci plnění požadavků PP, RP, PK, neohrožující SLDB 2011, hodnocení 2.
(3) Neshoda způsobující již snížení schopnosti plnění PP, RP, PK, neohrožující způsobilost pro úkoly SLDB 2011, hodnocení 3.
(4) Neshoda přímo ohrožující nebo může způsobit ohrožení plnění cílů SLDB 2011, hodnocení 4.

Výsledné hodnocení HA zohledňuje zjištěnou úroveň internalit i externalit. Je vypočítáno jako vážený průměr dílčích ukazatelů hodnocení internalit a externalit (HA_i a HA_e). Váhy pro tento výpočet stanovil tým expertů a v tomto případě bylo doporučeno zvolit váhu externalit ku internalitám v poměru 2:1.

V tomto textu není cílem uvedení podrobné struktury, obsahu a formy dílčích ukazatelů (indikátorů) kvality SLDB 2011. Pro tento účel byl zpracován samostatný projekt Indikátory kvality SLDB 2011, s využitím matematicko-statistických metod dle stanovených požadavků EUROSTATu, pro rok 2011. Dílčí hodnocení a vypočtené ukazatele externalit (HA_e), internalit (HA_i) a celkový ukazatel (HA) byly zaznamenávány do tabulky (ukázka viz obrázek č. 2) hodnocených 154 pracovišť (z celkového počtu cca 700 SM). Ukazatel HA_e byl stanovován takto:

$$HA_e = \left(1 - \frac{P_{e1} \cdot 0 + P_{e2} \cdot 1 + P_{e3} \cdot 2 + P_{e4} \cdot 3}{P_{ec} \cdot 3}\right) \cdot 100[\%]$$

kde:
P_{ec} ... je celkový počet externalit, hodnocených v rámci ověřování,
P_{e1} ... počet externalit, hodnocených v rámci daného ověřování stupněm 1,
P_{e2} ... počet externalit, hodnocených v rámci daného ověřování stupněm 2,
P_{e3} ... počet externalit, hodnocených v rámci daného ověřování stupněm 3,
P_{e4} ... počet externalit, hodnocených v rámci daného ověřování stupněm 4.

Ukazatel HA_i byl stanovován obdobným způsobem ze zjištěných internalit.

Celkový ukazatel hodnocení ověřování HA byl počítán podle následujícího vztahu:

$$HA = w_e \cdot HA_e + w_i \cdot HA_i[\%]$$

Kde:

w_e ... váha neshod typu externality, $w_e = 0,666$,
w_i ... váha neshod typu internality, $w_i = 0,333$,
HA_e ... dílčí ukazatel hodnocení externalit ověřovaného pracoviště,
HA_i ... dílčí ukazatel hodnocení internalit ověřovaného pracoviště.

Výsledný ukazatel HA, který vyjadřuje stupeň plnění požadavků hodnoceným pracovištěm, resp. míru jeho způsobilosti k plnění těchto požadavků byl následně vyhodnocován podle této stupnice:

$(95 - 100\rangle\%$ Audit prokázal způsobilost ověřovaného místa.
$(80 - 95\rangle\%$ Audit prokázal sníženou míru způsobilosti ověřovaného místa.
$(50 - 80\rangle\%$ Audit prokázal výrazněji sníženou míru způsobilosti ověřovaného místa neohrožující plnění úkolů SLDB 2011.
$\langle 0 - 50\rangle\%$ Audit prokázal nedostatečnou míru způsobilosti ověřovaného místa, která způsobuje nebo může způsobit ohrožení SLDB 2011.

4. **Závěr**

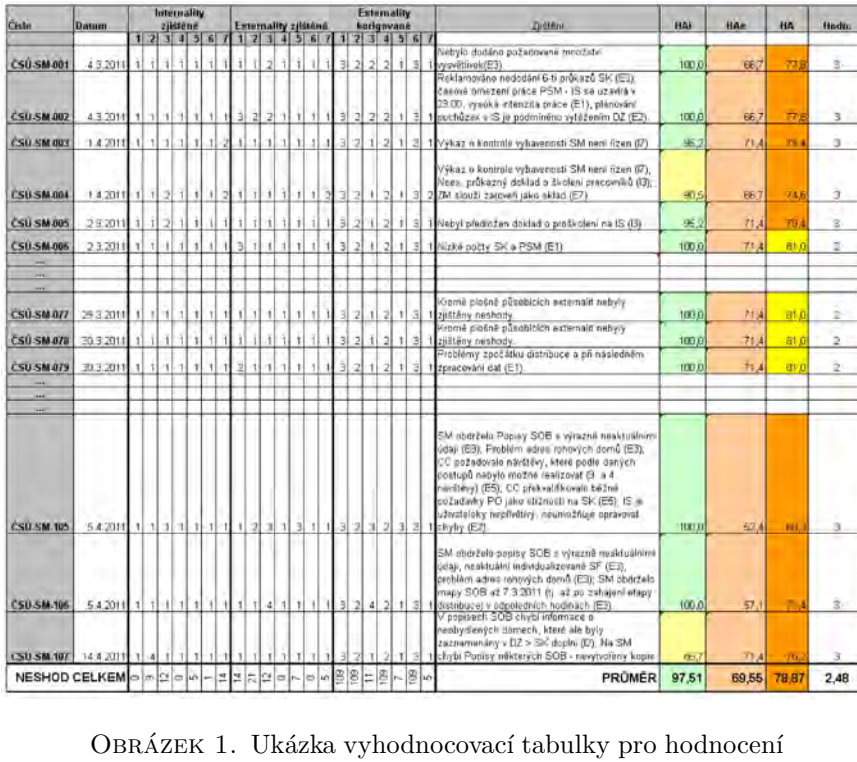
Na námitku, že takto pojatý úkol představoval nadměrné množství práce a nepřiměřené finanční nároky, lze uvést, že problematiku kvality řešil a zabezpečil tým ŘTK o cca 10 pracovnících, s využitím soudobých poznatků praxe a především statistických metod. Nákladově projekt kvality realizovaný ŘTK SLDB 2011 představuje částku cca desetiny procenta celkových nákladů v Kč.

Tímto zjištěním je jednoznačné, že uvedený realizovaný model pro ověřování kvality, kvantifikaci procesů a vytváření podmínek pro inovaci je nákladově únosný, metoda MRM je účinným a funkčním nástrojem.

Je nutno rovněž zdůraznit, že MRM při řešení a garantování projektů větších složitostí, objemů finančních prostředků je velmi významná pro zadavatele i dodavatele, neboť je nezbytné dokumentovat ověřené dosažené výsledky, metody a formy řízení a doložit důslednou a účinnou regulaci procesů systému managementu.

Seznam použitých zkratek:

CC	Call centrum SLDB	PP	Pracovní postupy
ČSÚ	Český statistický úřad	PSM	Pracovník SM
DP	Dislokované pracoviště ČSÚ	RP	Realizační projekt
DOD1	Zabezpečování dokladů sčítání	ŘTK	Řešitelský tým kvality
DOD2	Dodavatel IT služeb	SF	Sčítací formulář
DZ	Dodávací záznamy	SK	Sčítací komisař DOD1
IS	Informační systém dokladů	SLDB	Sčítání lidu, domů, bytů
KOS	Krajské oddělení sčítání ČSÚ	SM	Sběrné místo DOD1
MRM	metoda relačních matic	SOB	Sčítací obvod
PO	Povinná osoba	VSM	Vedoucí SM



Číslo	Datum	Internality zjištěné	Externality zjištěné	Externality korigované	Zjištění	HAi	HAe	HA	Hodin
		1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7					
ČSÚ-SM-001	4.3.2011	1 1 1 1 1 1 1	1 1 2 1 1 1 1	3 2 2 2 1 3 1	Nebylo dodáno požadované množství vysvětlivek(E3)	100,0	66,7	77,8	3
ČSÚ-SM-002	4.3.2011	1 1 1 1 1 1 1	3 2 2 1 1 1 1	3 2 2 2 1 3 1	Reklamováno nedodání 6-ti průkazů SK (E2); časové omezení práce PSM - IS se uzavírá v 23:00, vysoká intenzita práce (E1), plánování počtu/úsek v IS je podmíněno vytěžením DZ (E2)	100,0	66,7	77,8	3
ČSÚ-SM-003	1.4.2011	1 1 1 1 1 1 2	1 1 1 1 1 1 1	3 2 1 2 1 3 1	Výkaz o kontrole vybavenosti SM není řízen (I7)	96,2	71,4	79,4	3
ČSÚ-SM-004	1.4.2011	1 1 2 1 1 1 2	1 1 1 1 1 1 3	3 2 1 2 1 3 2	Výkaz o kontrole vybavenosti SM není řízen (I7); Neex. průkazný doklad o školení pracovníků (I3); ZM slouží zároveň jako sklad (E7)	90,5	66,7	74,6	3
ČSÚ-SM-005	2.3.2011	1 1 2 1 1 1 1	1 1 1 1 1 1 1	3 2 1 2 1 3 1	Nebyl předložen doklad o proškolení na IS (I3)	96,2	71,4	79,4	3
ČSÚ-SM-006	2.3.2011	1 1 1 1 1 1 1	3 1 1 1 1 1 1	3 2 1 2 1 3 1	Nízké počty SK a PSM (E1)	100,0	71,4	81,0	2
...									
...									
...									
ČSÚ-SM-077	29.3.2011	1 1 1 1 1 1 1	1 1 1 1 1 1 1	3 2 1 2 1 3 1	Kromě plošně působících externalit nebyly zjištěny neshody.	100,0	71,4	81,0	2
ČSÚ-SM-078	30.3.2011	1 1 1 1 1 1 1	1 1 1 1 1 1 1	3 2 1 2 1 3 1	Kromě plošně působících externalit nebyly zjištěny neshody.	100,0	71,4	81,0	2
ČSÚ-SM-079	30.3.2011	1 1 1 1 1 1 1	3 1 1 1 1 1 1	3 2 1 2 1 3 1	Problémy započátku distribuce a při následném zpracování call (E1)	100,0	71,4	81,0	2
...									
...									
...									
ČSÚ-SM-105	5.4.2011	1 1 1 1 1 1 1	1 2 3 1 3 1 1	3 2 3 2 3 3 1	SM obdržela Popisy SOB s výrazně neaktuálními údaji (E3). Problém adres rohových domů (E3), CC požadovalo návštěvy, které podle daných postupů nebylo možné realizovat (3. a 4. neúčinný) (E5), CC překvalifikovalo běžné požadavky PO jako důležité na SK (E5). IS je uživatelsky nepřátelský, neumožňuje opravovat chyby (E7)	100,0	57,1	66,7	3
ČSÚ-SM-106	5.4.2011	1 1 1 1 1 1 1	1 1 4 1 1 1 1	3 2 4 2 1 3 1	SM obdržela popisy SOB s výrazně neaktuálními údaji, neaktuální individualizované SF (E3), problém adres rohových domů (E3); SM obdržela mapy SOB až 7.3.2011 (I), až po zahájení etapy distribuce v odpoledních hodinách (E3)	100,0	57,1	71,4	3
ČSÚ-SM-107	14.4.2011	1 4 1 1 1 1 1	1 1 1 1 1 1 1	3 2 1 2 1 3 1	V popisech SOB chybí informace o nenabydených domech, které ale byly zaznamenány v DZ > SK doplní (D). Na SM chybí Popisy některých SOB - nevytvořený kopie	95,7	71,4	76,2	3
NESHOD CELKEM		0 3 12 0 5 1 14	14 21 12 0 7 0 5	189 11 189 7 189 5	PRŮMĚR	97,51	69,55	78,87	2,48

Obrázek 1. Ukázka vyhodnocovací tabulky pro hodnocení SM v rámci Projektu SLDB 2011

VYLEPŠENÍ MODELU PRO PREDIKCI VÝSLEDKU PHADIATOP TESTU PŘEDZPRACOVÁNÍM DAT

HOW TO IMPROVE PREDICTION OF THE PHADIATOP TEST VIA PRE-PROCESSING

Pavlína Kuráňová

Adresa: VŠB-TUO, Fakulta elektrotechniky a informatiky, Katedra aplikované matematiky;
pavlina.kuranova@vsb.cz

Abstrakt: Phadiatop test byl vyvinut pro měření závažnosti alergických onemocnění, zejména atopie. Provedení Phadiatop testu je velmi nákladné, tedy cílem našeho výzkumu je přibližně předpovědět výsledky Phadiatop testu pomocí pravděpodobnostního modelu, který vyhodnocuje údaje z osobní anamnézy pacienta. Zde prezentujeme logistickou regresní techniku pro předpovídání výsledků klasifikace pacientů patřících do dvou skupin Phadiatop testu. Naše databáze je založena na pacientech, kteří podstoupili vyšetření Phadiatop testu na Klinice pracovního a preventivního lékařství, Fakultní nemocnice Ostrava. Pravděpodobnostní predikční model byl ověřen na konzistentní databázi přibližně 400 pacientů. Poté byl model zjednodušen jen na statisticky významné proměnné. V důsledku tohoto procesu zjednodušení, konečný model závisí pouze na dvou nezávislých proměnných (astma a alergická rýma). Vyvinutý model je schopen předvídat přibližné výsledky Phadiatop testu s 70% pravděpodobností.

Abstract: The Phadiatop test was developed for screening of allergic sensitization. As the Phadiatop test is not cheap, the aim of our research is to approximately predict results of the Phadiatop test by a probalistic model, which evaluates data from a personal questionnaire. Our database is based on patients who underwent examinations of the Phadiatop test at the Clinic of Occupation and Preventive Medicine, University Hospital of Ostrava, Ostrava, Czech Republic. The probabilistic prediction model has been verified on a consistent database of approximately 400 patients. Then, the second sensitivity analysis study has been performed and the prediction model has been finally simplified. As a result of the model simplification process, the final prediction model depends only on two independent variables (asthma and allergic rhinitis). The developed model is able to predict the approximate results of the Phadiatop test with 70% probability, which may represent a significant financial saving for the public health sector.

Klíčová slova: Logistická regrese, medicínská data, Phadiatop test

Keywords: Logistic regression, medical data, atopy, Phadiatop test

1. Úvod

Atopie u obyvatel České republiky roste. Atopii můžeme chápat jako osobní nebo rodinnou predispozici, být přecitlivělým na normální expozici alergenů, obvykle proteinů. Tito jedinci jsou více citliví na typické příznaky jako např. astma, ekzém, atd. Phadiatop test se používá k měření atopie.

Původní predikční model, který byl vytvořen pro Phadiatop test, využíval dvě skupiny proměnných, konkrétně rodinnou a osobní anamnézu [4]. Testem věrohodnosti byla vyloučena rodinná anamnéza. Předpovědní model tedy uvažuje pouze čtyři proměnné osobní anamnézy [4],[5]. Podle závažnosti onemocnění, jsou výsledky Phadiatop testu rozděleny do šesti následujících skupin: Skupiny 0 a I pro žádnou nebo slabou formu atopie a zbývající skupiny (II, III, IV, V, VI) indikující rostoucí závažné formy atopických příznaků. Znalost výsledků Phadiatop testu je velmi důležitá zejména při diagnostice alergických dermatóz a také pro odbornou lékařskou péči o cestovatele [2], [3].

Informace získané z osobní anamnézy byly použity pro zjištění astmatu, alergické rýmy, ekzému nebo jiných forem alergie (kontaktní, potravinové alergie, atd. Osobní anamnéza každého pacienta byla spojena s výsledky skutečně naměřeného Phadiatop testu. Následně byl vytvořen a ověřen pravděpodobnostní klasifikační model pro zařazení pacienta do jedné ze dvou skupin Phadiatop testu.

Všechny výpočty byly provedeny v software Matlab a Statgraphics.

2. Formulace problému

Testovaná databáze medicínských dat pochází z Fakultní nemocnice v Ostravě, Kliniky pracovního a preventivní lékařství. Logistická regrese je použita za účelem odhadu výsledků Phadiatop testu. Lékařské databáze obsahuje informace o výsledku Phadiatop testu a osobní anamnézu 1027 pacientů.

Pacienti přiřazení do skupiny 0 mají výsledek Phadiatop testu 0 nebo I (tj. zdraví pacienti bez příznaků onemocnění), bez nutnosti jakékoliv léčby u těchto pacientů. Zbývající pacienti s výsledky měření Phadiatop testu II – VI jsou zařazeni do skupiny 1. U těchto pacientů je již nutné lékařské ošetření.

Máme tedy jednu závislou proměnnou Y = PhModel, která závisí na čtyřech nezávislých proměnných: astma, alergická rýma, ekzém a ostatní. Proměnná Y může nabývat hodnoty 0 nebo 1, podle naměřené hodnoty Phadiatop test, skupiny 0 – I nebo skupiny II – VI.

Hodnoty nezávislých proměnných byly získány od lékařských odborníků. Skóre odborných závažností je uvedeno v tabulce 1. Kategorie „ostatní“ zde představuje různé typy alergií (potravinové alergie, atd.).

Faktor	Astma	Alergická rýma	Ekzém	Ostatní
Hodnota	10	8	6	4

Tabulka 1: Expertní ohodnocení závažnosti onemocnění.

V návaznosti na kapitolu 2, má logistický regresní model tvar:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,$$

kde $\beta_i \in \mathbf{R},\ i = 0, 1, 2, 3, 4$ jsou logistické regresní koeficienty a vektor $\mathbf{x}$ má celkem 4 složky: $x_1 = \textit{astma}$, $x_2 = \textit{alergická rýma}$, $x_3 = \textit{ekzém}$, $x_4 = \textit{ostatní}$. Závislost $\pi(\mathbf{x})$ na $\mathbf{x}$ je tvaru:

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{1}$$

Zde $\pi(\mathbf{x})$ označuje pravděpodobnost výskytu onemocnění, $\pi(\mathbf{x}) = \textit{PhModel}$. Neznámé koeficienty $\beta_i,\ i = 0, 1, 2, 3, 4$ jsou získány metodou maximální věrohodnosti. V našem případě může být metoda maximální věrohodnosti pro náš model vyjádřena jako

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} . \big(1 - \pi(x_i)\big)^{1-y_i}$$

Odhad regresních parametrů β je prováděn maximalizací logaritmu z maximálně věrohodné funkce, což může být vyjádřeno takto:

$$\ln L(\beta) = \sum_{i=1}^{n} \left[y_i . \ln \pi(x_i) + (1 - y_i) . \ln \big(1 - \pi(x_i)\big) \right]$$

Nyní může být logistická regrese použita pro predikci, zda pacient s uvedenými příznaky atopie je členem vybrané skupiny Phadiatop testu. Kvalitu pravděpodobnostní predikci modelu lze charakterizovat pojmy: senzitivita, specificita, pozitivně a negativně předpovězené hodnoty [7].

Senzitivita (SE) vrací úspěšnost modelu správně rozpoznat pacienty s onemocněním, zatímco *specificita* (SP) ilustruje úspěšnost rozpoznání pacientů bez příznaků onemocnění (zdravých). Platí

SE = TP / (TP + FN)

SP = TN / (FP + TN)

kde TP = true positives, TN = true negatives, FP = false positives, FN = false negatives.

Mimo jiné, *pozitivně předpovězená hodnota* (PPV) je definována jako počet těch jedinců, které model vyhodnotil jako pozitivní a opravdu jsou nemocní, děleno součtem všech, které model zařadil jako nemocné:

PPV = TP / (TP + FP)

Nakonec *negativně předpovězená hodnota* (NPV) je definována jako počet těch, které model označil jako zdravé a skutečně jsou bez příznaků onemocnění děleno počtem všech, které model označil jako zdravé pacienty:

NPN= TN / (FN + TN).

V ideálním predikčním modelu, bychom neměli falešně negativní nebo falešně pozitivní hodnoty. Tedy

SE = SP = PPV = NPV =100 %

pro ideální predikční model.

3. **Predikční výsledky logistického modelu: celá databáze**

Jako první krok, byl vytvořen predikční model pro Phadiatop test, s využitím všech dat, tedy kompletní database zahrnující 1027 záznamů. Nalezený model byl otestován s využitím stejných dat. Získali jsme následující globální predikční model:

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = -1,4856 + 0,1528x_1 + 0,3156x_2 + 0,0765x_3 + 0,1549x_4$$

Náš model má jednu závislou proměnnou (PhModel) a čtyři nezávislé proměnné: astma, alergická rýma, ekzém a ostatní.

Predikční výsledky globálního modelu (7) jsou shrnuty v Tabulce 2. Například, je zde hodnota 166 TP tedy pacienti s pozitivním Phadiatop testem (PhTest=1), kteří jsou správně zařazeni predikčním modelem do skupiny PhModel=1. Podobně je zde 641 TN pacientů, tzn. pacientů s negativním výsledkem Phadiatop testu (PhTest=0) správně zařazených predikčním modelem do skupiny PhModel=0.

		PhModel=1	PhModel=0
PhTest=1		166 (TP)	165 (FN)
PhTest=0		55 (FP)	641 (TN)
SE	SP	PPV	NPV
0,50	0,92	0,75	0,80

Tabulka 2. Predikční výsledky logistické regrese: globální model, 1 027 pacientů.

Negativně předpovězená hodnota z tabulky 2 je velmi dobrá, říká, že pacient nebude mít atopické symptomy (NPV = 80%) a správně identifikuje 92%, z těch, kteří nemají žádný atopický symptom (specificity).

4. **Predikční výsledky logistického modelu: zmenšená databáze**

Všechny neúplně vyplněné záznamy v lékařské databázi náležely pacientům s PhTest=0. V sekci 4, byly tyto neúplné záznamy nahrazeny nulami, protože bylo očekáváno, že všechny pozitivní alergické příznaky budou v databázi vyplněny. Na rozdíl od předchozí části, tato část přináší výsledky predikce případu, kdy byly všechny neúplně vyplněné databázové záznamů odstraněny. Tímto předzpracování procesu, tj. odstranění těchto neúplně vyplněných záznamů, byla databáze zmenšena na konečných 376 záznamů. Pro tyto data, byl vytvořen nový predikční model - Model 2:

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = -0,9765 + 0,1251x_1 + 0,2675x_2 + 0,0113x_3 + 0,0623x_4$$

Predikční výsledky tohoto modelu jsou prezentovány v Tabulce 3. Například zde máme hodnotu 166 TP pacientů. To znamená, že nemocní pacienti

s výsledkem PhTest = 1 byli správně rozpoznáni v předpovědním modelu jako PhModel = 1. Dále, počet pacientů s pozitivním PhTest je 166 + 49 = 215.

		PhModel=1	PhModel=0
PhTest=1		166 (TP)	49 (FN)
PhTest=0		55 (FP)	106 (TN)
SE	SP	PPV	NPV
0,77	0,66	0,75	0,68

Tabulka 3. Predikční výsledky regresního modelu (8) pro 376 pacientů.

V důsledku toho, že senzitivita je 166/215 = 0,77, tedy model správně rozpozná 77% skutečné pozitivní pacientů – tj. 77% nemocných je správně rozpoznáno jako nemocní. Predikční model také potvrzuje nemocné jedince pomocí pozitivně předpovězené hodnoty (PPV = 75%).

Procenta z modelu senzitivity ukazují, že predikční výsledek je mírně lepší než predikce výsledků pro kompletní databázi 1 027 pacientů. Byly odstraněny neúplně naplněné záznamy, tzn., jsou analyzovány pouze plně relevantní údaje.

Poznamenejme, že i když databáze byla významně snížena, výsledky odpovídající sloupci „PhModel = 1“ z tabulky 2 a tabulky 3 jsou stejné.

Predikční model snížené databáze (Model 2) byl také předmětem statistických testů, viz tabulky 4 a 5.

Source	Variance	Df	P-Value
Model	84.8205	4	0.0000
Residual	482.644	371	0.0207
Total (corr.)	513.464	375	

Tabulka 4. Model 2: Analysis of deviance.

Factor	Chi-Square	Df	P-Value
Astma	13.9485	1	0.0002
Alergická rýma	56.2333	1	0.0000
Ekzém	0.0356	1	0.8503
Ostatní	0.6814	1	0.4091

Tabulka 5. Test významnosti proměnných pro Model 2.

Tabulka 5 ukazuje, že proměnná ekzém a proměnná Ostatní jsou statisticky nevýznamné a mohly by být z modelu vyloučeny. Po vyloučení těchto proměnných získáváme tento zjednodušený predikční model -– Model 3:

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = -0,7927 + 0,1189x_1 + 0,2538x_2$$

I tento předpovědní model byl předmětem testování, zařazení pacienta do jedné ze dvou Phadiatop skupin. Na základě těchto výsledků lze konstatovat,

že zjednodušený model (Model 3) funguje úplně stejné jako v případě, kdy má čtyři proměnné (Model 2).

5. Závěr

Phadiatop test je důležitá, ale velmi drahá technika vyvinutá pro monitorování alergických onemocnění. Z tohoto důvodu byl vytvořen pravděpodobnostní model, který vyhodnocuje údaje z osobní anamnézy získané od pacientů s podezřením na alergické onemocnění. V předpovědním modelu, je zahrnuta jen osobní anamnéza pacienta s ohodnocením alergických onemocnění. Z těchto onemocnění se zdají být významné faktory pro predikční model Phadiatop testu jen astma a alergická rýma, protože pro všechny zbývající atributů bylo prokázáno, že jsou statisticky nevýznamné a proto byly odebrány z předpovědního modelu. Předpovědní model má zajímavé vlastnosti: statistické testy skutečné databáze pacientů udávají, že 77% nemocných lidí bylo správně rozpoznáno jako nemocní. Tedy, přibližně jen každý čtvrtý nemocný pacient byl predikčním modelem nesprávně zařazen jako zdravý.

Literatura

[1] Hosmer, D.W., Lemeshow S.: *Applied Logistic Regression*. NewYork: Wiley–Interscience, 2000

[2] Hajduková, Z., Vantuchová, Y., Klimková, P., Makhoul, M., Hromádka R.: *Atopy in patiens with allergic contact dermatitis*, Jurnal of Czech Physicians: Occupational therapy, No.2, 2009, str. 69–73.

[3] Hajduková, Z., Pólová, J., Kosek V. (2005) *The importance of Atopy Investigation in the Department of Travel Medicine, Allergies*: Magazine for continuous education in allergy and clinical immunology, No.2, 2005, str. 106–109.

[4] Kuráňová P., Hajduková Z., Praks P. (2010) *Logistic regression as a tool for athopy investigation*, Mendel- 16th International Conference on Soft Computing 2010, str. 187–190.

[5] Kuráňová P., The processing of the medical data with the use of logistic regression, Reliability & Risk Analysis: Theory & Applications, Journal of international group on reliability, ISSN 1932-2321.

[6] Rabasová M., Briš R., Kuráňová P. (2010) *Modified logistic regression as a tool for discrimination*. Reliability, Risk and Safety:Theory and Applications. R. Briš, C. Guedes Soares, S. Martorell (eds.); Vol 3, 2010, pp.1967-1972

[7] Mangrulkar Rajesh S., Gruber S. (2012) *Patients and Populations: Medical Decision-Making First Steps Towards Lifelong Learning*. University of Michigan,

Poděkování: Příspěvek vznikl za podpory Evropského fondu regionálního rozvoje v IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070) and the FEECS VŠB – Technická univerzita Ostrava (Projekt č. SP2012/180).

MCMC PERFECT SAMPLING

David Legát
Adresa: MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8
david.legat@gmail.com

Abstract: Distribution of many statistical procedures is often so complex, that exact formula for parameter estimate is difficult or even impossible to derive. Simulation techniques such as rejection sampling or MCMC (Markov chain Monte Carlo) are usually employed in such situation and characterization of the distribution is derived from the sample. MCMC simulation generates Markov chain with desired stationary distribution, so that member of generated sequence, which is sufficiently far from the beginning, has approximately the distribution we need. However, there are several questions associated with this approach like: "How long sequence should we produce to be close enough to demanded distribution?" or "Is it possible to generate stationary Markov chain with given stationary distribution?" This paper describes technique which allows to generate initial member of stationary Markov chain with desired distribution along with supporting theory and demonstration on Ising model sample generation.

Keywords: MCMC simulations, perfect sampling, Ising model, Markov chain
Acknowledgement: Research was supported by grant GAČR 201/09/0755.

1. Introduction

Consider a finite set S and the probability distribution $\boldsymbol{\pi}$ given by probabilities $\pi_s,\ s \in S$. In this paper, we will try to generate sequence of random variables $X_i, i \in \mathbb{N}_0$, taking values in the set S, so that values of the sample X_i can be used to describe distribution $\boldsymbol{\pi}$. In particular, we will be interested in sequences where marginal law $\mathcal{L}(X_i)$ of random variables $X_i,\ i \in \mathbb{N}_0$ is given by probability $\boldsymbol{\pi}$.

Generation of independent identically distributed (iid) random variables X_i with distribution $\boldsymbol{\pi}$ is the most desirable in this situation. When there is no pseudo-random generator available for simulation of $\boldsymbol{\pi}$ we can apply algorithm called *rejection sampling*, which generates every instance of sample X_i in two steps.

(1) Generate $s^\star \in S$ from probability distribution $\widetilde{\boldsymbol{\pi}} = \left\{\widetilde{\pi};\ s \in S\right\}$ where $\widetilde{\boldsymbol{\pi}}$ is the distribution we are able to quickly generate random variable from, and for which there exists a constant $K > 0$ such that

$$\pi_2/\widetilde{\pi}_s \leq K, \qquad \forall s \in S.$$

(2) Accept the suggested value $s^\star$ from the first step with probability $p = \pi_{s^\star}/K\widetilde{\pi}_{s^\star}$ and proceed to generate next element of the sample X_i. Return back to the first step and generate new candidate $s^\star$ when the actual one is rejected.

Remark 1.1:

The constant K always exist for distributions on finite set S. However, the algorithm would reject too often (and would be too slow) if the distribution $\widetilde{\boldsymbol{\pi}}$ does not approximate well the distribution $\boldsymbol{\pi}$ and the constant K is too large. Thus, the usability of reject sampling depends on the existence of distribution $\widetilde{\boldsymbol{\pi}}$ which is close enough to the target distribution $\boldsymbol{\pi}$ and we are able to generate a sample from it. □

In some circumstances it could be more efficient to generate a Markov chain with desired stationary distribution instead of using iid random variables. Markov chain Monte Carlo (MCMC) methods are often used for this purpose. Especially the Metropolis-Hastings algorithm is one of MCMC methods which is similar to reject sampling. The algorithm described by Hastings in [4] generates a Markov chain with transitional probabilities forming a matrix $\boldsymbol{P}$ satisfying two conditions:

- Reversibility condition
$$\pi_r P_{r,s} = \pi_s P_{s,r}, \quad \forall r, s \in S$$
- Transitional probabilities could be rewritten as a multiplication
$$P_{r,s} = Q_{r,s}\alpha_{r,s},$$
where $\boldsymbol{Q} = \{Q_{r,s} : r, s \in S\}$ are transitional probabilities such that we can generate random variables with distribution given by probabilities $Q_{r,s}, s \in S$ for every $r \in S$, and constants $\alpha_{r,s}$ belong to interval $< 0, 1 >$.

Reversibility condition guarantees that $\boldsymbol{\pi}$ is a stationary distribution of generated chain, while the form of transitional probabilities enables generation of chain in two steps similarly to rejection sampling.

Let us concentrate on this point in detail. Select (at random or deterministically) starting point of generated chain $x^{(0)} \in S$. Assuming that first t elements $s^{(0)}, \ldots, s^{(t-1)}$ of the chain are generated, we will perform following two steps to generate the next element $s^{(t)}$.

1. Generate s^* from the distribution given by the transitional probabilities $Q_{s_{t-1},s}$, $s \in S$.
2. Set $s_t = s^*$ with probability α_{s_{t-1},s^*}, otherwise keep old value, i.e. set $s_t = s_{t-1}$.

To fulfill the reversibility condition, we must select acceptance probabilities $\boldsymbol{\alpha}$ of new candidate in the form

$$\alpha_{r,s} = \min\left(1\ ,\ \frac{\Pi_s Q_{s,r}}{\Pi_r Q_{r,s}}\right), \quad r, s \in S \tag{1}$$

legat/

Set $\alpha(r,s) = 1$ when $\Pi_r Q_{r,s} = 0$.

Following section will provide necessary theoretical background including the theorems showing that Metropolis-Hastings algorithm generates a sequence $\boldsymbol{X}_i$ with distribution converging to the target distribution $\boldsymbol{\pi}$.

2 Markov chain fundamentals

The Metropolis-Hastings algorithm generates a sequence $X_i, i \in \mathbb{N}^0$, so that both steps performed to obtain new member X_{n+1} uses only the information about the value of the actual state X_n. So X_{n+1} is conditionally independent on any random variable $X_{n-k}, k > 0$ given the value of X_n. This property of random processes is usually called *Markov property* and the process with this property is called *Markov process (chain)*. Precise definition follows.

Definition 2.1:

A sequence of random variables $X_i, i \in \mathbb{N}^0$, with values in finite set of states S is called Markov chain if it satisfies

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n),$$

for all $k \in \mathbb{N}^0$ and $x_0, \ldots, x_{n+1} \in S$ such that $P(X_n = x_n, \ldots, X_0 = x_0) > 0$. Markov chain is called *homogeneous* if conditional probabilities

$$P(X_{n+1} = x_{n+1} | X_n = x_n) = p(x_n, x_{n+1})$$

does not depend on n. Probabilities $p(x_n, x_{n+1})$ are called *transitional probabilities* and they form so called *transitional matrix* $\boldsymbol{P}$. Moreover, Markov chain is *stationary* if distribution of random vectors $P(X_{n+k} = x_{n+k}, \ldots, X_k = x_k)$ does not depend on $k \in \mathbb{N}^0$ for every $n \in \mathbb{N}^0$. □

We will be interested in the homogeneous Markov chains only. Its distribution is given by the distribution of the initial variable $\boldsymbol{\pi}_0 = \{P(X_0 = x), x \in S\}$ and transitional probabilities $p(x_1, x_2),\ x_1, x_2 \in S$. Markov chain generated by the Metropolis-Hastings algorithm is stationary if, and only if the initial state X_0 is generated from desired distribution $\boldsymbol{\pi}$. We can simply check, that the chain started in a fixed state $x_0 \in S$ is not stationary, because the distribution of the first variable X_0 is concentrated into one state x_0, while the distribution of the second member X_1 is given by suitable line of transition matrix $p(x_0, \cdot)$. However, under mild conditions on the structure of transition matrix P, we can show that marginal distributions $\mathcal{L}(X_n)$ converges as $n \to \infty$. This limit distribution is called stationary distribution.

Definition 2.2:

Let us consider homogeneous Markov chain $\{X_i\}_{i\in\mathbb{N}^0}$ with transitional probabilities $p(s,t),\ s,t\in S$. Probability distribution $\boldsymbol{\pi}$ on the set of states S which satisfies

$$\pi_t=\sum_{s\in S}\pi_s p(s,t)\qquad \forall t\in S$$

is *stationary distribution* of Markov chain $\{X_i\}_{i\in\mathbb{N}^0}$. □

Existence and unicity of stationary distribution is foremost given by the accessibi of the states among themselves.

Definition 2.3:

A state $t\in S$ is said to be *accessible* from state $s\in S$ if there exists an integer $n\in\mathbb{N}$ such that

$$P(X_n=t|X_0=s)>0.$$

Markov chain is said to be *irreducible* if t is accessible from s for every pair of states $s,t\in S$.
Each state $s\in S$ has its *period* given by

$$\Gamma_s=GCD(\{n:P(X_n=s|X_0=s)>0\})$$

where function GCD is the greatest common divisor. A state $s\in S$ is called *aperiodic* if $\Gamma_s=1$, otherwise it is called *periodic with period* Γ_s.
Let the random variable T_s is first return time (recurrence time) to state $s\in S$

$$T_s=\inf\{n\geq 1:X_n=s|X_0=s\}$$

and $f_{ss}^{(n)}$ is the probability that the chain returns to the state s for the first time after n steps.

$$f_{ss}^{(n)}=P(T_s=n).$$

The state $s\in S$ is said to be *transient* if

$$P(T_s<\infty)=\sum_{n=1}^{\infty}f_{ss}^{(n)}<1$$

otherwise the state s is said to be *persistent*. The persistent state $s\in S$ is *non-null* if the mean recurrence time is finite

$$M_s=ET_s=\sum_{n=1}^{\infty}nf_{ss}^{(n)}<\infty.$$

Otherwise the state is called *null* persistent. □

An accessibility condition says that a system started in state $r \in S$ has a non-zero probability of transitioning into state $s \in S$ at some time. It's equivalent to condition that there exist an integer $n \in \mathbb{N}$ and a sequence of states $x_1, \ldots, x_{n-1} \in S$ so that

$$p(r, x_1)p(x_1, x_2) \ldots p(x_{n-2}, x_{x-1})p(x_{n-1}, r) > 0.$$

A state $r \in S$ is said to communicate with a state $s \in S$ if both s is accessible from r and r is accessible from s. It can be shown that communication in this sense is an equivalence relation and thus it forms equivalence classes often called communication classes. The definition of irreducible Markov chain could be also reformulated to the condition that all states form one communication class.

In most of the literature dealing with Markov chain theory we can find following two lemmas which are crucial for the MCMC methods. Both lemmas were taken from the lecture notes [9] (in Czech language). Another example of monography about Markov chain theory containing necessary background is [8].

Lemma 2.4:

An irreducible Markov chain with a finite set of states S has all states non-null persistent. □

Lemma 2.5:

An irreducible chain has a stationary distribution $\boldsymbol{\pi}$ if and only if all of its states are non-null persistent. In that case, $\boldsymbol{\pi}$ is unique and is related to the mean recurrence time

$$\pi_s = \frac{1}{M_s}, \quad \forall s \in S$$

Furthermore, if all states of the chain are aperiodic, then for any $r, s \in S$

$$\lim_{n\to\infty} p_{rs}^{(n)} = \pi_s$$

□

According to these lemmas, if some method generates irreducible Markov chain with transition matrix $\boldsymbol{P}$, such Markov chain has unique stationary distribution $\boldsymbol{\pi}$ and, moreover, the marginal distribution of the chain converges to the $\boldsymbol{\pi}$

$$\boldsymbol{\nu}\boldsymbol{P}^n \longrightarrow \boldsymbol{\pi}$$

where $\boldsymbol{\nu}$ denotes a distribution of starting point X_0. Note that there is no assumption on the starting distribution. Thus the chain converges to the stationary distribution regardless of where it begins.

Markov chain Monte Carlo simulation methods often generate a chain which fulfills the reversibility condition. (e.g. Metropolis-Hastings or Gibbs sampler generates such chains) It means that there exists distribution $\bar{\boldsymbol{\pi}}$ such that

$$\bar{\pi}_r p(r,s) = \bar{\pi}_s p(s,r) \quad \forall r, s \in S.$$

Summing the equation over r gives

$$\begin{aligned} \sum_{r\in S} \bar{\pi}_r p(r,s) &= \sum_{r\in S} \bar{\pi}_s p(s,r) \\ &= \bar{\pi}_s \sum_{r\in S} p(s,r) = \bar{\pi}_s \end{aligned}$$

This could be rewritten as an equation $\bar{\boldsymbol{\pi}} = \bar{\boldsymbol{\pi}}\boldsymbol{P}$ from the definition of stationary distribution. It shows that Metropolis-Hastings algorithm generates a Markov chain with target stationary distribution.

3 Functional representation of Markov chain

We know from the previous section that theoretical value of Markov chain in the infinity has target distribution $\boldsymbol{\pi}$, but we are not able to observe this value. Therefore we will try to generate Markov chain from minus infinity (or some point sufficiently far in the past) and observe the value of chain in some finite time (e.g. zero). Because we are not able to decide at which point we started the chain in the minus infinity, we have to consider all possible starting points $s \in S$. For this purpose we need to select different representation of Markov chain. We will work with "both-sided" Markov chain in the reminder of this section, i.e. $\{X_i\}_{i\in\mathbb{Z}}$ instead of positive indices only. Moreover, we will represent Markov chains in terms of random maps.

Definition 3.1:

Let Φ denotes a set of all maps from S to itself

$$\Phi = \{\varphi : S \to S\} \equiv S^S.$$

Consider random variable $\varphi = (\Phi, \mathbb{P})$ with values in set Φ and probability distribution $\mathbb{P}$. We call φ a *random map* with respect to the transition matrix $\boldsymbol{P}$ if it satisfies

$$\mathbb{P}(\{\varphi : \varphi(s) = t\}) = p(s,t), \quad \forall s, t \in S. \tag{2}$$

Recall from Section 2 that transitional probabilities $p(s,t),\ s,t \in S$, are elements of matrix $\boldsymbol{P}$. □

Remark 3.2:

Given a matrix $\boldsymbol{P}$, a distribution fulfilling Condition (2) always exists. A simple example of such distribution is the one given by $\mathbb{P}(\{\varphi\}) = \prod_{s\in S} \boldsymbol{P}(s,\varphi(s))$. Rows $\boldsymbol{P}(s,\cdot)$ of the transitional matrix $\boldsymbol{P}$ represents probability distribution on the set S. Let $S = \{s_1, \cdots, s_i, \cdots\}$, then we can write

$$\begin{aligned}
\mathbb{P}(\Phi) &= \sum_{s_0'\in\S} \mathbb{P}(\{\varphi(s_1) = s_1'\}) = \sum_{s_1'\in\S}\sum_{s_2'\in\S} \mathbb{P}(\{\varphi(s_1) = s_1', \varphi(s_2) = s_2'\}) \\
&= \sum_{s_1'\in\S}\sum_{\cdots}\sum_{s_i'\in\S}\sum_{\cdots} \mathbb{P}(\{\varphi(s_1) = s_1', \cdots, \varphi(s_i) = s_i', \cdots\}) \\
&= \sum_{s_1'\in\S}\sum_{\cdots}\sum_{s_i'\in\S}\sum_{\cdots}\prod_{i\in\mathbb{N}} \boldsymbol{P}(s_i, s_i') = \sum_{s_1'\in\S} \boldsymbol{P}(s_1, s_1') \sum_{\cdots}\sum_{s_i'\in\S} \boldsymbol{P}(s_i, s_i') \sum_{\cdots} \\
&= \prod_{i\in\mathbb{N}} 1 = 1.
\end{aligned}$$

Thus $\mathbb{P}$ is probability measure. Similarly, choosing $k \in \mathbb{N}$, we then derive

$$\begin{aligned}
\mathbb{P}(\{\varphi(s_k) = s_k'\}) &= \sum_{s_1'\in\S}\sum_{\cdots}\sum_{s_{k-1}'\in\S}\sum_{s_{k+1}'\in\S}\sum_{\cdots}\prod_{i\in\mathbb{N}} \boldsymbol{P}(s_i, s_i') \\
&= \boldsymbol{P}(s_k, s_k') \prod_{i\neq k} 1 \\
&= \boldsymbol{P}(s_k, s_k').
\end{aligned}$$

This assures that $\mathbb{P}$ fulfills Condition (2). □

Random map with respect to transition matrix $\boldsymbol{P}$ mimic one step of Markov chain. Given the state $x_n \in S$ of Markov chain with transitional probabilities $p(x_n, x_{n+1})$ (from matrix $\boldsymbol{P}$) we can generate new member of the chain so that we select a random map $\varphi_{n+1} \in \Phi$ using the distribution $\mathbb{P}$ and subsequently we set $x_{n+1} = \varphi_{n+1}(x_n)$. Similarly, we can mimic m step transition of the Markov chain so that we select independently m maps $\varphi_i \in \Phi,\ i = n+1, \ldots, n+m$ and set $x_{n+m} = \varphi_{n+1}^{n+m}(x_n)$ where $\varphi_{n+1}^{n+m} = \varphi_{n+m} \circ \cdots \circ \varphi_{n+1}$. Symbol $\varphi \circ \psi$ denotes usual composition of maps given by $\varphi \circ \psi(x) = \varphi(\psi(x))$. Let us now consider sequences of random maps.

The space $\Omega = \Phi^{\mathbb{Z}}$ consists of "both-sided" sequences

$$\bar{\varphi} = \{\varphi_j\}_{j\in\mathbb{Z}} = (\ldots, \varphi_{j-1}, \varphi_j, \varphi_{j+1}, \ldots)$$

Let us consider a product probability measure $\bar{\mathbb{P}}$ on space Ω such that for every finite subset $I \subset \mathbb{Z}$ holds

$$\bar{\mathbb{P}}(\{\bar{\varphi} \in \Omega : \varphi_i = \psi_i,\ i \in I\}) = \prod_{i\in I} \mathbb{P}(\psi_i).$$

The probability space $(\Omega, \bar{\mathbb{P}})$ is often called *stochastic flow*. Given the condition (2) The process $X_n = x$, $X_{n+k} = \varphi_{n+1}^{n+k}(x)$ is a Markov chain starting at state $x \in S$ with

transitional probability $\boldsymbol{P}$. Hence a stochastic flow is a common representation of Ma chains starting at all initial states and all times.

Consider now a sequence $\bar{\varphi} \in \Omega$. We say that there is *complete coalesce* at tin if there exists a constant $k \in \mathbb{N}^0$ and state $\omega \in S$ such that $\varphi_{n-k}^n(x) = \omega$ for e $x \in S$. This condition says that all chains X_n given by the stochastic flow $\bar{\varphi}$ star at time $n-k$ will finish in state ω at time n no matter what the starting point Going further backward in time with the starting point does not change anything bec $\varphi_{n-k-l}^n(x) = \varphi_{n-k-l}^{n-k-1} \circ \varphi_{n-k}^n(x) = \omega$ for all $l \in \mathbb{N}$. Let us denote F_n a set of all stoch flows with complete coalesce at time n and $F = \bigcap_{n\in\mathbb{Z}} F_n$. We will be mainly intere in stochastic flow which satisfies

$$\mathbb{P}(F) = 1.$$

The Condition (3) is often called *almost sure complete coalesce in finite time*. We check when this condition holds and provide a methodology to construct a stochastic with this property later in the text.

Further denote

$$W_n(\bar{\varphi}) = \begin{cases} \lim_{m\to-\infty} \varphi_m^n(x)\ x \in S & \bar{\varphi} \in F_n \\ \omega_0 & \text{otherwise,} \end{cases}$$

where $\omega_0 \in S$ is arbitrarily selected state to make $W_n(\bar{\varphi})$ well defined on whole spac

Theorem 3.3:

Under conditions (2) and (3) the random process $\{W_n\}_{n\in\mathbb{Z}}$ is stationary homogeneous Markov chain with transition matrix $\boldsymbol{P}$. □

Proof: Let us verify the Markov property. Since

$$W_{n+1}(\bar{\varphi}) = \lim_{m\to-\infty} \varphi_m^{n+1}(x) = \varphi_{n+1}\left(\lim_{m\to-\infty} \varphi_m^n(x)\right) = \varphi_{n+1}\left(W_n(\bar{\varphi})\right)$$

for every $\bar{\varphi} \in F$ and $n \in \mathbb{N}$, $W_k(\bar{\varphi})$ depends on $\varphi_m,\ m \le n$ and therefore is independent with φ_{n+1} we can write

$$\begin{aligned} &P\left(W_{n+1} = s_{n+1}, W_n = s_n, \ldots, W_{n-k} = s_{n-k}\right) = \\ &\quad = P\left(\varphi_{n+1}(s_n) = s_{n+1}, W_n = s_n, \ldots, W_{n-k} = s_{n-k}\right) = \\ &\quad = P\left(\varphi_{n+1}(s_n) = s_{n+1}\right) P\left(W_n = s_n, \ldots, W_{n-k} = s_{n-k}\right) = \\ &\quad = p(s_n, s_{n+1}) P\left(W_n = s_n, \ldots, W_{n-k} = s_{n-k}\right) \end{aligned}$$

The last equation comes from the Condition (2). Finally, conditional probability is

$$P\left(W_{n+1} = s_{n+1} | W_n = s_n, \ldots, W_{n-k} = s_{n-k}\right) = p(s_n, s_{n+1}).$$

Thus, $\{W_n\}_{n\in\mathbb{Z}}$ is homogeneous Markov chain with transitional probability $\boldsymbol{P}$. The process $\{W_n\}_{n\in\mathbb{Z}}$ is also stationary, because sequences of random maps $\{\varphi_k\}_{k<m}$ have shift invariant distribution which does not depend on m. □

Theorem 3.3 offers a simulation method of random variable with exactly target distribution using Markov chain generation. Suppose, we have strictly positive distribution $\boldsymbol{\pi}$ on the set of states S and we are able to generate an irreducible Markov chain with kernel $\boldsymbol{P}$ so that $\boldsymbol{\pi}$ is its stationary distribution. Suppose further, that we can construct random maps so that it satisfies conditions (2) and (3). Than the random variable W_n is simulated so that we pick random maps $\varphi_0, \varphi_{-1}, \dots$ (starting at time 0 and continuing backwards) until complete coalesce of composed map $\varphi_k^0 = \varphi_0 \circ \varphi_{-1} \circ \cdots \circ \varphi_k$. Than the value $\omega = \varphi_k^0(s)$ which does not depend on $s \in S$ is good representative of random variable W_n with distribution $\boldsymbol{\pi}$. Let us demonstrate this approach on a simple example.

Example 3.4:

Let S={0,1} be a two point state space. Transitional probabilities

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}, \quad 0 < \alpha, \beta < 1$$

generates a Markov chain with stationary distribution $(\pi_1, \pi_2) = (\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta})$. Set of all maps S^S contains 4 elements:

$$\begin{array}{ll} \varphi^{(11)}: & \varphi^{(11)}(1) = 1, \quad \varphi^{(11)}(2) = 1 \\ \varphi^{(12)}: & \varphi^{(12)}(1) = 1, \quad \varphi^{(12)}(2) = 2 \\ \varphi^{(21)}: & \varphi^{(21)}(1) = 2, \quad \varphi^{(21)}(2) = 1 \\ \varphi^{(22)}: & \varphi^{(22)}(1) = 2, \quad \varphi^{(22)}(2) = 2 \end{array}$$

Maps $\varphi^{(11)}$ (resp $\varphi^{(22)}$) will coalesce both states into state 1 (resp. 2), while $\varphi^{(12)}$ will keep the states and $\varphi^{(21)}$ will switch each state to the other. Distribution of maps given by probabilities

$$\begin{array}{ll} P(\varphi^{(11)}) = (1-\alpha)\beta & P(\varphi^{(12)}) = (1-\alpha)(1-\beta) \\ P(\varphi^{(21)}) = \alpha\beta & P(\varphi^{(22)}) = \alpha(1-\beta) \end{array}$$

fulfills condition (2). Now we can pick a random map φ_0 from this distribution. If the selected map is $\varphi^{(11)}$ (or $\varphi^{(22)}$) we will assign $W_0 = 1$ (or $W_0 = 2$). Otherwise, we will continue and pick a random map φ_{-1} from the same distribution. We will continue until one of the maps $\varphi^{(11)}$ and $\varphi^{(22)}$ is selected. Probability of (complete) coalesce at finite time is

$$\sum_{k=0}^{\infty} \left(P(\varphi^{(11)}) + P(\varphi^{(11)})\right) \left(P(\varphi^{(21)}) + P(\varphi^{(12)})\right)^k = 1$$

Process $\{W_k\}_{k\in\mathbb{Z}}$ is stationary. Value W_{-1} is generated by the same principle like W_0 if $\varphi^{(12)}$ or $\varphi^{(21)}$ is selected in the first step. Therefore probabilities $P_1 = P(W_0 = 1) = P(W_{-1} = 1)$ and $P_2 = P(W_0 = 2) = P(W_{-1} = 2)$ have to solve the system of linear equations

$$\begin{array}{rcl} P(W_0 = 1) & = & P(\varphi^{(11)}) + P(\varphi^{(12)})P(W_{-1} = 1) + P(\varphi^{(21)})P(W_{-1} = 2) \\ P(W_0 = 2) & = & P(\varphi^{(22)}) + P(\varphi^{(12)})P(W_{-1} = 2) + P(\varphi^{(21)})P(W_{-1} = 1) \end{array} \quad (4)$$

Summation of these equations confirms that its solution is probability distribution

$$P_1 + P_2 = \frac{P(\varphi^{(11)}) + P(\varphi^{(22)})}{1 - P(\varphi^{(12)}) - P(\varphi^{(21)})} = 1$$

We can simply check that probabilities π_1 and π_2 are the only solution of equations (4) and therefore the algorithm generates a random variable with stationary distribution suitable to kernel $\boldsymbol{P}$.

Conversely, the forward coalesce approach does not generate a stationary distribution $\boldsymbol{\pi}$. Suppose that we generate two independent Markov chains $\{X_i^{(1)}\}_{i\in\mathbb{N}}$ and $\{X_i^{(2)}\}_{i\in\mathbb{N}}$ each starting in another state. Stop the Markov chains at the moment when $X_i^{(1)} = X_i^{(2)}$ for the first time. More precisely, let us denote random variable $Z = X_T^{(1)}$ where T is a random time such that $X_i^{(1)} \neq X_i^{(2)}$ for $i < T$ and $X_T^{(1)} = X_T^{(2)}$. Given that chains did not coalesce until time t we know that $X_t^{(1)} \neq X_t^{(2)}$. The probability that we will coalesce at time $t + 1$ is $(1 - \alpha)\beta + \alpha(1 - \beta)$ and it composes from the probability of coalesce into state 1 (one chain switch from state 2 to 1, while the other stay in 1)

$$P(Z = 1|T = t + 1) = (1 - \alpha)\beta$$

and the probability of coalesce into state 2

$$P(Z = 2|T = t + 1) = \alpha(1 - \beta).$$

Since

$$P(Z = z) = \sum_{t=1}^{\infty} P(Z = z|T = t)P(T = t)$$

and probabilities $P(Z = z|T = t)$ does not depend on t the distribution $\boldsymbol{\pi}_Z$ of random variable Z is given by the vector

$$\boldsymbol{\pi}_Z = \left(\frac{(1 - \alpha)\beta}{(1 - \alpha)\beta + \alpha(1 - \beta)}, \frac{\alpha(1 - \beta)}{(1 - \alpha)\beta + \alpha(1 - \beta)} \right)$$

which in general (except the case $\alpha = \beta = 1/2$) differs from the stationary distribution $\boldsymbol{\pi}$. □

In the Example 3.4 we worked with 2 elements state space S and the set of all maps S^S consisted of four elements only. However typical application of MCMC simulation operates on much larger spaces. We will describe a general construction of random maps which are related to transition matrix $\boldsymbol{P}$ by Condition (2) and that is easy to use for simulations even with more complex state spaces.

Let us consider a function $f : S \times \Theta \to S$ and independent identically distributed random variables $U_i, \in \mathbb{Z}$ taking values in Θ. Then $\varphi_i = f(\cdot, U_i)$ is a random map. Common choice is Θ to be interval $< 0, 1 >$ and random variables U_i with uniform distribution on that interval. Now, let us further consider that $s_1, \cdots, s_n \in S$ are elements of state space. Since the set S is finite, we can imagine the domain of function $f(s, u)$ like n lines going from 0 to 1 and the function $f(x, u)$ assigns values from state space S to the elements of those lines. We can divide the line corresponding to the state $s_i \in S$ proportionally to i-th line of transitional matrix $\boldsymbol{P}$. For specific s_i we assign

$$\begin{array}{ll} f(s_i, u) = s_1 & \text{if } u \in (0, p(s_i, s_1)) \\ f(s_i, u) = s_2 & \text{if } u \in (p(s_i, s_1), p(s_i, s_1) + p(s_i, s_2)) \\ \vdots & \vdots \\ f(s_i, u) = s_n & \text{if } u \in (1 - p(s_i, s_n), 1). \end{array}$$

We can see that $P(\varphi(s_i) = s_j) = P(f(s_i, U) = s_j) = p(s_i, s_j)$, which means that random maps constructed in the way described above fulfills the Condition (2).

Verification of the Condition (3) could be also more complicated with larger set of states S than the Example 3.4 presents. Several equivalent forms of this condition are stated in the next theorem. But prior to this theorem we will define random variable representing time of coalesce a we will show its sub-multiplicativity property.

Definition 3.5:

Let $\bar{\varphi} \in \Omega$ is a stochastic flow. Then a random variable

$$T_n(\bar{\varphi}) = sup\{m \leq n : \exists \omega \in S \text{ such that } \varphi_m^n(s) = \omega \ \forall s \in S\}$$

is called *time of the latest coalesce* before n. □

Time of the latest coalesce has so called property of sub-multiplicativity.

Lemma 3.6:

Let $n, m < 0$ be a negative integer. Then

$$\bar{\mathbb{P}}(T_0 \leq m + n) \leq \bar{\mathbb{P}}(T_0 \leq m)\bar{\mathbb{P}}(T_m \leq m + n) = \bar{\mathbb{P}}(T_0 \leq m)\bar{\mathbb{P}}(T_0 \leq n)$$

□

Proof: Inequality $T_0(\bar{\varphi}) \leq m+n$ imply that neither φ^n_{m+n+1} nor φ^0_{n+1} maps all $s \in S$ into one state. Since random maps φ^n_{m+n+1} and φ^0_{n+1} are independent this proves the left inequality $\bar{\mathbb{P}}(T_0 \leq m+n) \leq \bar{\mathbb{P}}(T_0 \leq m)\bar{\mathbb{P}}(T_m \leq m+n)$. The equality $\bar{\mathbb{P}}(T_m \leq m+n) = \bar{\mathbb{P}}(T_0 \leq n)$ is a consequence of stationarity. □

Theorem 3.7:

Let $(\Omega, \bar{\mathbb{P}})$ be a space of stochastic flows. Then following conditions are all equivalent to Condition (3)

1) $\bar{\mathbb{P}}\{\bar{\varphi} \in \Omega : T_n(\bar{\varphi}) > -\infty\} = 1 \quad \forall n \in \mathbb{Z}$
2) $\exists \tau \in \mathbb{N}$ such that $\bar{\mathbb{P}}\{\bar{\varphi} \in \Omega : T_0(\bar{\varphi}) > -\tau\} > 0$
3) $\forall s,t \in S\ \exists n_{st} \in \mathbb{N}$ such that $\bar{\mathbb{P}}\{\bar{\varphi} \in \Omega : \varphi_1^{n_{st}}(s) = \varphi_1^{n_{st}}(t)\} > 0$

□

Proof: The first condition is another way of formulating Condition (3). So we need to proof equivalence of conditions presented in this theorem only. Let us start by the implication 3) ⇒ 2).
Denote $n_m = \max_{s,t\in S}\{n_{st} : \bar{\mathbb{P}}(\varphi_1^{n_{st}}(s) = \varphi_1^{n_{st}}(t)) > 0\}$ the maximal number of steps n_{st} from condition 3). Since coalesce of states $s,t \in S$ at n_{st} enforces also its coalesce at n_m (once coalesced states could not be divided by further maps) every two states $s,t \in S$ have positive probability of coalesce at n_m. Than the minimum $p_m = \min_{s,t\in S} \bar{\mathbb{P}}(\varphi_1^{n_m}(s) = \varphi_1^{n_m}(t))$ is positive. Therefore $|S| > |\varphi_1^{n_m}(S)|$ with probability greater then p_m (if $|S| \geq 2$), where $|S|$ is number of S elements and $\varphi(S)$ denotes an image of set S by map φ. Similarly, the probability of $|S| > |\varphi_1^{n_m}(S)| > |\varphi_{n_m+1}^{2n_m}(S)|$ is greater then p_m^2 because random maps $\varphi_1^{n_m}(S)$ and $\varphi_{n_m+1}^{2n_m}(S)$ are independent. Since S is finite we need to repeat this iteration $|S|-1$ times at most to reach $|\varphi_1^{(|S|-1)n_m}(S)| = 1$ with probability greater than $p_m^{|S|-1} > 0$. Shift of indexes from $1,\ldots,\tau = n_m(|S|-1)$ to $-\tau+1,\ldots,0$ gives the condition 2).
The implication 2) ⇒ 1) is the consequence of sub-multiplicativity property of T_0 and a stationarity of $T_n,\ n \in \mathbb{Z}$. More precisely,

$$\bar{\mathbb{P}}(T_0 > -\infty) = 1 - \lim_{k\to\infty} \bar{\mathbb{P}}(T_0 \leq -k\tau)$$

where τ is the one from condition 2) satisfying $\bar{\mathbb{P}}(T_0 > -\tau) > 0$. The sub-multiplicativity condition gives

$$\bar{\mathbb{P}}(T_0 \leq -k\tau) \leq \bar{\mathbb{P}}(T_0 \leq -\tau)^k = (1 - \bar{\mathbb{P}}(T_0 > -\tau))^k \underset{n\to\infty}{\longrightarrow} 0$$

Moreover, the times $T_n,\ n \in \mathbb{Z}$ depend on random maps $\{\ldots,\varphi_{n-1},\varphi_n\}$ and their distributions are shift invariant. Therefor the distribution of T_n does not

depend on $n \in \mathbb{Z}$ and the condition $\mathbb{P}(T_0 > -\infty)$ could be extended to $\bar{\mathbb{P}}(T_n > -\infty)$ for every $n \in \mathbb{Z}$.
The implications in opposite direction are trivial because $\bar{\mathbb{P}}(T_0 > -\infty) = 1$ implies $\bar{\mathbb{P}}(T_0 > -\infty) > 0$ and complete coalesce enforces pairwise coalesce. □

4 Time of the latest coalesce

Theorem 3.7 works with the random variable T_0 and shows the circumstances on wh this variable is almost sure finite. However, no estimate of method's time comple was provided. A general approach to time complexity estimation independent of spe transition matrix $\boldsymbol{P}$ choice and random map φ construction is outlined by follov theorem.

Theorem 4.1:

Let $(\Omega, \bar{\mathbb{P}})$ be a space of stochastic flows and $T_0(\bar{\varphi}),\ \bar{\varphi} \in \Omega$ is a time of the latest coalesce. Assume that there exists $\tau \in \mathbb{N}$, such that $\bar{\mathbb{P}}(T_0 \geq -\tau) = \epsilon > 0$. Then moments of T_0 exist and

$$\mathbf{E}(-T_0)^n \leq k^n \sum_{m=0}^{\infty} (1-\epsilon)^m ((m+1)^n - m^n)$$

Specifically,

$$-\frac{\tau}{e} \leq \mathbf{E}T_0 \leq -\tau(1-\epsilon).$$

and

$$\mathbf{Var}\ T_0 \leq \tau^2 \left(\frac{1}{\epsilon} + \frac{2(1-\epsilon)}{\epsilon^2} - (1-\epsilon)^2 \right).$$

□

Proof: Let T be a discrete random variable with values in $\mathbb{N}_0$. If the moments of T exist, they are given by

$$\mathbf{E}T^n = \sum_{i=0}^{\infty} ((i+1)^n - i^n) P(T > i)$$

Derivation of this formula follows.

$$\begin{aligned} \mathbf{E}T^n &= \textstyle\sum_{i=1}^{\infty} i^n P(T=i) = \\ &= \textstyle\sum_{i=1}^{\infty} P(T=i) \sum_{j=1}^{i} (j^n - (j-1)^n) \end{aligned}$$

Summation over i and j could be switched if the moment exist.

$$\begin{aligned}\mathbf{E}T^n &= \textstyle\sum_{j=1}^{\infty}(j^n-(j-1)^n)\sum_{i=j}^{\infty}P(T=i) = \\ &= \textstyle\sum_{j=1}^{\infty}(j^n-(j-1)^n)P(T\geq j) = \\ &= \textstyle\sum_{j=0}^{\infty}((j+1)^n-(j)^n)P(T>j)\end{aligned}$$

Moreover, let T has the property of sub-multiplicativity and there exists $\tau\in\mathbb{N}$, such that $P(T\leq\tau)=\epsilon>0$. Then

$$P(T>j)\leq(1-\epsilon)^{[\frac{j}{\tau}]}$$

where $[x]$ denotes an integer part of x. Let us continue in derivation

$$\begin{aligned}\mathbf{E}T^n &\leq \textstyle\sum_{j=0}^{\infty}((j+1)^n-(j)^n)(1-\epsilon)^{[\frac{j}{\tau}]} = \\ &= \textstyle\sum_{k=0}^{\infty}\sum_{j=k\tau}^{(k+1)\tau-1}(1-\epsilon)^k\Big((j+1)^n-(j)^n\Big) = \\ &= \textstyle\sum_{k=0}^{\infty}(1-\epsilon)^k\Big((k+1)^n\tau^n-k^n\tau^n\Big) = \\ &= \tau^n\textstyle\sum_{k=0}^{\infty}(1-\epsilon)^k\Big((k+1)^n-k^n\Big)\end{aligned}$$

The sum converges for every $\epsilon>0$ and $n\in\mathbb{N}$, which gives the existence of all moments. Random variable T_0 has negative values, but $-T_0$ fulfills all necessary so that we can apply derived expression. This gives the first inequality to be proved. Specifically for $n=1$ the lower bound of expected value is

$$\mathbf{E}(-T_0)\leq\tau\textstyle\sum_{k=0}^{\infty}(1-\epsilon)^k=\frac{\tau}{\epsilon}$$

Choosing $n=2$ gives the upper bound for second moment

$$\begin{aligned}\mathbf{E}T_0^2 &\leq \tau^2\textstyle\sum_{k=0}^{\infty}(1-\epsilon)^k(2k+1) = \\ &= \tau^2(\frac{1}{\epsilon}+2\textstyle\sum_{k=1}^{\infty}k(1-\epsilon)^k) = \\ &= \tau^2(\frac{1}{\epsilon}+2\textstyle\sum_{l=1}^{\infty}\sum_{k=l}^{\infty}(1-\epsilon)^k) = \\ &= \tau^2(\frac{1}{\epsilon}+2\textstyle\sum_{l=1}^{\infty}(1-\epsilon)^l\sum_{k=0}^{\infty}(1-\epsilon)^k) = \\ &= \tau^2(\frac{1}{\epsilon}+\frac{2}{\epsilon}\textstyle\sum_{l=1}^{\infty}(1-\epsilon)^l = \\ &= \tau^2(\frac{1}{\epsilon}+\frac{2(1-\epsilon)}{\epsilon^2})\end{aligned}$$

Since

$$P(T_0<-i)\geq(1-\epsilon),\text{for } i<\tau$$

and therefore

$$-\mathbf{E}T_0>\sum_{i=0}^{\tau-1}\bar{\mathbb{P}}(-T_0>i)\geq\tau(1-\epsilon)$$

the upper bound for expected value is

$$\mathbf{E}T_0<-\tau(1-\epsilon).$$

We can finalize upper estimate of variance

$$\mathbf{Var}\ T_0 = \mathbf{E}T_0^2 - \mathbf{E}(T_0)^2 \leq \tau^2 \left(\frac{1}{\epsilon} + \frac{2(1-\epsilon)}{\epsilon^2} - (1-\epsilon)^2 \right).$$

□

5 Partially ordered Markov kernel

Checking whether random map φ_k^0 coalesce all states into one is impossible in common application where MCMC methods are used. In some circumstances we can simplify this check to comparison of extremal values of state space S only.

Definition 5.1:

A *partial order* on a set S is a relation $s \preceq t$ between elements $s, t \in S$ with the two properties

- **reflexivity** - $s \preceq s\ \forall s \in S$
- **transitivity** - $r \preceq s$ and $s \preceq t$ implies $r \preceq t$

□

Note that total order meets additional requirement that every pair of elements $s, t \in S$ is comparable, ie. $s \preceq t$ or $t \preceq s$ for all $s, t \in S$. If there exist elements $\underline{m}, \overline{m} \in S$ such that $\underline{m} \preceq s \preceq \overline{m},\ \forall s \in S$ then $\underline{m}$ is called a minimum of S and $\overline{m}$ is its maximum.

Definition 5.2:

Let us consider partially ordered set of states $(S, \preceq)$ and a set of all maps $\Phi = S^S$. Than a map $\varphi \in \Phi$ is *order preserving* if it fulfills

$$x \preceq y \Rightarrow \varphi(x) \preceq \varphi(y) \quad \forall x, y \in S$$

□

Denote $\Gamma = \{\varphi \in \Phi : x \preceq y \Rightarrow \varphi(x) \preceq \varphi(y), \forall x, y \in S\}$ a set of all order preserving maps. We say that random map $\varphi = (\Phi, \mathbb{P})$ is almost sure order preserving if

$$\mathbb{P}(\Gamma) = 1. \tag{5}$$

Theorem 5.3:

Suppose that $\boldsymbol{P}$ is a kernel of irreducible Markov chain on a partially ordered set of states $(S, \preceq)$ such that there exist maximal and minimal states $\underline{m}, \overline{m} \in S$ and

$$s \preceq t \text{ and } t \preceq s \Leftrightarrow s = t, \quad \forall s, t \in S. \tag{6}$$

A space $(\Phi, \mathbb{P})$ of random maps fulfill Conditions (3) and (5). Let $(\Omega, \bar{\mathbb{P}})$ be a space of stochastic flows related to random maps Φ. (ie. $\Omega = \Phi^{\mathbb{Z}}$ and $\bar{\mathbb{P}}$ is a product measure $\mathbb{P}^{\mathbb{Z}}$) Then coalesce of $\underline{m}, \overline{m}$ enforces complete coalesce. □

Proof: Composition of order preserving maps is also order preserving map. Thus,

$$\mathbb{P}\left(\varphi_k^0(\underline{m}) \preceq \varphi_k^0(s) \preceq \varphi_k^0(\overline{m})\right) \geq \mathbb{P}\left(\varphi_i \in \Gamma, i \in \{k, \ldots, 0\}\right) = \prod_{i=k}^{0} P(\varphi_i \in \Gamma) = 1$$

for any negative $k \in \mathbb{Z}$ and $s \in S$. Suppose that k is sufficiently large so that there exists $\omega \in S$ such that $\varphi_k^0(\underline{m}) = \varphi_k^0(\overline{m}) = \omega$ (ie. extremes coalesced). Then the property (6) gives

$$\mathbb{P}(\varphi_k^0(s) = \omega, \forall s \in S) = \mathbb{P}(\omega \preceq \varphi_k^0(s) \preceq \omega) = 1$$

and the almost sure complete coalesce is assured. □

6 Conclusion

The paper deals with MCMC method called perfect sampling and covers many practical issues like random map representation and generation, complete coalesce preconditions and recognition, and also estimates of method's time complexity.

MCMC simulation are widely used for image restoration and suppression of image noise, in particular. Therefore we decided to test the method on Ising model sampling, since Ising model is the simplest one for Black&White image used in MCMC image restoration algorithms.

Black&White image is a set of intensities $\boldsymbol{y} = \{y_{(i,j)} \in \{-1, 1\} : \ (i,j) \in T\}$ on a rectangular grid of pixels $T = \{(i,j) : 1 \leq i \leq M, 1 \leq j \leq N\}$. Ising model is a probability distribution on a set of all images which is given by following formula

$$P(\boldsymbol{y}) = Z^{-1} exp\left(\beta \sum_{(i,j)\sim(k,l)} y_{(i,j)} y_{(k,l)}\right)$$

where $(i,j) \sim (k,l)$ denotes neighboring pixels (ie. pixels where $|i-k| = 1$ and $j = l$ or vice versa), $\beta \in \mathbb{R}$ is model parameter and Z is normalizing constant. Thus, concordant neighbors are preferred with positive value of parameter β, while discordant neighbors are preferred for negative value. Figure 1 displays examples of images picked from Ising model for specific choices of parameter β.

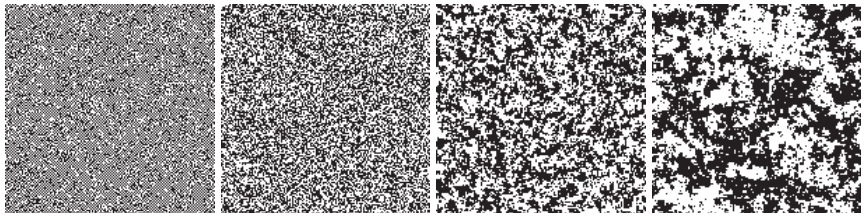


Figure 1: Ising model with varying values of parameter β (from left to right: -0,4; 0; 0,3 a 0,4.)

References

[1] Antoch, J., Hušková, M., Jarušková, D. (1998): Change point problem po deseti letech. ROBUST'98, 1-42, JČMF Praha, J. Antoch and G. Dohnal eds.

[2] Antoch, J., Hušková, M., (1999): Estimators of changes. Asymptotics, Nonparametrics and Time Series. Marcel Dekker, Basel, 533-577.

[3] Gilks, W. R., Richardson, S., Spiegelhalter, D. J. (1995): Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC, London.

[4] Hastings, W. K. (1970): Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97-109.

[5] Janžura, M., (1990): O jednom pravděpodobnostním algoritmu pro optimalizační metody. ROBUST'90, JČSMF Praha, J. Antoch and G. Dohnal eds.

[6] Legát, D. (2004): Metody MCMC. Diplomová práce MFF UK, Praha.

[7] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller,A.H., Teller, E. (1953): Equations of state calculations by fast computing machines. J. Chem. Phys., 21, 1087-1092.

[8] Pardoux, E. (2008): Markov Processes and Application. John Wiley & Sons.

[9] Prášková, Z., Lachout, P. (2005): Základy náhodných procesů. Karolinum, Praha.

[10] Robert, Ch. P., Casella, G. (2005): Monte Carlo Statistical Methods, Springer. Heidelberg.

[11] Volf, P., (1996): Bayesovský odhad parametrů modelu metodami MCMC s aplikací. ROBUST'94, 273-283, J(Č)MF Praha, J. Antoch and G. Dohnal eds.

INDEXY A CHÝBAJÚCE ÚDAJE V BATÉRII ORDINÁLNYCH PREMENNÝCH

INDICIES AND MISSING VALUES IN BATERY OF ORDINAL VARIABLES

Ján Luha

Adresa: ÚLBGKG LF UK a UN Bratislava; jan.luha@fmed.uniba.sk

Abstrakt: V príspevku prezentujeme konštrukciu indexov pre batériu ordinálnych premenných a riešenie problému chýbajúcich údajov na konkrétnom príklade výskumu verejnej mienky.

Klíčová slova: index batérie ordinálnych premenných, chýbajúce údaje, príklad výskumu verejnej mienky

Abstract: In this article we present construction of indexes for battery ordinal variables and splving problems with missing data on concrete example from public public opinion research.

Keywords: indexes for battery ordinal variables, missing data, public opinion research

1. Úvod

V príspevku sa zaoberáme konštrukciou simultánneho indexu batérie ordinálnych premenných s rovnakou škálou hodnôt a možnosťou riešenia chýbajúcich údajov. Riešenia prezentujeme na dátach konkrétneho výskumu verejnej mienky s láskavým dovolením koordinátora grantu za Slovenskú republiku prof. L. Macháčka z UCM Trnava. Na prezentáciu sme použili reálne, ale anonymizované dáta za SR z projektu: MYPLACE (Memory, Youth, Political Legacy And Civic Engagement) Grant agreement no: FP7-266831), podrobnejšie informácie o projekte sú na stránke: http://www.fp7-myplace.eu/index.php. Na výskume sa zúčastnili mladí ľudia vo veku 16 až 25 rokov z okresov Trnava a Rimavská Sobota. Údaje sme anonymizovali – slúžia na ilustráciu konštrukcie simultánneho indexu a riešenia problému chýbajúcich údajov.

2. Indexy pre ordinálne premenné

Teoretické aspekty konštrukcie indexov pre batériu ordinálnych premenných možno nájsť napríklad v prácach: [1] až [9].

Stručne definujeme konštrukciu indexov pre ordinálne premenné s rovnakou množinou hodnôt. Uvažujme ordinálny znak s r prvkovou množinou hodnôt. Škála odpovedí je kódovaná numericky kódmi: $1, 2, \ldots, r-1, r$.

Predpokladáme, že škála odpovedí je „usmernená“ tak, že najmenšia hodnota reprezentuje najmenšiu úroveň znaku a najväčšia zase najväčšiu úroveň

znaku. Rekódujeme škálu odpovedí tak, že najmenšia hodnota bude označená kódom 0 a najväčšia kódom $r-1$. Množina možných hodnôt znaku potom bude: $0, 1, \ldots, r-2, r-1$. Požadujeme aby index mal „normovanú“ množinu hodnôt – obyčajne od 0 po 1.

Index pre jednu otázku (premennú) vytvoríme jednoducho vydelením kódov maximálnou hodnotou, čo je po rekódovaní , hodnota $r-1$. Potom je množina možných hodnôt „individuálneho“ indexu:

$$0, 1/(r-1), 1/(r-1), ..., (r-2)/(r-1), 1 = (r-1)/(r-1).$$

Transformácie nad množinou hodnôt skúmaných premenných sú lineárne a jedno–jednoznačné, čím sa nemení ich rozdelenie pravdepodobnosti – transformujú sa iba charakteristiky polohy (napríklad priemer) a rozptýlenia (napríklad smerodajná odchýlka).

Z uvedeného vyplýva, že predpokladáme ekvidistantné škály odpovedí, naviac neuvažujeme diferencované váhy skúmaných premenných.

3. **Indexy pre ordinálne premenné – príklady**

Skúmajme batériu 6 otázok venovanú v hore uvedenom výskume verejnej mienky názorom mladých ľudí vo veku 16 až 25 rokov z okresov Trnava a Rimavská Sobota na otázku „Ako často diskutuješ politické otázky s nasledujúcimi ľuďmi?“: s otcom, mamou, súrodencom, starými rodičmi, priateľmi-partnermi a najlepším priateľom/priateľkou. Pôvodnú škálu odpovedí: 1=vždy, 2=často, 3=občas, 4=zriedka, 5=nikdy sme rekódovali priamo do škály od 0 po 1 postupom uvedeným v predošlej kapitole. Keďže popisky (labele) škály odpovedí v SPSS neumožňujú kódovanie s desatinnými miestami budú pri výsledkoch prezentované iba 0 najmenšia hodnota indexu a 1 najväčšia hodnota indexu. Nové hodnoty pôvodných kódov sú: 0=nikdy, 0.25=zriedka, 0.50=občas, 0.75=často a 1=vždy. Výsledky základnej štatistickej analýzy individuálnych indexov môžeme prezentovať pomocou frekvenčných tabuliek ale tiež oveľa prehľadnejšie pomocou priemerov aj v prípadoch, keď indexy nemajú normálne rozdelenie. Výsledky za frekvenčné tabuľky prezentujeme v tabuľke 1. a prehľadnejšie výsledky v tabuľke 2.

Tabuľka 1. Ako často diskutuješ politické otázky s nasledujúcimi ľuďmi?

Q4_1: S Tvojím otcom					
		Frequency	%	Valid %	Cumul. %
Valid	.00 Nikdy	309	25.8	27.7	27.7
	.25	339	28.3	30.4	58.2
	.50	310	25.8	27.8	86.0
	.75	141	11.8	12.7	98.7
	1.00 Vždy	15	1.3	1.3	100.0
	Total	1114	92.8	100.0	
Missing	System	86	7.2		
Total		1200	100.0		

Q4_2:	S Tvojou mamou				
		Frequency	%	Valid %	Cumul. %
Valid	.00 Nikdy	373	31.1	31.8	31.8
	.25	395	32.9	33.6	65.4
	.50	299	24.9	25.5	90.9
	.75	89	7.4	7.6	98.5
	1.00 Vždy	18	1.5	1.5	100.0
	Total	1174	97.8	100.0	
Missing	System	26	2.2		
Total		1200	100.0		

Q4_3:	S bratom alebo sestrou, čo je Ti najbližší/a				
		Frequency	%	Valid %	Cumul. %
Valid	.00 Nikdy	453	37.8	47.0	47.0
	.25	278	23.2	28.9	75.9
	.50	173	14.4	18.0	93.9
	.75	51	4.3	5.3	99.2
	1.00 Vždy	8	0.7	0.8	100.0
	Total	963	80.3	100.0	
Missing	System	237	19.8		
Total		1200	100.0		

Q4_4:	So starými rodičmi, ktorí sú Ti bližší				
		Frequency	%	Valid %	Cumul. %
Valid	.00 Nikdy	373	31.1	37.3	37.3
	.25	303	25.3	30.3	67.7
	.50	235	19.6	23.5	91.2
	.75	68	5.7	6.8	98.0
	1.00 Vždy	20	1.7	2.0	100.0
	Total	999	83.3	100.0	
Missing	System	201	16.8		
Total		1200	100.0		

Q4_5:	S priateľkou, priateľom, partnero				
		Frequency	%	Valid %	Cumul. %
Valid	.00 Nikdy	317	26.4	38.6	38.6
	.25	238	19.8	29.0	67.5
	.50	185	15.4	22.5	90.0
	.75	71	5.9	8.6	98.7
	1.00 Vždy	11	0.9	1.3	100.0
	Total	822	68.5	100.0	
Missing	System	378	31.5		
Total		1200	100.0		

Q4_6: S Tvojim najlepším priateľom/kou					
		Frequency	%	Valid %	Cumul. %
Valid	.00 Nikdy	431	35.9	39.0	39.0
	.25	348	29.0	31.5	70.6
	.50	255	21.3	23.1	93.7
	.75	54	4.5	4.9	98.6
	1.00 Vždy	16	1.3	1.4	100.0
	Total	1104	92.0	100.0	
Missing	System	96	8.0		
Total		1200	100.0		

Tabuľka 2. Základné štatistické charakteristiky individuálnych indexov

Index	N	Minimum	Maximum	Mean	Std.Dev.
Q4_1	1114	0.00	1.00	0.3236	0.2619
Q4_2	1174	0.00	1.00	0.2836	0.2495
Q4_3	963	0.00	1.00	0.2100	0.2387
Q4_4	999	0.00	1.00	0.2645	0.2572
Q4_5	822	0.00	1.00	0.2631	0.2592
Q4_6	1104	0.00	1.00	0.2455	0.2434
N (listwise)	661				

Pri reálnych výskumoch sa žiaľ často stáva, že respondent neodpovie na všetky otázky skúmanej batérie a je potrebné, okrem konštrukcie simultánneho indexu, riešiť aj problém chýbajúcich údajov. Z výsledkov v tabuľkách 1. a 2. vidno, že respondenti pri niektorých otázkach je podiel chýbajúcich údajov značný. Ak by sme pri mnohorozmerných analýzach vylúčili chýbajúce údaje „listwise“ – čiže, ak sa v danom zázname za batériu otázok nachádza chýbajúca hodnota – vylúčime celý záznam, zostalo by nám na analýzu iba 55.08% údajov. Podľa toho ako sú údaje „zviazané“ existujú možnosti náhrady (imputácie) chýbajúcich údajov. Jednou z úlohou štatistiky je „racionálne“ využiť čo možno najviac informácií v dátach obsiahnutých. Pri skúmaní batérie otázok s rovnakou, ordinálnou škálou odpovedí môžeme na imputáciu chýbajúcich údajov využiť konštrukciu simultánneho indexu.

4. **Simultánny index pre batérie ordinálnych premenných**

Skúmame batériu m otázok s rovnakou škálou odpovedí, ktoré charakterizujú určitú záujmovú oblasť. Pri všetkých otázkach je škála odpovedí zhodne usmernená a škály sú definované ako indexy s „normovanou“ množinou hodnôt od 0 po 1.

V citovaných prácach autora sú podrobne skúmané problémy spojené s tým, že sa snažíme viacrozmerný problém redukovať na jednorozmernú

charakteristiku. Zároveň sú v týchto prácach skúmané množiny možných výsledkov.

Simultánny index môžeme jednoducho definovať ako priemer m indexov danej batérie otázok. Keďže v menovateli individuálnych indexov je rovnaká hodnota je táto definícia je korektná. Množina možných hodnôt simultánneho indexu je:

$$0, 1/(m(r-1)), 2/(m(r-1)), \ldots (m(r-2))/(m(r-1)), 1$$

Aby simultánny index dobre meral simultánnu úroveň batérie otázok musí byť reliabilita tejto batérie otázok meraná Cronbachovým alfa aspoň 0,7 a ich korelačná matica musí byť nezáporná.

Tieto podmienky sú empiricky zistené zo skúmaní položkových anakýz a relaibilyty simultánnych škál.

Tieto podmienky zlepšia možnosť vyjadriť viacrozmernú charakteristiku pomocou jednorozmerného indexu a minimalizujú „podivné“ kombinácie (konfigurácie), ktoré majú rovnakú hodnotu simultánneho indexu. Najznámejšie zo simultánnych hodnotení je priemerná známka žiaka za „batériu“ predmetov. Konštrukciu simultánneho indexu a riešenie problému chýbajúcich údajov ilustrujeme na už spomínanom konkrétnom výskume verejnej mienky.

Aby sme preskúmali možnosti transformácie mnohorozmerného problému na jednorozmerný definujeme množiny možných „výsledkov“ najprv za všetky možné konfigurácie, ktoré by mohli nastať, ich redukciu až po množinu možných hodnôt simultánneho indexu. Označme vektor hodnôt individuálnych indexov $h = (h_1 = 0 = 0/(r-1),\ h_2 = (1/(r-1),\ \ldots,\ h_r = 1$. Pre každý objekt o zo skúmanej množiny $O = o_1, o_2, ..., o_N$ definujeme veličiny:

$$z_{ij} = \begin{cases} 1, & \text{ak} Z_i(o) = h_{ij}, \\ 0, & \text{inak.} \end{cases}$$

Pre každý objekt $o \in O$ môžeme zostaviť tabuľku:

Tabuľka 3. Matica hodnôt respondenta

Znak	Hodnota						Spolu
	1	2	...	j	...	r	
Z_1	z_{11}	z_{12}	...	z_{1j}	...	z_{1r}	1
Z_2	z_{21}	z_{22}	...	z_{2j}	...	z_{2r}	1
⋮	⋮	⋮	...	⋮	...	⋮	⋮
Z_i	z_{i1}	z_{i2}	...	z_{ij}	...	z_{ir}	1
⋮	⋮	⋮	...	⋮	...	⋮	⋮
Z_m	z_{m1}	z_{m2}	...	z_{mj}	...	z_{mr}	1
Spolu	x_1	x_1	...	x_1	...	x_1	m

Zo základných vlastností vyjadrenia ordinálnych znakov v uvedenom tvare zrejme platia vzťahy: $zJ'_r = J_m, J'_m z = x$ a $xJ'_r = m$, kde $z = (z_{ij})$ je $m \times r$ matica núl a jednotiek určujúca výsledky daného objektu a J_m je $m \times 1$ vektor jednotiek a J_r je $r \times 1$ vektor jednotiek. Potom

$$Z = \{z : z_{ij} = 0 \text{ alebo } 1, \quad zJ'_r = J_m, \quad J'_m z J_r = m\}$$

je množina všetkých matíc z, ktoré môžu reprezentovať výsledky dosiahnuté pre objekt $o \in O$. Čiarka nad vektorom (maticou) označuje jeho transpozíciu.

Bude užitočné definovať ďalšiu množinu, ktorá charakterizuje istú triedu všetkých možných výsledkov dosiahnutých objektmi pomocou veličín $x_j,\ j = 1, \ldots, r$:

$$X = \{(x_1, x_2, \ldots, x_r) : x' J_r = m, \quad 0 \Leftarrow x_j \Leftarrow m,\ x_j \text{celé číslo}\}.$$

Všeobecnejšie definujeme simultánny index vzťahom:

$$I(z) = \frac{J'_m.z.h}{m.(r-1)} = \frac{x.h'}{(m.(r-1))} = I(x),$$

pre $z \in Z$, resp. $x \in X$. Množinu možných hodnôt indexu $I(z)$, resp. $I(x)$ označme II.

$$II = \{I : I = I(z); \quad z \in Z\} = \{I : I = I(x), \quad x \in X\}.$$

a definujme množiny: $Z(x) = \{z \in Z : z' J_m = x\}$ pre $x \in X$. Zrejme platí $Z = \bigcup_{x \in X} Z(x)$, kde symbolom $\bigcup$ označujeme zjednotenie množín. Pre kardinalitu výsledkových množín platia vzťahy (pozri napr. [1], [6]):

(a) $\text{Kard}(Z) = r^m$,
(b) $\text{Kard}(X) = (m + r - 1)!/[m!(r-1)!]$,
(c) $\text{Kard}(Z(x)) = m!/[x_1!x_2!...x_r!]$,
(d) $\text{Kard}(II) = m.(r-1) + 1$.

Z definície indexu platí: $I(z) = I(x)$ pre $z \in Z(x)$.

Množinu X vieme jednoducho generovať, napr. pomocou aritmografického usporiadania. Potom máme

$$X = \{(m, 0, \ldots, 0), (m-1, 1, 0, \ldots, 0), \ldots, (x_1, x_2, \ldots, x_r), \ldots, (0, \ldots, 0, m)\}.$$

5. Simultánny index pre batérie ordinálnych premenných – príklad

Simultánne indexy môžeme počítať tromi spôsobmi:

1. priemery za všetky získané odpovede pomocou SPSS príkazu:

```
COMPUTE In_Q4 = mean(Q4_1_i to Q4_6_i)}.
```

2. priemery za odpovede respondentov, kde nechýba ani jeden údaj (v SPSS ide o vylúčenie missing value „listwise“)

```
COMPUTE Ix_Q4=(Q4_1_i+Q4_2_i+Q4_3_i+Q4_4_i+Q4_5_i+Q4_6_i)/6.
```

3. Na odhad simultánneho indexu použijeme `In_Q4` za odpovede, kde respondent odpovedal aspoň na polovicu otázok batérie.

Množina možných hodnôt simultánneho indexu batérie Q4 je:

$$0, 1/24, 2/24, ..., 23/24, 1 = 24/24.$$

Pre kardinalitu výsledkových množín ilustračného príkladu $m = 6,\ r = 5$ platí:

(a) $\text{Kard}(Z) = 5^6 = 15625$,

(b) $\text{Kard}(X) = (10)!/[6! * 4)!] = 210$,

(c) $\text{Kard}(Z(x)) = 6!/[x_1!x_2! \ldots x_r!]$, pre $x \in X$, čo je 210 hodnôt, pre $x = (6, 0, 0, 0, 0, 0), (5, 1, 0, 0, 0, 0), \ldots (0, 0, 0, 0, 0, 6)$, napríklad:

x_1	x_2	x_3	x_4	x_5	Kard$(Z(x))$	$I(x)$
6	0	0	0	0	1	0
5	1	0	0	0	6	0.041667
4	2	0	0	0	15	0.08333
3	3	0	0	0	20	0.125
2	4	0	0	0	15	0.1667
1	5	0	0	0	6	0.20833
0	6	0	0	0	1	0.25
5	0	1	0	0	6	0.08333
4	1	1	0	0	30	0.125

(d) $\text{Kard}(II) = 6 * 4 + 1 = 25$.

Tabuľka 2. Základné štatistické charakteristiky individuálnych indexov

Index	N	Minimum	Maximum	Mean	Std.Dev.
`In_Q4`	1196	0.00	1.00	0.2671	0.2068
`Ix_Q4`	661	0.00	0.96	0.2728	0.2056
`In_Q4_miss`	1148	0.00	1.00	0.2647	0.2054
Valid N (listwise)	661				

Prvý spôsob konštrukcie simultánneho indexu využíva všetky údaje, keď respondent uviedol aspoň jednu odpoveď na 6 otázok skúmanej batérie. Ak napríklad respondent odpovedal iba na jednu otázku z batérie 6 otázok je odhad simultánneho indexu práve touto hodnotou zrejme riskantný – potenciálna odchýlka od skutočnej hodnoty môže byť veľká. Pri druhom výpočte indexu `Ix_Q4` strácame príliš veľa hodnôt. Ak respondent odpovedal aspoň na polovicu z batérie otázok, čo v našom prípade značí 4 a viac odpovedí respondenta, tak využijeme podstatne viac informácií. Ako ukážeme v ďalších kapitolách pri dodržaní podmienky vysokej reliability a nezápornej korelačnej matici individuálnych indexov je riziko veľkej odchýlky pri tomto spôsobe odhadu simultánneho indexu minimalizované.

6. Riešenie problému chýbajúcich údajov

Najprv vypočítame `COMPUTE In_Q4_miss=In_Q4`, čo je vlastne prvý spôsob výpočtu simultánneho indexu. V tomto indexe budeme riešiť problém chýbajúcich (missing) hodnôt. Aby sme zistili počet chýbajúcich odpovedí vypočítame pomocnú premennú, pomocou ktorej vylúčime odpovede, kde viac ako polovica sú chýbajúce odpovede respondenta.

SPSS príkazy musia najprv redefinovať missingy, vypočítať pomocnú premennú a zase definovať missingy:

```
RECODE
Q4_1_i Q4_2_i Q4_3_i Q4_4_i Q4_5_i Q4_6_i  (SYSMIS=99)
EXECUTE
```

Potom našu pomocnú premennú počítame pomocou príkazu:

```
COMPUTE a_Q4 = SUM(Q4_1_i  to Q4_6_i)/99
EXECUTE
```

Aby sme zachovali chýbajúce hodnoty musíme ešte realizovať príkaz:

```
RECODE
Q4_1_i Q4_2_i Q4_3_i Q4_4_i Q4_5_i Q4_6_i  (99=SYSMIS)
EXECUTE
```

Pomocná premenná `a_Q4` približne vypočíta počet chýbajúcich odpovedí respondentov v batérii otázok (po zaokrúhlení dostaneme presný počet missingov). Potom v indexe `In_Q4_miss` vylúčime hodnoty indexu, kde je počet chýbajúcich údajov 4 a viac – buď mechanicky po sortovaní pomocnej premennej, alebo nájdeme potrebný SPSS príkaz.

Teraz využijeme hodnoty `In_Q4_miss` na imputáciu príslušných chýbajúcich údajov – čiže chýbajúce hodnoty pre záznamy s 3 a viac platnými hodnotami nahradíme missingy hodnotou `In_Q4_miss`.

Poznámka: Priemer indexov danej batérie po imputácii je rovnaký ako pôvodný index `In_Q4_miss`.

7. Reliabilita a korelačná matica pred a po imputácii chýbajúcich údajov

Imputácia chýbajúcich údajov spôsobuje zväčšenie množiny možných hodnôt individuálnych indexov. Reliabilita a korelačnej matice z pôvodných dát a po imputácii chýbajúcich hodnôt spĺňajú podmienky, pri splnení, ktorých môžeme skonštruovať simultánny index batérie otázok.

Ako vidno aj z poznámky v tabuľke 5, procedúra výpočtu reliability a korelačnej matice vylučuje chýbajúce údaje „listwise“ a tak sú tieto štatistiky počítané iba z 55,1% záznamov. Napriek tomu je reliabilita šestice otázok pre simultánny index vysoká až 0.894 a korelačné koeficienty v tabuľke 6 sú všetky kladné.

Tabuľka 5. Reliabilita batérie 6-tich otázok Q4 pred imputáciou

Case Processing Summary

Cases	N	%
Valid	661	55.1
Excluded*	539	44.9
Total	1200	100.0

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standard.Items	Number of Items
0.894	0.894	6

* Listwise deletion based on all variables in the procedure

Tabuľka 6. Korelačná matica batérie 6-tich otázok Q4 pred imputáciou

Index	Q4_1	Q4_2	Q4_3	Q4_4	Q4_5	Q4_6
Q4_1	1.000	0.723	0.618	0.556	0.555	0.561
Q4_2	0.723	1.000	0.643	0.575	0.584	0.590
Q4_3	0.618	0.643	1.000	0.513	0.618	0.566
Q4_4	0.556	0.575	0.513	1.000	0.484	0.498
Q4_5	0.555	0.584	0.618	0.484	1.000	0.698
Q4_6	0.561	0.590	0.566	0.498	0.698	1.000

Výsledky výpočtov z dát po imputácii chýbajúcich údajov sú v tabuľkách 7. a 8. Podstatné je, že sme pri týchto výpočtoch mohli využiť až 95,7% záznamov. Reliabilita sa mierne zvýšila na hodnotu 0.910 a korelačné koeficienty zostali kladné a všetky sa mierne zvýšili.

Tabuľka 7. Reliabilita batérie 6-tich otázok Q4 po imputácii

Case Processing Summary

Cases	N	%
Valid	1148	95.7
Excluded*	52	4.3
Total	1200	100.0

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standard.Items	Number of Items
0.910	0.910	6

* Listwise deletion based on all variables in the procedure

Tabuľka 8. Korelačná matica batérie 6-tich otázok Q4 po imputácii

Index	Q4_1	Q4_2	Q4_3	Q4_4	Q4_5	Q4_6
Q4_1	1.000	0.739	0.623	0.605	0.625	0.569
Q4_2	0.739	1.000	0.667	0.624	0.660	0.599
Q4_3	0.623	0.667	1.000	0.575	0.680	0.598
Q4_4	0.605	0.624	0.575	1.000	0.578	0.546
Q4_5	0.625	0.660	0.680	0.578	1.000	0.732
Q4_6	0.569	0.599	0.598	0.546	0.732	1.000

Aj pomocou kontingenčných tabuliek môžeme vyjadriť závislosť medzi individuálnymi indexmi. Nebudeme uvádzať všetkých 15 kontingenčných tabuliek skúmanej batérie otázok, ktoré vykazujú taktiež silnú závislosť. Na doplnenie uvádzame maticu Cramerovho V kontingenčného koeficianta v tabuľke 9.

Množina možných hodnôt individuálnych indexov sa po imputácii značne zväčší, pretože imputáciou doplníme chýbajúce údaje nahradením priemermi, ktoré nadobúdajú aj iné hodnoty ako pôvodné individuálne indexy. V tomto prípade možno považovať individuálne premenné za numerické a počítať korelačné koeficienty.

Tabuľka 9. Cramerovo V batérie 6-tich otázok Q4 pred imputáciou

Index	Q4_1	Q4_2	Q4_3	Q4_4	Q4_5	Q4_6
Q4_1	1.000	0.548	0.359	0.340	0.358	0.323
Q4_2		1.000	0.378	0.361	0.373	0.351
Q4_3			1.000	0.319	0.405	0.342
Q4_4				1.000	0.287	0.309
Q4_5					1.000	0.521
Q4_6						1.000

8. **Analýza odchýlok**

Pri imputácii chýbajúcich údajov je riziko, že nemusíme odhadnúť správnu odpoveď respondenta, ak by ju uviedol. Hodnota simultánneho indexu sa môže pohybovať od 0 po 1, pričom „skoky“ „susedných“ hodnôt závisia aj od počtu premenných v skúmanej batérii.

Napríklad pri počte 6 otázok v batérii, ako je to v ilustračnom príklade sú potenciálne odchýlky pri 5 chýbajúcich odpovediach respondenta od 0 (vo vzácnom prípade, ak „smieme“ predpokladať, že všetky chýbajúce odpovede budú rovnaké, ako tá jedna, čo dal respondent).

Pri 5-tich chýbajúcich odpovediach je maximálna odchýlka v absolútnej hodnote 0,833, v prípade 4 chýbajúcich údajov, čiže máme dve platné odpovede daného respondenta je „riziko“ veľkej odchýlky taktiež vysoké. Nulová odchýlka je taktiež vo vzácnych prípadoch, ak by chýbajúce odpovede „dávali rovnaký priemer ako platné odpovede“, maximálna odchýlka v absolútnej hodnote je v tomto prípade 0,667.

Pri 3 chýbajúcich údajoch a teda aj troch platných odpovediach respondenta je zase minimálna odchýlka 0 a maximálna 0,5.

Presnú analýzu odchýlok nepoznáme, pokladáme za dôležitú otázku ako využiť čo možno najviac údajov na odhad simultánneho indexu. Odvážnejší môžu za „dobrý“ odhad simultánneho indexu považovať odhad, ktorý využije polovicu a viac platných odpovedí respondenta na skúmanú batériu otázok

„Riziko“ veľkých odchýlok klesá samozrejme s počtom platných odpovedí respondenta. Pri dvoch chýbajúcich odpovediach máme 4 platné odpovede respondenta a odchýlky môžu byť v absolútnej hodnote od 0 po 0,333. A nakoniec pri jednej chýbajúcej hodnote je 5 platných odpovedí a odchýlky od 0 po 0,167.

V kapitole 5. sú uvedené kardinality výsledkových množín skúmaného simultánneho indexu batérie , množina všetkých možných konfiguráciu má mohutnosť 15625, tá sa redukuje na menšiu množinu s mohutnosťou 210 a nakoniec je iba 25 rôznych hodnôt indexu, v situácii keď nenahrádzame chýbajúce údaje. Ak by sme skúmali konfigurácie vrátane chýbajúcich údajov, tak je teoreticky možných až 46656 =6*6*6*6*6*6 konfigurácií. V reálnych situáciách pracujeme s výsledkami kde je oveľa menej záznamov. V ilustračnom príklade pracujeme máme 1200 záznamov – konfigurácií, niektoré z nich sa vyskytujú viackrát, takže počet rôznych konfigurácií je reálne menší. V tabuľke 10. je ukážka niekoľko konfigurácií individuálnych indexov vrátane hodnôt simultánnych indexov a početnosť konfigurácií, vrátane chýbajúcich údajov, ktoré potrebujeme imputovať.

V ukážke sme vypočítali vo vyznačených záznamoch `In_Q4_miss`, pomocou označených hodnôt tohto indexu nahradíme v príslušných záznamoch chýbajúce hodnoty.

Tabuľka 10. Ukážka konfigurácií vrátane chýbajúcich údajov

por. č.	Q4_1	Q4_2	Q4_3	Q4_4	Q4_5	Q4_6	In_Q4	Ix_Q4	In_Q4 _miss	n
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	79
2	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	32
3	0.00	0.00	0.00	0.00		0.00	0.00		0.00	31
4	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	27
5	0.00	0.00				0.00	0.00		0.00	13
6	0.00	0.00	0.00	0.25	0.00	0.00	0.04	0.04	0.04	13
7	0.25	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.04	11
8	0.25	0.25	0.00	0.00	0.00	0.00	0.08	0.08	0.08	11
9	0.00	0.00		0.00	0.00	0.00	0.00		0.00	10
10	0.00	0.00	0.00	0.00	0.25	0.25	0.08	0.08	0.08	10
11	0.25	0.25	0.00	0.25	0.00	0.00	0.13	0.13	0.13	10
12	0.00	0.00		0.00		0.00	0.00		0.00	9
13	0.25	0.00	0.00	0.00		0.00	0.05		0.05	9
14	0.00	0.00	0.00	0.00	0.00	0.25	0.04	0.04	0.04	8
15	0.25	0.00	0.00	0.25	0.00	0.00	0.08	0.08	0.08	8
16	0.00	0.00					0.00			6
17	0.00	0.00	0.00				0.00		0.00	6

Podmienky – vysoká reliabilita, Cronbachovo alfa aspoň 0,7 a to, že korelačná matica má iba kladné korelačné koeficienty – minimalizujú možnosť „divokých“ kombinácií (konfigurácií) aj pri chýbajúcich údajoch a tak môžeme použiť navrhovaný postup odhadu simultánneho indexu a imputácie chýbajúcich údajov.

Ak by v riadku záznamu č. 3 bola hodnota `Q4_6`=1, čo je najnepriaznivejšia možnosť – bola by odchýlka nášho odhadu od tejto možnosti 0.166667,

ale vzhľadom na uvedené podmienky je táto možnosť málo očakávaná. Pri odhade v riadku 5 a 12 je teoreticky najväčšia odchýlka ak by skutočné hodnoty namiesto chýbajúcich boli rovné 1 a odchýlka by bola až 0,5.

9. Závery

Za predpokladov o reliabilite a kladných koreláciách medzi premennými batérie možno skonštruovať simultánnu charakteristiku, ktorá dobre charakterizuje celú batériu premenných. Tým sa analýzy zjednodušia, pretože môžeme využiť štatistické metódy analýzy numerických premenných.

Literatúra

[1] Luha J.(1989) *Simultánne charakteristiky skupiny ordinálnych znakov*. Kandidátska dizertačná práca. MFF UK Bratislava 1989.

[2] Luha J., Kevická R. (1990) *Unifikácia indexov pomocou pravdepodobnostných modelov*. SOCIOLÓGIA č.1, 1990.

[3] Luha J. (1991) *Index for Ordinal Variables*. PROBASTAT91, Bratislava 1991.

[4] Luha J. (1996) *Indexy pre ordinálne znaky*. Zborník príspevkov Finančno-ekonomické analýzy, EKOMSTAT´96 3. 6. – 7. 6. 1996 Trenčianske Teplice, SŠDS.

[5] Luha J. (2003) *Meranie úrovne objektov charakterizovaných ordinálnymi znakmi*. Forum metricum Slovacum, Tom VII. SŠDS Bratislava 2003. ISBN 80-88946-30-1.

[6] Luha J. (2004) *Meranie úrovne objektov charakterizovaných ordinálnymi znakmi*. EKOMSTAT 2004, Štatistické metódy v praxi. SŠDS Trenčianske Teplice 2004.

[7] Luha J. (2004) *Niektoré otázky záverečného hodnotenia štátnych zamestnancov 1*, STATIS 1 / 2004.

[8] Luha J. (2004) *Niektoré otázky záverečného hodnotenia štátnych zamestnancov 2*, STATIS 1 / 2004.

[9] Luha J. (2006) *Meranie úrovne súboru objektov*. FORUM STATISTICUM SLOVACUM 4/2006. SŠDS Bratislava 2006. ISSN 1336-7420.

[10] Kubanová, J. (23008) *Statistické metody pro ekonomickou a technickou praxi*. STATIS, Bratislava 2008. Vydání třetí – doplněné. ISBN 978-80-85659-47-4.

[11] Linda, B. (2010) *Pravděpodobnost*. Univerzita Pardubice, Pardubice 2010. ISBN 978-80-7395-303-4. pp 168.

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

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

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

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

[16] Řezanková H.(2007) *Analýza dat z dotazníkových šetření*. Proffessional Publishing, Praha 2007. ISBN 978-80-86946-49-8.

[17] Stankovičová I., Vojtková M. (2007) *Viacrozmerné štatistické metódy s aplikáciami*. IURA EDITION, Bratislava 2007, ISBN 978-80-8078-152-1.

[18] Internetová stránka projektu MYPLACE: http://www.fp7-myplace.eu/index.php.

NÁVRH METODIKY PRO ANALÝZU STATISTICKÉ NESTABILITY PROCESU S VYUŽITÍM DOSTUPNÝCH STATISTICKÝCH PROGRAMŮ

DESIGN OF METHODOLOGY FOR ANALYSIS OF PROCESS STATISTICAL INSTABILITY USING AVAILABLE STATISTICAL SOFTWARE

Darja Noskievičová

Adresa: Katedra kontroly a řízení jakosti, Fakulta metalurgie a materiálového inženýrství, VŠB-TU Ostrava, 17. listopadu 15, 708 33 Ostrava – Poruba; darja.noskievicova@vsb.cz

Abstrakt: Hlavním výstupem tohoto článku je návrh metodiky pro analýzu statistické nestability procesu s využitím dostupných statistických programů. Návrh je postaven na předchozí analýze různých souborů pravidel pro testování nenáhodných seskupení a komplexním rozboru vybraných statistických programů obsahujících testy nenáhodných seskupení.

Abstract: The main output of this paper is the design of the methodology for the analysis of the process statistical instability using available statistical software. The design is based on the previous analysis of the different sets of the rules for the identification of the nonrandom patterns (run tests) and on the complex analysis of the selected statistical software containing these tests.

Klíčová slova: Statistická regulace procesu, vymezitelné příčiny, variabilita procesu, testy nenáhodných seskupení, počítačová podpora analýzy nenáhodných seskupení

Keywords: Statistical Process Control, Assignable Causes, Process Variability, Run Tests, Computer Aided Analysis of Nonrandom Patterns

1. Úvod

Statistická regulace procesů (SPC) představuje v praxi široce využívaný přístup k řízení výrobních i nevýrobních procesů. Je to především nástroj pro pochopení a analýzu variability procesů [1]. SPC je založeno na tzv. Shewhartově koncepci variability procesů, která rozlišuje mezi variabilitou způsobenou obvykle působícími běžnými příčinami (proces je považován za statisticky stabilní) a variabilitou způsobenou neobvyklými (speciálními, vymezitelnými) příčinami (proces není považován za statisticky stabilní). K rozlišení působení těchto dvou typů příčin variability se používá regulační diagram. Body zaznamenávané do regulačního diagramu jsou uspořádány náhodně (přirozeně) či tvoří nenáhodná (nepřirozená) seskupení. Regulační diagram je konstruován

tak, aby umožnil sledovat chování procesu v čase a odhalit jakékoliv nepřirozené seskupení bodů [2], považované za symptom působení neobvyklé příčiny a tedy za signál k analýze procesu s cílem stanovit konkrétní příčinu nestability procesu, aby bylo dále možné stanovit vhodné opatření ke stabilizaci a případnému zlepšení procesu. Základním kritériem pro rozhodnutí o statistické stabilitě či nestabilitě procesu jsou regulační meze. Jsou stanoveny tak, že je velmi malá pravděpodobnost výskytu bodu mimo regulační meze (při standardně používaných $\pm 3\sigma$ mezích a za předpokladu normálního rozdělení regulované veličiny je tato pravděpodobnost 0.0027). Regulační diagram se pak interpretuje následovně: pokud leží body uvnitř regulačních mezí a jsou rozloženy náhodně, lze považovat proces za statisticky stabilní. Pokud se nějaký bod či vice bodů vyskytuje mimo regulační meze nebo body uvnitř mezí vytvářejí nenáhodné seskupení, nelze proces považovat za statisticky stabilní a je nutné vyhledat působící neobvyklou (vymezitelnou) příčinu a její působení omezit či plně odstranit.

2. **Definice náhodných a nenáhodných seskupení**

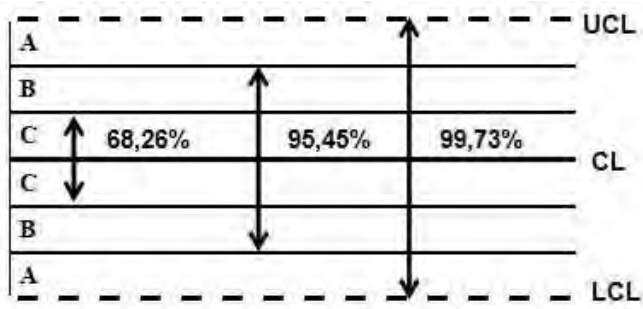
Přirozené seskupení bodů je takové seskupení, kde body uvnitř regulačních mezí jsou rozmístěny náhodně.Takové uspořádání bodů lze charakterizovat následovně [3]: body náhodně kolísají; většina bodů leží blízko střední přímky; pouze malý podíl se vyskytuje dále od střední přímky; velmi zřídka leží bod mimo regulační meze. Je-li kterýkoliv z těchto atributů porušen, je seskupení bodů klasifikováno jako nenáhodné seskupení. V regulačních diagramech se může vyskytovat mnoho takových nenáhodných seskupení, specifických pro různé procesy. Jejich rozpoznávání může být velmi obtížené. Proto bylo identifikováno několik generických nenáhodných seskupení, která lze aplikovat na většinu procesů. Nejčastěji uváděná nenáhodná seskupení (viz např. [3], [4], [5], [6], [7]) s jejich typickými projevy v Shewhartových regulačních diagramech jsou shrnuta v Tabulce 1. I přes existenci těchto generických nenáhodných seskupení může být jejich rozpoznávání komplikované. Pro standardizaci a zjednodušení tohoto procesu byly proto formulovány tzv. testy nenáhodných seskupení (run tests). Jde o pravidla, založená na kvantifikaci generických nenáhodných seskupení.

3. **Testy nenáhodných seskupení**

Testy nenáhodných seskupení doplňují důkaz o existenci nenáhodného seskupení, daný pozicí bodů vůči regulačním mezím a střední přímce, důkazem založeným na statistické teorii tzv. runs (řad, sekvencí, běhů) [8]. Testy jsou založeny na výpočtu pravděpodobnosti výskytu bodů blízko střední přímky, blízko regulačních mezí, atd. Pravidla jsou založena na rozdělení pásma mezi regulačními mezemi na 2 x 3 stejně široké zóny A, B, C, odpovídající svojí šíří 1σ při předpokladu normálního rozdělení regulované veličiny (viz Obrázek 1).

Tabulka 1. Popis nenáhodných seskupení

Č.	***Nenáhodné seskupení***	***Popis seskupení***	***Symptom v regulačním diagramu***
1	***Velké posuny (strays, freaks)***	Náhlá velká změna	Body blízko regulačních mezí nebo mimo ně
2	***Menší postupný posun***	Postupná menší změna	Řada bodů na jedné straně od střední přímky
3	***Trend***	Postupná změna v jednom směru	Postupně rostoucí nebo klesající řada bodů
4	***Stratifikace***	Malé rozdíly mezi hodnotami v dlouhé řadě bodů, absence bodů blízko regulačních mezí	Dlouhá řada bodů blízko střední přímky (na obou stranách)
5	***Mix***	Efekt "Saw-tooth", absence bodů blízko střední přímky	Řada po sobě následujících bodů na obou stranách střední přímky, všechny daleko od ní
6	***Systematická variabilita***	Řada hodnot pravidelně alternujících nahoru a dolů	Dlouhá sekvence bodů alternujících nahoru a dolů
7	***Cyklus***	Periodicky se opakující hodnoty	Cyklicky se opakující sekvence bodů



Obrázek 1. Rozdělení pásma mezi regulačními mezemi

V Tabulce 2 jsou uvedena nejčastěji používaná pravidla testování nenáhodných seskupení. (Za lomítkem ve sloupci s číslem pravidla je uveden typ nenáhodného seskupení, se kterým je pravidlo spojeno.)

Tabulka 2. Nejčastěji používaná pravidla nenáhodných seskupení

Pravidlo/ seskupení.	Popis pravidla
1/1	Jeden či více bodů je mimo zóny A
2/1	2 ze 3 po sobě následujících bodů leží v zóně A nebo za ní
3/2	4 z 5 po sobě následujících bodů leží v zóně B nebo za ní
4/2	8 po sobě jdoucích bodů leží na jedné straně od střední přímky
5/3	6 po sobě jdoucích bodů roste nebo klesá
6/4	15 po sobě jdoucích bodů leží nad nebo pod střední přímkou (v zónách C)
7/6	14 po sobě jdoucích bodů střídavě nahoře a dole
8/5	8 bodů v řadě za sebou leží na obou stranách střední přímky, ale žádný v C

Pravidla 1 – 4 byla definována pro rychlé rozpoznání seskupení spojených s posuny v procesu. Pravidlo 5 je spojeno s trendy, pravidla 6 a 8 se seskupeními vyvolanými nesprávným způsobem tvorby logických podskupin (výběrů) a pravidlo 7 s abnormální oscilací. Uvedená pravidla byla různými autory modifikována a vznikla řada souborů pravidel. V Tabulce 3 jsou shrnuty nejznámější soubory (Shewhartovo pravidlo(Sh.) [9], pravidla Western Electric (WE) [3], Nelsonovy testy [10], testy uvedené v ČSN ISO 2859 [11]). Dále jsou uvedeny méně známé soubory vytvořené v konkrétní firmě (Boeing [12], AIAG [13]) a nejnovější soubor publikovaný Trietschem [2]. V tabulce je patrné prolínání souborů (viz barevná políčka). Současně jsou patrné rozdíly mezi jednotlivými soubory testů, a to v míře pokrytí různých nenáhodných seskupení, v pořadí jednotlivých pravidel (viz číslo v závorce) a v definici délky sekvence bodů u daného pravidla. Historicky první pravidlo pro identifikaci vymezitelných příčin vytvořil W. A. Shewhart (v Tabulce 3 je to pravidlo 1). V souboru Western Electric pravidel (WE) přibyly k Shewhartovu kritériu testy nenáhodných seskupení (run tests) kvantifikující seskupení spojená s menšími či postupnými změnami (tzv. zonální testy), smíšená seskupení, stratifikaci a systematickou variabilitu. Pravidla 2, 3 a 4 byla navržena pro rychlejší odhalení posunu parametru procesu ve srovnání s pravidlem 1. (tzv. výstražné indikátory).

4. **Stanovení délky nenáhodného seskupení**

Jak je patrné z Tabulky 3, u pravidel č. 4, 6, 7 a 8 se soubory liší v definici délky sekvence bodů u daného pravidla. Nejvíce rozdílů se objevuje u pravidla 4. Při stanovení délky sekvence bodů se vychází z rozdělení pásma mezi regulačními mezemi na 6 stejně širokých zón (viz Obrázek 1). Za předpokladu normálního rozdělení pak lze očekávat, že v zónách C bude 68.27% hodnot,

TABULKA 3. Soubory testů nenáhodných seskupení

Testy	**Shewhart**	WE	Nelson N	ČSN ISO 8258	AIAG	Boeing ASQ	Trietsch T
Prav. 1	(1)	(1)	(1)	(1)	(1)	(1)	(1)
Prav. 2		(2)	(5)	(5)		(2)	(5)
Prav. 3		(3)	(6)	(6)		(3)	(6)
Prav. 4		(4) 8 bodů	(2) 9 bodů	(2) 9 bodů	(2) 7 bodů	(4)	(2) 9 bodů
Prav. 5			(3)	(3)	(3)		(3)
Prav. 6			(7) 15 bodů	(7) 15 bodů			(7) 13 bodů
Prav. 7			(4) 14 bodů	(4) 14 bodů			(4) 13 bodů
Prav. 8		(5) 8 bodů	(8) 8 bodů	(8) 8 bodů			(8) 5 bodů

v zónách B 27.18% a v zónách A 4.28% hodnot (viz Obrázek 1), pokud je proces statisticky stabilní. Pak lze pro různé délky sekvence bodů r daného seskupení (při výskytu určitého počtu bodů r v určitých zónách, odpovídající danému nenáhodnému seskupení) stanovit pravděpodobnost výskytu tohoto seskupení délky r v regulačním diagramu. Pomocí výpočtu této pravděpodobnosti se přímo či nepřímo získá hodnota rizika zbytečného signálu α pro dané pravidlo. Nejvhodnější délka sekvence bodů u každého pravidla je pak stanovena tak, aby pravděpodobnost rizika zbytečného signálu, spojeného s daným nenáhodným seskupením, nebyla příliš vzdálená od hodnoty rizika zbytečného signálu spojeného s Shewhartovým kritériem (pravidlo 1), které se při standardně používaných $\pm 3\sigma$ regulačních mezích a za předpokladu normálního rozdělení rovná hodnotě 0.0027 (počítáno vůči oběma mezím). V Tabulce 4 jsou uvedeny výsledky výpočtu rizika zbytečného signálu pro pravidla 4, 6, 8 pro délky r použité v různých souborech pravidel (viz Tabulka 3).Výpočty jsou provedeny nejprve klasickým způsobem a po té přesnějším postupem doporučeným Trietschem [2].

a) Při klasickém způsobu výpočtu rizika zbytečného signálu lze použít následující obecný vzorec:

$$\alpha_i = p^r, \tag{1}$$

kde
α_i ... je riziko zbytečného signálu i-tého pravidla;
p ... je pravděpodobnost výskytu bodu v jedné nebo více zónách (viz Obrázek 1)
r ... je délka sekvence bodů.
Pro pravidlo 4 pak tento obecný vzorec je třeba upravit následovně:

$$\alpha_4 = p^r.2 \;\; pro \;\; p = 0.5, \tag{2}$$

U pravidla 6 platí:

$$\alpha_6 = p^r \;\; pro \;\; p = 0.6826 \tag{3}$$

Tabulka 4. Hodnoty rizika zbytečného signálu

r/prav.	4		6		8	
Výpočet	klasický	dle T	klasický	dle T	klasický	dle T
5					**0.0032 T**	**0.0022 T**
7	0.0156 AIAG	0.0079 AIAG				
8	0.0078 WE	0.0039 WE			0.0001 N	0.00007 WE N
9	**0.0039 N, T**	**0.0020 N,T**				
13			0.00698 T	**0.0020 T**		
15			**0.00325 N**	0.001 N		

a pro pravidlo 8 pak platí:

$$\alpha_8 = p^r \quad pro \quad p = 0.3174. \tag{4}$$

b) Trietsch (T) doporučuje používat k výpočtům pravděpodobností výskytu určité sekvence bodů v regulačním diagramu následujícího výpočtu, který vede k přesnějším výsledkům [2]:

$$\alpha_i = p^r + p^{r+1}, \tag{5}$$

Z analýzy, jejíž závěry jsou shrnuty v Tabulce 4, plyne, že optimální hodnota délky sekvence bodů r u pravidla 4 je jednoznačně 9, u pravidla 8 je to jednoznačně hodnota 5. U pravidla 6 lze použít jak délku nadefinovanou Nelsonem (r = 15), tak i hodnotu 13 podle Trietscha (rozdíl oproti hodnotě 0.0027 je téměř stejný).

5. **Simultánní aplikace více testů**

Simultánní aplikaci více pravidel je třeba důkladně zvážit a vzít v potaz tyto skutečnosti:

- některé testy jsou vhodné při zahájení implementace SPC (pravidla 6 a 8);
- některé testy jsou vhodné zejména ve fázi ověřování a zabezpečování statistické stability procesu (pravidlo 1 a 4);
- některé testy jsou vhodné až ve fázi dlouhodobé regulace procesu (pravidla 2 nebo 3, 7);
- čím vice testů je aplikováno najednou, tím je vyšší riziko zbytečného signálu. Tuto skutečnost vyjadřují následující vzorce pro výpočet celkového rizika zbytečného signálu.

Riziko zbytečného signálu pro nezávislé testy se vypočte ze vztahu:

$$\alpha = 1 - \prod_{i=1}^{k}(1 - \alpha_i). \tag{6}$$

Testy 1, 2, 3 a 4 však mohou být mezi sebou pozitivně korelovány (seskupení, které aktivuje jeden test, bude pravděpodobně aktivovat rovněž jiné testy). Riziko zbytečného signálu pro závislé testy pak lze stanovit dle jednoduchého vzorce [2]:

$$\alpha \leq \sum_{i=1}^{k} \alpha_i. \tag{7}$$

6. **Analýza vybraných softwarových produktů**

Na trhu je dostupná řada statistických programů, které podporují identifikaci nenáhodných seskupení. Pokud jsou tyto metody správně aplikovány, mohou přispět ke zvýšení efektivnosti SPC jako procesu řešeni problému. Naopak rutinní aplikace testů nenáhodných seskupení může vest ke snížení efektivnosti implementace SPC. V této kapitole bude pozornost věnována produktu Statgraphics Plus (v tabulkách 5 a 6 značeno S v.15) [15], Statgraphics Centurion (v tabulkách 5 a 6 značeno S Cent.) [16], Minitab (v tabulkách 5 a 6 značeno M v.15)[17], které jsou využívány v rámci výuky na katedře kontroly a řízení jakosti na Fakultě metalurgie a materiálového inženýrství VŠB-TU Ostrava, a produktu Statistica (v tabulkách 5 a 6 značeno Sca v.16) [18], který je hojně používán na řadě českých univerzit. Analýza těchto programů je zaměřena na jejich podporu identifikace, případně další analýzy nenáhodných seskupení v regulačních diagramech. Analýza je vedena z pohledu faktorů a závěrů diskutovaných v předchozích kapitolách. Obecně všechny analyzované softwarové produkty pracují s malými odchylkami se souborem Nelsonových testů. Všechny Nelsonovy testy včetně Shewhartova pravidla 1 obsahuje pouze Minitab. Ostatní analyzované programy mají pravidlo 1 zahrnuté do základní analýzy regulačních diagramů mimo testy nenáhodných seskupení. V programech Statgraphics je mírně odlišně definováno pravidlo 5 a 8. Komplexní hodnocení jednotlivých produktů je uvedeno v Tabulce 5 a 6.

Předchozí analýza vybraných statistických produktů shrnutá v Tabulce 5 a 6 vede k závěru, že nejlépe jsou pro aplikaci nenáhodných seskupení vybaveny programy Minitab a Statistica. Avšak i ony mají několik nedostatků, které mohou způsobit, že méně zkušení uživatelé uplatní testy nenáhodných seskupení nesprávně, což může vést k nesprávným závěrům ohledně stability procesu. To pak může vyústit v situaci, kdy uživatel přestane aplikovat testy nenáhodných seskupení, či dokonce v selhání celého systému SPC. Proto byla navržena jednoduchá univerzální metodika aplikace testů nenáhodných seskupení, která je popsána v následující kapitole.

TABULKA 5. Komplexní analýza vybraných softwarových produktů

Vlastnosti/ SW	S v.15	S Cent.	M v.15	Sca v.16
Termín pro nenáhdné seskupení	Unusual pattern	Unusual sequence	Nonrandom pattern	Systematic pattern
Termín pro testy (pravidla)	Runs tests	Runs tests	Tests for special causes	Runs tests
Komplexnost	6 z 8	7 z 8	8 z 8	7 z 8
Statistické základy testů	ne	ne	ne	ano
Počet preddefinovaných odlišně od Nelsonových testů	2	2	0	0
Možnost výběru pravidel	ano	ano	ano	ano
Možnost změnit délku seskupení	ano	ano	ano	ano
Možnost předefinování zóny	ano	ano	ne	ano
Indikace nenáhodného seskupení v regulačním diagramu	ano	ano	ano	ano
Záznam možných vymezitelných příčin do regulačního diagramu	ne	ne	ne	ne
Záznam zásahů do regulačního diagramu	ne	ne	ne	ano
Upozornění na rizika simultánního použití více pravidel najednou	ne	ne	ano	ano
Upozornění na potřebu mít hluboké znalosti o procesu	ne	ne	ano	ano
Typy regulačních diagramů (S-Shewhart, J- jiné)	S + J	S + J	S	S

TABULKA 6. Komplexní analýza vybraných softwarových produktů - pokračování

Vlastnosti/ SW	S v.15	S Cent.	M v.15	Sca v.16
Stejný soubor pravidel přednastavený pro různé regulační diagramy	ano	ano	ano	ano
Informace, na které typy diagramů jsou jednotlivá pravidla aplikovatelná	ano (nepřesná)	ano (nepřesná)	částečná	částečná
Definování potenciálních obecných vymezitelných příčin	ne	ne	ne	ano
Popis pravidel	ano	ano	ano	ano
Interpretace pravidel	ne	ne	ne	ano

	Použitá nenáhodná seskupení			
Velký posun	ano	ano	ano	ano
Menší přetrvávající posun	ano	ano	ano	ano
Trendy	ano	ano	ano	ano
Stratifikace	ano	ano	ano	ano
Smíšené seskupení	ano	ano	ano	ano
Systematická variabilita	ne	ano	ano	ano

7. Metodika aplikace testů nenáhodných seskupení

- Před aplikací testů nenáhodných seskupení je třeba ověřit, které testy jsou u zvolených regulačních diagram přednastaveny.

- Nikdy se nemají aplikovat rutinně všechny testy, i když jsou v použitém programu přednastaveny.
- Je třeba ověřit nastavenou délku sekvence bodů r u jednotlivých vybraných testů a v případě potřeby nastavit optimální hodnoty r.
- Dále je třeba ověřit, zda zvolené regulační diagramy mají přibližně symetrické meze. Pokud ano, platí následující závěry:
  - Všechny testy (pravidla 1 - 8) mohou být aplikovány na diagram pro průměry a individuální hodnoty (předpokládá se normální rozdělení regulované veličiny);
  - Pravidla 1 – 4 mohou být aplikována na diagram R nebo s bez jakékoliv modifikace, je-li rozsah podskupiny $n \geq 5$;
  - Všechna pravidla mohou být aplikována bez omezení na regulační diagramy np a c za předpokladu, že platí aproximace normálním rozdělením a meze jsou rozumně symetrické.
  - Totéž platí pro diagram p a u s konstantními mezemi;
- Pravidla 5 a 7 fungují dostatečně dobře při jakémkoliv spojitém rozdělení, neboť jde o neparametrické testy;
- Z hlediska sekvence a souběhu aplikace testů nenáhodných seskupení se mají testy aplikovat podle následujícího schématu:
  - Pravidla 6 a 8 se mají aplikovat při zahájení implementace SPC s cílem verifikace správnosti tvorby logických podskupin;
  - Ve fázi ověřování a zajišťování statistické stability procesu se mají aplikovat jako první pravidla 1 a 4;
  - Je-li potřeba dalšího zvýšení citlivosti regulačního diagramu na změny parametrů procesu, může se přidat pravidlo 2 nebo 3, případně obě;
  - Jestliže se znalost procesu během předchozí aplikace SPC prohloubila, je možné aplikovat zbývající běžná pravidla.
  - Pravidlo 5 se má aplikovat samostatně (jeho přínos v podobě zvýšení citlivosti regulačního diagramu na skutečné změny v procesu je menší než zvýšení rizika zbytečného signálu ([14], p. 137).
- Pravidla 1 – 4 je třeba aplikovat na obě poloviny regulačního diagramu, ale samostatně.

8. **Závěr**

Na základě teoretických východisek zpracovaných v kapitole 1 a 2 se tento článek zabýval analýzou a srovnáním několika souborů pravidel pro identifikaci nenáhodných seskupení v regulačních diagramech, a to Shewhartova kritéria, souboru Western Electric, Nelsonových testů, testů z normy ČSN ISO 2859, souboru pravidel Boeingu a AIAG a souboru publikovaného Trietschem. Závěry této analýzy byly použity pro následnou analýzu vybraných softwarových produktů se zaměřením na testy nenáhodných seskupení. Výsledky obou analýz byly dále promítnuty do návrhu metodiky pro

analýzu statistické nestability procesu s využitím dostupných statistických programů.

Literatura

[1] Stapenhurst, T.: *Mastering Statistical Process Control*. Oxford, Elsevier, 2005, 460 stran.

[2] Trietsch, D. *Statistical Quality Control. A Loss Minimization Approach*. London, World Scientific, 1999, 387 s.

[3] Western Electric Company, Inc.: *Statistical Quality Control Handbook*, 2^{nd} ed. Easton, Mack Printing Co., 1958, 328 s.

[4] Montgomery, D. C.: *Introduction to Statistical Quality Control*. New York, J. Wiley & Sons, 2001, 796 s.

[5] Griffifth, G. K.: *Statistical Process Control Methods for Long and Short Runs*. Milwaukee, Wisconsin, ASQC Quality Press, 1996, 250 s.

[6] Evans, J. R., Lindsay, W. M.: *Managing for Quality and Performance Excellence*. 7th ed. Mason, OH, Thomson, 2008, 783 s.

[7] Besterfield, D. H.: *Quality Control*. 8^{th} ed. New Jersey, Prentice Hall, 2009, 552 s.

[8] Grant, E. L., Leawenhworth, R. S.: *Statistical Quality Control*. New York, McGraw Hill, 1996, 764 s.

[9] Shewhart, W. A.: *Economic Control of Quality of Manufactured Product*. New York, Van Nostrand, 1931. 501 s.

[10] Nelson, L.S.: The Shewhart Control Chart – Tests For Special Causes. *Journal of Quality Technology,* 1984, sv.16, č. 4, ss. 337-339.

[11] ČSN ISO 8258: *Shewhartovy regulační diagramy*. Praha, ČSNI, 1994.

[12] D1-9000: ASQ Boeing Company, 2000.

[13] AIAG SPC-3, 2005.

[14] Wheeler, D.: *Advanced Topics in Statistical Process Control*. Knoxville: SPC Press, Inc., 2004.

[15] STATGRAPHICS PLUS v. 5.1.: Manugistics Inc., 2000.

[16] STATGRAPHICS CENTURION, v XV.: StatPoint Technologies, 2006.

[17] MINITAB v. 16: Minitab Inc., 2010.

[18] STATISTICA v. 10,: StatSoft, Inc., 2010.

Poděkování: Tento článek byl zpracován v rámci projektu specifického výzkumu č. SP2012/42,podporovaného MŠMT ČR, který byl řešen na Fakultě metalurgie a materiálového inženýrství na VŠB-TU Ostrava.

CALCULATION OF FLEXIBILITY AND REWORK AVAILABILITY IN AN ASSEMBLY LINE WITH TWO TYPES OF PARTS

Israel Fabian Oropeza Peña

Address: Robert Bosch, České Budějovice; oropeza.fabian@gmail.com

Abstract: There are a lot of facts about the flexibility and their implementation into a production line; most theories suggest that it should be implemented as quickly possible and with the lowest investment. Due the continue increasing of the production and the huge demand of the customer for new designs, the automotive industries should include into their production lines, changeable process, fixtures, work stations, conveyors, etc; production procedure should very adjustable, allowing to assembly not only new parts with new design, but also parts with complex construction and big impact into the following production steps, and these new assembly parts should keep the standards of the cycle productions. Therefore companies should adopt a flexible production line since the beginning of construction of the lines to keep molding and readjustment them according to internal and external development of the automotive industry. This paper will focus on the Flexibility of the lines their activities and the availability of rework assembled components.

Keywords: Process availability, automobile production

1. Introduction

The automobile industry is exposed to a high level of competition and improvements in their design. Increase of productivity and meet the client's needs are the principal goal of the companies. Automotive manufacturing, demand that their suppliers meet the clients new requests, keeping production line standards, and subcontractors should be able to deliver in time, the new parts with same or best quality, in where flexibility role is a key performing that automotive company has to keep in mind. In the past years the automotive production has had a significant solid increase in central Europe and the Asian countries, where the companies can offer more products to their clients, reducing the production cycle, and prepare the lines for further new generations. Also and very important factor, is to meet the worlds strictest limits for nitrogen-oxide emissions, efficient exhaust-gas treatment systems are mandatory. For the further years good political of production can lead to have new request from the customers, developers can created new and complex design, where the automotive companies will have the task to fulfill their requirements, therefore the best solution is to increased the flexibility into their production lines, work on new generation investment plan the for new production lines including changes into their work stations. In this paper,

product flexibility as well as volume flexibility of one assembly line in a manufacturing system is examined. The mathematical treats product flexibility, volume flexibility and rework. For this study has only considerate the production lines for dosing module. The assembly line consists of workstations. As automatically assembly is predominantly applied, the material handling system, ancillary functions and layout of the factory. Assembly line have the task to assemble two different connecting-piece into a body, where the body has the same function for all the types, but the connecting-piece have different dimensions and parameters. The purpose of the study at the company is to analyze the flexibility of the assembly line in order to be able to understand the problems, their principal cost created by introduction of new models variants and reworking process impact.

2. **Problem description**

New design for the connecting piece, different dimensions and assembly parameters, readjustment of the production line, fixture, production time, control test and high risk at the time of assembly.

3. **Flexibility**

Flexibility is defined in Webster's Dictionary as an adverbial of flexible meaning, "capable of responding or conforming to changing or new situation" [1]. There are several definitions of flexibility in the literature in the area of manufacturing of which a few are given below [2].

- Mix flexibility: Processing at any time a mix of different parts which are loosely related to each other in some way such as belonging to the same family.
- Parts flexibility: Adding parts to and removing parts from the mix over time.
- Volume flexibility: Handling shifts in volume for a given part. A more detailed investigation of flexibility is given in [3]. In this paper, based on the extent of automation and the diversity of the parts produced, is necessary to considerate, the following aspects of flexibility:
- Machine flexibility: The ease of making the changes required to produce a given set of part types. The main issue is set-up time for the machines.
- Process flexibility: The ability to produce a given set of part types, each possibly using different material, in several ways. This flexibility is inversely dependent on the set-up costs.
- Product flexibility: The ability to changeover to produce a new (set of) product(s) very economically and quickly. Mandelbaum defines this as, "acction flexibility, the capacity for taking new action to meet new circunstances". [4]. This flexibility heightens a company's potential responsiveness to competition and/or market changes. Product

flexibility can be measured by the time required to switch from one part mix to another, not necessarily of the same part types (compare with Gerwinâ's design-change flexibility).

There are other numerous definitions that have not been discussed here because of the scope of the subject. In their paper of, "an examination of rooting flexibility problems for small batch assembly of dosing modules", Taylor and Graves analyzed introduction of new processing flexibility in printed circuit board assembly [5]. The introduction of routing flexibility and its limitations and benefits is analyzed in the paper based on experimental results. The same authors also treat sensitivity analysis for a three level flexible control strategy in an assembly environment [6]. The control strategy in this paper includes product mix flexibility at the system level, product routeing flexibility at the cell level, and product sequencing flexibility at the machine level. Product flexibility of a manufacturing system comprise the capability of "transformaton" of an existing manufacturing system to manufacture either new variants of existing models or completely new products. This can require exchange and/or introduction of new machines, extension or alteration of handling and/or inter-linking devices, control systems, and control programs and ancillary functions [7].

4. **Rework and reworkable**

Rework is to change an item in order to improve it or make it more suitable for a particular purpose, e.g. to rework a defective product into one that exhibits the quality required for acceptance. In other words, Rework is the work done to correct defects that bring a product back into 100% conformance to requirements. Reworkable is the capability of being worked again, deal with, handled or anew, capable of being put into effective operation, process, practicable or feasible. The need for the ability to rework component has, in recent years, resulted in developmental efforts to formulate components that can easily be reworked as well as provide requisite product reliability. The implementation of reworkable for components involves: choice of proper material, development of an acceptable process, and a verification of reliability.

5. **Calculations of Flexibility**

The flexibility for production line can be calculated using the combination of three types of flexibilities, which play out in different time frames: Mix flexibility (long-term), volume flexibility (mid-term) and emergency service flexibility (short term), which can all be categorized as provider-side flexibilities Nilchiani and Hastings [8]. Mix flexibility is defined as the strategic ability to offer a variety of products with the given manufacture installed. Volume flexibility is the ability to respond to drastic changes in demand. Emergency products flexibility is the tactical ability of the system to provide

emergency (non-scheduled) products to independent work stations. The overall provider side flexibility metric is then obtained by taking the weighted average of the above flexibilities. [9]

6. **Manufacturing system components**

As stated in the sekce1, one assembly line is treated in this study where is assembly different types of products. The assembly line, consist of workstations, which perform certain types of assembly operations. Production line consists of seven stations arranged one after another, to be operated by one person. Single stations of the line are data interconnected due to tracing of production process. The data are archived by the server, the product when is completed is packing into a box and loaded into a pallet waiting to be transported.Fig. 1.

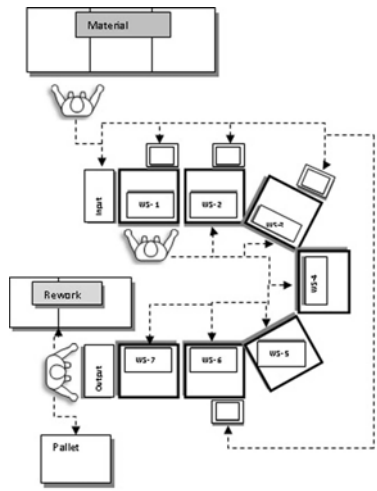


FIGURE 1. Assembly sequences at the line

In the production line there are three operators: Loader, Worker, and Packer. Loader picks the material loading requested work stations, worker start from the first work station $WS-1$ making circles, picking the assembled piece from one WS to the next WS, and Packer function is to close the full box and place into a pallet.

7. **Flexibility of a workstations**

A productivity matrix: Here, the direct cost (in terms of time) is given for each operation at each workstation (assembly station), corresponding to each element of the matrix. Number of rows will be equal to the number of workstations constituting the whole assembly process. The unit of the elements in the matrix is time in hours, measured in real working conditions. Workstation flexibility is explained in two matrices. s is the matrix that denotes the set-up costs required for each operation (or operation group) for every item treated at eaach the workstation. Thus, elements of s, aij; will denote set up costs for the ith operation for the jth independent item. It can include cost for special tools required for the operations, cost of other equipment, cost for education, etc. [10]

$$FM_{sf} = \begin{bmatrix} 6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 & 0 \\ 0 & 6 & 0 & 5 & 0 & 0 \\ 0 & 6 & 0 & 0 & 4 & 0 \\ 0 & 6 & 0 & 0 & 0 & 6 \end{bmatrix}$$

FWr is the cost for resetting up the workstation for each operation in terms of time. For this studied case, it is the time spent to set up the line for a new model or variant:

$$FW_s = [0,083 \;\; 0,007 \;\; 0,007 \;\; 0,013 \;\; 0,005 \;\; 0,005 \;\; 0,005].U,$$

where U is a unit matrice of size $n \times n, n = 7$. If an operation needs to be set-up for a new item, that element in the ST will be one, if not, it will be zero. Let us set $OS = ST = U$.

The sum of these times spent for each assembly operation constitutes the total assembly time

$$P_t = [0,40 \;\; 0,40 \;\; 0,42 \;\; 0,45 \;\; 0,50 \;\; 0,23 \;\; 0,55]^T.$$

Manufacturing characteristics can be expressed in the following formulas: Resetting time $RT = trFW_r^T OS$, tr is the trace operator, i is the interest rate (in our calculations taken as 10%), b is the batch size. In our case it is

$$RT = [9 \;\; 13,5 \;\; 10,5 \;\; 7,5 \;\; 7,5 \;\; 6 \;\; 9]^T$$

Total processing time $TPT = P^T * OS^T * I$, where I is a $n \times 1$ vector of ones. This gives 177 seconds since the part enters to the assembly line until is completed. Total setting costs $TSC = FW_r^T ST$, 42 EUR.

8. **Calculation of Reworking Availability**

Rework operations represent a major technical difficulty because of the relatively long time, turning this time into idle time. This issue is even more critical at the time when rework for new parts with new design and different

condition of assembly.
Fig 2, shows the availability at each station, (b) is the rework with release and (b) is the rework with disassembly, green check mark, means that is possible to send the part to this WS and red mark means that no rework at this WS.
Availability of reworking: two matrices are used for this purpose: the feasibi-

FIGURE 2. (a) Release, (b) dissembled availability

lity to rework in each station is show in the following matrix. Where value 1 means the part can be direct release and rework with no need of dissembled R_d if is not possible will be zero. We have $R = U$.

Individual time for each work station:

$$T_W = [0,42 \;\; 0,47 \;\; 0,48 \;\; 0,50 \;\; 0,53 \;\; 0,57 \;\; 0,60]^T.$$

Total release time per part for each station is:

$$T_R = [1,25 \;\; 1,40 \;\; 1,45 \;\; 2,00 \;\; 1,60 \;\; 1,13 \;\; 1,20]^T.$$

Total releasing time for one part is the sum of T_d for this case is 10:02 min 602 sec.

Time for dissembled and release per part at each work station:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{1}$$

Where $T_D = D * T_W = [1,25 \;\; 1,87 \;\; 2,42 \;\; 2,50 \;\; 3,20 \;\; 1,13 \;\; 0,60]^T$.

Total releasing time for one part is the sum of T_d for this case is 13:54 min 834 sec.

To calculate the probability of release and dissembled a piece using the historical data from the past 3 years of production.(Fig. 3)

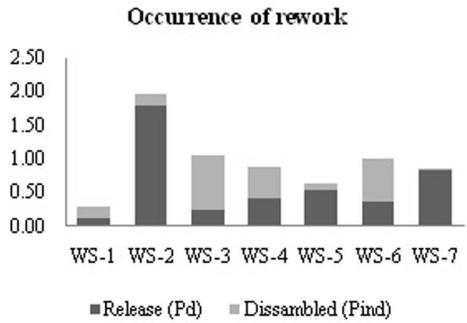


FIGURE 3. Probability of rework

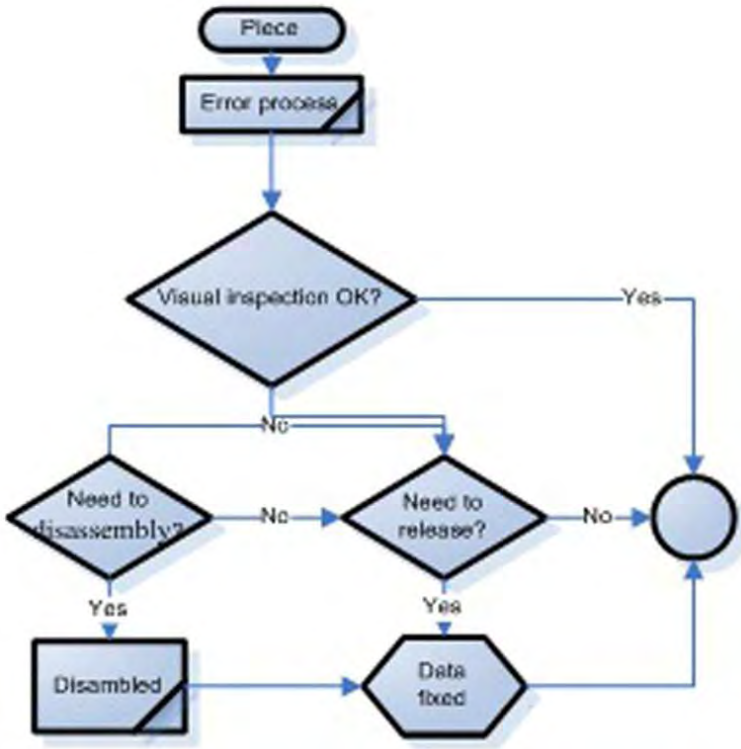


FIGURE 4. Logical sequence of rework

For $P_d = [0,13 \;\; 1,80 \;\; 0,24 \;\; 0,42 \;\; 0,54 \;\; 0,38 \;\; 0,84]^T$ we have

$$P_{RW} = R * P_D = [0,39 \;\; 5,40 \;\; 0,72 \;\; 1,68 \;\; 1,62 \;\; 0,76 \;\; 1,68]^T.$$

Same calculation for dissembled parts gives us

$$P_{rel} = [0,17 \;\; 0,16 \;\; 0,81 \;\; 0,47 \;\; 0,10 \;\; 0,64 \;\; 0,01]^T.$$

Therefore the probability of recurrence for dissembled parts is:

$$P_{DW} = D * P_{rel} = [0,51 \;\; 0,96 \;\; 4,05 \;\; 2,35 \;\; 0,60 \;\; 1,28 \;\; 0,01]^T.$$

To calculate rework availability time, for type X at each work station including time for release we shall use the following formula:

$$AR_{tI} = R * T_I$$

We have $T_I = [0,67 \;\; 0,17 \;\; 0,42 \;\; 0,00 \;\; 0,03 \;\; 0,03 \;\; 0,00]^T$. Similar calculations are carried out for the type Y, which is set up to manufacture a new product at the same line. Then $T_{II} = [0,83 \;\; 0,25 \;\; 0,67 \;\; 0,00 \;\; 0,03 \;\; 0,03 \;\; 0,00]^T$. Total dissembling time for one part is the sum product of $R_{d-d} * T_d$, for this case is 109 seconds.

Rework availability time, for type Y has the next time matrix:

$$AR_{tII} = R * T_{II} = [3,33 \;\; 1,00 \;\; 2,67 \;\; 0,00 \;\; 0,13 \;\; 0,17 \;\; 0,00]^T.$$

From this analyze can be identified that in the first three work stations takes more time than other and for the WS1 and WS3 to rework and release a part, it means that if for this stations a new design of component can significant increase idle time. Figure 5 shows a development and increasing of the time to take to change over.

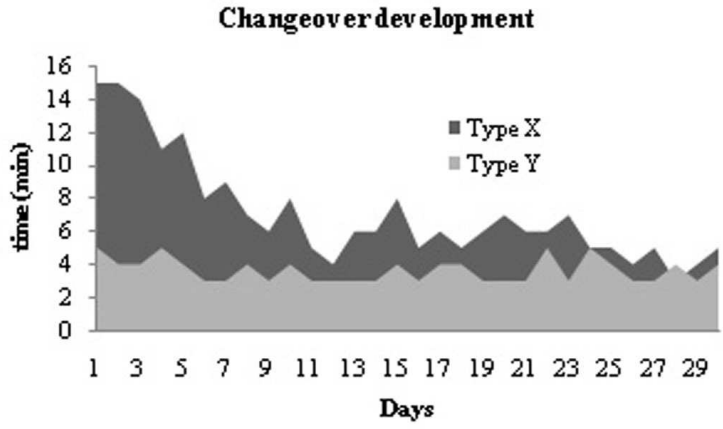


FIGURE 5. Change over from type X to type Y

Variation of the time between new and current production type, calculated from the idle time per work station Fig.6.

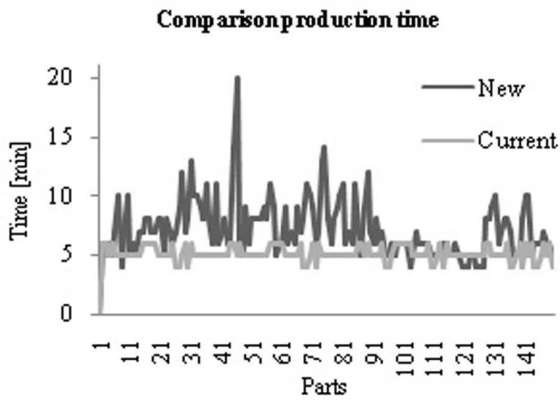


FIGURE 6. Comparison of idle time

9. **Calculation of Flexibility**

In this research the mix flexibility is defined as the ratio of profit resulting from adding more product types, taking into account additional cost incurred by the necessary changes in design, to the profit with same production line. Mix flexibilities larger than 1.0 indicate a system that increases in value by offering multiple products types. For this purpose, mix flexibility is defined by Equation 1.

$$F_m = \frac{S_m - E_m}{S - E}.$$

Where F_m is Mix flexibility, E represents total system cost over the production process, S total revenue over the lifetime, m multiple types of products (X, Y).

Volume flexibility is the ability to respond to drastic changes in demand. In the production line it is defined as the value of the product for the provider over the range of market uncertainties, determined by the value of the product for the provider at currently projected demand. Values equal to or larger than 1.0 indicate that the expected value of the system over the range of market uncertainties is higher than its value at currently projected demand, indicating system flexibility with regard to demand change. Thus,

$$f_v = \frac{\int_0^E e^{-} r t_m (S - E) p(S) dS}{I_{Risk} - free}.$$

Emergency service flexibility is the tactical ability of the system to provide emergency (non-scheduled) products assembled parts. It can be defined as the

capacity of the product (maximum product capacity) divided by the current level of product per shift.

$$f_E = \frac{Cap_{max}}{Cap_{current}}.$$

Product flexibility is then obtained by taking the weighted average of the above flexibilities. This ensures that the flexibility preferences of the provider are taken into consideration.

The weight coefficients are determined by the provider, based on the priorities specified in the production line mission objectives. For instance, for a client base consisting mostly of commercial parts, volume flexibility and mix flexibility have a higher weight than emergency flexibility, whereas for new design parts, emergency flexibility and mix flexibility have higher weight coefficients than volume flexibility. Different metrics are explored for objectives with different weights, and all possible changes are ranked in a trade, based on the resulting flexibility, value and performance metrics. Thus service flexibility can be defined as:

$$f = \frac{\sum_{i=M,V,E}^{W_i} * f_i}{\sum_{i=M,V,E}^{W_i}}.$$

Combined product flexibility can be defined by a weighted sum of the above flexibilities.

As Figures 7 show, adding the flexibility dimension changes the Pareto type X in terms of the assembled part. In this particular instance, type X, is replaced by type Y. This shows how taking into account flexibility can affect the choice of type in a way that the system becomes somewhat more expensive, but can handle volume-level, service-mix and emergency-service uncertainties and fluctuations inherent into the production line.

10. **Calculation of assembling new type**

Calculation of the assembly processes time for new type, according to current standard T_{np}.

$$T_{np} = \sum_{0}^{n_p} * T_p + \sum_{WS-1}^{WS/n} * A_m * n_p,$$

where A_m is the sum of the total percentage of the accumulation of readjustment per machine. S_{td} is the Standard of the total production time divided into the total amount of assembled parts per shift, which means only clean production time (440 min).

$$S_{td} = \frac{\sum_0^n N_p}{\sum_0^N * P_p}$$

Standard for piece production N_p,

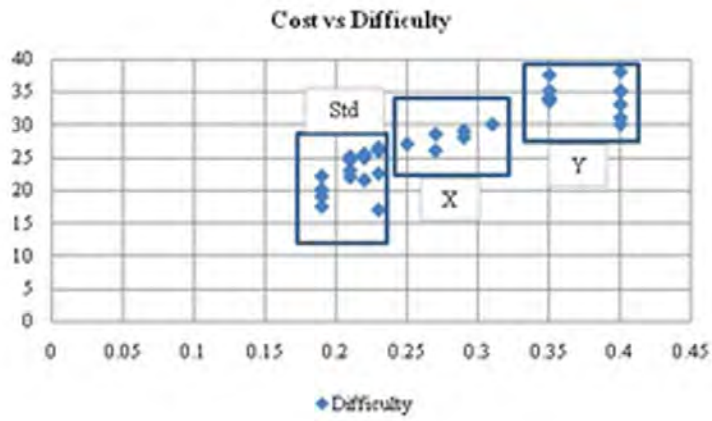


FIGURE 7. Cost and difficulty

Minutes for single part is the accumulated time per part, is the required time for only one part to be assembled.

$$T_{sp} = N_p * S_{td}$$

Batch production time, is the sequentially time for the new parts in minutes, assembled one after one sequentially with no pause after each part

Characteristics	Std	Real	New
Difficulty	20%	25%	40%
Standard	1,1	1,1	1,1
Pieces	150	150	150
Minutes for single part	660	777	1082
Minutes for all parts	165	194,25	270

11. **Conclusion**

This paper investigates product flexibility and rework availability of two products assembled in the same production line. Product flexibility is expressed by two matrices, one matrix for setting-up of the line and another for resetting of the line. An equation is obtained using these matrices for each type expressing the total manufacturing costs of the new products, and thereby the flexibility of rework is obtained. These equations are also used to investigate the optimum assembly volumes and corresponding costs for each type. The analysis in this paper shows how the calculation of flexibility can help decision-making for components assembled into a production line with new components integrated into the production. These results could help for

planning and logistic department, at the time of calculation the cost and time for new products, also for the development engineering department, to simulate the impact for new component into the production line.

References

[1] Webster. Springfield, MA, USA: G.&C. Merriam Company, 1976.
[2] Gerwin D. Do's and don'ts of computerized manufacturing. Harv Bus Rev, 1982.
[3] Brown J., et al. Classification of flexible manufacturing systems.FMS Mag, 1984.
[4] Mandelbaum M. Flexibility in decision-making: an exploration and unification. PhD dissertation, Department of Industrial Engineering, University of Toronto, Ont., Canada, 1978.
[5] Taylor D, Graves JR. An examination of routeing flexibility for small batch assembly of printed circuit boards. Int J Prod Res, 1990.
[6] Taylor D, Graves JR. Integrated decision making in a flexible assembly system: sensitivity analysis and extended testing. Prod Planning Control, 1991.
[7] De Toni, A., and Tonchia, S., Lean organization, management-by-process and performance measurement, International Journal of Operations and Production Management, 16 (2) 1998.
[8] Nilchiani, Roshanak and Hastings, Daniel E Measuring flexibility in design of an orbital transportation network, Presented at the AIAA Space 2003 Conference, Long Beach, California, September, 2003.
[9] A. Kurtoglu, Flexibility analysis of the assembly lines. ABB Automation Technology Products, Lugnagatan, V. aster, October, 2003.
[10] Chryssolorouris G. Flexibility, its measurement. Annals of CIRP,Keynote Paper, Vol. 45(2), 1996.

Acknowledgement: This research conducted under short term research grant which contributed by Czech University of Prague – CTU.

INOVOVANÝ PŘÍSTUP MĚŘENÍ ELEKTRICKÉHO ODPORU ELIMINUJÍCÍ PROBLÉM KONTAKTNÍCH ODPORŮ

AN INNOVATIVE APPROACH TO ELECTRICAL RESISTANCE MEASUREMENT ELIMINATING CONTACT RESISTANCE PROBLEM

Veronika Šafářová, Luboš Hes, Jiří Militký

Adresa: Technická univerzita v Liberci, Fakulta textilní, Studentská 2, 461 17 Liberec; veronika.safarova@tul.cz

Abstrakt: Elektrická vodivost je jedním z rozhodujících parametrů pro zlepšení odolnosti vůči elektromagnetickému smogu, snížení tendence k hromadění elektrostatického náboje a konstrukci inteligentních textilií obsahujících vodivé dráhy. Elektricky vodivým textiliím se v současné době dostává zvýšené pozornosti, a to zejména kvůli jejich flexibilitě a nízké hmotnosti. Znalost elektrických vlastností textilního materiálu např. ve formě vlákna, či příze prostřednictvím měření elektrického odporu je velmi důležitá a to zejména pro využití za účelem predikce elektrické vodivosti celého systému „vlákno – příze – textilie“ a následného návrhu výrobku pro konkrétní použití. Pro měření elektrického odporu délkových textilních útvarů je v současné době používáno velké množství metod, mezi které patří např. metoda ampérmetru a voltmetru, či metoda Wheatsonova můstku. Uspořádání vzorku při měření je také významné. Nejčastěji je používáno k upnutí lineárních textilních útvarů kovových svorek. Problém vytváří kontaktní odpor vznikající na styku mezi měřeným materiálem a kovovou svorkou. Výsledkem je vznik chyb, jimiž je snižována přesnost a reprodukovatelnost měření. Hlavním cílem této práce je představení inovované metodiky měření, která je založena naa měření elektrického odporu na nejméně dvou definovaných úsecích délkového textilního útvaru a následný výpočet elektrického odporu, který není zatížen chybou způsobenou kontaktním odporem.

Abstract: Electric conductivity is one of dominant parameters for improving electromagnetic shielding resistivity, reducing electrostatic charge accumulation and creating intelligent textile structures containing conductive tracks. Conductive fabrics have obtained increased attention mainly due to their desirable flexibility and lightweight. Knowledge of electrical properties of textile material (for example in fiber or yarn form) is very important especially for conductivity prediction of whole system “fiber – yarn – fabric” and subsequent design of product for specific application. Variety methods of measurement (ammeter and voltmeter, Wheatstone-bridge method) and many different arrangements of the material to be tested can be used. Bulldog clips are used most often to hold the yarn between electrodes. The contact resistance generated at measured material/metal bulldog clip interface creates

measurement errors. This phenomenon leads to lower accuracy and reproducibility of measurements. The main aim of this work is introduction of innovated method eliminating contact resistance problem. This method is based on measurement of two and more sections of yarns and then electrical resistance calculation which is not loaded with measurement errors.

Klíčová slova: elektrická vodivost, metodika měření, kontaktní odpor

Keywords: Electric conductivity, measurement method, contact resistance

1. Úvod

Znalost elektrických vlastností textilního materiálu např. ve formě vlákna, či příze prostřednictvím měření elektrického odporu je velmi důležitá a to zejména pro využití za účelem predikce elektrické vodivosti celého systému „vlákno – příze – textilie“ a následného návrhu výrobku pro konkrétní použití.

Elektrické vlastnosti materiálů se obecně hodnotí dle měrného odporu ρ [Ω.m]. Pro délkové textilní útvary (stejně tak jako u mechanických vlastností) je vhodnější založit definici na lineární hustotě materiálu neboli jemnosti. Hmotnostní rezistivita R_S neboli hmotnostní specifický odpor je veličina vyjadřující elektrický odpor mezi konci vzorku 1 m dlouhého o hmotnosti 1 kg; hlavní jednotku je [Ω.kg/m^2]. Pro vyjádření obou veličin je nutná znalost velikosti elektrického odporu na definovaných úsecích délkového útvaru, kterou je možno zjistit experimentálně.

2. Problematika kontaktních odporů

Pro měření elektrického odporu délkových textilních útvarů je v současné době používáno velké množství metod, mezi které patří např. metoda ampérmetru a voltmetru, či metoda Wheatsonova můstku. Uspořádání vzorku při měření je také významné. V současné době je používáno různých přístupů. Elektrický odpor může být měřen podél jednotlivých vláken, podél paralelně uspořádaných vláken, mezi konci příze, či mezi konci paralelně uspořádaných přízí. Nejčastěji je používáno k upnutí lineárních textilních útvarů kovových svorek, nepříliš četná je pak tvorba kontaktu pomocí vodivého nátěru či lepení [1].

Problém vytváří kontaktní odpor vznikající na styku mezi měřeným materiálem a kovovou svorkou. Kontaktním odporem je označován poměr potenciálního rozdílu mezi dotýkajícími se plochami k intenzitě proudu kontaktem procházejícího. Tento odpor nemusí být vždy zanedbatelný a to z těchto příčin:

- Plochy kontaktu nebývají vždy ideálně hladké, tudíž se nedotýkají ve všech bodech.
- Plochy kontaktu nebývají nikdy ideálně čisté, vždy jsou pokryty vrstvou kysličníku nebo jiných sloučenin, jejichž vodivost je nepatrná.
- Kontaktní odpor je silně závislý také na tlaku.

Výsledkem je vznik chyb, jimiž je snižována přesnost a reprodukovatelnost měření. Z výše uvedených důvodů je vhodné kontaktní odpor při měření vyloučit.

3. Inovovaná metodika měření

Podstata metodiky řešení spočívá v měření elektrického odporu na nejméně dvou definovaných úsecích délkového textilního útvaru a následný výpočet elektrického odporu, který není zatížen chybou způsobenou kontaktním odporem.

Pro hodnocení elektrické vodivosti klasických lineárních textilních útvarů, pro které platí, že elektrický odpor je lineárně závislý na upínací délce je dostačující měření na dvou definovaných úsecích. Pro hodnocení tzv. hybridních délkových textilních útvarů obsahující ve své struktuře vodivou komponentu ve staplové formě (vodivá vlákna konečné délky), je pro popis závislosti elektrického odporu na upínací délce nutno měřit elektrický odpor minimálně na třech definovaných úsecích. A to proto, že závislost elektrického odporu na upínací délce je pro hybridní příze možno popisovat nelineární rostoucí funkcí jak bylo zjištěno a ověřeno např. v [2, 3].

Každý měřený úsek příze je zatížen chybou, kterou způsobují dva kontaktní odpory v místě upnutí vzorku. Princip inovovaného řešení spočívá úpravě standardní metodiky měření, pomocí níž je možno eliminovat kontaktní odpor způsobený uchycením délkového textilního útvaru elektrodami.

Jak je patrno z dalšího popisu, elektrický odpor je pomocí nové metodiky experimentálně měřen na úseku délkového textilního útvaru o délce L_0. Tento odpor je označen R_{tot0} a jedná se o celkový odpor úseku příze L_0, vč. sumy kontaktních odporů. Následně je experimentálně změřen úsek příze o délce L_1. Tato hodnota elektrického odporu je označena R_{tot1} a jedná se o celkový odpor úseku příze L_1 vč. sumy kontaktních odporů, viz obr. 1. Dle jednoduchého vztahu (viz níže) je možno určit elektrický odpor úseku příze L_0, který není zatížen chybou způsobenou kontaktními odpory.

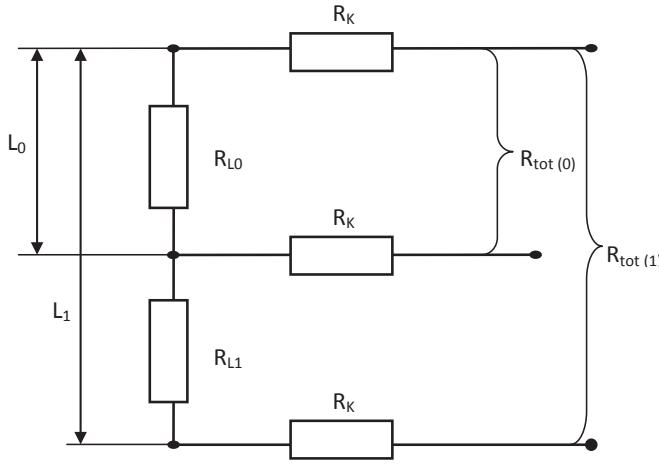
Pro lineární vzrůst elektrického odporu se vzrůstající upínací délkou platí:

$$R_{L1} = R_{L0}\frac{L_1}{L_0} \tag{1}$$

$$R_{tot(L)} = 2R_K + R_{L0}\frac{L_1}{L_0} \tag{2}$$

$$R_{tot(0)} = 2R_K + R_{L_0} \tag{3}$$

Rozdílem naměřených celkových odporů je možno získat elektrický odpor délkového textilních útvaru o upínací délce L_0, který není zařízen chybou kontaktních odporů:



Obrázek 1. Schématický nákres odpovídajícího elektrického obvodu reprezentující měření elektrického odporu délkových textilních útvarů, kde: R_{L0}, R_{L1} [Ω] - elektrický odpor délkového textilního útvaru o upínací délce L_0, resp. L_1 (neznámý), L_0, L_1 [m] - upínací délka textilního útvaru (měřitelné hodnoty), R_K[Ω] - kontaktní odpor (neznámý) a $R_{tot(0)}, R_{tot(1)}$ - celkový elektrický odpor úseku příze o délce L_0, resp. L_1 vč. sumy kontaktních odporů (měřitelná hodnota.)

$$R_{tot(L)} - R_{tot(0)} = R_{L0}\frac{L_1}{L_0} - R_{L0} \tag{4}$$

$$R_{tot(L)} - R_{tot(0)} = R_{L0}\left(\frac{L_1}{L_0} - 1\right) \tag{5}$$

$$\frac{R_{tot(L)} - R_{tot(0)}}{\frac{L_1}{L_0} - 1} = R_{L0} \tag{6}$$

Pro $L_1 = 2L_0$ platí:

$$R_{L0} = R_{tot(2L)} - R_{tot(1L)} \tag{7}$$

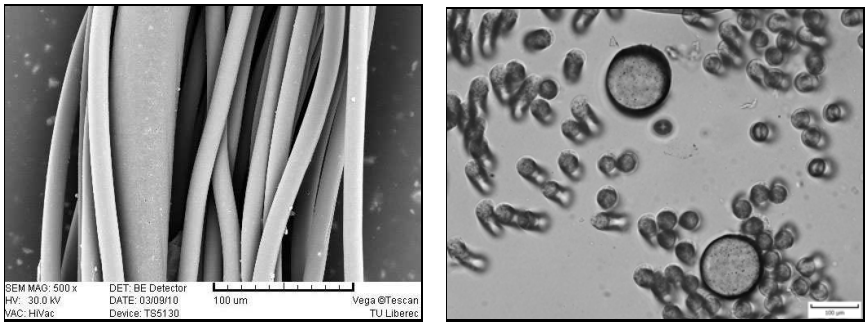
Pro nelineární vzrůst elektrického odporu platí:

$$R_{L1} = \frac{R_{L0}}{n+1}\left(\frac{L_1}{L_0}\right)^{n+1} \tag{8}$$

$$R_{L0} = \frac{\left(R_{tot(2L)} - R_{tot(1L)}\right)(n+1)}{2^{n+1} - 1} \tag{9}$$

4. **Experimentální ověření**

Studovány byly elektrické vlastnosti příze se zvýšenou elektrickou vodivostí, označovány jako antistatické – příze označená Resistat. Tento typ příze je založen na použití bikomponentních vláken. Příze Resistat obsahuje ve své struktuře bikomponentní vlákno Resistat F9601, tj. polyamidové vlákno obsahující na svém povrchu vodivou uhlíkovou vrstvu. Jemnost délkové textilie je 11 tex. Materiálové složení a zastoupení jednotlivých komponent je následující: 81dtexf35PESh+25dtexf1 Resistat F9601. Na obrázku 2 je zobrazen mikroskopický snímek podélného a příčného pohledu na antistatickou přízi Resistat. Je zřejmé, že vodivá komponenta je umístěna kontinuálně v ose celé příze.



(a) (b)

Obrázek 2. Mikroskopické snímky příze Resistat: (a) podélný pohled; (b) příčný řez. Vodivá komponenta Resistat F9601 obsahuje na povrchu PA vlákna uhlíkovou vrstvu.

Připravený vzorek délkové textilie se upevní pomocí kovových svorek do elektrodového systému. Zaznamenávány jsou hodnoty rezistence na upínací délce L_0. Upínací délka L_0 může být např. 50 mm. Po změně upínací délky na např. dvojnásobnou velikost (100 mm) se proměří elektrický odpor totožného vzorku. Pro jednotlivé upínací délky délkové textilie je nutno proměřit alespoň 10 vzorků kvůli statistickému zpracování naměřených dat. Elektrický odpor vzorku délky 50 mm nezatížený chybou způsobenou kontaktními odpory je možno stanovit dle vzorce (7), který je platný pro délkového textilní útvary s lineárním vzrůstem elektrické vodivosti se vzrůstající upínací délkou. Naměřené hodnoty elektrického odporu vzorku antistatické příze při odlišné upínací délky jsou uvedeny v tabulce 1.

Elektrický odpor vzorku délky 50 mm nezatížený chybou způsobenou kontaktními odpory je dle vzorce (7) 1.45E+06 Ω. Je zřejmé, že je hodnota elektrického odporu vzorku nezatíženého chybou (1.45E+06 Ω) při upínací délce 50 mm odlišná od hodnoty, která byla zjištěna přímo (1.61E+06 Ω), tzn.

TABULKA 1. Naměřené hodnoty elektrického odporu v závislosti na upínací délce vzorku.

	$\mathbf{L_1 = 50\ mm}$	$\mathbf{L_2 = 100\ mm}$
$R_{toti}[\Omega]$	1.54E+06	3.16E+06
	1.63E+06	3.02E+06
	1.58E+06	2.96E+06
	1.56E+06	3.05E+06
	1.72E+06	3.03E+06
	1.57E+06	3.10E+06
	1.61E+06	3.16E+06
	1.54E+06	3.02E+06
	1.63E+06	3.09E+06
	1.71E+06	3.04E+06
$R_{tot}[\Omega]$	1.61E+06	3.06E+06
Sm. odch. [Ω]	6.47E+04	6.53E+04

obsahuje chybu způsobenou přítomností kontaktních odporů ve styku svorky se vzorkem ve výši cca 11 %.

5. Závěr

Na základě rozsáhlých experimentů bylo zjištěno, že kontaktní odpor může zvýšit chybu měření až o 10%, což je prakticky velmi omezující. Inovovaná metodika měření elektrického odporu délkových textilních útvarů (založená na měření elektrického odporu na nejméně třech definovaných úsecích délkového textilního útvaru a následný výpočet elektrického odporu nezatíženého chybou kontaktních odporů) předchází problémy způsobené kontaktními odpory a poskytuje měření bez chyb způsobených neideálními kontakty v místě styku vzorku se svorkou s vyšší přesností a reprodukovatelností.

Poděkování: Tato práce vznikla za podpory projektu MPO FR-TI1/122 – Textilie se zvýšeným komfortem odolné vůči elektromagnetickému záření.

Literatura

[1] Morton, W.E., Hearle, J.W.S. *Physical Properties of Textile Fibres.* London: Woodhead Publishing, 2008. 796 s. ISBN 1845692209.

[2] Militký, J., Šafářová, V. Anomalous Electrical Resistance of Hybrid Yarns Containing Metal Fibers. *Proceedings of Fiber Society 2011 Spring Conference.* Hong Kong, 2011.

[3] Šafářová, V., Militký, J. A Study of Electrical Conductivity of Hybrid Yarns Containing Metal Fibers *Journal of Materials Science and Engineering B*, 2(2), 2012, pp. 197-202. ISSN 2161-6221.

ODHAD ORIENTACE VLÁKENNÝCH SYSTÉMŮ ESTIMATION OF FIBRE SYSTEMS ORIENTATION

Maroš Tunák, Jiří Kula, Jiří Chvojka

Adresa: Technická univerzita v Liberci, Fakulta textilní, Studentská 2, 461 17 Liberec; maros.tunak@tul.cz

Abstrakt: Při analýze textilních materiálů se často setkáváme s hodnocením strukturní anizotropie nebo směrového uspořádání textilních objektových systémů. Objekty jako důležité součástí obrazu nás z hlediska dalšího zpracování zajímají a odpovídají konkrétním objektům zobrazovaného světa. Zpracování obrazových dat dovoluje porozumět obsahu a provést kvantitativní nebo kvalitativní popis objektů zájmu v obraze. Příspěvek se věnuje popisu metod založených na použití obrazové analýzy pro odhad směrového rozložení objektů v obraze a vizualizaci odhadu směrové orientace nebo strukturní anizotropie vlákenných systémů.

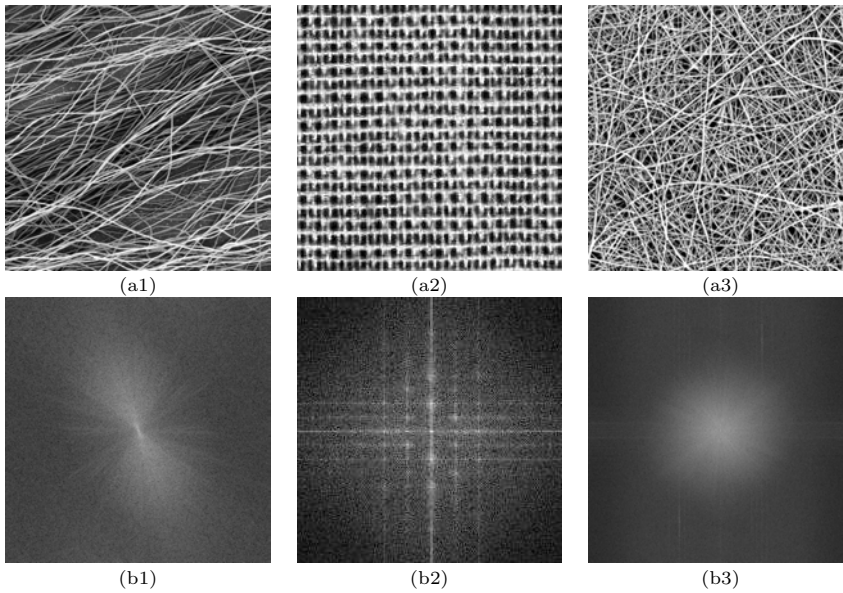
Abstract: Textile materials analysis generally includes measurement of structure anisotropy or directional orientation of textile object systems. Objects represent real-world objects and as an important part of an image are significant for following processing. Processing of image data allows understand of image content and perform quantitative and qualitative description of objects of interest. This contribution deals with the description of methods based on image analysis for estimation of fibre systems orientation and visualisation of such estimation.

Klíčová slova: Vlákenný systém, digitální obraz, Fourierova transformace, momenty obrazové funkce

Keywords: Fibre system, digital image, fourier transform, moments of image function

1. Úvod

V současné době jsou textilní vlákenné materiály používané v mnoha oblastech průmyslu a aplikacích např. materiály pro účely oděvní, účely technické (automobilový průmysl, kompozity, geotextilie, textilie ve stavebnictví atd.), speciální účely (materiály pro medicínu, tkáňové inženýrství atd.). Vlastnosti textilních materiálů v mnohém závisí od vlastností jednotlivých vláken a uspořádání nebo struktury, které jednotlivá vlákna formují. Uspořádání vláken ovlivňuje mechanické vlastnosti délkových a plošných textilií, ve vlákenných porézních materiálech orientace vláken ovlivňuje vlastnosti jako prodyšnost, propustnost a absorpce kapalin atd. Z tohoto důvodu je měření směrové orientace nebo odhad strukturní anisotropie objektových systémů na základě digitálních obrazů důležitou součástí kvantitativního měření v textilní metrologii. Objekty rozumíme ty části obrazu, které nás z hlediska



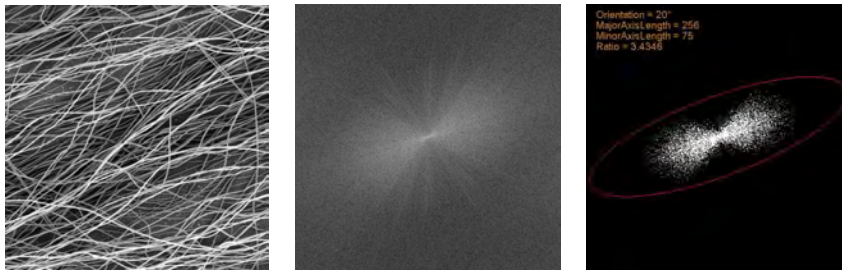
Obrázek 1. (a*) Obrazy textilních objektových systémů, (b*) odpovídající výkonová spektra.

dalšího zpracování zajímají a odpovídají konkrétním objektům zobrazovaného světa. Objekty by měly být v kontrastu s pozadím obrazu (gradient obrazové funkce na hranici objektu a pozadí). V textilní praxi můžeme za objekty považovat vlákna, příze, řezy vláken a pod., systémy obsahující objekty mohou být rouna, vlákenné vrstvy, tkaniny, pleteniny, netkané textilie a např. nanovlákenné vrstvy. V současné době je zkoumání směrových vlastností prováděno převážně manuálně nebo s použitím specializovaného softwaru, kde odhad orientace je zatížen subjektivním pohledem.

2. **Orientace vlákenných systémů**

Charakteristikou anizotropie je úhlová hustota délek nitě $f(\alpha)$ směřujících do úhlového rozmezí $\alpha \pm \alpha/2$. Funkce $f(\alpha)$ se označuje jako směrová růžice. Experimentální grafická metoda pro odhad $f(\alpha)$ je popsaná v práci *Rataje a Saxla* [1]. Metoda využívá síť úhlů $\alpha_1, \ldots, \alpha_n$, umístěných na povrch sledovaného systému k sestrojení průsečíkové růžice (stanovené z počtu průsečíků sítě a vlákenných objektů). Směrová růžice je pak získána z průsečíkové růžice pomocí grafické konstrukce Steinerova kompaktu.

Techniky využívající nástrojů obrazové analýzy založené na spektrálním přístupu, které nejprve převedou texturní obrazy do frekvenční oblasti, jsou vhodné pro popis směrovosti periodických nebo téměř periodických vzorů



(a) (b) (c)

Obrázek 2. (a) Originální obraz, (b) výkonové spektrum, (c) oblast zájmu.

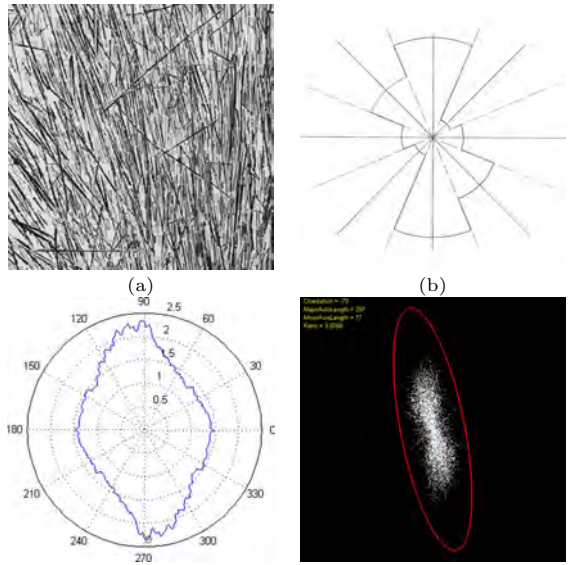
v monochromatických obrazech textury. Tyto techniky jsou založené na vlastnostech Fourierova spektra a popisují globální periodicitu úrovní šedi obrazu. Směr rozložení vysokých hodnot frekvenčních komponent ve frekvenční oblasti odpovídá převažujícím směrům objektů v obraze v prostorové oblasti. Naproti tomu náhodná textura způsobuje, že vysoké hodnoty frekvenčních komponent v obraze spektra jsou rozloženy isotropně a tvoří přibližně kruhový tvar. Výzkumem v oblasti směrové orientace založené na spektrálním přístupu se zabývali i jiní autoři, například *Josso et al.* [2],*Liu* [3], *Holota a Němeček* [4], *Tonar et al.* [5], *Kula* [6].

Nechť $f(x, y)$ je dvojrozměrná obrazová funkce, kde $x = 0, 1, 2, \ldots, m-1$ a $y = 0, 1, 2, \ldots, n-1$ jsou prostorové souřadnice a $f(x, y)$ je úroveň šedi obrazových bodů obrazu o velikosti mxn. Pro takový obraz je dvojrozměrná diskrétní Fourierova transformace (2DFT) dána vztahem [9]

$$F(u,v) = \sum_{x=0}^{m-1} \sum_{y=0}^{n-1} f(x,y)\, e^{-j\,2\,\pi(ux/m+vy/n)}, \tag{1}$$

kde, $u = 0, 1, 2, \ldots, m-1$ a $v = 0, 1, 2, \ldots, n-1$ jsou frekvenční proměnné. $F(0, 0)$ představuje počátek frekvenční oblasti. Jestliže $f(x, y)$ je reálná funkce, její transformace je funkce komplexní. Z důvodu vizuální analýzy transformace je vhodné vypočíst její spektrum $|F(u, v)|$ a zobrazit jako obraz. Výkonové spektrum je definované jako druhá mocnina $|F(u, v)|$, tj. $P(u, v) = |F(u, v)|^2$. Pro účely vizualizace je vhodné zredukovat dynamický rozsah koeficientů logaritmickou transformací $Q(u, v) = log(1 + P(u, v))$. Příklad textilních objektových systémů ve formě šedotónových obrazů a jejich odpovídající výkonová spektra jsou zobrazeny na obrázcích 1(a*)-(b*).

Metoda pro odhad směrového rozložení objektů založená na spektrálním přístupu je uvedena v práci *Tunáka a Linky* [7]. Odhad anisotropie je vypočten jako suma frekvenčních komponent směrového vektoru v celém rozsahu úhlů. Odhad souřadnic směrového vektoru je proveden pomocí DDA algoritmu. Protože transformace reálné funkce $f(x, y)$ je komplexní číslo, jsou



Obrázek 3. (a) Brodatzova textura (D15), (b) směrová růžice, (c) polární diagram, (d) orientace podle elipsy.

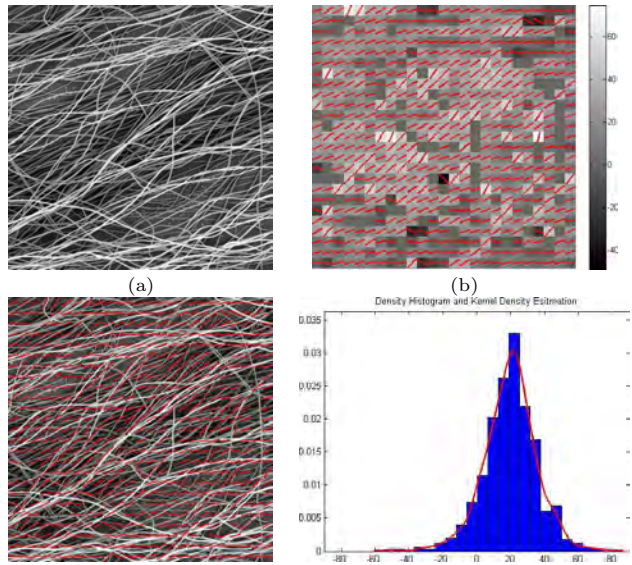
sečteny koeficienty Fourierova spektra $|F(u,v)|$ a vyneseny do polárního diagramu. Výhodou metody je její rychlost a sledování orientace s úhlovým krokem 1°. V práci jsou uvedeny příklady odhadu směrového rozložení pro simulované obrazy a monochromatické obrazy textilních struktur.

3. **Momenty obrazové funkce**

Jak už bylo zmíněno, směr frekvenčních komponent ve frekvenční oblasti koresponduje ze směrem hran objektů v prostorové oblasti. Další metoda pro odhad anisotropie vychází z transformace výkonového spektra do binárního obrazu prahováním, tím dojde k odsegmentování významných frekvenčních komponent. V takovýchto binárních obrazech uvažujeme shluk bílých pixelů jako oblast zájmu. Pro jednoduchý popis objektů nebo oblasti zájmu v segmentovaném obrazu je možné využít obrazové momenty. 2D obecný moment řádu $(p+q)$ obrazové funkce $f(x,y)$ je dán [8]

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y). \tag{2}$$

Hodnoty obecných momentů určují zajímavé charakteristiky objektů v obraze, například moment m_{00} je pro binární obraz plocha objektu, podíly momentů prvního řádu m_{10} a m_{01} s momentem m_{00} určují těžiště objektu.



Obrázek 4. (a) Originální obraz, (b) šedotónová mapa směrů, (c) vektory směrů, (d) histogram a jádrový odhad hustoty distribuce směrů.

Dalším typem jsou centrální momenty řádu $(p+q)$

$$\mu_{pq} = \sum_x \sum_y (x-\overline{x})^p (y-\overline{y})^q f(x,y), \tag{3}$$

kde

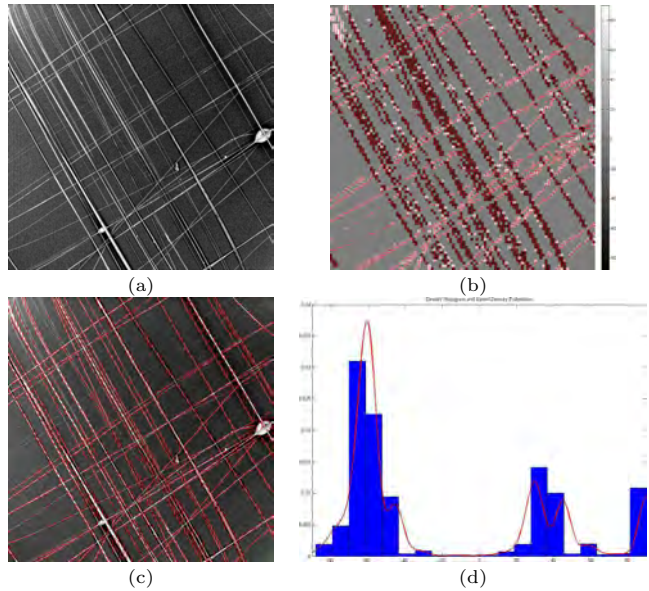
$$\overline{x} = \frac{m_{10}}{m_{00}}, \qquad \overline{y} = \frac{m_{01}}{m_{00}}. \tag{4}$$

Centrální momenty prvního a druhého řádu jsou vhodné pro popis orientace objektu nebo oblasti zájmu, orientace objektu nebo oblasti zájmu mezi osou x a hlavní polosou elipsy je dána [8]

$$\theta = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right). \tag{5}$$

V binárních obrazech je možné určit vlastnosti jako délku hlavní, vedlejší osy a orientaci elipsy (úhel ve stupních v intervalu -90 až 90° mezi osou x a hlavní osou elipsy), která má stejný normalizovaný druhý centrální moment jako oblast zájmu. Orientace koresponduje s převládajícími směry objektů v prostorové oblasti (viz obrázek 2, výkonové spektrum je pootočené o $\pi/2$).

Porovnání výše zmíněných metod je uvedeno na obrázcích 3(a)-(d). Obrázek 3(a) zobrazuje Brodatzovu texturu (D15), obrázek 3(b) představuje směrovou růžici odhadnutou experimentální metodou dle *Rataje a Saxla*, na



Obrázek 5. (a) Originální obraz, (b) šedotónová mapa směrů, (c) vektory směrů, (d) histogram a jádrový odhad hustoty distribuce směrů.

obrázku 3(c) je odhad směrové růžice ve formě polárního diagramu podle *Tunáka a Linky*. Elipsu zobrazenou červenou barvou, délku hlavní a vedlejší osy elipsy, a orientaci elipsy je možné vidět na obrázku 3(d). Z obrázků je patrná shoda v převládajícím směru.

Obě navržené metody provádí odhad směrového rozložení objektů pro monochromatický obraz jako celek. Ukazuje se, že pro textilní vlákenné systémy např. netkané textilie nebo nanovlákenné vrstvy by byla vhodnější podrobnější analýza. Myšlenka je založena na rozdělení obrazu na menší části a provedení analýzy pro takovéto oblasti. Obraz viskózových vláken o velikosti 500x500 pixelů se zřejmou preferencí směrů uvedený na obrázku 4 (a) je rozdělen na menší podokna určité velkosti. Analýza směrového uspořádání je pak provedena pro každé podokno. Převládající orientace objektů pro každé podokno je reprezentována směrovým vektorem zobrazeným červenou barvou (při podmínce, že poměr hlavní a vedlejší osy elipsy je větší než 2). Kromě toho, orientace ve stupních je zobrazena jako mapa v šedé škále. Obrázek 4(b) představuje šedotónovou mapu orientace pro podokna velikosti 20x20. Na obrázku 4(c) jsou vyznačeny vektory směrů v originálním šedotónovém obraze. Odpovídající distribuce směrů je uvedena ve formě histogramu a jádrového odhadů hustoty na obrázku 4(d). Výsledky ukazují, že menší velikost podoken poskytuje přesnější výsledky.

Příklad nanovlákenné vrstvy, kde jsou vidět dva preferované směry nanovláken je uveden na obrázku 5. Je zřejmé, že oblasti zobrazené střední šedou, které neobsahují vektory směrů, reprezentují oblast bez preferovaného směru (poměr hlavní a vedlejší poloosy elipsy je menší než 2).

Metodu pro hodnocení anizotropie nebo směrové orientace vlákenných nebo jiných objektových systémů za pomoci 2DFT je možné využít pro hodnocení plošných textilních struktur z pohledu jejich homogenity, vad a náhodných odchylek od struktury.

4. Závěr

Obsahem příspěvku jsou metody obrazové analýzy pro monitorování struktury textilních útvarů. Obrazová analýza má v oblasti monitorování kvality velmi důležité místo, protože poskytuje informaci o geometrii, povrchu, defektech, o úpravách povrchu výrobku a jiných charakteristikách. Získané výsledky dokumentují, že uvedené postupy lze použít pro monitorování strukturní anizotropie nebo směrové orientace vlákenných a jiných objektových systémů. U délkových a plošných textilních útvarů je nezbytně nutné zajistit stabilitu kvalitativních charakteristik těchto struktur. K tomuto účelu je nutné z těchto struktur odhadnout rozdělení směrového uspořádání vlákenného materiálu v ploše. V příspěvku uvedené postupy umožňují tyto kvalitativní charakteristiky monitorovat.

Literatura

[1] Rataj J., Saxl I.. Analysis of Planar Anisotropy by Means of Steiner Compact: A Simple Graphical Method. *Acta Stereologica* 7/2: 1988, 107–112.

[2] Josso B., Burton R., Lalor J. Texture Orientation and Anisotropy Calculation by Fourier Transform and Principal Component Analysis. *Mechanical Systems and Signal Processing*, 19(2), 2005, pp. 1152–1161.

[3] Liu, Z. Q. Scale Space Approach to Directional Analysis of Images. *Applied Optics*, 30(11), 1991, pp. 1369–1373.

[4] Holota, R. and Němeček, S. Recognition of Oriented Structures by 2D Fourier Transform. In: *Applied Electronics 2002*. Plzeň, Czech Republic, 2002.

[5] Tonar Z. et al. Microscopic Image Analysis of Elastin Network in Samples of Normal Atherosclerotic and Aneurismatic Abdominal Aorta and its Biomechanical Implications. *Journal of Applied Biomedicine* 1:149–159.

[6] Kula, J. Segmentace objektů z obrazu vlákenné struktury na základě jejich orientace. In: *Workshop pro doktorandy FS a FT TUL*. Rokytnice nad Jizerou, Czech Republic, 2011.

[7] Tunák M., Linka A. Planar Anisotropy of Fibre System by Using 2D Fourier Transform. *Fibers and Textiles in Easter Europe* 15/5-6:86–90.

[8] Kilian J. *Simple Image Analysis by Moments*. Cranfield University, UK, 2001.

[9] Gonzales R., Woods R.E., Eddins, S. *Digital Image Processing using MATLAB*. Pearson Prentice-Hall, 2004.

Poděkování: Příspěvek vznikl za podpory projektu 1M06047 Centrum pro jakost a spolehlivost výroby.

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Sokolovská 83, 186 00 Praha 8. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in bulletin are published in English, Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toho číslo bylo vytištěno s laskavou podporou JČMF, konference ROBUST
a s podporou projektu OPVK Klimatext č. CZ.1.07/2.3.00/20.0086.