INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 29, číslo 3, září 2018

STATISTICKÁ ANALÝZA ZÁVISLOSTÍ MEZI RŮZNÝMI TYPY DOKUMENTŮ PODANÝCH NA OBECNÍ ÚŘADY V ČESKÉ REPUBLICE

STATISTICAL ANALYSIS OF DEPENDENCIES AMONG SUBMISSIONS TO MUNICIPALITIES IN THE CZECH REPUBLIC

Anna Pidnebesna[1,2,3], Kateřina Helisová[1], Jakub Staněk[4]

Adresa: [1]FEL ČVUT v Praze, Technická 2, 166 27 Praha 6; [2]ÚI AV ČR, Pod Vodárenskou věží 271/2, 182 07 Praha 8; [3]NÚDZ, Topolová 748, 250 67 Klecany; [4]MFF UK v Praze, Sokolovská 83, 186 75 Praha 8

E-mail: pidnebesna@cs.cas.cz, heliskat@fel.cvut.cz, stanekj@karlin.mff.cuni.cz

Abstrakt: V článku je prezentována statistická analýza závislostí mezi různými typy dokumentů podaných na obecní úřady v České republice. Cílem je popsat chování jednotlivých typů dokumentů (elektronické, podané osobně, zaslané poštou atd.), nalézt model jejich závislostí a prostudovat vliv změn v zákonech České republiky na jejich počet. K tomu jsou použity metody pro analýzu vícedimenzionálních časových řad, konkrétně lineární a korelační analýza a metody k nalezení bodu změny (change point detection). Dále pak je aplikován model gamma rozdělení na počet podaných dokumentů za měsíc. Získané výsledky jsou stručně vysvětleny a okomentovány.

Klíčová slova: Bod změny (change point), gamma rozdělení, obecní úřad, podaný dokument, vícedimenzionální časová řada.

Abstract: A statistical analysis of submissions (forms, emails, letters etc.) to municipalities in the Czech Republic is presented. The aim is to describe the behaviour of different types of the submissions (electronic, personal, sent by post etc.), provide a model of dependencies among them and study the influence of changes in laws in the Czech Republic to them. Methods for multiple time series are chosen as the main tool, namely linear and correlation analysis, and methods for detecting change points are used. Further, the number of submissions per month is modelled by gamma distribution. The obtained results are briefly commented and explained.

Keywords: Change Point Detection, Gamma Distribution, Multiple Time Series, Municipality, Submission.

1. Introduction

Effective planning and logistic are important parts of administrative processes in municipalities. The knowledge of expected amount of such processes can make the administration more efficient. The presented paper focuses on this problem. It concerns modelling of numbers of submissions to municipalities in the Czech Republic. By submissions, we mean all documents delivered to the municipalities, for example emails, several letters like notifications or complaints, several forms like application forms or forms for tax return submitted by data boxes or personally etc. The submissions are divided into several groups according to the way of their processing, so the forecast of evolution of their amount can help the municipalities to organise their inner structures.

The presented analysis is based on data containing the numbers of different types of submissions in the time period from January 2002 to December 2011, see Section 2. The data were first analysed in [13], where the importance of such analysis for public administration is explained in details. The authors analyse evolution of the municipality communications in time finding trends and seasonal components. Then in [14], spatial distribution and interactions of municipalities are studied. Further in [16], the data are considered to be spatio-temporal point patterns and methods for spatio-temporal point processes are used. However, the authors noticed that the data can be modelled neither in the continuous domain nor by classical methods on a grid, moreover, spatial and temporal coordinates of the process are not independent and number of the submissions cannot be described by Poisson distribution. So finally, a discrete spatio-temporal model based mainly on empirical distributions was fitted. The authors have also found a natural similarity in behaviour in specific days (Mondays and Wednesdays with usually longer office hours have similar numbers of submissions and, of course, the numbers differ from the other days), and a periodicity by months. The number of submissions per month is the main variable in the present paper, too.

Here, we focus on different groups of the submission according to the way of their processing and study their mutual dependencies as well as the influence of a change in law which order to all municipalities to keep data boxes. This is an important intervention to the system, which may have changed the submission evolution. Based on this fact, we decided to apply change point methods (see e.g. [2]) to detect whether significant changes in evolution of numbers of different submission types occurred, together with correlation analysis (see e.g. [21]) and linear regression methods (see e.g. [7], [8]). Finally, we suggest a probability distribution for the number of submissions

per a given time period. Comparing to [16] where these numbers follow their empirical distributions, here we noticed that they can be modelled by gamma distribution.

The paper is organised as follows. The data are described in Section 2. Section 3 introduces the methodology consisting of change point detection (Subsection 3.1), correlation analysis and linear regression (Subsection 3.2), and fitting gamma distribution (Subsection 3.3). Application of the methods to the data is shown in Section 4. The results are summarised in Section 5.

The numerical results and the plots introduced in this paper are processed in Microsoft Excel and Wolfram Mathematica.

2. Data description

Municipalities play the role of local administrative units of the Czech Republic. One of their main tasks is processing of different submissions like emails, letters, forms submitted personally or through data boxes etc. Here, we consider the submissions to be divided into five groups according to the way of their processing as follows:

- type 1: documents submitted electronically signed by an advanced electronic signature (containing mainly data boxes),
- type 2: documents submitted electronically without advanced electronic signature,
- type 3: physical documents sent by post,
- type 4: physical documents submitted in person,
- type 5: others.

Moreover, the procedures presented below are applied to received and sent submissions separately.

The original data, created with electronic records management systems as arrays of submission records for selected municipalities, include the information about the date of sending or receiving the submission, addressee of sender, agenda and way of communication. We have grouped the submissions due to the month of sending or receiving, respectively, since as mentioned in [16], one month is a suitable unit for modelling the number of submissions. The agenda and the way of communication then determine the type of submission listed above. Thus, the data are represented as multivariate time series $\{t_1(i), t_2(i), t_3(i), t_4(i), t_5(i),\ i = 1, \ldots, n\}$, where $t_1(i), \ldots, t_5(i)$ denote the numbers of submissions of type 1,..., type 5, respectively, in the

i-th month, $i = 1, \ldots, 120$, corresponding to the months from January 2002 to December 2011.

Since different months have different numbers of working days and moreover, the numbers of submissions on Mondays and Wednesdays, which are usually longer office days at municipalities, differ from the numbers of submissions on remaining days, we apply a standardisation. The standardisation is based on the assumption that every month has 20 working days consisting of 4 Mondays, 4 Tuesdays, 4 Wednesdays, 4 Thursdays and 4 Fridays. We calculate $\overline{OD}$ the mean number of submissions on the longer office days (Monday and Wednesday) and $\overline{RD}$ the mean number of submissions in the remaining working days. Then we consider the number of submissions in the given month as $8 \times \overline{OD} + 12 \times \overline{RD}$.

The analysis is based on the collection of 36 datasets, more precisely on data from 18 municipalities, each containing both the sent and the received submissions. For presentation of the results, the received submissions of one concrete municipality are chosen as an example (called "Municipality 1" in the sequel). Fig. 1 shows temporal evolution of the numbers of particular submission types per month for Municipality 1. Note that the behaviour of the remaining time series is very similar.

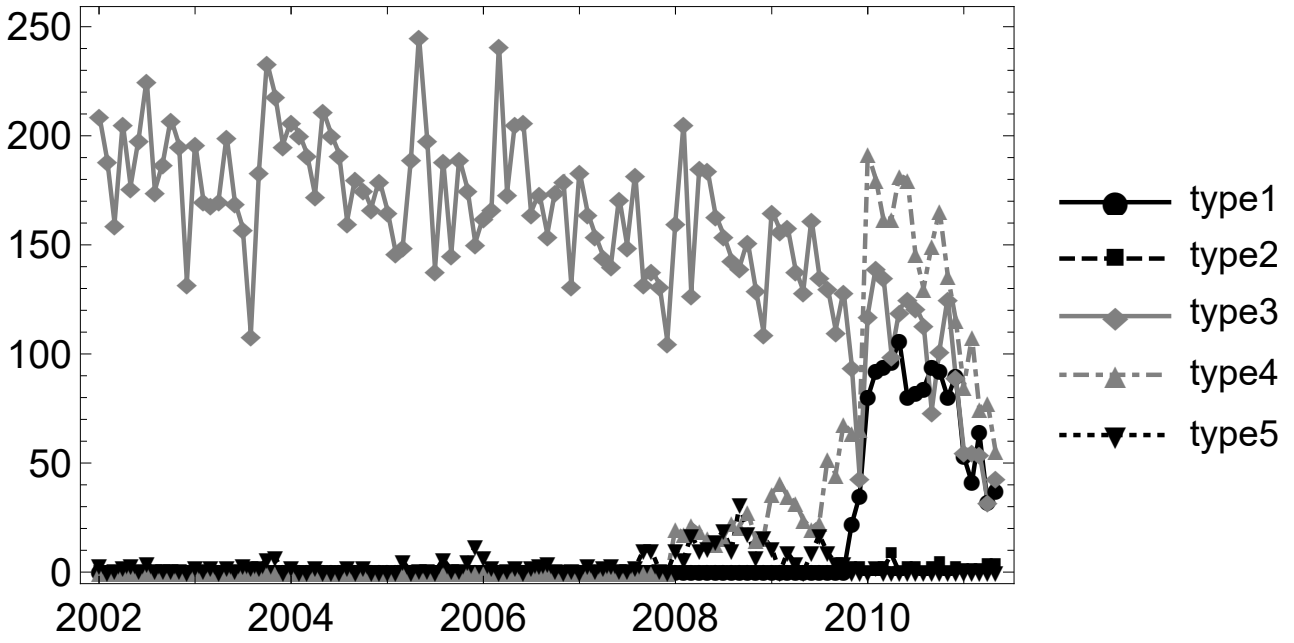


Figure 1: Temporal evolution of the numbers of particular submission types per month in the time period from January 2002 to December 2011 in the municipality whose results are introduced in this paper in details (called "Municipality 1" in the text).

3. Methodology

The main idea is the following. Since in the last years, there were changes in laws, especially the obligation to keep data boxes from the beginning of 2010, first, we detect whether there are significant changes in behaviour of the time series $\{t_1(i), t_2(i), t_3(i), t_4(i), t_5(i),\ i = 1, \ldots, n\}$. Then we focus on the series after the last change and using linear regression methods, we find a model of their mutual dependences. Finally, we fit the model for the number of submissions per month. Comparing to [16], where the authors conclude that the number of submissions cannot be modelled by Poisson distribution as expected and deal with empirical distributions, the present study suggest to apply the gamma distribution.

The theoretical background of the methodology is introduced in the following subsections.

3.1. Change points

There exists a wide class of change point detection methods and approaches (see e.g. [2], [10], [15]). In our study, we test whether the parameters of linear regression are changed during the observed period. We use a flexible change point detection technique. The method assumes that the number of submissions in time has partly linear trends. Thus, we test the null-hypothesis that there are no changes in the linear regression parameters against the hypothesis that there exists one change. The advantages of such an approach is simplicity of the method and easy calculations while the assumption of the linear trend is not too binding, because we care of the change points only and not of the trend itself.

In general, the null hypothesis is formulated as

$$H_0 : Y_i = a + bx_i + e_i, \quad i = 1, \ldots, n,$$

while the alternative hypothesis is

$$\begin{aligned} H_A : \exists m \in \{2, \ldots, n-2\} : Y_i &= a + bx_i + e_i, \quad i = 1, \ldots, m, \\ Y_i &= a^0 + b^0 x_i + e_i, \quad i = m+1, \ldots, n, \end{aligned}$$

where $a \neq a^0$ or $b \neq b^0$ are the coefficients of linear regression, $Y = \{Y_1, \ldots, Y_n\}$ is the dependent variable, $x = \{x_1, \ldots, x_n\}$ is the explanatory variable and the noise $e = \{e_1, \ldots, e_n\}$ is formed by i.i.d. random variables with the mean $\mathrm{E}e_i = 0$ and the variance $\mathrm{var}\, e_i = \sigma^2 > 0$. In our case, the dependent variable $Y = t_j$, $j = 1, \ldots, 5$, is used for the numbers of particular submission types and the explanatory variable x is time [in months], i.e. $x_i = i$,

$i = 1, \ldots, n$. Note that we cannot expect e_i, $i = 1, \ldots, n$, to be normal, because our study shows that Y_i, $i = 1, \ldots, n$, are rather gamma distributed (see below). However, due to convergence of gamma distribution to normal distribution and with respect to our purposes, the assumption of strict normality is not needed.

Denote

$$X_k = \begin{pmatrix} 1 & \cdots & x_1 \\ \ldots & \ldots & \ldots \\ 1 & \cdots & x_k \end{pmatrix}, \quad X_k^o = \begin{pmatrix} 1 & \cdots & x_{k+1} \\ \ldots & \ldots & \ldots \\ 1 & \cdots & x_n \end{pmatrix}$$

and

$$\chi_k^2 = \frac{1}{\sigma^2}(\widehat{a}_k - \widehat{a}_k^0, \widehat{b}_k - \widehat{b}_k^0)\big((X_k^T X_k)^{-1} + (X_k^{0\,T} X_k^0)^{-1}\big)^{-1}(\widehat{a}_k - \widehat{a}_k^0, \widehat{b}_k - \widehat{b}_k^0)^T,$$

where $\widehat{a}_k$, $\widehat{a}_k^0$, $\widehat{b}_k$ and $\widehat{b}_k^0$ are estimates obtained by least squares under the alternative with $m = k$, and A^T denotes transposition of matrix A. Then under the null-hypothesis, for the maximum-type test statistics, it holds (see [2]) that

$$P\left(\max_{2 \le k \le n-2} \{\chi_k^2\} > \left(\frac{x + b_n}{a_n}\right)^2\right) \approx 1 - \exp\{-\mathrm{e}^{-x}\}, \quad x \in \mathbb{R}^1,$$

where

$$a_n = \sqrt{2 \log \log n} \quad \text{and} \quad b_n = 2 \log \log n + \frac{1}{2} \log \log \log n.$$

In the case that we reject the null-hypothesis, we are interested in estimation of the change point m and the parameters of linear regression. The estimate of the change point m is defined as

$$\widehat{m}_{\mathrm{regr}} = \arg\min \left\{ \sum_{i=1}^{k} (Y_i - \widehat{a}_k - \widehat{b}_k x_i)^2 + \sum_{i=k+1}^{n} (Y_i - \widehat{a}_k^0 - \widehat{b}_k^0 x_i)^2;\ k = 1, \ldots, n \right\},$$

where $\widehat{a}_k$, $\widehat{b}_k$ and $\widehat{a}_k^0$, $\widehat{b}_k^0$ are least squares estimates of a, b and a^0, b^0 based on $Y_1, \ldots, Y_n$.

Note that this approach is suitable for detecting and estimating parameters of a model with not more than one change point. However in our case, we would like to obtain more change points, because we focus on whether there is a change at the beginning of 2010, independently on whether the change is

the most important. There exist methods that allow to detect more change points in the time series (see e.g. [2], [12] or [9]), but they are based on very time-consuming calculations. Therefore, we use the following simplification: for all time series, we detect three change points using the basic algorithm three times – for whole observed period, and for the parts before and after the first detected change point. We are aware that we eliminate the situations when both the second and the third change points lie before or after the first change point, respectively, but this simplification is satisfactory for our purposes, because we are looking for the change points just as a rough guess.

3.2. Analysis of type dependencies

The second task is to find dependencies among the particular types. For this purpose, we first calculate the correlations and consequently, we try to make linear analysis in which different types are employed. All calculations are done for two time periods: for the whole time period and after the beginning of 2010, when all municipalities had to establish data boxes.

The correlation coefficient for each pair of types is calculated by the Spearman rank correlation (see [21]) instead of classical Pearson correlation because the data do not have the normal distribution and may include outliers. The Spearman rank correlation is defined as the Pearson correlation coefficient between the ranked variables:

$$r_s = \frac{\mathrm{cov}(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}},$$

where X and Y are random variables, rg_X and rg_Y are corresponding ranked variables, $\mathrm{cov}(rg_X, rg_Y)$ and $\sigma_{rg_X}, \sigma_{rg_Y}$ are covariance and standard deviations, respectively, of the ranked variables.

Finally, we provide a model of dependencies among the types. As mentioned above, we use the linear regression. Since the most important changes came with the obligation of data boxes, which belong to the type 1, we focus on modelling dependence of the type 1 on the other types while we observe its dependence on the most correlated types, see Section 4.

3.3. Modelling the number of submissions by gamma distribution

As mentioned above, the authors of [16] try to model the number of submissions in a given period by Poisson distribution, but the hypothesis of Poisson

distribution is rejected. However in practice, there are other distributions used for describing numbers of events. For example in insurance companies, gamma distribution is used to model the number of insurance claims. In some sense, the behaviour of the number of submissions can be compared to the behaviour of the number of insurance claims. Therefore, we try to approximate the distribution of the number of submissions per month by gamma distribution $\Gamma(\alpha, \beta)$.

The density of gamma distribution $\Gamma(\alpha, \beta)$ is

$$f(x) = \frac{x^{\alpha-1}\mathrm{e}^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \quad x > 0,$$

where $\alpha > 0$ and $\beta > 0$ are the shape and the scale parameters, respectively. The estimation of the parameters is quite complicated because it requires usage of numerical methods. More information about parameters estimation can be found e.g. in [5].

An easy way is to use the moments method, but its efficiency is low. Another method is the maximum likelihood estimate (MLE). The equations for MLE estimates of α and β are

$$n^{-1}\sum_{i=1}^{n}\log X_i = \log\widehat{\beta}_{\mathrm{MLE}} + \psi(\widehat{\alpha}_{\mathrm{MLE}}), \quad \overline{X} = \widehat{\alpha}_{\mathrm{MLE}}\widehat{\beta}_{\mathrm{MLE}},$$

or equivalently

$$R_n = \log\widehat{\alpha}_{\mathrm{MLE}} - \psi(\widehat{\alpha}_{\mathrm{MLE}}), \quad \widehat{\beta}_{\mathrm{MLE}} = \overline{X}/\widehat{\alpha}_{\mathrm{MLE}},$$

where $\overline{X}$ is the average of the data, $R_n = \log(\overline{X}/\widetilde{X})$, $\widetilde{X}$ is the geometric mean of the sample data, and $\psi(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\log\Gamma(\alpha)$. However, it is known that MLE of the shape parameter is significantly biased, especially for small samples (see e.g. [6], [11], [1]). Therefore, we use a modification of MLE.

There exist various modifications of MLE (see e.g. [18], [17], [20]). We use the approach described in [22], which reduces the bias and improves the efficiency of the estimate. The parameters α and β are estimated as

$$\widehat{\alpha}^{*}_{\mathrm{MLE}} = \frac{n}{n+4.6}\widehat{\alpha}_{\mathrm{MLE}} + \frac{0.54}{n+4.6}, \quad \widehat{\beta}^{*}_{\mathrm{MLE}} = \widehat{\beta}_{\mathrm{MLE}},$$

while the values 4.6 and 0.54 are obtained as the regression coefficients from the equation

$$\mathrm{E}(\widehat{\alpha}_{\mathrm{MLE}}) - \alpha \approx \frac{a\alpha}{n} + \frac{b}{n}$$

for minimising the bias of the estimate $\widehat{\alpha}$.

The last task we have to solve is the problem of outliers, which is needed for cleaning the data before testing the hypothesis of gamma distribution. Here, we use the method based on the interquartile range, so called Tukey's test (see [19]), considering the point x to be an outlier if

$$x \notin \big[Q_1 - c(Q_3 - Q_1), Q_3 + c(Q_3 - Q_1)\big],$$

where Q_1 and Q_3 are the lower and upper quartiles, respectively, and c is a non-negative constant, while we use $c = 1.5$ proposed in [19]. When a point is detected to be an outlier, we simply skip it from the dataset.

4. Numerical results

Recall that for transparency, most numerical results are introduced only for Municipality 1, while the results for the remaining municipalities are very similar.

4.1. Change points

Change point detection is applied to all types of submissions. The results for Municipality 1 for the most important types, i.e. for the type 1 and the type 3, in the whole time period from January 2002 to December 2011 are shown in Fig. 4.

Further in this section, we deal only with the time period from January 2007 to December 2011, since we have not enough data for the types 1, 2, 4 and 5 till December 2006 (see Fig. 1). For each type, we observe the proportions of municipalities (in %) in which the change point occurred in a given quarter, while we consider all municipalities together. As seen from Table 1 and Fig. 2, the most frequent change point appears in the third quarter of 2009 for the types 1 and 3. It can be explained by establishing data boxes, because in this period, electronic submissions signed by an advanced electronic signature (type 1) started to repress the classical post (type 3) and personal submissions (type 4).

Table 1: The proportions of municipalities (in %) in which a change point occurred in the given quarter (for all municipalities together) in the time period from January 2007 to December 2011.

Changes ratio	Type 1	Type 2	Type 3	Type 4	Type 5
1st quarter 2007	0 %	0 %	5 %	13 %	10 %
2nd quarter 2007	14 %	7 %	14 %	12 %	0 %
3rd quarter 2007	17 %	13 %	0 %	6 %	0 %
4th quarter 2007	0 %	0 %	13 %	7 %	17 %
1st quarter 2008	0 %	0 %	0 %	13 %	0 %
2nd quarter 2008	14 %	0 %	17 %	6 %	8 %
3rd quarter 2008	20 %	12 %	8 %	0 %	0 %
4th quarter 2008	0 %	7 %	0 %	0 %	0 %
1st quarter 2009	11 %	6 %	8 %	10 %	0 %
2nd quarter 2009	0 %	%0	12 %	0 %	0 %
3rd quarter 2009	100 %	6 %	37 %	5 %	8 %
4th quarter 2009	11 %	5 %	13 %	17 %	7 %
1st quarter 2010	21 %	9 %	3 %	4 %	8 %
2nd quarter 2010	7 %	0 %	6 %	3 %	0 %
3rd quarter 2010	16 %	0 %	9 %	13 %	0 %
4th quarter 2010	12 %	0 %	15 %	14 %	12 %
1st quarter 2011	12 %	4 %	12 %	3 %	0 %
2nd quarter 2011	3 %	0 %	6 %	6 %	7 %
3rd quarter 2011	5 %	7 %	0 %	0 %	13 %
4th quarter 2011	0 %	0 %	0 %	0 %	0 %

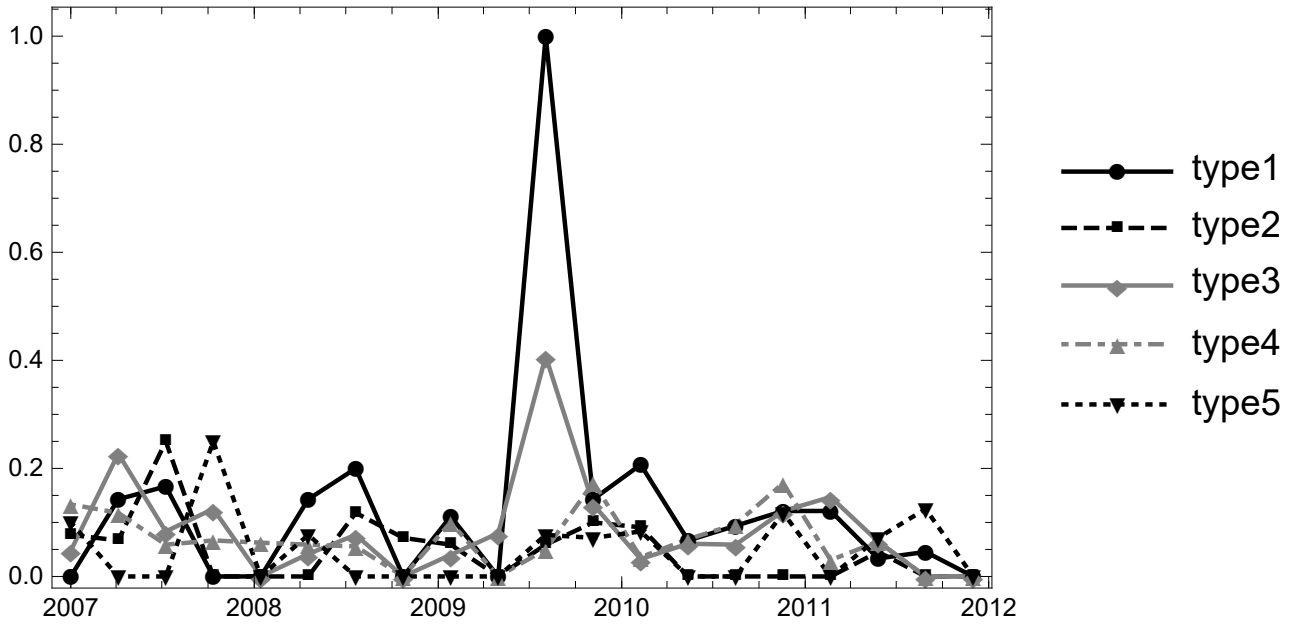


Figure 2: The proportions of municipalities in which a change point occurred in the given quarter (for all municipalities together) in the time period from January 2007 to December 2011.

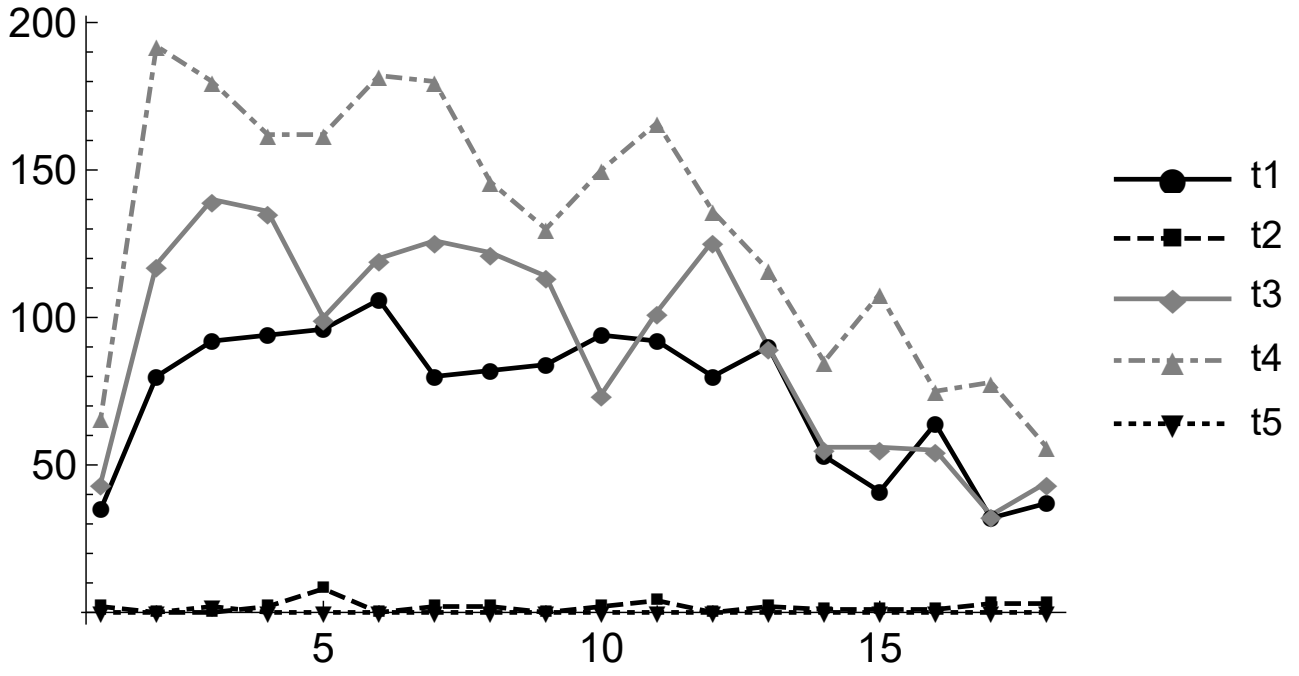


Figure 3: The numbers of all submission types per month from January 2010.

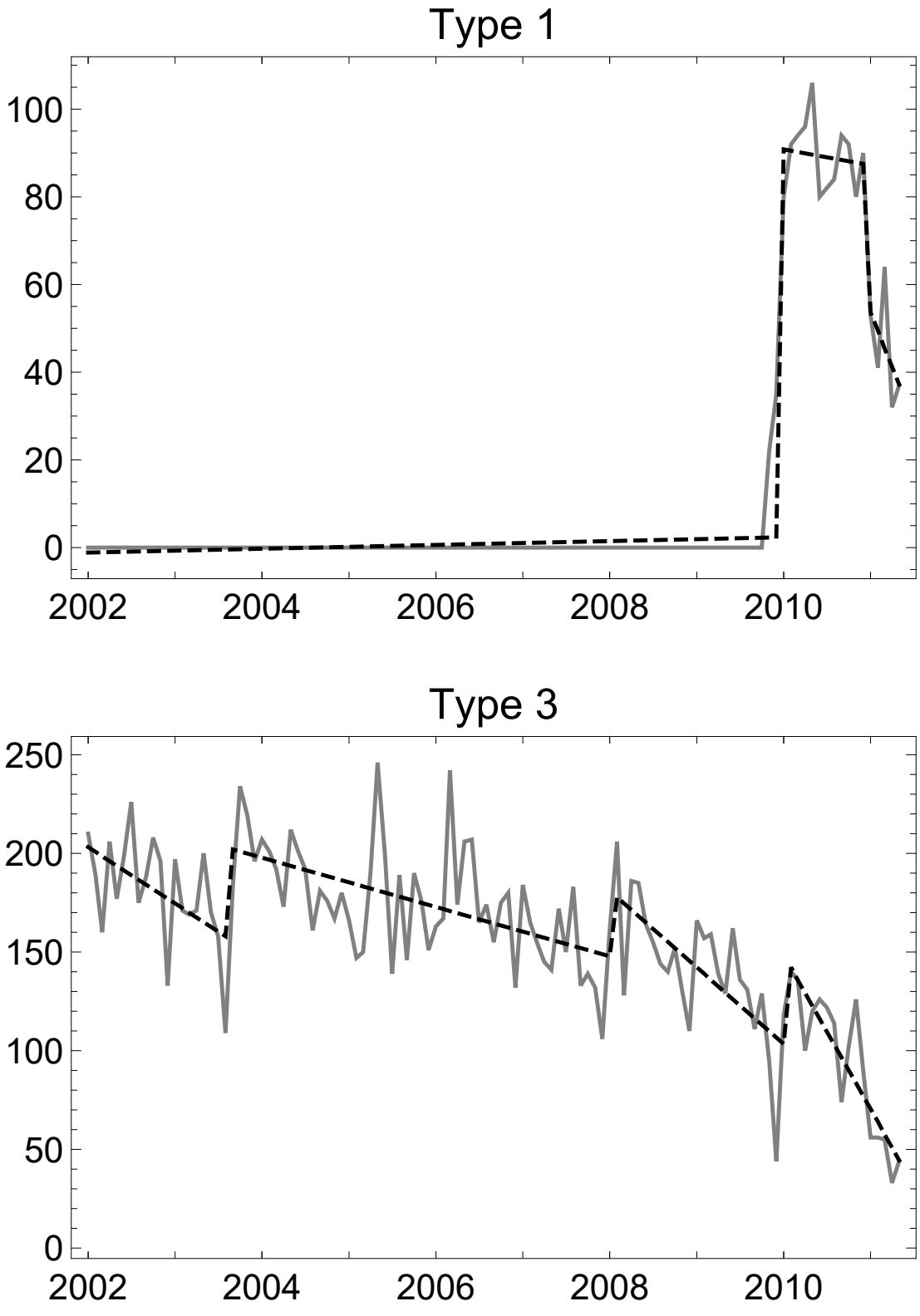


Figure 4: Linear approximation (black dashed line) of the number of submission of the types 1 and 3 (gray solid line) after application of change point detection in the time period from January 2002 to December 2011 in Municipality 1.

4.2. Analysis of type dependencies

The result from Subsection 4.1 led us to the idea that the type 1 and type 3 are dependent. However, it may lead to other dependencies among the types, too. Therefore, we first calculate the correlation coefficients. The coefficients for the whole time period from January 2002 to December 2011 are introduced in Table 2. It is seen that most of them are statistically significant according to the Spearman rank correlation test at 1% level. Indeed, the coefficient of correlation between the types 1 and 3 is negative which confirm our idea that establishment of data boxes caused increase of the number of electronic submissions signed by an advanced electronic signature and decrease of the number of physical submissions. The coefficients for the time period from January 2010 to December 2011, i.e. for the period after establishing data boxes, are different, see Table 3. In this period, the significant correlation coefficients are only that ones between the types 1 and 3, 1 and 4, and 3 and 4. Moreover, comparing to the whole time period, the most of the correlation coefficients are positive which implies similar behaviour of the mentioned types.

For better imagination of the behaviour, we draw the scatter plots of significantly correlated pairs of types of submissions, see Fig. 5. Since the type of dependence is not very clear, we try to fit the model of linear dependence. We focus on modelling the dependence of the type 1 on the type 3, further on the dependence of the type 1 on the types 3 and 4, and finally, we compare it to the model of dependence of the type 1 on all the types 2–5, i.e. we fit the models $t_1 = at_3 + b$, $t_1 = at_3 + bt_4 + c$ and $t_1 = at_2 + bt_3 + ct_4 + dt_5 + e$, respectively. The fitted models are

$$t_1 = 0.55t_3 + 23.24, \tag{1}$$

$$t_1 = 0.23t_3 + 0.45t_4 + 14.13, \tag{2}$$

$$t_1 = 1.30t_2 + 0.31t_3 + 0.25t_4 - 3.58t_5 + 9.98. \tag{3}$$

However, according to 95%-level confidence intervals, only the first model has the coefficients significantly different from zero, while the corresponding adjusted coefficient of determination is $R^2 = 0.68$. Comparing of the model (1) to the data is graphically shown in Fig. 6.

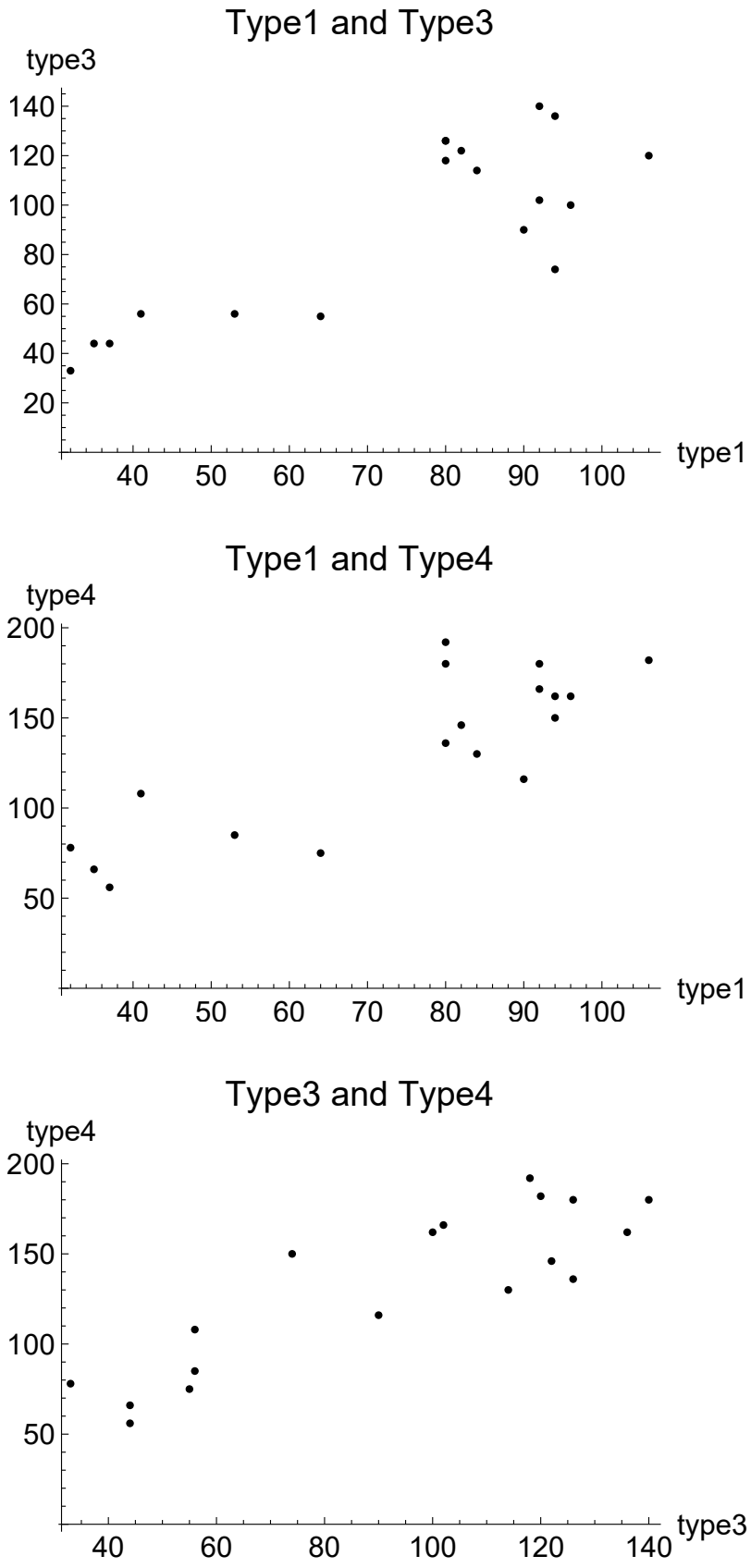


Figure 5: Scatter plots of all combinations of the correlated types 1, 3 and 4 in the time period January 2010 to December 2011.

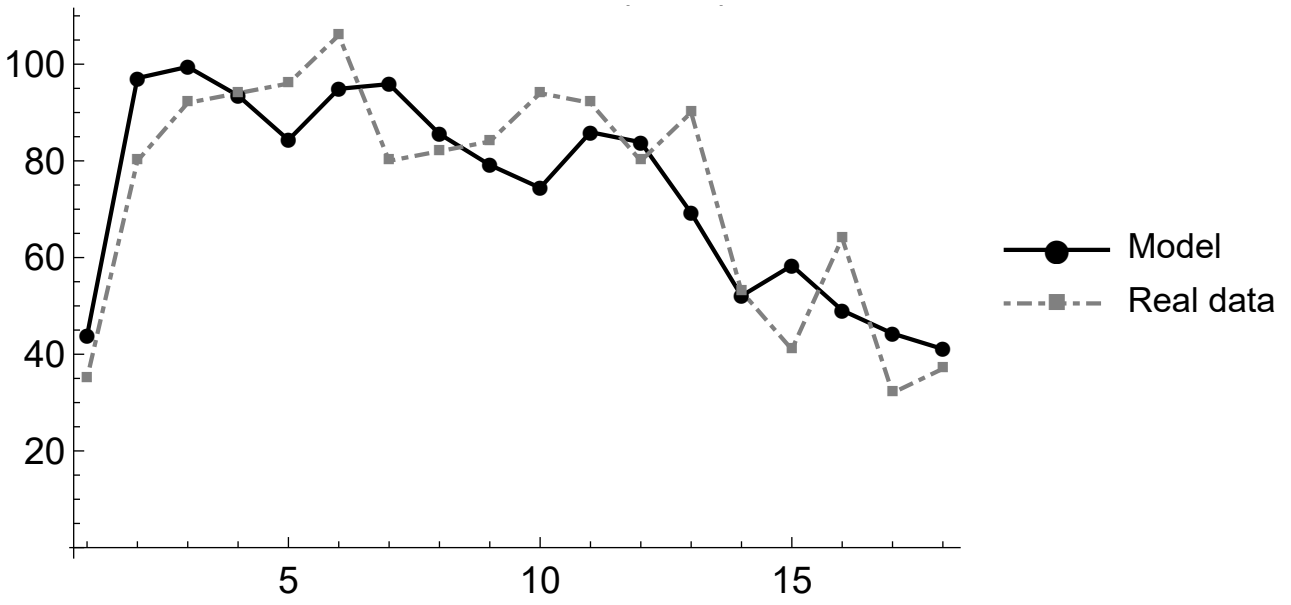


Figure 6: Comparing the data (gray dashed line) and the fitted model (black solid line) of dependence of the type 1 on the type 3 in the time period January 2010 to December 2011.

Table 2: Correlations of submission types in the time period from January 2002 to December 2011 (significant correlations are bold).

	Type 1	Type 2	Type 3	Type 4	Type 5
Type 1	1				
Type 2	**0.619381**	1			
Type 3	**−0.57369**	**−0.526174**	1		
Type 4	**0.952893**	**0.594155**	**−0.600362**	1	
Type 5	**−0.26551**	**−0.197074**	0.00176309	−0.152891	1

Table 3: Correlations of submission types in the time period January 2010 to December 2011 (significant correlations are bold).

	Type 1	Type 2	Type 3	Type 4	Type 5
Type 1	1				
Type 2	0.0113816	1			
Type 3	**0.803786**	−0.244707	1		
Type 4	**0.839609**	−0.00916543	**0.852376**	1	
Type 5	0.211319	−0.228509	0.333468	0.378921	1

4.3. Modelling the number of submissions by gamma distribution

Finally, we study the distribution of the number of submissions per one month while we focus on all types of submissions together. Even when there are trends in evolutions of different types (see Fig. 1), we can suppose that their sum has approximately constant trend, so the distribution is supposed to be independent on time.

For each dataset, we estimate the corresponding parameters and using the Pearson Chi-square test, we test the null-hypothesis about the gamma distribution. Taking into account the number of tested hypothesis, the threshold for the p-value with the Bonferoni correction (see e.g. [3], [4]) is 0.001389. The results are shown in Table 4. As seen from the table, the hypothesis of gamma distribution is not rejected in the most cases of both obtained and sent submissions. Thus, we can conclude that the number of submissions can be modelled by gamma distribution instead of empirical distribution suggested in [16].

5. Conclusion

In the presented analysis, we have obtained three main results.

First, the change in law which has been in force from January 2010 and obliges the municipalities to keep data boxes caused changes in behaviour of the numbers of different types of submissions. Especially, the number of electronic submissions signed by an advanced electronic signature has increased while the number of physical submissions sent by classical post has decreased. Therefore, in future analyses we recommend to deal only with data after January 2010.

Secondly, we have found dependencies among the numbers of submissions of different types. Considering the data after 2010, there are strong dependencies among electronic submissions signed by an advanced electronic signature (type 1), physical documents sent by post (type 3) and physical documents submitted in person (type 4), while the documents submitted electronically without advanced electronic signature (type 2) and other submissions (type 5) have relation neither to the types 1, 3 and 4, nor to each other. Just note that it is surprising that all the significant correlation coefficients are positive. Since there are documents which must be submitted by many people but the way of submitting is optional (for example tax return which can be submitted either through data box or personally), we would rather expect that increasing number of the electronic submissions signed by an advanced

electronic signature (type 1) would reduce the number of physical submissions (types 3 and 4), and vice versa. Although this phenomenon can be slightly seen in Fig. 3, we have recorded a decrease of the numbers of all these types whose influence is stronger to the correlation than the expected behaviour. Further, we can state that linear regression model applied to the most dependent types 1, 3 and 4 seems to be appropriate while the type 4 can be omitted.

Finally, we fit gamma distribution to the number of submissions and apply a goodness-of-fit test which shows that gamma distribution is suitable approximation.

Table 4: Numbers of submissions per month obtained by ("In") and sent from ("Out") the analysed municipalities fitted by gamma distribution and tested by Pearson χ^2 test.

Dataset	In; p-value	In; $\widehat{\alpha}$	In; $\widehat{\beta}$	Out; p-value	Out; $\widehat{\alpha}$	Out; $\widehat{\beta}$
Municipality 1	0.0018	48.34	3.51	0.0000	641.42	0.50
Municipality 2	0.0904	11.08	27.45	0.0835	3.25	139.26
Municipality 3	0.0934	4.76	2.98	0.3726	2.12	2.81
Municipality 4	0.0004	34.89	7.49	0.0195	16.42	15.93
Municipality 5	0.0385	12.68	4.37	0.0069	4.81	5.58
Municipality 6	0.0042	51.95	82.04	0.1158	33.88	115.34
Municipality 7	0.0133	14.78	1.70	0.1769	4.35	3.32
Municipality 8	0.0000	1.23	20.60	0.4159	1.28	9.36
Municipality 9	0.0012	1.70	25.19	0.2964	1.95	6.26
Municipality 10	0.1131	10.17	2.86	0.0764	3.96	4.23
Municipality 11	0.1215	26.16	41.82	0.0000	921.49	0.50
Municipality 12	0.0000	66.20	1.37	0.6763	0.93	8.69
Municipality 13	0.0206	4.27	5.43	0.0000	8.35	0.50
Municipality 14	0.0000	122.01	2.19	0.0000	32.85	10.93
Municipality 15	0.0164	35.10	42.75	0.0000	99.96	18.25
Municipality 16	0.0258	44.53	3.65	0.2276	7.45	42.52
Municipality 17	0.1636	26.58	3.08	0.1588	17.89	4.69
Municipality 18	0.0619	45.28	7.38	0.0582	1.84	54.48

Acknowledgements

This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/083/OHK3/1T/13.

References

[1] Anderson C. W., Ray W. D.: Improved maximum likelihood estimators for the gamma distribution. *Communications in Statistics,* 4(5), 437–448, 1975.

[2] Antoch J., Hušková M., Jarušková D.: Change point detection. *5th ERS IASC Summer School*, 2000.

[3] Armstrong R. A.: When to use the bonferroni correction. *Ophthalmic and Physiological Optics,* 34(5), 502–508, 2014.

[4] Bonferroni C. E.: *Teoria statistica delle classi e calcolo delle probabilit.* R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

[5] Bowman K. O., Shenton L. R.: *Properties of Estimators for the Gamma Distribution.* Taylor Francis Inc, United States, 1988.

[6] Choi S. C., Wette R.: Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics,* 11(4), 683–690, 1969.

[7] Cohen J., Cohen P., West S. G., Aiken L. S.: *Applied multiple regression/correlation analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ, Lawrence Erlbaum Associates, 2003.

[8] Darwin C.: *The Variation of Animals and Plants under Domestication,* 1868.

[9] Eichinger B., Kirch C.: A MOSUM procedure for the estimation of multiple random change points. *Bernoulli,* 24(1), 526–564, 2018.

[10] Hawkins D. M., Deng Q.: A nonparametric change-point control chart. *Journal of Quality Technology,* 42(2), 2010.

[11] Johnson N. L., Kotz S., Balakrishnan N.: *Continuous Univariate Distributions,* Volume 1, 2nd ed. Wiley, 1994.

[12] Khaleghi A., Ryabko D.: Nonparametric multiple change point estimation in highly dependent time series. *Theoretical Computer Science,* 620, 119–133, 2016.

[13] Lechnerová R., Lechner T.: Aplikace bodových procesů při analýze veřejné správy v ČR. *Information Bullentin of the Czech Statistical Society,* 22(3), 81–88, 2010.

[14] Lechnerová R., Lechner T.: Analýza časových řad formální komunikace obcí. *Information Bulletin of the Czech Statistical Society,* 24(3–4), 63–70, 2013.

[15] Mahmoud M. A., Parker P. A., Woodall W. H., Hawkins D. M.: A change point method for linear profile data. *Quality and reliability engineering international,* 23, 247–268, 2007.

[16] Pidnebesna A., Helisová K., Dvořák J., Lechnerová R., Lechner T.: Statistical analysis and modelling of submissions to municipalities in the Czech Republic. *Information Bulletin of the Czech Statistical Society,* 27(4), 1–18, 2016.

[17] Pradhan B., Kundu D.: Bayes estimation and prediction of the two-parameter gamma distribution. *Journal of Statistical Computation and Simulation,* 81(9), 1187–1198, 2011.

[18] Stacy E. W.: Quasimaximum likelihood estimators for two-parameter gamma distributions. *IBM Journal of Research and Development,* 17(2), 115–124, 1973.

[19] Tukey J. W.: *Exploratory Data Analysis.* Addison-Wesley Publishing Company, 1977.

[20] Yanagimoto T.: The conditional maximum likelihood estimator of the shape parameter in the gamma distribution. *Metrika,* 35(1), 161–175, 1988.

[21] Zar J. H.: Spearman rank correlation. In: Armitage P., Colton T. (Eds.), *Encyclopedia of Biostatistics.* 2nd ed., Vol. 7, 5095–5101, 2005.

[22] Zhang J.: Reducing bias of the maximum likelihood estimator of shape parameter for the gamma distribution. *Computational statistics,* 28(4), 1715–1724, 2013.

ZASEDÁNÍ PŘEDSTAVITELŮ NÁRODNÍCH STATISTICKÝCH SPOLEČNOSTÍ SKUPINY V7 POPRVÉ V POLSKU

MEETING OF THE REPRESENTATIVES OF THE V7 STATISTICAL SOCIETIES: FOR THE FIRST TIME IN POLAND

Hana Řezanková

E-mail: hana.rezankova@vse.cz

Ve čtvrtek 12. července 2018 v odpoledních hodinách se ve Varšavě konalo zasedání představitelů národních statistických společností skupiny V7, kterého se zúčastnilo sedm zástupců pěti národních společností (z České republiky, Polska, Rumunska, Slovenska a Slovinska). Termín byl vybrán ve dnech oslav stého výročí polské statistiky, kdy se statistikové z Polska i z jiných zemí sešli ve Varšavě na II. kongresu polské statistiky organizovaném statistickým úřadem a statistickou společností. Účastníci zasedání skupiny V7 byli pozváni do prezidentského paláce na slavnostní předání vyznamenání a vědecký seminář. Na odpoledním jednání se představitelé statistických společností navzájem informovali o pořádaných vědeckých aktivitách a diskutovali o významu Federace evropských národních statistických společností (FENStatS), v níž jsou všechny společnosti skupiny V7 zastoupeny. Podrobnější informace lze nalézt v článku Mach P.: 14. stretnutie štatistických společnosti V7 vo Varšave. *Forum Statisticum Slovacum,* 14(1), 61–63, 2018, http://www.ssds.sk/casopis/archiv/2018/fss0118.pdf#page=63.

PADESÁTÉ VÝROČÍ SLOVENSKÉ ŠTATISTICKÉ A DEMOGRAFICKÉ SPOLOČNOSTI A SLAVNOSTNÍ KONFERENCE

THE 50TH ANNIVERSARY OF THE SLOVAK STATISTICAL AND DEMOGRAPHIC SOCIETY AND THE CEREMONIAL CONFERENCE

Hana Řezanková, Jitka Langhamrová

E-mail: hana.rezankova@vse.cz, jitka.langhamrova@vse.cz

V letošním roce uplynulo již 50 let od založení společnosti, která sdružuje statistiky a demografy především ze Slovenska a jejímiž členy jsou též kolegové z České republiky. Původní název z roku 1968 zněl Slovenská demografická a štatistická spoločnosť. Ten byl v roce 1990 změněn na současný název Slovenská štatistická a demografická spoločnosť (SŠDS). Ve dnech 18. až 20. června 2018 se u příležitosti tohoto významného výročí konala v obci Častá – Papiernička slavnostní konference, na kterou navázaly konference PES 2018 (Pohledy na ekonomiku Slovenska) a FERNSTAT 2018 (Finance, ekonomie a statistika). Zájemci se tak vlastně na jednom místě během tří dnů mohli účastnit tří konferenci, které byly prezentovány pod názvem „troj-konference".

Slavnostní akci zahájila předsedkyně SŠDS Iveta Stankovičová. Poté jednání pozdravili místopředsedkyně České statistické společnosti Hana Řezanková, předsedkyně České demografické společnosti Jitka Langhamrová a ředitel odboru vysokoškolského vzdělávání Ministerstva školství, vědy, výzkumu a sportu SR, pan Andrej Piovarči. Podpředseda SŠDS Peter Mach přečetl pozdravný list od předsedy Rady slovenských vědeckých společností při SAV, Viktora Milaty.

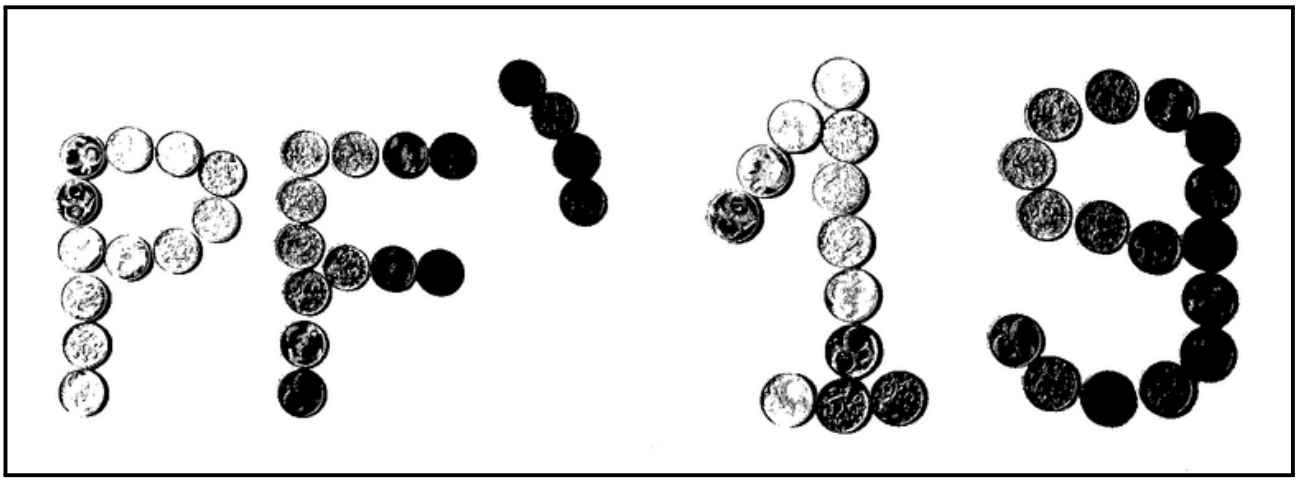
V odborném programu jako první vystoupila Silvia Szabová, ředitelka pracoviště Statistického úřadu Slovenské republiky (ŠÚ SR) v Bratislavě. Její příspěvek informoval o Bratislavském samosprávném kraji. Následovaly historicky zaměřené příspěvky, které přednesli Prokop Závodský z Vysoké školy ekonomické v Praze (Meziválečná Československá statistická společnost), Peter Mach (História SŠDS – 50 rokov od vzniku Spoločnosti), místopředseda ŠÚ SR František Bernadič (25 rokov samostatnej štátnej štatistiky v SR) a Branislav Bleha z Univerzity Komenského v Bratislavě (Slovenská demografia po roku 1993). V závěru této první části slavnostní konference byli vyznamenáni zasloužilí členové SŠDS.

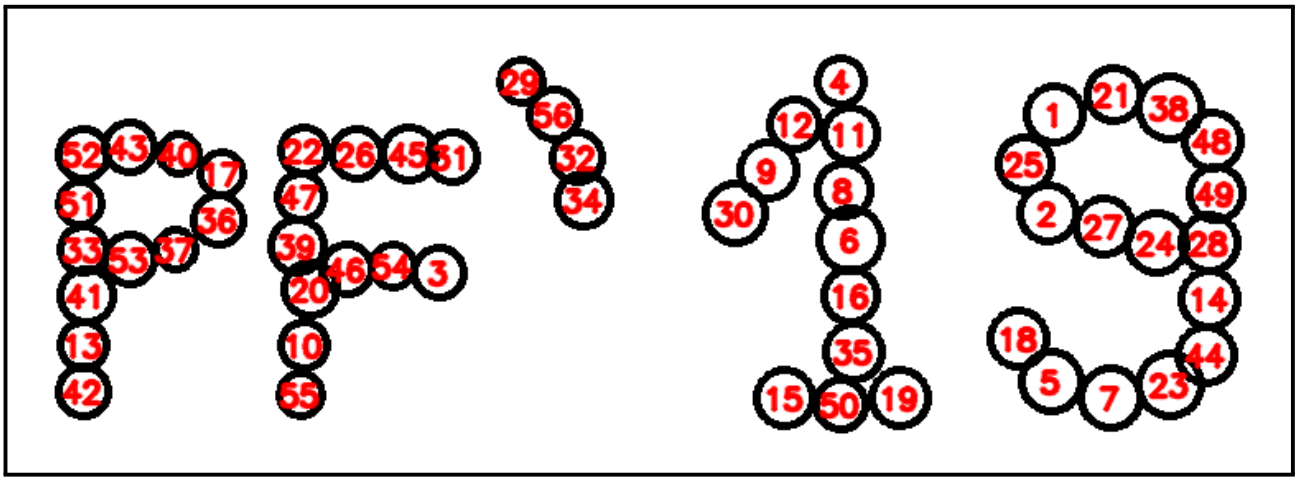
Obsahem druhé části bylo téma výuky statistiky a demografie. O průzkumu postojů studentů ke statistice informoval Tomáš Želinský z Technické univerzity v Košicích. Současných problémů s výukou statistiky, potřeby analýzy rozsáhlých souborů dat a významu datové vědy se dotkla Iveta Stankovičová z Univerzity Komenského. Následující panelová diskuze byla zaměřena na výuku statistiky a demografie ve Slovenské republice a České republice, na jejímž úvodu vystoupily Hana Řezanková a Jitka Langhamrová z Vysoké školy ekonomické v Praze.

Organizačnímu týmu pod vedením předsedkyně SŠDS Ivety Stankovičové patří velký dík za kvalitní odborný program a příjemnou atmosféru na nádherném místě, stejně jako za společenský program, jehož součástí byla vycházka na hrad Červený kameň.

~ ~ ~

Kolik neplatných mincí má redakce k dispozici do nového roku?
Řešení je na další straně. Tip: vyfiltrujeme pozadí...






```
#!/usr/bin/env python3
# Spuštění: python3 segment-bulletin.py
# Inspirace z: http://blog.christianperone.com/2014/06/simple-and-effective-
#    coin-segmentation-using-python-and-opencv/
import numpy as np # Scipy.org, test s verzí 1.15.0
import cv2         # OpenCV, test s verzí 3.4.2 na Pythonu 3.6.7

font=cv2.FONT_HERSHEY_SIMPLEX # Základní řez písma
textscale=0.6 # Velikost písma
textline=2    # Síla linky písma
image=cv2.imread("mince.png") # Načtení vstupního obrázku, předfiltr byl
    zrealizován v programu GIMP, byť OpenCV to taky umí...
gray=cv2.cvtColor(image, cv2.COLOR_BGR2GRAY) # Filtr obrázku: do stupňů šedi
circles = cv2.HoughCircles(gray, cv2.HOUGH_GRADIENT, 1, 20, param1=7,
    param2=25, minRadius=10, maxRadius=30) # Nalezení kružnic
image[:][:]=(255,255,255) # Zobrazit ve finále jen výsledek, nikoliv mince

if circles is not None: # Vykreslení a popsání kružnic, je-li tam alespoň 1
  circles = np.round(circles[0,:]).astype("int")
  # Zobrazit prvně jen kruhy
  for (x, y, r) in circles:
    cv2.circle(image, (x,y), r, (0,0,0), 4)
  # A na kruhy dát čísla, kvůli nechtěnému překreslování cifer kružnicemi
  for c,(x, y, r) in enumerate(circles): # Vystředění textu do středu kruhu
    text=str(c+1) # Pořadové číslo kruhu, v Pythonu od 0, v reálu od 1
    textsize=cv2.getTextSize(text, font, textscale, textline) # Změření textu
    textX=int(x-textsize[0][0]/2) # Posun o polovinu délky v ose x.
    textY=int(y+textsize[0][1]/2) # V ose y. +, (0,0) je v levém horním rohu
    cv2.putText(image, text, (textX,textY), font, textscale, (0,0,255),
     textline, cv2.LINE_AA) # Napsání čísla do kruhu
  print("Počet kruhů:", len(circles)) # Vypsání počtu kruhů na terminál
  cv2.imwrite("mince-vysledek.png", image) # Uložení výsledku do souboru
  cv2.imshow("Output", image) # Zobrazení výsledného obrázku na obrazovku
  cv2.waitKey(0) # Vyčkání na klávesu od uživatele, konec programu
  # Řešení je 56, po písmenech: 12 (P), 12 (F), 4 (apostrof), 12 (1) a 16 (9)
```

POZVÁNKA NA ČLENSKOU SCHŮZI
INVITATION TO THE ANNUAL MEETING

Výbor České statistické společnosti

Milé kolegyně, milí kolegové,

dovolte, abychom Vás pozvali na **členskou schůzi** společnosti, která se uskuteční **ve čtvrtek 31. ledna 2019, od 13 hodin, v budově Českého statistického úřadu v místnosti 226.**

Po členské schůzi Vás zveme na přednášku na téma **Mezinárodní výzkumy ve vzdělávání, včetně výzkumu vědomosti a dovednosti dospělých,** s níž vystoupí paní **doc. RNDr. Jana Straková, Ph.D.,** z Pedagogické fakulty Univerzity Karlovy v Praze.

S přáním hezkého dne,

výbor ČStS

Obsah

Vědecké a odborné články

Zprávy a informace

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 Na padesátém 81, 100 82 Praha 10. 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 the journal are published in English, Czech and Slovak languages.

Předseda společnosti: RNDr. Marek Malý, CSc., Státní zdravotní ústav, Šrobárova 48, Praha 10, 100 42, e-mail: mmaly@szu.cz.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.