UDK 31 ISSN 0039-05-34

# STATISTIČKA REVIJA

GODINA LVII Broj 1-2 ZA 2008. GOD.

## SADRŽAJ

### PREGLEDNI NAUČNI ČLANCI



### STRUČNI ČLANCI



### PRIKAZI KNJIGA



IZDAJE: STATISTIČKO DRUŠTVO SRBIJE — BEOGRAD

# S T A T I S T I Č K A  R E V I J A

IZLAZI TROMESEČNO - PUBLICATION TRIMESTRIELLE - QUARTERLY





Adresa redakcije:
Statističko društavo Srbije, 11000 Beograd, Milana Rakića 5
Žiro-račun br. 205-99554-89 kod Komercijalne banke
Tiraž: 200


---

SARADNICI U OVOM BROJU:

*Prof. dr Katarina Čobanović*
*Prof. dr Zagorka Lozanov-Crvenković*
*mr Emilija Nikolić-Đorić*
*Edin Šabanović*
*Vesna Bucevska*
*Olga Melovski*
*Božidar Popović*
*Nataša Cvetković*
*Marija Panović*
*Jelena Kovijanić*
*Elizabeta Dumik*
*Arijana Amina Muhić*
*Asmira Hotović*
*Slobodan Nićin*
*Predrag Čanović*


**Predato u štampu: novembar 2008. godine**
**Izašlo iz štampe: novembar 2008. godine**

---

Štampa: Republički zavod za statistiku Srbije, Beograd, Milana Rakića 5

# PREGLEDNI NAUČNI ČLANCI

*Dr Katarina Čobanović, Department of Agricultural Economics and Rural Sociology, Faculty of Agricultural, University of Novi Sad*
*Dr Zagorka Lozanov-Crvenković, Department of Mathematics and Informatics, Faculty of Science and Mathematics, University of Novi Sad*
*Mr Emilija Nikolić-Đorić. Department of Agricultural Economics and Rural Sociology, Faculty of Agricultural, University of Novi Sad*

# THE APPLICATION OF PERIODIC REGRESSION

## ABSTRACT

*Periodic regression is seldom included in syllabus of statistical courses. However, data following periodic or cyclic behavior are often encountered, especially in agriculture. Therefore, it is proposed that this type of regression should be thought to students of agriculture, economy and biology even in basic courses of statistics. In this paper the way of teaching periodic regression through examples usually encountered in practice is recommended. The analysis of data is based on the graphical interpretation, providing the visual display of the investigated problems as well. The model of periodic regression could be successfully used in the research work in the analysis of seasonal variations. The periodic regression technique has found the use in the researches in agriculture, economy, biology and medicine.*

***Key words:*** *cyclic variations, seasonal variations, model of periodic regression, periodic components, and stepwise regression procedure*

## 1. INTRODUCTION

Many biological and agricultural series are characterized by seasonal variations. Periodic phenomena are close primary to biological data, but, also, to non-biological data. Periodic or cyclic phenomena are characteristic for many different types of data, which are synchronized with daily, lunar, or annual changes (Bliss, 1970). Many kinds of agricultural data tend to fluctuate up and down in regular time intervals showing periodic character (Little, Hills, 1978). Periodic or cyclic character of many phenomena in natural environment is expressed in time and in space. Many cyclical and seasonal variations of biological and other data are occurring depending on daily, monthly, yearly or other cycle's changes. In biology and agriculture, too, periodic and cyclic variations could be identified in a number of investigated problems (Bliss, 1970; Little, Hills, 1975; Marko, Nikolić, 1982).

Biological cycles could be divided in two categories: those which do not strictly depend on physical environment and those which strictly depend on physical environment (Bliss, 1970). In the first category cycles are determined physiologically. In the second category cycles depend on many potential factors. The cycle period in a natural environment, concerning animals or plants, may be of different length.

The authors have had in mind that "Periodic type is a curve that relates some variable to time and is repeated at fixed time intervals. It is known in mathematical texts as a Fourier curve and is useful for any kind of data that tends to fluctuate up and down at regular intervals. Very few statistics texts discuss fitting data of this kind, but we have found it so useful for many kinds of agricultural data, that we will give a brief outline of the general method." (Little, Hills, 1978). Using the advantages of computers techniques today, it is very easy to apply periodic regression, which is of type of polynomial curve. In such situation, in teaching statistics, it would be simple to apply regression methods on complicated trigonometric functions, expressing wave forms, having in mind that the method of fitting a Fourier curve is very similar to the method for fitting a polynomial regression (Little, Hills, 1978). Many statistical courses include basic regression methods, which could be a solid base for introduction to analysis of the cyclical and seasonal variations of time series data in the teaching process. Thus, students could easily apply their already acquired knowledge in specific case of periodic regression.

Some authors suggest other methods in the analysis of cyclical and seasonal variations of explored data, like the use of dummy variable (Hebden, 1981; Čobanović, Lučić, 1992). But, the authors in this paper have found that simplified trigonometric functions for data series with periodicity could be useful in the process of teaching Statistics.

There are some interesting research works in medicine and biology which are using periodic regression. For example, periodic regression was used to detect cyclical patterns, e.g. lunar cycles, in a study with marine fishes. Lunar cycles are typical in the movement, feeding and reproduction of marine fishes (de Bruyn et al. 2001).

A periodic regression model was used in analyzing the annual numbers of human births with regard to an 11-year cycle. Eleven-year rhythms were found and compared with solar and geophysical variables. The authors of the study emphasized that the correlation of conceptions with the 11-year solar cycle may be a potential guide in the selection of further variables for the control and regulation of the rhythms in human conceptions (Randall, Moos, 1993).

Experimental study with the bioluminescence of Amphiphollis squamata, which was assessed from freshly collected individuals for 16 successive months, showed that the bioluminescence varied seasonally. The seasonal variations of bioluminescence were analyzed by periodic regression model (Deheyn et al., 2000).

In economy, fluctuations of prices and parity of prices may also be successfully modeled by application of periodic regression ( Marić, 1980; Nikolić-Đorić, et al. 1993; Kabić, 1997).

At the Faculty of Agriculture, University of Novi Sad, at the study group of Agriculture Economy, the data with periodic character which need to be analyzed are often encountered. Therefore, in this paper we wish to emphasize the need of using some specific methods of analyzing the cyclical and seasonal data usually encountered in agricultural practice. One of these methods is the use of periodic regression based on Fourier curve. We propose the way of teaching periodic regression through examples with real data from agriculture. Data series in examples are related to daily, monthly and yearly variations. Daily variations present data of the value of UV Index (ultraviolet) radiation during the

period of three years; monthly variations represent the values of price index, concerning the prices of agricultural products during the period of five years; yearly variations represent the values of wheat yield (t/ha) during the period of 57 years.

## 2. METHODS

In the analysis of the periodicity the most important is the definition of three main parameters: the length of the cycle or fundamental period; its amplitude or the range from the minimal to the maximal response and the phase angle or angular point in time during the cycle when the response is maximal (Bliss, 1970). These parameters can be easily estimated using any statistical software. Also, F-test and t-test for testing the goodness of fit can be performed Here we use STATISTICA 7.0 (University license, University of Novi Sad).

A time series $Y_t\,(t=1,...,N)$ observed at equal intervals of time and may be expressed as: $Y_t = \hat{Y}_t + \varepsilon_t$, where $\hat{Y}_t$ is unobserved fixed value at time $t$ and $\{\varepsilon_t\}$ is a sequence of random errors identically and independently distributed with expectation 0 and variance $\sigma^2$. To determine whether the variability of the time series has periodic components, the series is approximated by finite Fourier series of the form, if the number of data is even: $N = 2n$,

$$\hat{Y}_t = A_0 + 2\sum_{m=1}^{n-1}(A_m \cos 2\pi m f_1 t + B_m \sin 2\pi m f_1 t) + A_n \cos 2\pi n f_1 t, \quad \text{or}$$

$$\hat{Y}_t = A_0 + 2\sum_{m=1}^{n-1}(A_m \cos 2\pi m f_1 t + B_m \sin 2\pi m f_1 t),$$

if the number of data is odd: $N = 2n$ -1.

Here $R_m = \sqrt{A_m^2 + B_m^2}$ is the amplitude, and $\phi_m = arctg(B_m / A_m)$ is the phase of the *i*-th component. The function $\hat{Y}_t$ is a linear combination of sinus and cosinus functions with frequencies proportional to fundamental frequencies $f_1 = 1/N$, so it is linear multiple regression with sinus and cosinus functions as repressors. Since $\frac{1}{N}\sum_{t=1}^{N} Y_t^2 = R_0^2 + 2\sum_{m=1}^{n-1} R_m^2 + R_n^2$, a contribution of i-th harmonical component to the mean of the total sum of squares of time series is equal to $R_i^2$. By decomposing the mean of the total sum of squares it is possible to single out harmonical components which describe the series well.

With additional assumption that errors are normally distributed, the estimated periodic regression model may be tested by F-test and particular estimates of parameters with *t* –test.

## 3. RESULTS

**Example 1.** The first examined time series contained daily measurements of maximal UV index, made in period 25.04.2003. –29.06.2005. at Department of Physics, University of Novi Sad in the Vojvodina Province, Serbia. Line plot of time series suggested cyclic behavior with the constant periodicity of approximately one year ($f_1$=1/365). The periodic regression may be considered as special case of multiple regression with time series ($Y_t$, t=1,...,615) as dependent and two regressors $X_{1t} = sin(2\pi f_1 t)$ and $X_{21t} = cos(2\pi f_1 t)$ as independent variables. Ordinary least squares estimates of unknown regression coefficients, standard errors of coefficients, t-values, F-value, coefficients of multiple correlation, determination, adjusted coefficient of determination and standard error of the regression were calculated by means of program STATISTICA 7 and presented in Table 1. Original data and estimated model are presented at Figure 1. Special attention should be given to low values of UV Index during the summer day, due to cloudy days. It should be pointed out to students that better fit could be obtained by including the dummy variable.

***Table 1.*** *Results of periodic regression of UV data*

Regression Summary for Dependent Variable: Y
R= .90421869 R²= .81761143 Adjusted R²= .81701539
F(2,612)=1371.7 p<0.0000, Std.Error of estimate: 1.2218

| | Beta | Std.Err. of Beta | B | Std.Err. | t(612) | p-level |
|---|---|---|---|---|---|---|
| Intercept | | | 3.920087 | 0.050101 | 78.24385 | 0.0000 |
| sin(2*pi*t*f1) | 0.789134 | 0.017342 | 3.272730 | 0.071922 | 45.50395 | 0.0000 |
| cos(2*pi*t*f1) | 0.372651 | 0.017342 | 1.490214 | 0.069350 | 21.48823 | 0.0000 |

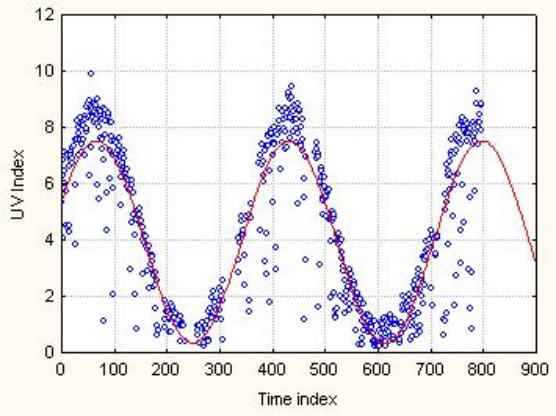



***Figure 1.*** *Periodic regression of UV data*

**Example 2**. As an example of monthly data the time series of agricultural production indices of prices in period 1999-2004. with base period (2004=100) for Serbia, was considered. Line diagram of time series suggested seasonal oscillations of indices around quadratic trend. So the adequate model for trend and seasonal movements is linear combination of polynomial of second degree and periodic components. It may be shown that fitting of model will be improved by introduction of trigonometric components with period of 6 months. See Table 2 and Figure 2.

***Table 2.*** *Results of periodic regression of agricultural production indices of prices*

Regression Summary for Dependent Variable: Y
R= .97903602 $R^2$= .95851153 Adjusted $R^2$= .95468182 F(6,65)=250.28
F(6,65)=250.28 p<0.0000 Std.Error of estimate: 5.8190

| | Beta | Std.Err. of Beta | B | Std.Err. | t(65) | p-level |
|---|---|---|---|---|---|---|
| Intercept | | | 10.36715 | 2.123943 | 4.8811 | 0.000007 |
| t | 2.25428 | 0.102516 | 2.94431 | 0.133896 | 21.9895 | 0.000000 |
| t2 | -1.38139 | 0.102449 | -0.02395 | 0.001776 | -13.4837 | 0.000000 |
| sin(2*Pi*t*1/12) | 0.07824 | 0.025472 | 3.00363 | 0.977816 | 3.0718 | 0.003106 |
| cos(2*Pi*t*1/12) | -0.02447 | 0.025287 | -0.93935 | 0.970723 | -0.9677 | 0.336790 |
| sin(2*Pi*t*1/6) | -0.05781 | 0.025309 | -2.21922 | 0.971600 | -2.2841 | 0.025641 |
| cos(2*Pi*t*1/6) | 0.06112 | 0.025280 | 2.34621 | 0.970383 | 2.4178 | 0.018426 |

**Example 3.** Time series of yearly wheat data (t/ha) in the period 1947-2002 is characterized by cyclic variations around third degree polynomial trend. The dominant frequency was obtained by means of periodogram applied on detrended time series. The dominant frequency in this case is 0.2857 (periodicity 3.5 years). The OLS coefficients of regression model that includes variables selected by application of stepwise regression are given in Table 3. The model is presented at Figure 3.

***Table 3.*** *Results of polynomial-periodic regression of wheat data*

Regression Summary for Dependent Variable: Y
R= .94889712 $R^2$= .90040575 Adjusted $R^2$= .89476834
F(3,53)=159.72 p<0.0000, Std.Error of estimate: 0.33750

| | Beta | Std.Err. of Beta | B | Std.Err. | t(53) | p-level |
|---|---|---|---|---|---|---|
| Intercept | | | 1.046239 | 0.090675 | 11.5384 | 0.000000 |
| t2 | 4.42319 | 0.261654 | 0.004633 | 0.000274 | 16.9047 | 0.000000 |
| t3 | -3.78387 | 0.261708 | -0.000072 | 0.000005 | -14.4584 | 0.000000 |
| sin(2*Pi*f1*t) | -0.15030 | 0.043417 | -0.217534 | 0.062839 | -3.4616 | 0.001069 |

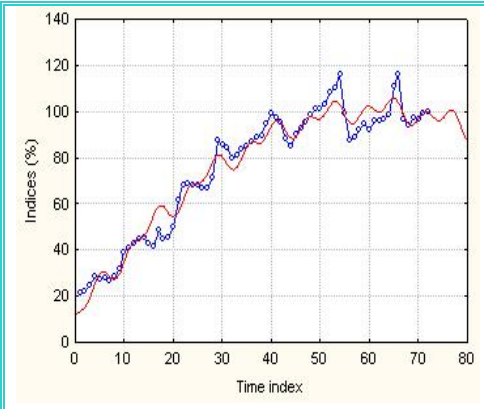



***Figure 2.*** *Periodic regression of agricultural production indices of prices data*

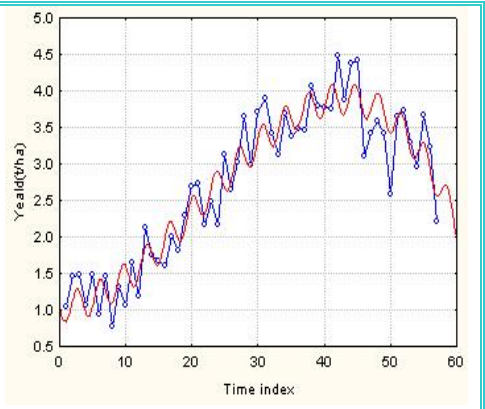



***Figure 3.*** *Polynomial-periodic regression of wheat data*

## 4. CONCLUSION

This paper illustrates that periodic regression is simple and flexible method for detection of cyclical and seasonal pattern in time series and should be included in statistical courses for students of agriculture, economy and biology. This kind of model can represent time series with small number of parameters which is of great importance especially in the case when time series is not monotonic and stationary and contains nonlinear trend, cyclical and seasonal components with different periodicity. As periodic regression is special case of multiple regression, it can be performed at any commonly used statistical software. This method is not adequate for time series with varying amplitudes, frequencies and phases. Time series of sun spot numbers (http://www.spaceweather.com/glossary/ sunspotnumber.html) is the example of time series with non-symmetric cycles with variable amplitudes.

## ACKNOWLEDGEMENTS

The paper titled "Periodic regression" was presented at the 7-th International Conference on Teaching Statistics (ICOTS7), held in Salvador, Bahia, Brazil, 2-7 July 2006. The Conference motto was: "Working Cooperatively in Statistics Education". This paper is a little revised and extended form of the previous paper. Previous paper is given in the Proceedings of ICOTS7 (Editors: Allan Rossman, Beth Chance,2006).

## REFERENCES

[1] Bliss, C.I., (1970), *Statistics in Biology*, Vol.II, McGraw-Hill Book Company, New York.

[2] de Bruyn, A.M.H., Jessica J. Meeuwig (2001), *Detecting lunar cycles in marine ecology: periodic regression versus categorical ANOVA,* Marine Ecology Progress Series, Vol. 214, pp. 307-310.

[3] Čobanović, K., Lučić, Đ. (1992), *The investigation of seasonality in milk production using dummy variable,* Zbornik radova, Savez inženjera i tehničara Jugoslavije, Beograd

[4] Deheyn, D., Mallefet, J., Jangoux, M. (2000), *Evidence of seasonal variation in bioluminescence of Amphiphollis squamata (Ophiuroidea,Echinodermata): effects of environmental factors,* Journal of Experimental Marine Biology and Ecology, Volume 245, Issue 2, pp.245-264.

[5] Hebden, J. (1981), *Statistics for Economists,* Philip Allan, Publishers Limited, Oxford.

[6] Kabić, Z. (1997), *Modeliranje eksperimentalnih rezultata pomoću Furijeovih redova,* Statistička revija, 3-4, Beograd, str. 68-78.

[7] Little, T.M., Hills, F.J. (1978), *Agricultural Experimentation Design and Analysis*, John Wiley and Sons, Inc.

[8] Marko, J., Nikolić, E. (1982), *Characteristics of Yield Development of the Main Field Crops in SAP of Vojvodina,* Contemporary Agriculture, Vol.30, No-2, Novi Sad, pp.87-98.

[9] Marić, N. (1981), *Regresiona analiza vremenskih serija, Teorija i primena reresione analize* (Redakcija: Dr Stevan Hadživuković), Institut za ekonomiku poljoprivrede, Poljoprivredni fakultet, Novi Sad, str.129-136.

[10] Nikolić-Đorić Emilija, Novković N., Rodić Vesna, Aleksić Lj. (1993), *Izbor adekvatnog modela u predviđanju pariteta cena svinje-kukuruz,* Agroekonomika 22, Institut za ekonomiku poljoprivrede, Poljoprivredni fakultet, Novi Sad, str.111-122.

[11] Randall,W., Moos,W.S. (1993) *The 11-year cycle in human births*, International Journal of Biometeorology, Volume 37, Number 2, pp.72-77.

[12] STATISTICA 7.0 (2005), Stat Soft, Inc. University License, University of Novi Sad
[13] STAT. YEARB.SCG.2003, (2003), Statistical Office of Serbia, Belgrade.

*Katarina Čobanović, Zagorka Lozanov-Crvenković, Emilija Nikolić-Đorić*

# PRIMENA PERIODIČNE REGRESIJE

## REZIME

*U nastavnim programima iz statistike retko se koristi model periodične regresije. Međutim, mnoge pojave u prirodi ispoljavaju periodičnost, ili ciklično ponašanje, što se naročito javlja u poljoprivredi. Autori rada predlažu da se ovaj model regresije uvede u nastavni program iz statistike za studente poljoprivrede, ekonomije i biologije. U radu je data procedura periodične regresije koja se može primeniti u praktičnom radu. Dati su primeri sa grafičkim ilustracijama. Autori naglašavaju da se tehnika periodične regresije može uspešno primeniti u naučnim istraživanjima u poljoprivredi, biologiji i medicini pri ispitivanju sezonskih varijacija.*

***Ključne reči:*** *ciklične varijacije, sezonske varijacije, model periodične regresije, sezonska komponenta, procedura stepwise regresije*

# STRUČNI ČLANCI

*Edin Šabanović, MSc. Agency for Statistics of Bosnia and Herzegovina,*
*Sarajevo School of Economics and Business, Sarajevo University, Sarajevo*
*Bosnia and Herzegovina*
*e-mail: edinsabanovic@efsa.unsa.ba*

# DEALING WITH SAMPLE FRAME IMPERFECTIONS: SURVEY EXPERIENCES FROM BOSNIA AND HERZEGOVINA AND FUTURE CHALLENGES

## ABSTRACT

*Household based surveys are sample surveys and the quality of their results depends inter alia on the sample design. The representative samples are drawn from sample frames that usually are population or household registers, population censuses or their updates. In absence of these data sources, surveys deal with sample frame imperfections which increase the non-sampling errors and have an impact on estimates. Bosnia and Herzegovina was in a situation which required sensitive solutions. This article illustrates the characteristics of the sample frames and sample designs in Bosnia and Herzegovina as well as the constraints in sample selection and the procedures applied in the last seven years*

***Key words:*** *sample frame, master sample, sample design.*

## 1. BACKGROUND

After the end of the war in the country, the statistical institutes in Bosnia and Herzegovina planned to conduct household surveys in order to provide most recent basic inputs for different statistical sectors (price statistics, national accounts, labor and trade statistics, etc.) and to surpass the gap in statistical data production due to the war. Three basic household surveys were given priority: living standard measurement survey, household budget and labour force survey. In the preparation of survey methodologies, institutes were confronted with the serious problem of household sample selection. This problem arose due to the lack of population data and consequently, the lack of an appropriate and updated frame for designing representative samples for household based surveys.

## 2. EXISTING POPULATION STATISTICS

Every sampling attempt in the post-war Bosnia and Herzegovina faced the frame problem at the very begining of the survey preparation activities. Bosnia and Herzegovina in the post-war period did not have neither the population nor the households registers. The

postal address registers, if available, were incomplete and the same situation was with the elections lists. The last population census was conducted in 1991 and its data were out-of-date in a country where population movements during and after the war, as well as new housing construction were considerable.

The statistical system in the country had quite good organisation of the area units registers which consisted of almost 20000 enumeration areas grouped in settlements, municipalities, entities and District according the newest administrative structure of the country (Lynn 2004). But, all existing data on them were from 1991 and far from the real situation because of the considerable demographic changes in the period 1992-1995. So, it was reasonable to have a lot of empty enumeration areas or areas with only few households in the region were people left their houses (mostly in rural areas), or to have enumeration areas with more households than before the war (mostly in urban areas).

Statistical institutes in Bosnia and Herzegovina defined household based surveys as their priorities but they did not have any financial support for conducting them. From other point of view, their resources and experiences in sampling methods were very limited. Owing to the considerable international community support in capacity building, statistical system in Bosnia and Herzegovina was established and first household surveys were conducted. Following chapters described more detailed sample frame organizations and sample designs adopted through survey methodologies.

## 3. EXISTING AGRICULTURAL STATISTICS

The situation with the agricultural statistics was even worse in comparison with the population statistics. There weren`t any agricultural holdings registers in the country. During the population census 1991 a set of data on agricultural households was recorded but it has never been processed. The agricultural statistics relied mainly on estimates which were made by municipality agricultural experts. The quality of this data was often questionable and it wasn`t any additional data source to compare and to check them. One more problem in agrucultural data collection were millions of mines in the fields. Finally, the agricultural census and surveys weren`t set as priorities in the programmes of the statistical institutes. All these facts had a negative impact on the production of the agricultural data and they stayed very poor for years.

## 4. CONSTRUCTION OF THE FIRST SAMPLE FRAME

The first sample frame in the post-war period in Bosnia and Herzegovina was constructed in 2000 for the purpose of Living Standard Measurement Survey (LSMS) 2001. The sample frame was created as a „master sample“ from which the sample of households could be selected. The frame construction process and sample design consisted of the following stages:

a) All 146 municipalities in Bosnia and Herzegovina were split into six strata (2 entities as main geographical strata and 3 area type: urban, rural and semi-urban) and 25

municipalities were selected as Primary Sampling Units with the probability proportional to size.

b) The maps for all selected enumeration areas were made by the entity geodesic institutes

c) Within selected municipalities all households were enumerated in the face-to-face interview and the basic demographic data on their members were collected. Exactly 6097 enumeration areas and about 380000 households were enumerated.

d) 450 enumeration areas were selected as Secondary Sampling Units with the probability proportional to size (number of enumerated households in the stage c).

e) Selection of 12 households with equal probability within each sampled enumeration area.

Hence, the sample design for LSMS 2001 was three stage stratified cluster sample which final size was 5402 households: 3002 in Federation B&H and 2400 in Republika Srpska. The overal assesment of the sample design for LSMS 2001 was made by prof. Peter Lynn from Essex University and it shown hudge design effect of 2.13 (Lynn 2004) due to differential selection probabilities and sample clustering. This result indicated that the effective sample size was only 5400/2.13= 2535 households which was considered as hudge loss in efficiency and resources.

## 5. NEW CONCEPT OF SAMPLE FRAME CONSTRUCTION

The survey experiances from the LSMS 2001 were very useful for the design of following surveys. The next household survey conducted in Bosnia and Herzegovina was Household Budget Survey 2004. Considering the imperfections of the frame and sample design for LSMS (Lynn 2004), the additional organizational and financial efforts were made in constructing the new frame which should be used not only for HBS 2004, but also for other household surveys in the nearest future. The main idea was kept: to construct the statistically representative master sample for multiple surveys in Bosnia and Herzegovina, but by using of the approach that would be simple and cost effective. The solution was found in the implementation of an area based approach in 2003. Sampling process consisted of three stages:

a) Selection of 710 enumeration areas as primary sampling units which were considered as appropriate number of EA`s in order to provide representative estimates on state and entity levels.

b) Enumeration of all households within the selected EA`s using the combined non-intrusive and semi-intrusive method. The total number of enumerated dwellings/households was 39081.

c) Selection of households as secondary sampling units.

The sample design for HBS 2004 was two-stage systematic sample with designed size of 9570 households in order to have 6000 interviewed SSU`s (considered non-response rate was 37.3% and was estimated from the pilot survey). The selection procedure provided both explicit and implicit stratification which resulted in reasonably accurate

survey estimates (See: B&H HBS: Methodology 2004). The expected design effect due to clustering was 1.18 (Lynn 2004).

## 6. THE MOST RECENT SAMPLE FRAME CONSTRUCTION

Whereas the statistical programme of the Agency for Statistics of Bosnia and Herzegovina proposed the sample surveys as activities of the highest priority and the sample frame from 2003 got out of date, the need for new one arised. Three surveys were planned to be conducted: Labour Force Survey (LFS) and Multiple Cluster Survey (MICS) in 2006 and Household Budget Survey (HBS) in 2007. Knowing this priorities and having in consideration of expected activities on population census, the dynamic of the demographic changes in Bosnia and Hercegovina as well as the expressed competence of donors and partner organisations, the idea to create double sized sample frame appeared. This frame was constructed at the beginning of 2006 and schould provide the representative household samples until 2009.

The main goal of the project was to select the representative master sample containing 1500 enumeration areas which maps will be updated and within these all existing households will be enumerated. The aim was to create the list of households containing their main socio-demographic variables. In comparison with the master sample from 2003 it was planned to record more household and household members caracteristics in order to facilitate the sample designs more adjusted to surveys requirements.

According the estimation from the previous experiences, it was expected to have the list of about 80000 households which schould be big enough for sample selection for planned surveys in the next three years.

This time, the sampling process consisted of following stages:

a) Selection of additional 790 EA`s in order to get the 1500 EA`s as primary sampling units (710 EA`s from master sample 2003 were used).

b) Enumeration of all households within 1500 selected EA`s in face-to-face interview in order to register the basic household and household members caracteristics using intrusive method. The total number of enumerated households was 79630.

c) Selection of households as secondary sampling units.

The sample design for LFS and MICS 2006 and for HBS 2007 were selected from this master sample. All of them were two-stage systematic samples. The sample size for LFS was 10000, for MICS 6000 and for HBS 9274 households (in order to get 7600 fully co-operating households; estimated non-response rate of 18% from HBS 2004 was used in determining the sample size). Because the sample designs were very similar to this one for HBS 2004, it is expected to have satisfactory survey estimates, especially for main survey variables.

## 7. CONCLUSION AND PROPOSAL

In the condition of absence of relevant population and agricultural statistics, statistical offices in Bosnia and Herzegovina deal with different kind of problem. One of them is connected with inappropriate sample frame. First household survey (LSMS) most suffered of sample frame imperfection but it constructed the initial basis for future master sample. In this way future household surveys could improve their methodologies using LSMS results and experiences.

After few surveys the situation regarding sample frame improved. The reason for that were not only in the cumulative effects of previous survey experiences, but also in the improving capacity building of statistical institutes in Bosnia and Herzegovina. Together with international support, institutes invested in sample frame constructions and staff education.

The survey practice in last seven years in Bosnia and Herzegovina solved a lot of initial problems, but not all of them. It is clear that the quality level of surves estimates, which depend of sample frame, will achieve european standard as the population census data were available. Until this, every survey will have more or less frame problems to solve. At the moment the frame infrastructure is on the level which does not require frequent up-dating and which provides modest survey result quality. One of advantages of described frame construction strategy is that the listing procedure can be carried out immediately prior to the field work for any particular survey, ensuring that the frame of households is as up-to-date as possible. The disadvantages of this frame construction are in cost of the updating procedures and in time limited value. Considering the current activities on population census (which is scheduled for 2011 and for which is expected to contain an agricultural module) and increasing needs for survey results, statistical institutes in Bosnia and Herzegovina certainly must construct one more master sample in 2009.

Our proposal is that this master sample should be improved both in quality and quantity. The goal will be not only to get the same and recent data on households, but to collect the basic agricultural data (data on agricultural holdings) in order to be able to conduct first agricultural survey in the period before the population census. Regarding the sample design for the agricultural survey, which will be based on the new sampling frame, we propose to use two-stage design with rural enumeration areas as primary and agrucultural households as secondary sampling units. Some experiences from the ad-hoc agricultural survey on individual agricultural producers in 2007 in the Federation BiH gave us positive feed backs for continuing this activity as proposed above. At the end, it is necessary to stress that this kind of master sample extension will be worth-wile only in the case that statistical offices in Bosnia and Herzegovina agreed about agricultural survey as priority in their short-term programmes of statistical activities.

## REFERENCES

[1] Lynn, Peter: Development of a sampling method for household surveys in post-war Bosnia and Herzegovina, Statistics in transition, June 2004, Vol. 6, No.6, p. 953-977.
[2] Bodin, Jette: Advice on Master Sample for Bosnia and Herzegovina, Statistics Sweden, International Consulting Office, March 2000.
[3] Šabanović, Edin: Two stage sampling and multiple frame, TAIEX, Follow up regional workshop on crop statistics, Zagreb, 18-19 September 2008, http://taiex.ec.europa.eu/Seminar%20Organisation/tools/presentations/24946/AGR%2024946_Sabanovic.ppt
[4] The B&H Household Budget Survey: Methodology, Agency for Statistics of Bosnia and Herzegovina, 2004.
[5] The B&H Household Budget Survey: Final results, Agency for Statistics of Bosnia and Herzegovina, 2004.
[6] Sample Frame Update for B&H, Final Project Report, UNDP B&H, 2006.

*Edin Šabanović*

# NESAVRŠENOSTI OKVIRA ZA UZORAK: ISKUSTVA ISTRAŽIVANJA U BOSNI I HERCEGOVINI I BUDUĆI IZAZOVI

## REZIME

*Istraživanja u domaćinstvima su istraživanja na uzorku i kvalitet njihovih rezultata zavisi, između ostalog, od izrade uzorka. Reprezentativni uzorci su izvučeni iz okvira za uzorak, a to su obično registri stanovništva ili domaćinstava, popisi stanovništva ili njihovi ažurirani podaci. U nedostatku ovih izvora podataka, u istraživanjima nailazimo na nesavršenosti koje povećavaju uticaj neuzoračkih grešaka i imaju uticaj na procene. Bosna i Hercegovina je bila u situaciji koja je zahtevala delikatna rešenja. Ovaj članak ilustruje karakteristike okvira za uzorak i različite izrade uzorka u Bosni i Hercegovini, kao i organičenja u izboru uzorka i procedure koje su primenjene u poslednjih sedam godina.*

*Associate Professor Vesna Bucevska, Ph.D*

# AN EMPIRICAL ESTIMATION OF THE DETERMINANTS OF FOREIGN DIRECT INVESTMENTS IN EU CANDIDATE COUNTRIES

## ABSTRACT

*This paper examines the potential implications of EU accession for the inflows of foreign direct investments (FDI) from Germany, as the biggest investor among the countries from EU, in the three EU candidate countries: Turkey, Croatia and Macedonia using an estimated gravity model. Once the main characteristics of the country donator and countries receivers of FDI have been recognized, we can forecast the future FDI inflows. In the paper we have identified the following key determinants: the size and the openness of the economy receiver, the risk of the country receiver, the labour costs in the country receiver and the distance between the country investor and the country receiver. The countries with low inflows of FDI can become more attractive for potential investors by focusing on some of the key determinants that are here identified.*

***Key words:*** *foreign direct investments, key determinants, candidate countries for EU, the gravity model.*

## 1. INTRODUCTION

Enlargement has been in the focus of the European Union over several decades. The Copenhagen European Council in December 2002 confirmed the European perspective of the countries of the Western Balkans, as potential candidates. The European Council in March 2003 reiterated that the future of the Western Balkans is within the European Union and pledged the Union's full support to the endeavours of the countries of the region to consolidate democracy, stability and to promote economic development.

The present agenda of EU enlargement covers the countries of the Western Balkans (Macedonia and Croatia) and Turkey. They are EU candidate countries, but are at various stages on their road toward the EU. Croatia and Turkey stared accession negotiations on 3 October 2005. In December 2005 the European Council granted Macedonia status of a candidate country; however accession negotiations have not still started. All other Western Balkan countries are considered to be potential candidate countries.

The motives for joining the European Union are partly political, and partly economic. The political motive is based on the conviction that the membership of a supranational organization, such as the EU, could eliminate the threat of war. The

economic motive rested on the belief that joining of the EU would create larger markets and promote competition, and thus lead to greater productivity and higher living standard and respond better to the challenges of globalization.

This paper elaborates the economic effects of the EU accession likely accession of candidate countries' to the European Union. However, it does not elaborate all economic implications, but rather focuses on the effect that the EU accession could have on the foreign direct investments (FDI) of the EU candidate countries. According to the economic theory, regional economic integration provides an important stimulus not only to trade, but also to FDI flows, thus enhancing economic growth and the catching-pup process. In this paper we will provide an econometric support of the above economic theory on the example of the EU candidate countries: Macedonia, Croatia and Turkey.

## 2. THE FDI INWARD STOCKS AND TRENDS IN EU CANDIDATE COUNTRIES

The EU candidate countries (Macedonia, Croatia and Turkey) have rich opportunities and potential for joining the EU and are experiencing significant progress in terms of economic reforms, economic growth and foreign direct investment. They have achieved significant progress. The EU candidate countries are deepening their mutual cooperation in a number of fields, for instance, through conclusion of bilateral free trade agreements and creating regional market of electricity. However the challenges in the EU candidate countries are still related to further improvement of investment climate, access to funds and better business environment, especially for small and medium-sized enterprises. Part of the efforts to attract more FDI inflows involves restructuring of the political and legal sectors and public administration, promoting growth, stability and sustainable development Common goal of the there EU candidate countries (Macedonia, Croatia and Turkey) is become members of the EU in order to overcome the obstacles for attracting FDI and improve the overall economic performance.

The main advantages of the EU candidate countries as FDI host countries are the size of the market, high skilled labor force, good productivity and competitive cots, improved infrastructure, bilateral agreements for mutual promotion and protection of investments, bilateral agreement for avoidance of double taxation, etc.

According to Eurostat data[1)], in terms of total FDI inflows denominated in Euros, the largest economy in the EU candidate countries, Turkey is on the top of the list of FDI inflows in EU candidate countries in 2005. However according to the more indicative ratios of FDI per capita and FDI per GDP Croatia is the best performer. The start of EU accession negotiations with Macedonia, Croatia and Turkey will further improve the FDI inflows to these there candidate countries, but the overall effect will be much mort obvious in Turkey.

According to the same source of data, FDI flows from EU-25 to candidate countries increased from EUR 5 billion in 2004 to EUR 10 bn in 2005. This increase is a result of higher FDI inflows to all EU candidate countries. However, Turkey was the most attractive destination for FDI flows from EU-25. It registered an increase of FDI flows from EUR 1,1 bn in 2004 to EUR 3,3 bn in 2005.

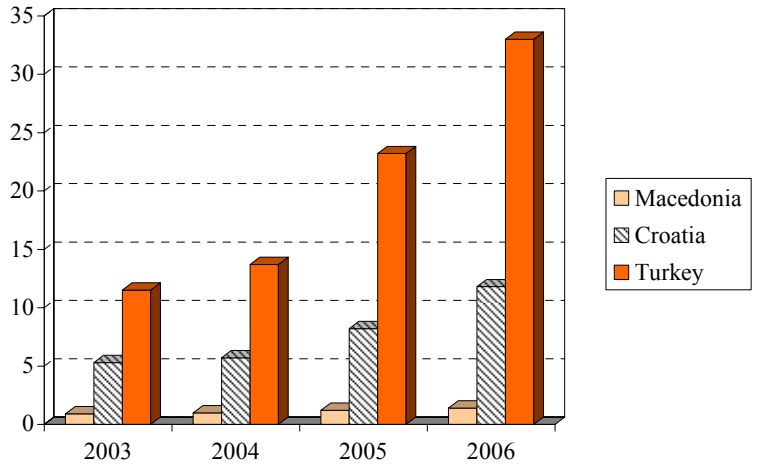
---

[1)] Eurostat pocketbooks, *European Union Foreign Direct Investment Yearbook 2007,* Data 2001-2005

In 2005, the FDI flows from Turkey to EU reached the amount of EUR 3,2 bn, and in Croatia EUR 1 bn. For comparison, the FDI flows from EU-25 to Macedonia in 2004 amounted EUR 1,2 bn, and in 2005 EUR 1,3 bn.[2)]

The FDI stocks from EU-25 to candidate countries have gradually increased in the period from 2001 to 2005, reaching the amount of EUR 42 bn at the end of 2005.

In the period 2001-2005, Turkey attracted most of the FDI inflows among the candidate countries which at the end of 2005 EUR amounted 16,6 bn.




*Source: Eurostat, Statistics in focus-68/2008, web site of the National Bank of the Republic of Macedonia (www.nbrm.gov.mk) and own calculations*

***Figure 1:** Foreign direct investment stocks from EU-25 to Macedonia, Turkey and Croatia in the period 2003-2006 in billion EUR*

In Figure 2 we rank the EU candidate countries according to the FDI percentage of GDP and GDP growth rate in 2005 in four main groups. The ranking is based on the following assumptions:

- GDP growth rate under 4% is considered relatively low and GDP growth rate above 4%-relatively high
- FDI/GDP ratio under 4% is considered relatively low and FDI/GDP ratio above 4%-relatively high.

[2)] Source: www.nbrm.gov.mk

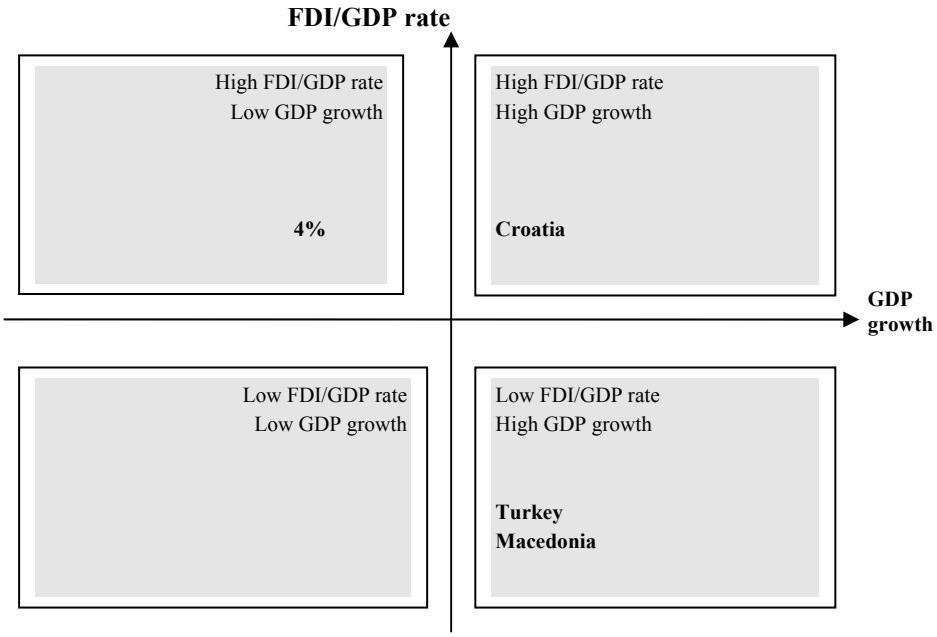



***Figure 2*** *FDI/GDP and GDP growth rate (2005)*

As can be seen from Figure 2, all candidate countries are enjoying high GDP growth rates. However, only Croatia in 2005 had high GDP growth rate and high FDI inflows as percent of GDP, while Macedonia and Turkey enjoyed high GDP growth rate, but low FDI inflows measured as percent of GDP.

## 3. THE DETERMINANTS OF FOREIGN DIRECT INVESTMENTS IN EU CANDIDATE COUNTRIES

There is a growing body of literature on determinants of foreign direct investment, yet literature which treats specifically the determinants of foreign direct investment in EU candidate countries is rather scarce. According to the theory, the main determinants of the inward FDI flows include:

- The size of the market;
- Economic prospects of the country;
- Level of income, measured by the GDP per capita;
- Openness of the economy, expressed as the ratio between trade (exports and imports) to GDP, but also by import tariffs of the host country;
- Business Climate and
- Labor market conditions[3)].

[3)] Botric, V. and Skuflic, L., Main Determinants of Foreign Direct Investment in the South East European Countries, paper prepared for the 2nd Euroframe Conference on Economic Policy Issues in the European Union

The above determinants can be grouped into two main groups: market related (which cover the GDP, GDP per capita and the GDP growth rate) and trade related-specifically, openness of the market. Besides mentioned traditional determinants, some economist use non-traditional determinants, such as human capital[4]. In our case of the EU candidate countries, we have chosen the following determinants which affect the FDI inflows from EU-25 to those countries:

1. The size of the market is a very important factor for attracting FDI. Economic variables, such as population, GDP, GDP per capita and growth rate of GDP can be used in order to determine the size of the market. The size of the market is an indicator of the domestic market sophistication. Thus economies with bigger markets, but also with other favorable FDI determinants, could attract more FDI flows. The size of the market is an important factor for attracting FDI flows, since it enables using of the economies of scope and the standard specialization of production factors, resulting in costs decline and market increase, thus increasing the supply side in the host country. For a foreign investor a bigger market means higher sales potential in the domestic market and higher sales profitability compared to exports to foreign markets which justifies the use of local resources. Such increased market provides the investor with more opportunities for sale and profits, and attracts more FDI. However, Edwards (1990) and Asidu (2002) show that the size of the market does not have significant impact on the FDI flows.
2. Second important determinant of FDI that can be assessed within the overall economic policies is the openness of the economy. According to Basar and Tosunoglu (2005) a country can attract more FDI if the ratio between the foreign trade (import and export) and GDP is higher, other authors, Caves (1996) and Singh and Jun (1995) doubt in the existence between the FDI and the openness of the economy. In this article we measure the openness of the economy as a ratio of the average export and import to GDP, as it is registered in the data base of Euro stat. The expected effects may differ by the type of investment regarding local market or export orientation, the host country's foreign exchange laws and applied capital taxation[5]. However, for our group if the three EU candidate countries (Macedonia, Croatia and Turkey), we expect that the openness of the economy will also indicate the level of integration of the economy in the regional economies. Therefore we expect the openness of the economy to have positive impact on FDI inward flows of EU candidate countries.
3. The labor costs are another important FDI determinant. This determinant is represented by the wages. Since the countries in our sample are with small markets (with exception of Turkey), we expect that one of the incentives for foreign investors to be the cheap labor force. Cheap labor is of particular interest for the EU-25 countries which wage level are high and which companies are look for reduction of costs by relocating production countries where resources are available at a lower cost.. Therefore, there could be also a positive correlation between FDI and labor costs.

"Trade, FDI and Relocation: Challenges for Employment and Growth in the European Union?", June 3rd, Vienna, Austria, 2005

[4] Ibid

[5] Ibid

4. Another important FDI determinant is the inflation rate. Higher return on investment boosts FDI and as a result of that the increase of prices of products in which the foreign investor invested, should be positively correlated to the FDI. However, very high inflation rates can represent a barrier to FDI, since it is clear sign of macroeconomic instability. Therefore we can not predetermine the expected sign of inflation rate.
5. A very important FDI determinant is also the country's risk rating. Deichman (2001) and Bevan and Estrin (2000) found a significant positive relationship between the country risk and FDI. This means that a healthy investment climate characterized by macroeconomic stability and political stability benefit the FDI of the host country. In this paper we use the Euromoney bi-annual country risk (dated September) to produce our annualized data. Euromoney country risk establishes an overall score for countries using nine weighted categories: Economic performance (25% weighting), Political Risk (25%), Debt indicators (10%), Debt in default or rescheduled (10%), Credit ratings (10%), Access to bank finance (5%), Access to short-term finance (5%), Access to capital markets (5%) and Discount on forfaiting (5%). The overall score is calculated as follows: the highest score in each category receives the full mark for the weighting; **the lowest receives 0**. In between, figures are calculated according to the formula: final score = (weighting / (maximum score-minimum score)*(maximum score-minimum score). The ranking shows the final scores after weighting.

## 4. ECONOMETRIC MODEL

The previous section pointed out that although there is growing body of literature covering foreign direct investment, yet there is not generally accepted econometric model which can serve as a basis for empirical research.

The objective of this section is to outline the econometric model for estimating the impact of EU accession on the growth of FDI in EU candidate countries. Neither general considerations nor the previous enlargement process allows for ready quantification of the impact of EU accession on FDI. One possible approach to quantify the effect of EU accession on FDI flows is to estimate a gravity model. This model has proved very successful in estimating trade flows and has been used by some authors (Di Mauro and Egger and Pfaffermayr) for estimation of the effect of EU integration on FDI flows. Due to data constraint, we have been forced to focus on shorter time series from 1998 to 2005 and to analyze the FDI inflows from only one country of EU-25, Germany to the three EU candidate countries (Macedonia, Croatia and Turkey). Why Germany? According to the Eurostat European Union Foreign Direct Investment Yearbook 2007[6)] the main investors among the EU member states were: Great Britain, France, Germany and Netherlands which accounted for 52% of the total EU outward FDI stocks at the end of 2004. Only Germany accounts for 11% of the total EU outward FDI stocks at the end of 2004. On the other hand Germany as the main holder of FDI stocks in candidate countries with 17% of EU FDI

[6)] See 1

stocks allocated in these three candidate countries. Out of these three countries, Turkey was the main investment destination for Germany.

The purpose of the econometric model, which will be specified below, is to establish a relationship, if any, between the economic integration and foreign direct investment. The estimation is based on a sample of 24 observations (8 years x 3 countries). We suggest the following model:

$$FDI_{ij} = \beta_0 + \beta_1 OPENESS_j + \beta_2 \ln GDP_j + \beta_3 WAGES + \beta_4 RANK + \beta_5 DISTANCE + u$$

whereas: *i*=Germany, and *j*=EU candidate countries (Macedonia, Croatia and Turkey).

*FDI*- the dependent variable is the FDI inward flow in the EU candidate countries from Germany at the end of the year measured in million Euros;

*OPENESS*- openess of the economy defined as a share of trade (import and export) in GDP[7)].

ln*GDP* - logarithm value of the gross domestic product of the host country measured in million Euros[8)].

*WAGES*- labour costs in each of the EU candidate economies. Data are provided as average gross monthly wages expressed in US dollars. In order to produce a bigger contrast between the labour costs of the investor country (Germany) and the labour costs of the host country, we have used the simple absolute value of the difference between the wages of those countries in order to demonstrate the magnitude of changes in the labour costs from the investor country to the host country[9)].

*RANK*- this variable denotes the ranking of the country according to the Euromoney country risk ranking.

*DISTANCE*- airline distance in kilometers between the capital city of Germany, Berlin and the capital cities of Macedonia, Croatia and Turkey (Skopje, Zagreb and Ankara). Data for the distance between the capital cities is obtained from www.indo.com/distance.

On the basis of the data used and by applying the econometric software package EViews 6, we obtain the following results:

Dependent Variable: LOG(FDI)
Method: Least Squares
Date: 01/30/08 Time: 10:20
Sample: 1 24
Included observations: 24

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -11.41405 | 2.008549 | -5.682734 | 0.0000 |
| OTVORENOST | 0.070563 | 0.020008 | 3.526778 | 0.0024 |
| LOG(GDP25) | 1.597297 | 0.122164 | 13.07505 | 0.0000 |
| PLATI | -1.89E-05 | 0.000125 | -0.151055 | 0.8816 |
| RATING | -0.010625 | 0.005382 | -1.974189 | 0.0639 |
| DIST | -0.000906 | 0.000369 | -2.459155 | 0.0243 |

7) Source: EUROSTAT, Pocketbook on Candidate and Potential Candidate Countries, 2007
8) Source:: EUROSTAT, Pocketbook on Candidate and Potential Candidate Countries, p. 46, 2007
9) Source: EUROSTAT, Pocketbook on Candidate and Potential Candidate Countries, p. 24, 2007

| | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| R-squared | 0.985820 | Mean dependent var | | 5.706229 |
| Adjusted R-squared | 0.981881 | S.D. dependent var | | 2.150848 |
| S.E. of regression | 0.289516 | Akaike info criterion | | 0.571107 |
| Sum squared resid | 1.508753 | Schwarz criterion | | 0.865620 |
| Log likelihood | -0.853282 | Hannan-Quinn criter. | | 0.649241 |
| F-statistic | 250.2819 | Durbin-Watson stat | | 1.885024 |
| Prob(F-statistic) | 0.000000 | | | |

How to interpret the above obtained results? They suggest that the estimated econometric model has a significant explanatory power, having in mind that value of the adjusted coefficient of determination R2 is equal to 98.18%. The value of the F-ratio (250.28) is big enough to conclude that there is a joint significance of the chosen independent variables. They are also individually significant (with exception of the wages that can excluded from the model in future), suggesting that there is no multicollinearity in the model.

The size of the market, expressed as coefficient of the variable lnGDP in the model 1,157 has proven to be the most important FDI determinant in our model (very high t ratio of 13.075 and statistical significance of 0.0000). This variable is in logarithmic form. The significance of the variable, even in logarithmic form conforms that the relationship between FDI and the size of the market is not a simple linear relationship, since the benefit of bigger markets grows, but at declining rate. The OPENNESS of the economy has also proven to be statistically significant at a level of significance of 0.0024 and it shows that if the share of import and export in the GDP increases for one percentage point, that will lead to increase of the volume for 0.07 million EUR for each country. Also the risk of investing in one country (the variable RANKING) is statistically significant and shows that if the country ranking improves for one place, that will contribute to increase of the FDI inflows in EU candidate countries for 0.005 million Euros. The distance (DIST) confirms the former expectations i.e. it affect inversely the level of FDI. In the estimated model, the coefficient of the variable DIST shows that if the distance between Berlin and one of the capital cities of the EU candidate countries increases for one kilometer, that will lead to decrease of the FDI inflow to EU candidate countries for 0.0009 million Euros.

## REFERENCES:


[1] Botric, V. and Skuflic, L., Main Determinants of Foreign Direct Investment in the South East European Countries, paper prepared for the 2nd Euroframe Conference on Economic Policy Issues in the European Union "*Trade, FDI and Relocation: Challenges for Employment and Growth in the European Union?"*, June 3rd, Vienna, Austria, 2005.

[2] Brenton, P., F. Di Mauro and M. Lucke, Economic integration and FDI: An empirical analysis of foreign investment in the EU an in Central and Eastern Europe, *Empirica*, 26, pp.95-121, 1999.

[3] Caves, R.E., *Multinational Enterprise and Economic Analysis (2nd Edition)*, Cambridge University Press, Cambridge and New York, 1996.

[4] Deutche Bundesbank, *Kapitalverflechtung mit dem Ausland, Statistische Sonderveroffentlichung 10*, various issues.

[5] Egger, P., An econometric view on the estimation of gravity models and the calculation of trade potentials, *The World Economy*, 25, pp. 297-312, 2002.

[6] Eurostat, Pocketbook on Candidate and Potential Candidate Countries, 2007.

[7] Eurostat Pocketbooks, *European Union Foreign Direct Investment Yearbook 2007, Data 2001-2005,* European Commission and Eurostat, 2007.

[8] Eurostat Pocketbooks, *European Union Foreign Direct Investment Yearbook 2008, Data 2001-2005*, European Commission and Eurostat, 2008.

[9] Eurostat, *Statistics in focus*-68/2008.


*Vesna Bucevska*

# EMPRIRIJSKA PROCENA DETERMINANTI DIREKTNIH STRANIH ULAGANJA U ZEMLJAMA KANDIDATIMA ZA ČLANSTVO U EU

## REZIME

*Ovaj rad ispituje moguće implikacije pristupa EU u vezi sa prilivom direktnih stranih investicija (DSI) iz Nemačke, kao najvećeg investitora među zemljema EU, u tri zemlje – kandidate za članstvo u EU: Tursku, Hrvatsku i Makedoniju, primenom modela procenjene gravitacije. Kada se prepoznaju i potvrde glavne karakteristike zemlje donatora i zemalja primalaca DSI, možemo da predvidimo budućnost priliva DSI. U ovom radu smo identifikovali i opisali sledeće ključne determinante: veličinu i ekonomsku otvorenost primaoca, stepen rizika zemlje primaoca, troškove rada u zemlji primaocu i udaljenost između zemlje investitora i zemlje primaoca. Zemlje sa niskim prilivom DSI mogu da postanu privlačnije za potencijalne investitore kroz fokusiranje na neke od ključnih determinanti koje su ovde opisane.*

*Olga Melovski Trpinac*
*Marija Panović*
*Republički zavod za statistiku*

# MESEČNA ANKETA MALIH INDUSTRIJSKIH PREDUZEĆA U SRBIJI[1)]

## REZIME

*U radu su ukratko prikazani metodologija i rezultati mesečne ankete malih preduzeća prerađivačkih delatnosti koju Republički zavod za statistiku sprovodi od januara 2004. Analiza serija indeksa prerađivačke proizodnje pokazuje da proizvodnja u malim preduzećima brže raste nego u velikim i da bi uključivanje proizvodnje malih u obračun mesečnog indeksa prerađivačke proizvodnje u proseku povećala indeks za 0,26 u periodu od januara 2004. do decembra 2007. godine. Dekompozicija serija je pokazala da serija desezoniranih indeksa proizvodnje malih preduzeća ima veći varijabilitet i veća kolebanja nego serija velikih preduzeća. Pri tumačenju tog varijabiliteta, osim ekonomskih faktora, treba uzeti u obzir da su indeksi malih preduzeća ocene na bazi slučajnog uzorka. Trend ciklus komponenta malih preduzeća uz rast izrazitije pokazuje i ciklične promene u datom periodu.*

***Ključne reči:*** *mala preduzeća prerađivačke delatnosti; serije mesečnh indeksa, desezoniranih indeksa i trend-ciklus komponenti prerađivačke proizvodnje.*

## 1. UVOD

Promene društveno-ekonomskih prilika krajem 80-tih godina prošlog veka omogućile su osnivanje preduzeća u privatnom vlasništvu i pokretanje procesa tranzicije planske u tržišnu ekonomiju. Statistika, koja je do tada pratila samo rad preduzeća u društvenoj i državnoj svojini, morala je da uvodi dopunska isrtaživanja kako bi obuhvatila i rad velikog broja novonastalih privatnih preduzeća. Tokom 90tih godina uvedene su ankete, istraživanja na bazi slučajnih uzoraka, za dopune redovnih istraživanja trgovine, zaposlenosti i ugostiteljstva. Vremenom diferencijacija na privatna i ostala preduzeća postaje manje važna od diferencijacije preduzeća na mala i velika. Sada se od Statistike zahteva ne samo da poboljša obuhvat istraživanja već i da posebno prati rad malih ekonomskih subjekata.

U istraživanjima industrije, rad malih preduzeća prerađivačke delatnosti prati se tek od januara 2004. godine, kada je uvedeno mesečno istraživanje na bazi uzorka. Ovo istraživanje dopunjuje redovno mesećno istraživanje IND-1 kojim se prati rad velikih preduzeća industrijske delatnosti. Podaci prikupljeni u anketi koriste se da bi se ocenio indeks proizvodnje malih preduzeća i obračunao indeks ukupne proizvodnje za prerađivačku industriju.

[1] Zahvaljujemo se Ljubici Živadinović, Jasmini Kostić-Simov i Nadi Đerić na savetima i pomoći prilikom izrade ovog rada. Posebnu zahvalnost dugujemo profesoru Miodragu Nikoliću.

Prema podacima završnih računa, mala preduzeća čine oko 87% od ukupnog broja preduzeća koja se bave prerađivačkom delatnošću, a njihovo učešće u ukupnom prihodu od prodaje proizvoda je oko 14%.

Cilj ovog rada je da na bazi podataka istraživanja malih industrijskih preduzeća za period od januara 2004. do decembra 2007. pruži odgovore na dva važna pitanja: "da li razvoj malih preduzeća teče isto kao i velikih i da li bi se njihovim obuhvatanjem promenili rezultati obračuna indeksa prerađivačke proizvodnje"?

Nakon kratkog prikaza metodologije mesečnog istraživanja malih preduzeća prerađivačke delatnosti, date su serije indeksa prerađivačke proizvodnje. Osim originalnih serija indeksa, po prvi put, su za mala preduzeća prikazane: serija sezonske komponente, desezonirana serija i serija trend-ciklus komponente (Hendersonova kriva). Poređenja radi, date su i odgovarajuće serije velikih preduzeća industrijske delatnosti. Dekompozicija serija je izvršena metodom X-11-ARIMA koja se koristi u Republičkom zavodu za statistiku.

## 2. METODOLOŠKE OSNOVE MESEČNOG ISTRAŽIVANJA INDUSTRIJE MALIH PREDUZEĆA

Cilj Mesečnog istraživanje na bazi uzorka (AIND) jeste da se prikupe podaci o ekonomskim aktivnostima skupa preduzeća koji nije obuhvaćen redovnim mesečnim istraživanjem, da se oceni mesečni indeks proizvodnje tog skupa i koriguje mesečni indeks industrijske proizvodnje.

Prikupljaju se podaci:
- o prihodu od prodaje proizvoda i usluga (u prethodnoj godini i tekućem mesecu);
- o broju zaposlenih – ukupno, i broju zaposlenih u proizvodnji;
- o pretežnoj delatnosti preduzeća i
- o stavovima izveštajne jedinice o kretanju u odnosu na prethodni mesec (lošije, isto, bolje) plasmana proizvoda, novih narudžbenica, nabavki sirovina, nabavki energije i goriva i ekonomske politike.

Osnovni skup istraživanja AIND čine preduzeća koja se bave aktivnostima u okviru prerađivačke industrije, koja imaju manje od 50 zaposlenih i koja nisu obuhvaćena redovnim istraživanjima u industriji.

Okvir za izbor uzorka za godinu *t* istraživanja formira se na bazi Statističkog poslovnnog registra (SPR) od 31.12. godine *t-1* i čini ga skup preduzeća koji ispunjava sledeće uslove:
- da im je status registracije aktivan;
- da imaju manje od 50 zaposlenih;
- da im je pretežna delatnost (statističkog preduzeća u SPR-u) iz oblasti prerađivačke industrije (oblasti 15-37 NACE rev. 1.1);
- da nisu uključena u mesečna i godišnja istraživanja industrije (IND-1 i IND-21);
- da su predala završni račun za godinu *t-2*.

Iz ovako izdvojenog skupa preduzeća isključuje se 5% najmanjih. Konačan okvir za izbor uzorka malih preduzeća sadrži oko 3,5 hiljada preduzeća. S druge strane, u redovnom mesečnom istraživanju industrije prati se rad oko 2 hiljade velikih preduzeća.

U istraživanjima AIND u 2004. i 2005. veličina preduzeća utvrđena je prema prihodu od prodaje proizovda i usluga iz završnih računa za 2002, odnosno 2003. godinu. U istraživanjima AIND u 2006. i 2007. veličina preduzeća utvrđena je, umesto prema prihodu od prodaje proizvoda i usluga, prema troškovima materijala preduzeća iskazanim u završnim računima za 2004, odnosno 2005. godinu, jer podaci o prihodu od prodaje proizvoda i usluga nisu postojali u tim završnim računima. Analiza podataka iz završnih računa za 2002. i 2003. pokazala je da za mala preduzeća postoji visoka korelacija između prihoda od prodaje proizvoda i usluga i troškova materijala. Time je opravdan izbor nove mere veličine.

Okvir je stratifikovan prema veličini na šest stratuma. Jedan stratum čine najveća među malim preduzećima. Ostala preduzeća su razvrstana u 5 jednako veličinskih stratuma: lista preduzeća se sortira po veličini preduzeća u rastući niz i granice stratuma se postavljaju tako da zbirovi mere veličine po stratumima budu približno jednaki.

Od 2006. godine stratfikacija okvira je izvršena, osim po veličini, i prema teritoriji na Centralnu Srbiju bez Beograda, Beograd i Vojvodinu. Na taj način se postiže veća homogenizacija preduzeća po stratumima i bolja kontrola alokacije uzorka po teritoriji.

U istraživanjima 2004, 2005, alokacija uzorka je *x*- optimalna (Sarnall, Swensson i Wretman, 1992, strana 107), a u istraživanjima 2006. i 2007, godine uzorak je alociran prema Bethel-ovom algoritmu (Bethel, 1989). Pri oba metoda alokacije korišćena je pomoćna varijabla - mera veličine (prihod od prodaje proizvoda, odnosno troškovi materijala). Za alokaciju Bethelovim algoritmom zadana je granica greške od 1,8% za ocenu totala troškova materijala za svaki od tri domena teritorije, Centralnu Srbiju bez Beograda, Beograd i Vojvodinu.

U tabeli 1 dati su podaci o ukupnom broju preduzeća prerađivačke delatnosti, o broju malih preduzeća, o broju preduzeća u okviru za izbor uzorka i obimu uzorak za period od 2004. do 2007. U tabeli 2 dati su podaci o broju preduzeća i alociranom uzorku po stratumima. Stratum 1 sadrži najmanja preduzeća prema meri veličine dok stratum 6 sadrži najveća mala preduzeća i ona se popisuju.

U 2004. i 2005. izbor uzorka unutar stratuma bio je sistematski, pri čemu su jedinice prethodno sortirane po delatnosti i teritoriji. Od 2006, uzorak unutar stratuma je prost slučajan, a izbor jedinica vrši se uz pomoć permanentnih slučajnih brojeva. Za izbor uz pomoć slučajnih brojeva svakoj jedinici okvira pridružuje se slučajan broj iz intervala (0,1]. Jedinice unutar svakog stratuma *h* sortiraju se u rastući niz po pridruženim slučajnim brojevima. Prvih $n_h$ jedinica sortirane liste, čiji su slučajni brojevi veći od početka $\alpha$, čine izabrani uzorak iz stratuma *h*. Ovakav izbor uzorka je jednostavan i omogućava koordinaciju uzroka u vremenu.

***Tabela 1.*** *Broj preduzeća prerađivačke delatnosti i broj preduzeća u okviru i uzorku malih preduzeća*

| Godina | Broj preduzeća prerađivačke delatnosti | | | |
|---|---|---|---|---|
| | ukupno | u skupu malih preduzeća | u okviru za izbor uzorka malih preduzeća | u planiranom uzorku malih preduzeća |
| 2004 | 15106 | 13214 | 3407 | 358 |
| 2005 | 16030 | 14127 | 3683 | 336 |
| 2006 | 16300 | 14209 | 3495 | 313 |
| 2007 | 16241 | 14256 | 3772 | 334 |

***Tabela 2****. Broj jedinica u okviru i uzorka u veličinskim stratumima po godinama istraživanja*

| | Broj preduzeća | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 2004 | | 2005 | | 2006 | | 2007 | |
| Stratum | okvir | uzorak | okvir | uzorak | okvir | uzorak | okvir | uzorak |
| 1 | 1559 | 56 | 2022 | 48 | 2005 | 84 | 2231 | 101 |
| 2 | 800 | 44 | 795 | 49 | 720 | 46 | 759 | 49 |
| 3 | 492 | 46 | 421 | 55 | 374 | 39 | 385 | 47 |
| 4 | 370 | 97 | 240 | 60 | 206 | 36 | 211 | 37 |
| 5 | 127 | 56 | 140 | 59 | 120 | 38 | 127 | 42 |
| najveći - 6 | 59 | 59 | 65 | 65 | 70 | 70 | 58 | 58 |
| Ukupno | 3407 | 358 | 3683 | 336 | 3495 | 313 | 3771 | 334 |

Tokom godine, jednom u 6 meseci, deo jedinica uzorka iz manjih veličinskih stratuma se zamenjuje. Kandidati za zamenu su jedinice koje se nisu nijednom odazvale u toku 6 meseci ili one koje su se odazvale 10 i više puta, i to u sukcesivnim mesecima. Zbog male realizacije uzorka, od oko 35% od planiranog obima uzorka, korišćen je postupak zamena, a ne planska rotacija jedinica uzorka. Slabom odazivu preduzeća doprinosi način anketiranja (putem pošte i telefonom) i relativno kratak vremenski periodu u kome je potrebno da se prikupe podaci.

Ocene totala, količnika i proporcija izvode se po standardnom postupku za stratifikovan prost slučajan uzorak. Ocena indeksa industrijske proizvodnje, za tekući mesec u odnosu na prosek prethodne godine, određuje se kao količnik ocene mesečnog prihoda od prodaje proizvoda i ocene prosečnog prihoda od prodaje proizvoda prethodne godine. Prethodno se mesečni prihodi jedinica realizovanog uzorka svode na stalne cene, odnosno prosečne cene prethodne godine u odgovarajućoj oblasti delatnosti po sledećem obrascu:

$$Prihodmd = \frac{Prihod}{I} \cdot 100$$

gde je:

*Prihodm* - Mesečni prihod ostvaren od prodaje proizvoda i usluga i izražen u tekućim cenama;

*I* - Indeks cena proizvođača industrijskih proizvoda za tekući mesec prema proseku prethodne godine;

*Prihodmd* - Mesečni prihod ostvaren od prodaje proizvoda i usluga i izražen u stalnim cenama (u cenama odgovarajućeg meseca prethodne godine).

Pri svođenju na stalne cene koriste se indeksi cena za tekući mesec u odnosu na prosek prethodne godine mesečnog saopštenja RZS ”Indeks cena proizvođača industrijskih proizvoda za Republiku Srbiju”.

Relativne greške ocena mesečnih indeksa industrijkske proizvodnje malih preduzeća za Srbiju su oko 5% u posmatranom periodu od januar 2004. do decembra 2007. Ove greške se postižu sa malom realizacijom uzorka, pa je moguća izvesna pristrasnost ocena.

Zvanično objavljeni podaci u mesečnom saopštenju 'Indeski industrijske proizvodnje' se odnose na proizvodnju velikih preduzeća. U napomeni Saopštenja daje se i ocena - indeksa ukupne prerađivačke industrije za Srbiju (integralni indeks) u čiji obračun ulazi i ocena indeksa malih preduzeća.

Integralni indeks prerađivačke proizvodnje određuje se sabiranjem, uz odgovarajuću ponderaciju, indeksa proizvodnje velikih preduzeća i indeksa proizvodnje u malim preduzećima

$$I = (1-\beta)\cdot I_v + \beta \cdot I_m$$

gde su $I$, $I_v$ i $I_m$ redom indeksi ukupne prerađivačke proizvodnje, indeks obračunat za velika preduzeća i indeks ocenjen na bazi ankete malih preduzeća, a $\beta$ je ponder za mala preduzeća. Za određivanje pondera za 2004. i 2005. godinu korišćeni su podaci završnih računa o prihodu od prodaje proizvoda i usluga. Ponder za mala preduzeća jednak je učešću malih preduzeća u ukupnom prihodu od prodaje proizvoda i usluga preduzeća prerađivačke delatnosti. U završnim računima za 2005. i 2006. godinu nisu postojali podaci o prihodu od prodaje proizvoda i usluga pa su učešća malih procenjena na osnovu učešća ovog skupa preduzeća u ukupnim troškovima materijala preduzeća prerađivačke delatnosti. U tabeli 3 data su učešća (ponderi) za određivanje integralnog indeksa prerađivačke industrije – ukupno za tekući mesec u odnosu na prosek prethodne godine.

***Tabela 3.*** *Ponderi za obračun integralnog indeksa*

| Godina | Velika preduzeća | Mala preduzeća |
|---|---|---|
| 2004 | 0,863 | 0,137 |
| 2005 | 0,863 | 0,137 |
| 2006 | 0,855 | 0,145 |
| 2007 | 0,872 | 0,128 |

## 3. REZULTATI ISTRAŽIVANJA

U tabeli 4 date su serije indeksa prerađivačke proizvodnje za velika preduzeća (zvanično objavljeni indeks), za mala preduzeća (indeks ocenjen u anketi) i serija indeksa ukupne prerađivačke industrije (indeks je ponderisana suma indeksa velikih i indeksa malih preduzeća). Date su i razlike integralnog indeksa i indeksa velikih preduzeća. Ove razlike predstavljaju vrednosti za koje bi se objavljeni indeks prerađivačke proizvodnje (mesec na prosek prethodne godine) promenio kada bi se u obračun uključila proizvodnja malih preduzeća. Prosečna vrednost razlika je 0,26 i standardna devijacija 1,12. Sa pouzdanošću od 95% razlike su u intervalu -1,93 do 2,46. Medijalna vrednost je 0,09 i ona je manja od prosečne vrednosti (0,24), tako da je raspodela razlika asimetrična udesno. To znači da ima više većih pozitivnih razlika nego većih negativnih razlika. Osim, toga broj negativnih razlika je najveći u 2004. kada ih je bilo 9, a opada sa vremenom tako da u 2007. ima samo jedna negativna razlika i to u januaru. Na kraju može da se zaključi da, iako uključivanje proizvodnje malih preduzeća u obračun indeksa industrijske proizvodnje ne menja indeks zna-

čajno, proizvodnja u malim preduzećima pokazuje tendenciju bržeg rasta nego što je to slučaj sa proizvodnjom u velikm preduzećima.

U tabeli 5 prikazani su podaci za mesece u kojima bi uključivanje proizvodnje u malim preduzećima u obračun indeksa prerađivačke proizvodnjE najviše smanjilo, odnosno povećalo objavljeni indeks. Tako bi u januru 2004. indeks prerađivačke proizvodnje bio manji za 3,23, a u julu 2007. veći za 2,16. Uočava se tendencija da je integralni indeks veći od ineksa proizvodnje velikih preduzeća.

***Tabela 4.*** *Serije indeksa prerađivačke proizvodnje i razlike integralnog indeka i indeksa velikih (mesec na prosek prethodne godine)*

| Godina i mesec | | Velika preduzeća | Ocena za mala preduzeća | Integralno za prerađivačku industriju | Razlika integralnog i indeksa velikih |
|---|---|---|---|---|---|
| 2004 | 1 | 78,7 | 55,14 | 75,47 | -3,23 |
| | 2 | 91,8 | 88,90 | 91,40 | -0,40 |
| | 3 | 116,2 | 106,34 | 114,85 | -1,35 |
| | 4 | 108,1 | 94,81 | 106,28 | -1,82 |
| | 5 | 112,6 | 99,63 | 110,82 | -1,78 |
| | 6 | 119,8 | 117,79 | 119,53 | -0,27 |
| | 7 | 106,8 | 103,65 | 106,37 | -0,43 |
| | 8 | 98,8 | 103,89 | 99,50 | 0,70 |
| | 9 | 115,1 | 114,38 | 115,00 | -0,10 |
| | 10 | 120,5 | 121,27 | 120,61 | 0,11 |
| | 11 | 122,5 | 120,77 | 122,26 | -0,24 |
| | 12 | 123,8 | 135,95 | 125,47 | 1,67 |
| 2005 | 1 | 73,1 | 76,46 | 73,56 | 0,46 |
| | 2 | 77,9 | 89,13 | 79,44 | 1,54 |
| | 3 | 95,9 | 105,41 | 97,20 | 1,30 |
| | 4 | 96,9 | 95,11 | 96,65 | -0,25 |
| | 5 | 96,6 | 93,17 | 96,13 | -0,47 |
| | 6 | 104,3 | 104,89 | 104,38 | 0,08 |
| | 7 | 95,3 | 101,85 | 96,20 | 0,90 |
| | 8 | 101,3 | 100,37 | 101,17 | -0,13 |
| | 9 | 107,6 | 107,32 | 107,56 | -0,04 |
| | 10 | 116,5 | 113,56 | 116,10 | -0,40 |
| | 11 | 114,8 | 104,06 | 113,33 | -1,47 |
| | 12 | 111,6 | 120,57 | 112,83 | 1,23 |
| 2006 | 1 | 75,5 | 90,34 | 77,65 | 2,15 |
| | 2 | 89,9 | 94,56 | 90,58 | 0,68 |
| | 3 | 104,9 | 104,29 | 104,81 | -0,09 |
| | 4 | 99,3 | 107,54 | 100,49 | 1,19 |
| | 5 | 108,8 | 116,54 | 109,92 | 1,12 |
| | 6 | 111 | 107,27 | 110,46 | -0,54 |
| | 7 | 105 | 112,96 | 106,15 | 1,15 |

| Godina i mesec | | Velika preduzeća | Ocena za mala preduzeća | Integralno za prerađivačku industriju | Razlika integralnog i indeksa velikih |
|---|---|---|---|---|---|
| | 8 | 104,3 | 108,14 | 104,86 | 0,56 |
| | 9 | 111,1 | 110,33 | 110,99 | -0,11 |
| | 10 | 118,7 | 113,68 | 117,97 | -0,73 |
| | 11 | 117,4 | 117,80 | 117,46 | 0,06 |
| | 12 | 118,9 | 113,41 | 118,10 | -0,80 |
| 2007 | 1 | 82,8 | 79,27 | 82,35 | -0,45 |
| | 2 | 88,7 | 94,68 | 89,47 | 0,77 |
| | 3 | 105,7 | 106,33 | 105,78 | 0,08 |
| | 4 | 102,4 | 104,98 | 102,73 | 0,33 |
| | 5 | 106,9 | 118,89 | 108,43 | 1,53 |
| | 6 | 107,3 | 115,51 | 108,35 | 1,05 |
| | 7 | 102,8 | 119,67 | 104,96 | 2,16 |
| | 8 | 105,8 | 118,47 | 107,42 | 1,62 |
| | 9 | 105,1 | 117,82 | 106,73 | 1,63 |
| | 10 | 116,1 | 120,9 | 116,71 | 0,61 |
| | 11 | 108,3 | 122,36 | 110,10 | 1,80 |
| | 12 | 111,3 | 121,29 | 112,58 | 1,28 |

***Tabela 5.*** *Najveće razlike integralnog indeksa i indeksa velikih preduzeća*

| Godina | Mesec | Velika preduzeća | Ocena za mala preduzeća | Integralno za prerađivačku industriju | Razlika integralnog i indeksa velikih |
|---|---|---|---|---|---|
| 2004 | 1 | 78,7 | 55,14 | 75,47 | -3,23 |
| 2004 | 4 | 108,1 | 94,81 | 106,28 | -1,82 |
| 2004 | 5 | 112,6 | 99,63 | 110,82 | -1,78 |
| 2005 | 11 | 114,8 | 104,06 | 113,33 | -1,47 |
| 2004 | 3 | 116,2 | 106,34 | 114,85 | -1,35 |
| 2007 | 9 | 105,1 | 117,82 | 106,73 | 1,63 |
| 2004 | 12 | 123,8 | 135,95 | 125,47 | 1,67 |
| 2007 | 11 | 108,3 | 122,36 | 110,10 | 1,80 |
| 2006 | 1 | 75,5 | 90,34 | 77,65 | 2,15 |
| 2007 | 7 | 102,8 | 119,67 | 104,96 | 2,16 |

Prosečna vrednost mesečnih indeksa predstavlja ocenu lančanog indeksa, tj. indeksa tekuće godine na prethodnu godinu. U tabeli 6 prikazani su lančani godišnji indeksi prerađivačke proizvodnje. Lančani indeksi malih preduzeća u periodu 2004. do 2007. pokazuju rast u tekućoj u odnosu na prethodnu godinu, a intenzitet rasta se kreće od 1% u 2005. u odnosu na 2004. do 11,7% u 2007. u odnosu na 2006. godinu. Lančani indeksi velikih pokazuju najveći rast od 9,6% u 2004. u odnosu na 2003. godinu, zatim pad od oko 0,7% u 2005. u odnosu na 2004. i manji rast od malih preduzeča u 2006, i 2007. godini. Uključivanje malih u obračun indeksa bi smanjilo indeks prerađivačke proizvodnje u 2004. a za 0,6, a najviše bi povećalo u 2007. godini, i to za 1.

***Tabela 6.*** *Lančani godišnji indeksi prerađivačke proizvodnje*

| Godina | Velika preduzeća | Ocena za mala preduzeća | Integralno za prerađivačku industriju | Razlika integralnog i indeksa velikih |
|---|---|---|---|---|
| 2004 /2003 | 109,56 | 105,21 | 108,96 | -0,60 |
| 2005 /2004 | 99,32 | 100,99 | 99,55 | 0,23 |
| 2006 /2005 | 105,40 | 108,07 | 105,79 | 0,39 |
| 2007 /2006 | 103,60 | 111,66 | 104,63 | 1,03 |

***Tabela 7.*** *Godišnji indeksi prerađivačke proizvodnje, baza 2005.*

| Godina | Velika preduzeća | Ocena za mala preduzeća | Integralno za prerađivačku industriju | Razlika integralnog i indeksa velikih |
|---|---|---|---|---|
| 2004 | 100,69 | 99,02 | 100,46 | -0,24 |
| 2005 | 100,00 | 100,00 | 100,00 | 0,00 |
| 2006 | 105,40 | 108,07 | 105,79 | 0,39 |
| 2007 | 109,19 | 120,67 | 110,68 | 1,49 |

Rast godišnje proizvodnje u malim preduzećima vidi se iz baznih indeksa (tabela 7). U 2004. proizvodnja velikih preduzeća bila je veća za 0,7% u odnosu na proizvodnju u 2005, dok je proizvodnja u malim preduzećima bila manja za 1%. U 2006. i 2007. i velika i mala preduzeća beleže rast proizvodnje u odnosu na proizvodnju u baznoj 2005. godini s tim što mala imaju veći rast. Uključivanje proizvodnje u malim preduzećima u obračun indeksa industrijske proizvodnje bi neznatno promenilo bazni indeks u 2004. i 2006, dok bi ga u 2007. povećalo za 1,5.

## 4. DEKOMPOZICIJA SERIJA PRERAĐIVAČKE PROIZVODNJE

Da bi se bolje sagledale tendencije razvoja serija indeksa industrijske proizvodnje malih i velikih preduzeća serije su svedene na stalnu bazu (baza 2005. godina) i predstavljene u formi multiplikativnog modela:

$$O(t) = TC(t) \cdot S(t) \cdot K(t) \cdot I(t)$$

gde je: *O(t)* originalna serija, *TC(t)* trend-ciklus komponenta, *S(t)* sezonska komponenta, *K(t)* kalendarska komponenta i *I(t)* slučajna (iregularna) komponenta.

Prosečni godišnji indeksi iz tabele 6 korišćeni su za preračunavanje mesečnih indeksa iz tabele 4 u serije na istoj bazi za ceo posmatrani period. Dekompozicija tih serija urađena je metodom X-11-ARIMA. Rezultati su prikazani na grafikonu 1 i u tabelama koje su date u Prilogu. Grafikon, osim originalnih serija indeksa proizvodnje velikih i malih preduzeća, prikazuje i serije sezonskih komponenata, serije desezoniranih indeksa i serije trend-ciklus komponenti.

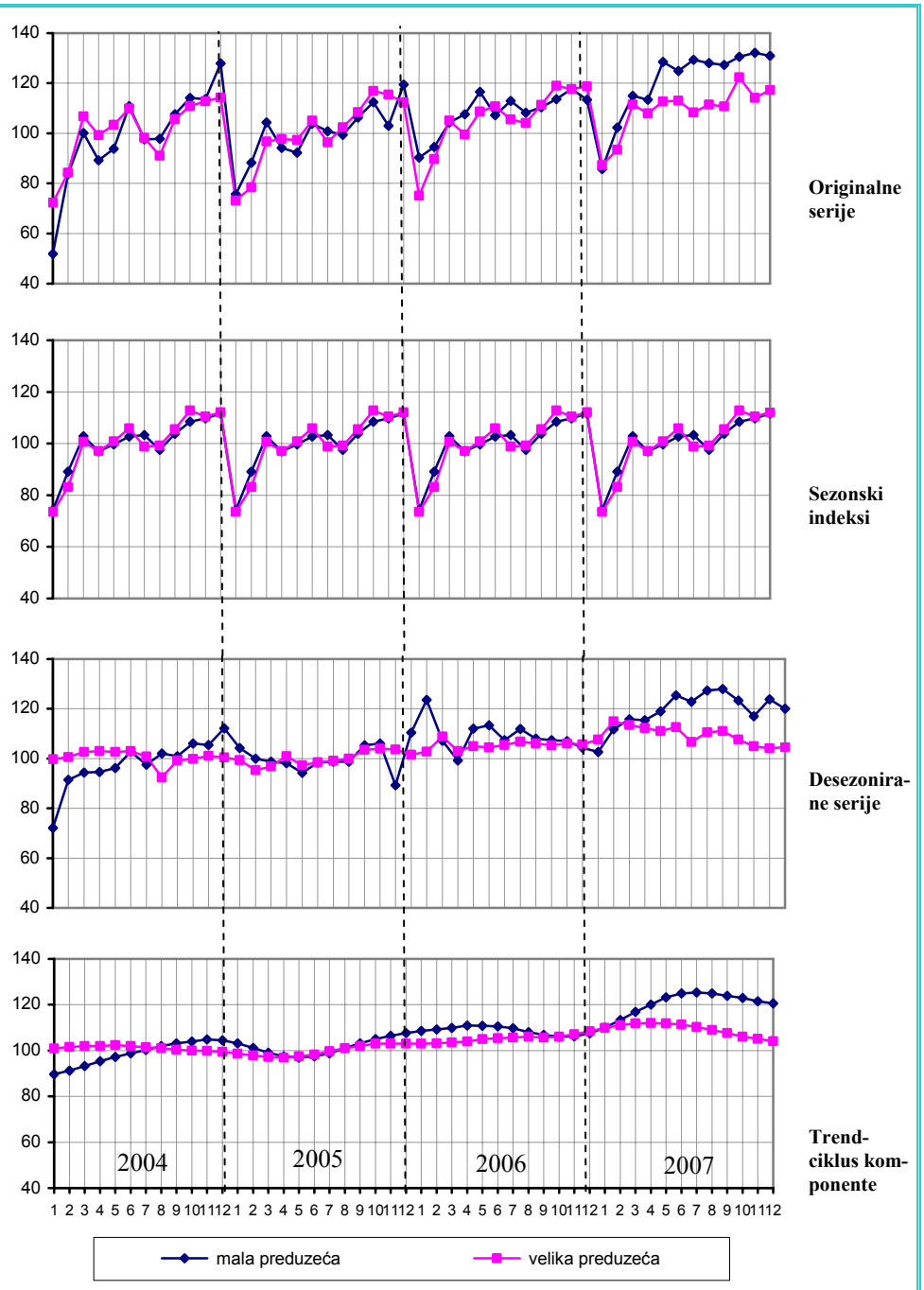
Sa grafikona 1 vidi se da vrednosti indeksa prerađivačke proizvodnje originalne serija malih preduzeća u nekim mesecima znatno odstupaju od odgovarajućih vrednosti indeksa velikih preduzeća, kao što je to slučaj, u januaru, aprilu, maju i decembru 2004, novembru 2005, januaru 2006. i od maja 2007, nadalje. Uprkos tome, periodi rasta, opadanja i stagnacija ne razlikuju se značajno.

Između serija sezonskih komponenti indeksa proizvodnje malih i velikih industrijskih preduzeća nema značajnijih razlika, te isti sezonski faktori deluju i na mala i na velika preduzeća prerađivačke proizvodnje.

Serija desezoniranih indeksa velikih preduzeća ima manja kolebanja u posmatramom periodu od serije malih preduzeća koja je nestabilnija sa više iznenadnih skokova i padova. Serija malih ima maksimume u decembru 2004, januaru, aprilu i maju 2006, a u januaru 2004. i novembru 2005. minimume. Od aprila do decembra 2007. proizvodnja u malim preduzećima konstantno ima ekstremni rast, koji je znatno veći od rasta proizvodnje u velikim preduzećima u tom periodu. Postoje slične tendencija razvoja velikih i malih preduzeća, ali postoje i tačke u kojima jedna serija menja pravac od rasta ka padu a druga od pada ka rastu, na primer, avgust 2004, novembar 2005 i januar 2006.

Trend-ciklus komponenta indeksa prerađivačke proizvodnje velikih u periodu od 2004. do decembra 2007. pokazuje blagi rast. Kriva trend-ciklusa malih preduzeća u posmatranom periodu pokazuje rast proizovdnje koji je, osim na početku posmatranog perioda, veći ili jednak rastu proizvodnje velikih preduzeća. Mada ciklična kolebanja nisu velika, postoje u obe serije. Znatno su izrazitija u seriji za mala preduzeća što može da ukaže na to da taj skup preduzeća oštrije reaguje na uzroke konjuktrurnih promena.

Pri tumačenju varijabiliteta serija indeksa malih preduzeća, osim ekonomskih faktora, treba uzeti u obzir da su indeksi malih preduzeća ocene na bazi slučajnog uzorka. Obim uzorka je oko 300 jedinica i realizacija je mala (oko 35%) te ocene sadrže i uzoračku grešku i grešku neodgovora. Osim toga, deo varijabiliteta može biti posledica ažuriranja okvira (koje se sprovodi jednom godišnje) kao i ažuriranja uzorka (koje se sprovodi jednom u šest meseci). Promene okvira i uzorka mogu u izvesnoj meri da promene pondere koji se koriste za određivanje ocena na bazi uzorka.




***Grafikon 1.*** *Mesečne varijacije indeksa proizvodnje malih i velikih preduzeća*

## 4. ZAKLJUČNE NAPOMENE

Rad sadrži metodološke osnove mesečnog istraživanja malih industrijskih preduzeća i deo rezultata koji se odnosi na indekse industrijske proizvodnje u tim preduzećima za period od januara 2004. do decembra 2007. Ovi rezultati se sada po prvi put publikuju.

Mala industrijska preduzeća iako znatno brojnija od velikih, učestvuju u industrijskoj proizvodnji sa oko 14%. Analiza serija indeksa prerađivačke proizvodnje malih preduzeća, velikih preduzeća i prerađivačke proizovodnje u celini pokazala je da:

- proizvodnja u malim preduzećima ne može značajno da utiče na kretanje mesečne prerađivačke proizvodnje. Uključivanje proizvodnje u malim preduzećima u obračun indeksa bi ga u proseku povećalo za 0,26, najviše smanjilo u januaru 2004 za 3,23, a najviše povećalo u julu 2007. za 2,17. Interval sa granicama -1,93 i 2,46 sa verovatnoćom od 95% sadrži razlike integralnog indeksa (obračunatog na osnovu proizvodnje velikih i malih preduzeća) i indeksa velikih preduzeća. Takođe, može da se uoči da u posmatranom periodu proizvodnja u malim preduzećima brže raste nego proizvodnja u velikim, te da obračunati integralni indeks prerađivačke proizvodnje ima tendeciju da bude veći nego indeks proizvodnje velikih preduzeća;
- uključivanje proizvodnje malih u obračun godišnjeg lančanog indeksa u periodu 2004. do 2007. najviše bi promenilo obračunat indeks za 1;
- bazni godišnji indeski malih preduzeća pokazuju tendenciju rasta u peridu 2004. do 2007;
- sezonska komponenta serije malih preduzeća ne razlikuje se značajno od sezonske komponente indeksa proizvodnje velikih preduzeća;
- serija desezoniranih indeksa malih preduzeća ima veći varijabilitet i veća kolebanja nego serija velikih preduzeća;
- trend ciklus komponenta malih preduzeća pokazuje rast koji je veći ili jednak rastu velikih preduzeća. Uz brži rast pokazuje i oštrije ciklične promene u datom periodu;
- Pri tumačenju varijabiliteta serija indeksa malih preduzeća, osim ekonomskih faktora, treba uzeti u obzir da su indeksi malih preduzeća ocene na bazi slučajnog uzorka.

## LITERATURA

**[1]** Bethel, J. "Sample allocation in multivariate surveys". *Survey methodology*, 15 (1989): 47-57, Statistics Canada.

[2] Brokwell P.J. and Davis R.A. (1996) *Introduction to Time Series and Forecasting* , Springer-Verlag, New York.

[3] Indeks fizičkog obima industrijske proizvodnje (2005). Metodologija i strandardi br 10. Republički zavod za statistiku, Beograd.

[4] Lohr L.S (1999) *Sampling: Design and Analysis,* Brooks/Cole Publishing Company

[5] Nikolić M. i Jevtić Ž. (1974) *Metod desezoniranja vremenskih serija,* Savezni zavod za statistiku.

[6] Nikolić M. i Marić N. (1982) *Desezonizacija Jugoslovenskih vremenskih serija 1971-1980, Studije Analize i Prikazi 109,* Savezni zavod za statistiku.

[7] Särndal, C.E., B. Swensson, J. Wretman. *Model Assisted Survey Sampling.* New York: Springer-Verlag, (1992).

# MONTHLY SURVEY OF SMALL INDUSTRIAL ENTERPRISES IN SERBIA

## ABSTRACT

*The monthly sample survey for small enterprises involved in manufacturing industry in Serbia (AIND) is carried out in order to supplement regular monthly survey for large industrial enterprises. Estimate of index of industrial production in the month of the current year with respect to the previous year's average for small enterprises is obtained from this survey. Index calculated for large enterprises is combined with this index, and as a weighted sum, a corrected monthly index of total manufacturing industrial production is derived. Analysis of production indices series shows that calculated indices for the total industrial production do not significantly differ from the indices for large enterprises in the period form January 2004 to December 2007. When compared to index of large enterprises, on the average, index for total industrial production is higher for 0.26. Seasonally adjusted series of production indices for small enterprises exhibits higher variability than the series for large enterprises with several extreme values. Trend-cycle components (Henderson curve) for small and large enterprises show that production in large and small enterprises grows in this period, though it is faster for small enterprises. In spite the differences, small and large enterprises have similar tendencies.*

***Key words:*** *small manufacturing industrial enterprises, seasonal adjusted series of indices for small and large industrial enterprises, Henderson curve*

## PRILOG

### Serije originalnih indeksa, desezonirane serije indeksa, trend-ciklus komponente i iregularne komponente

***Tabela 1.*** *Originala serija indeksa za mala preduzeća - baza prosek 2005.*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 51,9 | 83,7 | 100,1 | 89,2 | 93,8 | 110,9 | 97,6 | 97,8 | 107,7 | 114,1 | 113,7 | 128,0 |
| 2005 | 75,7 | 88,3 | 104,4 | 94,2 | 92,3 | 103,9 | 100,9 | 99,4 | 106,3 | 112,4 | 103,0 | 119,4 |
| 2006 | 90,3 | 94,6 | 104,3 | 107,5 | 116,5 | 107,3 | 113,0 | 108,1 | 110,3 | 113,7 | 117,8 | 113,4 |
| 2007 | 85,7 | 102,3 | 114,9 | 113,4 | 128,5 | 124,8 | 129,3 | 128,0 | 127,3 | 130,6 | 132,2 | 131,0 |

***Tabela 2.*** *Originala serija indeksa za velika preduzeća-baza prosek 2005.*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 72,3 | 84,2 | 106,7 | 99,2 | 103,4 | 109,9 | 98,2 | 91,0 | 105,6 | 110,8 | 112,8 | 114,3 |
| 2005 | 73,2 | 78,5 | 96,7 | 97,7 | 97,3 | 105,0 | 96,3 | 102,4 | 108,4 | 116,9 | 115,4 | 112,3 |
| 2006 | 75,0 | 89,7 | 105,1 | 99,5 | 108,7 | 110,7 | 105,5 | 104,1 | 111,3 | 118,9 | 117,6 | 118,7 |
| 2007 | 87,3 | 93,5 | 111,4 | 107,9 | 112,7 | 113,1 | 108,4 | 111,5 | 110,8 | 122,4 | 114,2 | 117,3 |

***Tabela 3.*** *Originala serija integralnog indeksa-baza prosek 2005.*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 69,6 | 84,3 | 105,9 | 98,0 | 102,2 | 110,2 | 98,1 | 91,7 | 106,0 | 111,2 | 112,7 | 115,7 |
| 2005 | 73,9 | 79,8 | 97,7 | 97,1 | 96,6 | 104,9 | 96,6 | 101,6 | 108,1 | 116,6 | 113,9 | 113,3 |
| 2006 | 77,7 | 90,6 | 104,8 | 100,5 | 109,9 | 110,5 | 106,2 | 104,9 | 111,0 | 118,0 | 117,5 | 118,1 |
| 2007 | 87,1 | 94,6 | 111,9 | 108,7 | 114,7 | 114,6 | 111,0 | 113,6 | 112,9 | 123,4 | 116,5 | 119,1 |

***Tabela 4.*** *Serija desezoniranih indeksa za mala preduzeća*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 72,1 | 91,5 | 94,4 | 94,6 | 96,2 | 102,8 | 97,5 | 102,0 | 100,9 | 106,0 | 105,4 | 112,2 |
| 2005 | 104,2 | 100,0 | 98,8 | 98,2 | 94,2 | 98,4 | 98,7 | 98,8 | 105,3 | 106,0 | 89,3 | 110,4 |
| 2006 | 123,6 | 107,1 | 99,3 | 112,0 | 113,3 | 107,4 | 111,9 | 108,0 | 107,4 | 107,0 | 104,4 | 102,7 |
| 2007 | 111,8 | 115,8 | 115,4 | 119,0 | 125,4 | 122,8 | 127,3 | 128,0 | 123,3 | 117,0 | 123,8 | 120,0 |

***Tabela 5.*** *Serija desezoniranih indeksa za velika preduzeća*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 99,7 | 100,6 | 102,7 | 103,0 | 102,6 | 103,0 | 100,8 | 92,4 | 99,3 | 99,9 | 101,1 | 100,5 |
| 2005 | 99,4 | 95,3 | 96,8 | 101,0 | 97,3 | 98,5 | 99,1 | 100,0 | 103,5 | 104,0 | 103,6 | 101,6 |
| 2006 | 102,8 | 108,9 | 103,0 | 105,0 | 104,5 | 105,4 | 106,8 | 106,0 | 105,3 | 106,0 | 105,7 | 107,6 |
| 2007 | 115,0 | 113,5 | 112,3 | 111,0 | 112,6 | 106,7 | 110,6 | 111,0 | 107,6 | 105,0 | 104,1 | 104,6 |

***Tabela 6.*** *Serija desezoniranih integralnih indeksa*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 94,3 | 97,5 | 101,7 | 101,0 | 102,3 | 103,9 | 100,9 | 93,8 | 99,8 | 101,0 | 102,1 | 101,6 |
| 2005 | 98,8 | 93,8 | 97,3 | 99,5 | 97,0 | 99,4 | 99,4 | 99,5 | 103,8 | 105,0 | 102,6 | 103,2 |
| 2006 | 104,1 | 106,5 | 102,5 | 106,0 | 105,6 | 106,8 | 107,7 | 107,0 | 105,8 | 106,0 | 106,5 | 107,7 |
| 2007 | 111,7 | 111,2 | 113,5 | 111,0 | 114,4 | 110,0 | 112,9 | 113,0 | 110,5 | 106,0 | 107,7 | 107,0 |

***Tabela 7.*** *Trend-ciklus komponenta - Hendersonova kriva za mala preduzeća*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 89,7 | 91,3 | 93,3 | 95,3 | 97,2 | 98,8 | 100,3 | 102,0 | 103,2 | 104,0 | 104,8 | 104,4 |
| 2005 | 103,1 | 101,1 | 99,1 | 97,5 | 96,9 | 97,4 | 98,8 | 101,0 | 103,1 | 105,0 | 106,5 | 107,6 |
| 2006 | 108,5 | 109,2 | 109,9 | 111,0 | 110,8 | 110,5 | 109,7 | 108,0 | 106,8 | 106,0 | 106,2 | 107,5 |
| 2007 | 110,0 | 113,3 | 116,9 | 120,0 | 123,2 | 124,9 | 125,4 | 125,0 | 123,9 | 123,0 | 121,5 | 120,6 |

***Tabela 8.*** *Trend-ciklus komponenta-Hendersonova kriva za velika preduzeća*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 100,9 | 101,5 | 102,0 | 102,0 | 102,4 | 102,0 | 101.4 | 101,0 | 100,4 | 100,0 | 99,9 | 99,4 |
| 2005 | 98,7 | 97,9 | 97,2 | 97,0 | 97,4 | 98,3 | 99,7 | 101,0 | 102,0 | 103,0 | 103,0 | 103,0 |
| 2006 | 103,0 | 103,2 | 103,6 | 104,0 | 105,0 | 105,4 | 105,6 | 106,0 | 105,7 | 106,0 | 107,1 | 108,4 |
| 2007 | 109,9 | 111,1 | 111,9 | 112,0 | 111,9 | 111,3 | 110,3 | 109,0 | 107,7 | 106,0 | 105,1 | 104,1 |

***Tabela 9.*** *Trend-ciklus komponenta-Hendersonova kriva za integralni indeks*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 98,0 | 99.1 | 100.2 | 101.2 | 101.9 | 102.1 | 101.9 | 101.6 | 101.3 | 101.1 | 100.7 | 100.2 |
| 2005 | 99.5 | 98.7 | 98.1 | 97.9 | 98.1 | 98.9 | 99.9 | 101.1 | 102.2 | 103.1 | 103.7 | 104,0 |
| 2006 | 104.2 | 104.4 | 104.8 | 105.4 | 106,0 | 106.4 | 106.5 | 106.5 | 106.5 | 106.7 | 107.4 | 108.6 |
| 2007 | 110,0 | 111.2 | 112.1 | 112.5 | 112.7 | 112.5 | 112,0 | 111.2 | 110.1 | 108.9 | 107.8 | 107.1 |

***Tabela 10.*** *Sezonska komponenta za mala preduzeca*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 74,3 | 89,1 | 102,8 | 97,0 | 99,7 | 102,7 | 103,3 | 97,6 | 103,8 | 108,4 | 109,8 | 111,6 |
| 2005 | 74,3 | 89,1 | 102,8 | 97,0 | 99,7 | 102,7 | 103,3 | 97,6 | 103,8 | 108,4 | 109,8 | 111,6 |
| 2006 | 74,3 | 89,1 | 102,8 | 97,0 | 99,7 | 102,7 | 103,3 | 97,6 | 103,8 | 108,4 | 109,8 | 111,6 |
| 2007 | 74,3 | 89,1 | 102,8 | 97,0 | 99,7 | 102,7 | 103,3 | 97,6 | 103,8 | 108,4 | 109,8 | 111,6 |

***Tabela 11.*** *Sezonska komponenta za velika preduzeca*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 73,6 | 83,1 | 100,6 | 97,0 | 100,8 | 105,8 | 98,8 | 99,3 | 105,6 | 112,8 | 110,5 | 112,1 |
| 2005 | 73,6 | 83,1 | 100,6 | 97,0 | 100,8 | 105,8 | 98,8 | 99,3 | 105,6 | 112,8 | 110,5 | 112,1 |
| 2006 | 73,6 | 83,1 | 100,6 | 97,0 | 100,8 | 105,8 | 98,8 | 99,3 | 105,6 | 112,8 | 110,5 | 112,1 |
| 2007 | 73,6 | 83,1 | 100,6 | 97,0 | 100,8 | 105,8 | 98,8 | 99,3 | 105,6 | 112,8 | 110,5 | 112,1 |

***Tabela 12.*** *Sezonska komponenta za integralni indeks*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 75,2 | 85,9 | 100,5 | 97,8 | 100,4 | 104,4 | 99,1 | 98,5 | 105,1 | 111,9 | 109,2 | 111,9 |
| 2005 | 75,2 | 85,9 | 100,5 | 97,8 | 100,4 | 104,4 | 99,1 | 98,5 | 105,1 | 111,9 | 109,2 | 111,9 |
| 2006 | 75,2 | 85,9 | 100,5 | 97,8 | 100,4 | 104,4 | 99,1 | 98,5 | 105,1 | 111,9 | 109,2 | 111,9 |
| 2007 | 75,2 | 85,9 | 100,5 | 97,8 | 100,4 | 104,4 | 99,1 | 98,5 | 105,1 | 111,9 | 109,2 | 111,9 |

***Tabela 13.*** *Iregularna komponenta za mala preduzeca*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 80.5 | 100.2 | 101.1 | 99.2 | 99,0 | 104,0 | 97.2 | 100.2 | 97.8 | 102,0 | 100.6 | 107.5 |
| 2005 | 101.1 | 98.9 | 99.8 | 100.8 | 97.1 | 101,0 | 99.9 | 97.9 | 102.2 | 100.9 | 83.9 | 102.6 |
| 2006 | 114,0 | 98.1 | 90.4 | 100.8 | 102.3 | 97.2 | 102,0 | 99.6 | 100.6 | 100.6 | 98.3 | 95.5 |
| 2007 | 101.7 | 102.2 | 98.8 | 99,0 | 101.8 | 98.4 | 101.5 | 102.8 | 99.5 | 95.2 | 101.9 | 99.5 |

***Tabela 14.*** *Iregularna komponenta za velika preduzeca*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 98.8 | 99.1 | 100.7 | 100.7 | 100.2 | 100.9 | 99.4 | 91.7 | 98.9 | 99.7 | 101.1 | 101.1 |
| 2005 | 100.7 | 97.4 | 99.6 | 103.7 | 100,0 | 100.2 | 99.4 | 99,0 | 101.5 | 100.9 | 100.6 | 98.6 |
| 2006 | 99.8 | 105.6 | 99.3 | 100.8 | 99.5 | 100,0 | 101.1 | 100,0 | 99.6 | 100.1 | 98.7 | 99.3 |
| 2007 | 104.7 | 102.1 | 100.3 | 98.5 | 100.6 | 95.9 | 100.3 | 101.5 | 99.9 | 98.9 | 99.1 | 100.5 |

***Tabela 15.*** *Iregularna komponenta za integralni indeks*

| God. | Mesec | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 2004 | 96.3 | 98.4 | 101.5 | 99.9 | 100.4 | 101.8 | 99,0 | 92.4 | 98.5 | 100.4 | 101.4 | 101.4 |
| 2005 | 99.3 | 95,0 | 99.2 | 101.6 | 98.8 | 100.5 | 99.5 | 98.5 | 101.6 | 101.7 | 99,0 | 99.2 |
| 2006 | 99.9 | 102,0 | 97.8 | 100.3 | 99.6 | 100.4 | 101.1 | 100.2 | 99.4 | 99.7 | 99.1 | 99.1 |
| 2007 | 101.5 | 99.9 | 101.3 | 98.3 | 101.5 | 97.7 | 100.8 | 102,0 | 100.4 | 97.7 | 99.9 | 100,0 |

**Svođenje indeksa na istu bazu (prosek 2005)**

Godišnji indeks, godina na prethodnu godinu, izračunat je kao aritmetička sredina mesečnih indeksa u godini. Tako su dobijeni godišnji indeksi : 2004/2003, 2005/2004, 2006/2005, 2007/2006.

Mesečne indekse iz 2004. godine treba pomnožiti sa faktorom ((2003/2004)*100)* *((2004/2005)*100).

Mesečne indekse iz 2005. godine treba pomnožiti sa faktorom (2004/2005)*100.

Mesečne indekse iz 2007. godine treba pomnožiti sa faktorom (2006/2005)/100.

*Bozidar Popovic, Statistical Office of the Republic of Serbia*
*e-mail: bpopovic@stat.gov.rs*

# TREATMENT OF OUTLIERS WITH EXAMPLE IN SBS SURVEY IN SERBIA

## ABSTRACT

*In this paper different methods for treatment of outliers are described. Quartile method was used for outlier detection in Quarterly Structural Business Statistics (SBS) Survey in Serbia 2007. Variable of interest was weighted income, realized from goods, products and services sale. Outliers detected using this method were analyzed. The results of the study are presented, and some conclusions are stated.*

## 1. INTRODUCTION

The aim of this paper is to give a short oversight of some methods for outlier treatment, with one practical example.

In section 2 general outliers theory is given.

In section 3 a review of sources of influential observations are presented.

Different methods for treatment of outliers are presented in section 4. Proofs of theorems and details are not given, they can be found in references.

In section 5 practical example is given. The 1st and 2nd quarter SBS 2007 income, realized from goods, products and services sale data were used.

Concluding remarks are given in section 6.

## 2. OUTLIERS IN SAMPLE SURVEYS

In statistical sample survey practice we often encounter observations that differ substantially from most observations in the sample. The statistical literature refers to these observations as outliers. Two distinct types of outliers are defined: *representative outliers*, which are correctly recorded and represent other population units similar in value to the observed outlier, and *non-representative outliers*, which are incorrectly recorded or unique in the sense that there is no other unit like them. Errors that lead to non-representative outliers should be detected and corrected at the editing stage.

Typically, in business surveys data is collected for a few very large units and many small, so that distributions of measured variables are skewed. In the phase of estimation and inference from such surveys, outlier issues often play an important role.

Outliers are of interest for at least two reasons. First, as implied by the terms 'representative' and 'non-representative' some analysts have a fundamental interest in identification of units that are so unique that they may not be of interest for certain classes of analyses.

Second, analysts often have special interest in population units that are influential in the sense that their inclusion or exclusion from an estimator may lead to substantial changes in the numerical value of the resulting estimate. The influence of an observation varies depending on the estimator used.

## 3. SOURCES OF INFLUENTIAL OBSERVATIONS

Influential observations generally arise from one of three sources. First, they may arise from observations that are unusually large or otherwise deviate in unusually extreme forms from the center of reference distribution.

Second, the observation may be associated with a unit that had an unusually low selection probability, and thus an unusually high probability weight.

Third, observation may have a weight that is very large due to problems with stratum jumping, large nonresponse adjustment factors arising from unusually low response rate within a given adjustment cell, unusual calibration-weighting effects, or other factors. Very often strata are defined by the kind of business and the measure of size, that is, the size of a known or estimated characteristic closely related to those being measured in the survey. When either of these characteristic changes in a unit, its true characteristics are no longer consistent with the sampling stratum initially assigned. In such cases, informally labeled 'stratum jumpers', can have influential values.

## 4. DEALING WITH OUTLIERS

There are two major steps in dealing with outliers, outlier detection and outlier treatment. Both steps can be preformed in many ways and most adequate procedures for the survey of interest need to be chosen.

Methods for treatment of outliers are based on changing the survey weight, survey value or both.

### 4.1 Changing the weight

A relatively common practice is to set the survey weight equal to 1. In an informal sense, this would be consistent with the idea that the identified outlier is 'non-representative' outlier. Also there are methods that reduce the weight to value different than 1. Changing the weight is one of the robust techniques of outlier treatment. The idea is reducing the weight of $k$ weights to $r < \frac{N}{n}$. Set $s_1$ denotes set of non-outliers, and set $s_2$

denotes the set of outliers. The HT total estimate, in case of SRS, is

$$\hat{T}_1 = \sum_{s_1} \frac{N}{n} \cdot x_i + \sum_{s_2} r \cdot x_i \,,$$

or rather, reduce the weights of the outliers and change the weights for non-outliers

$$\hat{T}_2 = \sum_{s_1} \frac{(N - r \cdot k)}{(n - k)} \cdot x_i + \sum_{s_2} r \cdot x_i$$

In case of stratified SRS, HT total estimate is

$$\hat{T}_{str} = \sum_{h=1}^{H} \left( \sum_{s_{h1}} \frac{(N_h - r_h \cdot k_h)}{(n_h - k_h)} \cdot y_i + \sum_{s_{h2}} r_h \cdot y_i \right)$$

Where $k_h$ is the number of outliers in strata $h$, $s_{h1}$ is set of non-outliers in strata $h$, $s_{h2}$ is set of outliers in strata $h$ .

In all pervious cases we need $r$, or $r_h$ depends on survey design. Rao(1971) and Chinnappa (1976) assumed that observed outliers are self-representative or non-representative but genuine (no unsampled outliers in the population) assign weight $r = 1$ to the outliers. Hidiroglou and Srinath (1981) determined the optimal weight of $r$ by minimizing the MSE of the estimator in (1).

### 4.2 Winsorization

The theory of winsorization currently deals with a population where the values are heavily skewed to the left that is with a long right tail to the distribution so that there are occasionally large values.

The general idea of winsorization is that if an observation exceeds a pre-specified cut-off value $K$, then the observation will be replaced with value $K$.

Under Type I Winsorization, we consider original observations $Y_i$ and an estimator of population total

$$\hat{T}_W = \sum_{i \in s} w_i Y_i^*$$

where is

$$Y_i^* = \begin{cases} K, & Y_i > K \\ Y_i, & Y_i \le K \end{cases}$$

Under Type II Winsorization, the same estimator of total is used, but we define

$$Y_i^* = \begin{cases} \dfrac{Y_i}{w_i} + K\dfrac{w_i - 1}{w_i}, & Y_i > K \\ Y_i, & Y_i \le K \end{cases}$$

Where $w_i$ is weight.

The second way of defining the winsorized contribution of a particular unit to the estimate of population total defined as sum of two parts- the first being the actual $Y$ value (unwinsorized) of the unit and second being its (winsorized) contribution to the estimated sum of $Y$ –values for the non-sampled population units. This second interpretation falls in naturally with the idea that all the outlying values in our sample data are correct and so outliers should be allowed to contribute their (unweighted) values to the estimate of population total. Their contribution to the population total of the non-sampled units, however, cannot be left unchanged, so their winsorized value is multiplied by their estimation weight minus one to define this contribution.

Another question is how to determine cut-off value *K*. Kokic and Bell (1994) dealt with this problem within the context of stratified random sampling. Under the assumptions that for each stratum h $\{x_{hi} : i=1,2,\ldots,N_h\}$ is a sequence of uncorrelated and identically distributed random variables, i.e.

$$E_m(x_{hi}) = \bar{y}_h$$
$$Var_m(x_{hi}) = \sigma_h^2 < \infty$$

they proved the theorem which shows that the overall mean squared error

$$MSE(\hat{t}) = E(\hat{t} - t)^2,$$

($E=E_mE_d$, $E_m$) denotes expectation over the model, $E_d$ denotes expectation over the design), of the stratified type II winsorized estimator is minimized if a value of $K$ within the stratum $h$ is of the form

$$K_h = \left(\frac{N_h}{n_h} - 1\right)^{-1} L + \bar{y}_h$$

Here $\bar{y}_h$ denotes the sample mean within the stratum *h*,

$$L = \sum_h N_h(1 - f_h)\{E_m(X_h J_h) - K_h E_m(J_h)\}$$

a constant chosen so that the bias of the stratified winsorized estimator is $-L$,

$$J_h = \begin{cases} 1, & x_{hi} \ge K_h \\ 0, & x_{hi} < K_h \end{cases} \text{ and } f_h = \frac{n_h}{N_h}$$

Clarke (1995) obtained a model-based extension of this result.

It is important to note that the cut-offs are optimum only at the level on which estimates are being formed.

### 4.3 M- Estimators

M-Estimators are a kind of robust estimators. The purpose of robust estimators is to produce an efficient estimator in presence of outliers, while minimizing bias. This is done by reducing the influence of the outliers on the estimator.

M-estimators first introduced by Huber (1964), where M stands for maximum likelihood. They are robustified maximum likelihood estimators that use a weight function to discount, or downweight, extreme values. In the univariate case, a robust M-estimator of location may be created as follows: for the observation $x_i$, and location estimate $A$, define the residual $r_i = x_i - A$. Define the function $\rho(x,A) = \rho(x_i - A)$, where $\rho(x,A)$ is symmetric (i.e. $\rho(x,A) = \rho(-x, A)$) with a unique minimum at zero. Typically, the role of this function is to decrease the influence of observations with large residuals, often resulting in a winsorized estimator.

Once this function has been defined, minimization is performed. For example, for some function $\rho(x,A)$, a univariate M-estimate of location, $A$, could be obtained by solving the equation:

$$Min_A \sum_{i=1}^{n} \rho(x_i - A)$$

Differentiating this expression with respect to $A$ yields:

$$\sum_{i=1}^{n} \rho'(x_i - A) = 0$$

Different $\rho(x,A)$, or its derivate, yield different M-estimators. Huber (1964) proposed the following function:

$$\rho'(x_i - A) = \begin{cases} -k, & x_i - A < -k \\ x_i - A, & -k < x_i - A < k \\ k, & x_i - A > k \end{cases}$$

where $k$ is constant called the tuning factor. This function determines the extent to which outliers are treated.

### 4.4 Some Multivariate Outlier Detection Methods

Let $X=\{x_1,...,x_i,...,x_n\}$ represent a set of $n$ data points from $R^p$, where *i-th* observation is $X_i^T = (x_{i1},...,x_{ip})$. If $X$ a random sample from multivariate normal distribution with mean vector $\mathbf{u}$ and covariance matrix $\mathbf{V}$ in $R^p$, a classical way of detecting outliers is to calculate Mahalonobis' distance for each observation using estimates of $\mathbf{u}$ and $\mathbf{V}$ as follows

$$D_i = (\mathbf{x}_i - \hat{\mathbf{u}})^T \hat{\mathbf{V}}^{-1} (\mathbf{x}_i - \hat{\mathbf{u}})$$

Mahalonobis' distance identifies observations that lie far away from the centre of the data cloud, giving less weight to variables with large variances or to groups of highly correlated variables (Joliffe 1986). This distance is often preferred to the Euclidian distance which ignores the covariance structure and treats all variables equally. Mahalanobis' distance is widely used in cluster analysis and other classification techniques. It is closely related to Hotelling's T-square distribution used for multivariate statistical testing.

## 5 OUTLIER TREATMENT IN SBS SERBIA 2007

For outlier treatment simulation the 1st and 2nd quarters of SBS data were used. The weighted income variable was treated. Using the quartile method 15 outliers were found for the first quarter, and 11 for the second quarter (Cvetkovic 2008…). Detection of outliers was by sections. The tolerance interval $(q_{0.5}-25d_u,\ q_{0.5}+25d_u)$, where is $q_{0.5}$ median of weighted income, $q_{0.75}$ the third quartile of weighted income and $d_u=q_{0.75}-q_{0.5}$, was constructed per sections. Constant $c_u=25$ was determined using boxplot diagrams. Namely, boxplots are constructed for values of $c_u$ from interval [10,40] with step size 5. The idea was to determine the interval $[c,40]$, $c \in [10,40]$, that includes values of constant $c_u$, for which the number of outliers will be almost equal. For those values of $c_u$ boxplots will be almost the same. In our case, number of outliers will be almost equal for interval [25,40]. That was the reason why the value of constant $c_u$ is 25. Any observation with income value which is greater or equal than upper bound of this interval was identified to be an outlier.

After the analysis, it was concluded that two units (in both quarters) should be weighted by 1. Those units are marked as extreme units. For the other 13 units in the first, and 9 units in the second quarter, weighted income were adjusted according to

$$w_i Y_i = \begin{cases} w_i Y_i, & w_i Y_i < q_{0.5} + 25d_u \\ q_{0.5} + 25d_u, & w_i Y_i \geq q_{0.5} + 25d_u \end{cases}$$

Estimates for income population total were based on HT estimator, and before outlier treatment were produced using CLAN. Its standard deviation was produced using CLAN as well.

After the treatment procedure, the same formula like in CLAN was used for variance estimation.

The result of this procedure is given in following table. It is expected that the values of estimated income and coefficient of variation would slightly decrease.

***Table 1*** *Results of total income estimation before and after outlier treatment (1st Quarter)*

| *Before treatment* | | *After treatment* | |
|---|---|---|---|
| *Estimated income (in thousands RSD)* | *CV* | *Estimated income (in thousands RSD)* | *CV* |
| *990 845 514.09* | *4.24* | *966 863 482.9* | *4.1* |

***Table 2*** *Results of total income estimation before and after outlier treatment (2nd Quarter)*

| *Before treatment* | | *After treatment* | |
|---|---|---|---|
| *Estimated income (in thousands RSD)* | *CV* | *Estimated income (in thousands RSD)* | *CV* |
| *1 139 399 024* | *3.88* | *1 133 810 294.7* | *3.77* |

***Table 3*** *Chain indices*

| *Before treatment* | *After treatment* |
|---|---|
| *114.99* | *117.27* |

## 6 CONCLUDING REMARKS

Dealing with outliers is not an easy task.. Statisticians that are involved in this topic should have a very good apprehension of survey subject, and they should be aware that outlier treatment can introduce bias. In repeated surveys, treatment of outliers must be performerd under the same rules. Further work should be to apply this procedure on other two quarters, and to check influence of outlier treatment on chain indices.

## REFERENCES


[1] Afifi, A.A, Azen, S.P (1972), *Statistical Analysis: a computer oriented approach,* New York, Academic Press, pp. 282

[2] Clarke, M (1995), *Winsorization Methods in Sample Surveys,* Master Thesis, Department of Statistics, Australia National University

[3] Chambers, R.L (1986), *Outlier Robust Finite Population Estimation,* JASA, Vol. 81, pp. 1063-1069

[4] Chinnappa B. N. (1976), *A preliminary note on methods of dealing with unusually large units in sampling from skewed populations*, Unpublished report , Ottawa. Statistics Canada.

[5] Cvetkovic, N (2008), *Detection of Outliers with Example in SBS Survey in Serbia,* The Third Balkan Summer School

[6] Eltinge, J.L, Cantwell, P.J (2006), *Outliers and Influential Observations in Establishment Surveys,* U.S. Census Bureau

[7] Hidiroglou M.A. and Srinath K.P. (1981), *Some Estimators of a Population Total from Simple Random Samples Containing Large Units*, Journal of the American Statistical Association, 76, 690-695.

[8] Huber, P.J (1964), *Robust Estimation of a Location Parameter*, Annals of Mathematical Statistics, Vol. 35, pp.73-101

[9] Joliffe, I.T (1986), *Principal Component Analysis,* New York, Springer Werlag

[10] Kokic, P.N, Bell, P.A (1994), *Optimal Winsorizing Cut-offs for Stratified Finite Population Estimates,* Journal of Official Statistics, Vol. 10, pp. 419-435

[11] Rao C.R. (1971), *Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations*, In V. P. Godambe D. A. Sprott (eds.), *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston.

*Božidar Popović*

# TRETMAN EKSTREMNIH VREDNOSTI NA PRIMERU STRUKTURNIH BIZNIS STATISTIKA U SRBIJI

## REZIME

*Dat je opis različitih metoda tretmana ekstremnih vrednosti (outliers). Detekciju i tretman ekstremnih vrednosti, pokazali smo na primeru ponderisane vrednosti prihoda iz istraživanja o strukturnim biznis statistikama u Srbiji 2007.godine. Ekstremne vrednosti detektovane su metodom kvartila. Data je ocena prihoda pre i posle tretmana, kao i odgovarajući koeficijenti varijacije.*

*Natasa Cvetkovic, Statistical Office of the Republic Serbia, Serbia*
*e-mail: ncvetkovic@stat.gov.rs*

# DETECTION OF OUTLIERS WITH EXAMPLE IN SBS SURVEY IN SERBIA

## ABSTRACT

*Quarterly Structural Business Statistics (SBS) Survey was introduced in the Statistical Office of the Republic of Serbia in 2007. This survey collects data on operating income, operating costs and expenses, stocks, number of employees, and investments. Some of the outlier detection methods were already applied and some outliers were detected. In this paper we discuss implementation of some new methods, and give results of comparison between different methods.*

## 1 INTRODUCTION

Outliers could arise for different reasons, they could be a result of an error, or they could arise from the variability of the dataset, or from other sources. The outliers more often appear in business surveys due to skewness of the target population. In this situation ordinary statistics, based on the least squares principle, are not suitable. The usual tool applied in this situation is size-stratification or probability proportional to size sampling. This is effective as long as the size measure is a good predictor of the actual survey values. However, the size measure becomes outdated very fast, and outliers appear regularly. Another problem is that the size measure could be good for some variables but not for the others. The ad-hoc treatment is the most usual way to treat the outliers, for several reasons. One is their surprising appearance, since that makes surveyors reluctant to systematic treatment of outliers. Another one is the complexity of techniques used to treat the outliers, or lack of proper variance estimators, or their unknown inferential properties.

According to Chambers (1986) outliers could be classified into two groups: representative outliers, corresponding to data values that are correctly recorded and represent other population units that are similar in value to the observed outlier, and non-representative outliers, corresponding to values that are actually errors introduced into the sample data at some stage prior to estimation or unique in the sense that there are no other units that are similar to them. Errors that lead to non-representative outliers should be detected and corrected at the editing stage.

In sample surveys outlier detection could be performed with or without the sampling weights. In that sense we could make distinction between extreme values and influential observations. Influential observations are those for which the combination of the reported value and the sampling weight has a large influence on the estimate. Extreme values may or may not be influential.

Outliers may be univariate or multivariate. Multivariate outliers are those observations that are inconsistent with the correlational structure of the dataset. Thus, while univariate outlier detection is performed independently on each variable, multivariate methods investigate the relationship between several variables.

Although most surveys collect multivariate data, univariate detection methods are usually favored because of their simplicity. But univariate methods fail to detect observations that violate the correlational structure of the dataset. This is the disadvantage if the variables are highly correlated.

## 2 DETECTION OF OUTLIERS

Most outlier detection methods use some measure of distance to evaluate how far away an observation is from the centre of the data.

Let $y_1, y_2, ..., y_n$ be the observations. The relative distance of $y_i$ is

$$d_i = \frac{|y_i - m|}{s}$$

where $m$ is the location estimate and $s$ is the scale estimate. If $d_i$ exceeds a predetermined cutoff value $c$, $d_i > c$, then the observation is considered to be an outlier. Or, a tolerance interval $(m - c_l \cdot s, m + c_u \cdot s)$ is used, where $c_l$ and $c_u$ are predetermined values. The bounds $c_1$ and $c_u$ can be chosen by examining past data or using past experience.

The sample mean and variance may be used to measure this distance, but since they are not robust to outliers they can mask the observations we want to detect. For that reason, robust scale and location estimators, which are resistant to outliers, are used instead. Therefore, many outlier detection methods use order statistics, like the median or quartile.

### 2.1 The quartile method

The most popular univariate outlier detection method for the survey data is the quartile method. The idea is to create the tolerance interval

$$(q_{0.5} - c_l \cdot d_l, q_{0.5} + c_u \cdot d_u)$$

where $d_l$ is the lower interquartile range, and $d_u$ is the upper interquartile range, i.e.

$$d_l = q_{0.5} - q_{0.25}$$

and

$$d_u = q_{0.75} - q_{0.5}\ .$$

Those data that fall outside of this interval are considered to be outliers. This method is often applied to weighted data. In general, this interval is not symmetric because of

skewness of data. The interval also can be made one-sided. This method is not only robust, but also simple and non-parametric.

## 3 QUARTERLY SBS SURVEY IN SERBIA

Quarterly SBS survey has been introduced as a pilot survey in Serbia in 2007. This survey collects data on operating income, operating costs and expenses, stocks, number of employees, and investments. A sample of 2150 enterprises in sections A to O excluding sections J (Financial intermediation) and L (Public administration and defence) was selected. The smallest enterprises (entrepreneurs) were excluded from the frame. Other very small enterprises according to turnover were also excluded . It is a stratified simple random sample. Stratification was done according to activity, number of employees and turnover. Thus we obtained 276 strata.

## 4 DETECTION OF OUTLIERS IN QUARTERLY SBS SURVEY IN SERBIA – EXISTING AND NEW PROCEDURES

In this paper we consider data from the first quarter of SBS survey in 2007. The variable we consider here is income realized from goods, products and services sale. Some outliers were detected in an iterative procedure using stratum variance. Those strata that have the significant share of stratum variance in the total variance were analysed in order to detect outliers. As a result 10 units with weights that are not equal to one or units from small strata were detected as outliers.

Another method was also analysed. Data were sorted by weighted income in descending order. Those units that have very big value of weighted income are considered to be outliers. Ten units were obtained as outliers.

The idea of this simulation study was to consider a quartile method as a new method for detection of outliers in this survey, and also to compare it with the existing procedure. This method was applied on the level of section. For each section the tolerance interval was created. The interval was one-sided because of skewness of data. The method was applied to weighted data. The median and the third quartile were calculated for each section, and then the upper bound was determined as $q_{0.5} + c_u \cdot d_u$. For $c_u$ the value 25 was taken, which was determined analysing boxplot diagrams for different values of $c_u$ (see Appendix). Several boxplots were constructed for $c_u$ taking values from the interval [10, 40] with step size 5. The idea was to determine the interval [$c$, 40], $c \in$ [10, 40], that includes values of constant $c_u$, for which the number of outliers will be almost equal. These values of $c_u$ will give very similar appearance of boxplots. In our case, number of outliers will be almost equal for interval [25, 40]. For that reason the value of constant $c_u$ was chosen to be 25. Each unit for which the value of the observed variable exceeded this upper bound of the interval was considered as outlier. At the end 15 units were detected as outliers.

Four same units were detected as outliers by all three methods. But these three methods are not quite comparable since they are applied on different levels.

Table1 shows differences in number of outliers by sections using these three different methods:

***Table 1.***

| Section | Number of outliers by old method - 1 | Number of outliers by old method - 2 | Number of outliers by quartile method |
|---|---|---|---|
| A | 1 | 0 | 1 |
| D | 0 | 1 | 3 |
| E | 1 | 2 | 1 |
| F | 1 | 0 | 1 |
| G | 5 | 4 | 1 |
| I | 2 | 3 | 8 |
| Total | 10 | 10 | 15 |

## 5 CONCLUSIONS

Detection of outliers in business surveys seems to be a very delicate issue. One reason for that is certainly their surprising appearance. That makes systematic approach to their detection and treatment very difficult, and that is something most surveyors would like to use. Instead, the ad-hoc treatment is the most usual way to deal with outliers. Another problem seems to be the variety of methods that could be used to handle outliers. The decision on which method to choose is not always easy. A lot of questions usually arise. In our example there are several questions that should be carefully considered. Should we use univariate or multivariate approach? On which level the quartile method should be applied? Should we use one-sided or two-sided interval? How to determine the values for $c_l$ and $c_u$? All these questions should be carefully considered before making any decisions.

## REFERENCES

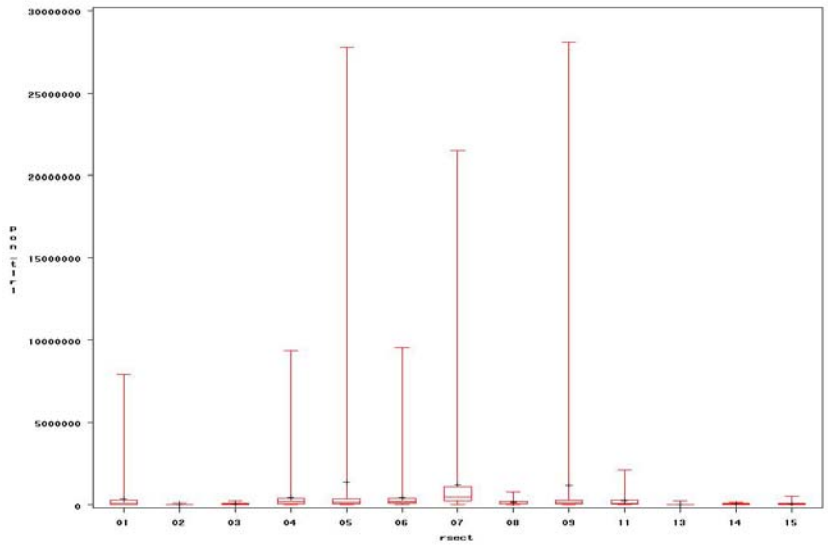
[1] Gwet, J.-P. and Lee, H., (2000). *An evaluation of outlier-resistant procedures in establishment surveys.* Proceedings of the Second International Conference on Establishment Surveys, June 17-21, 2000, Buffalo, New York, 707-716

[2] Chambers, R.L., Kokic P., Smith P. and Crudas M. (2000). *Winsorization for identifying and treating outliers in business surveys.* Proceedings of the Second International Conference on Establishment Surveys, June 17-21, 2000, Buffalo, New York, 717-726

[3] Franklin, Sarah F., Thomas, S. and Brodeur, M. (2000). *Robust multivariate outlier detection using mahalanobis' distance and modified Stahel-Donoho estimators.* Proceedings of the Second International Conference on Establishment Surveys, June 17-21, 2000, Buffalo, New York, 697-706

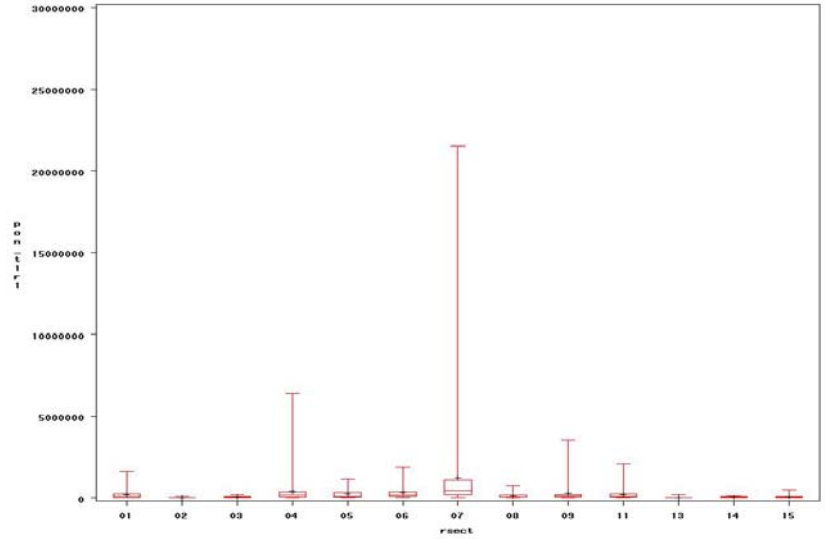
## APPENDIX

***Boxplot1.*** *Dataset before detection of outliers*



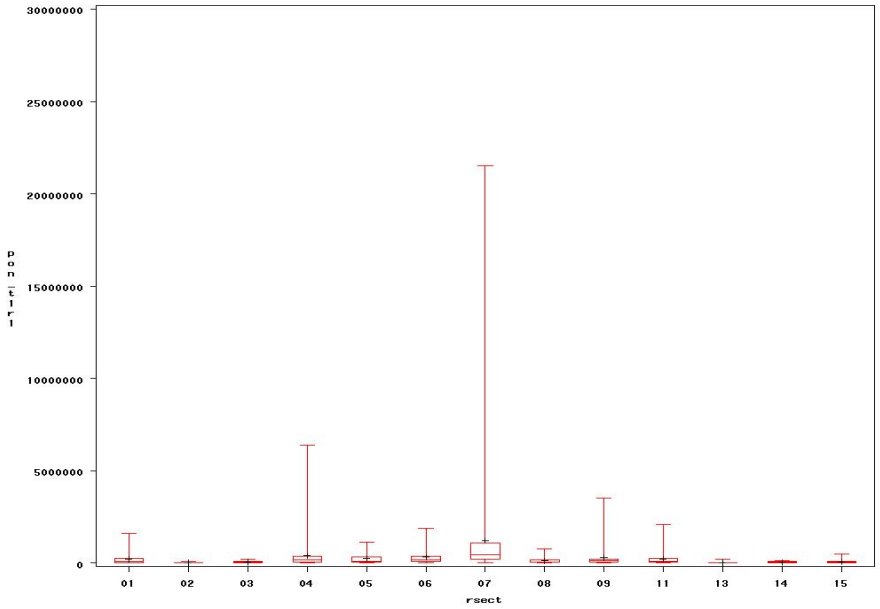

pon_t1r1 – weighted income, rsect - section

***Boxplot2***. *Dataset after excluding outliers detected setting* $c_u = 40$



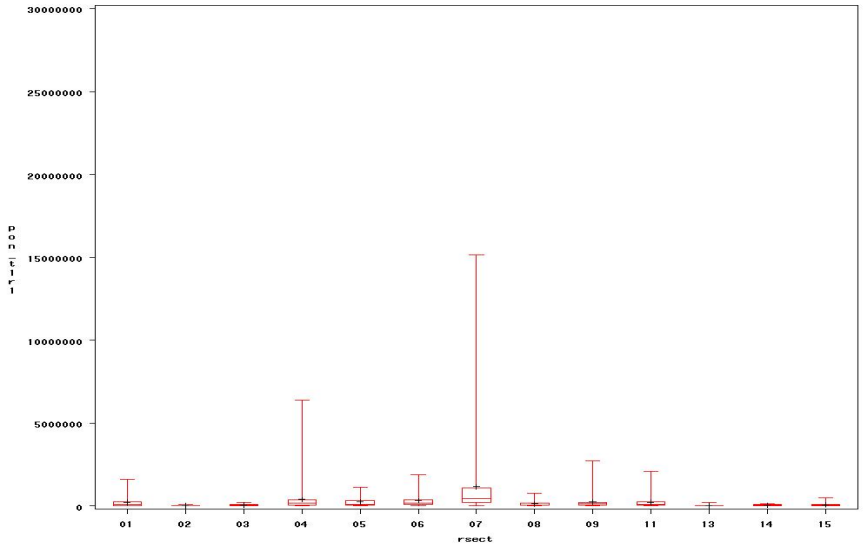

pon_t1r1 – weighted income, rsect - section

***Boxplot3***. *Dataset after excluding outliers detected setting* $c_u = 35$



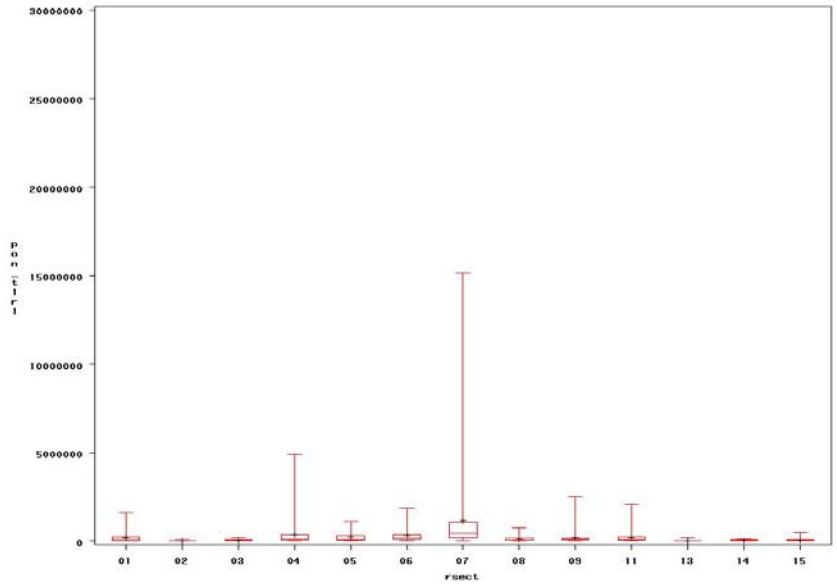

pon_t1r1 – weighted income, rsect - section

***Boxplot4.*** *Dataset after excluding outliers detected setting* $c_u = 30$



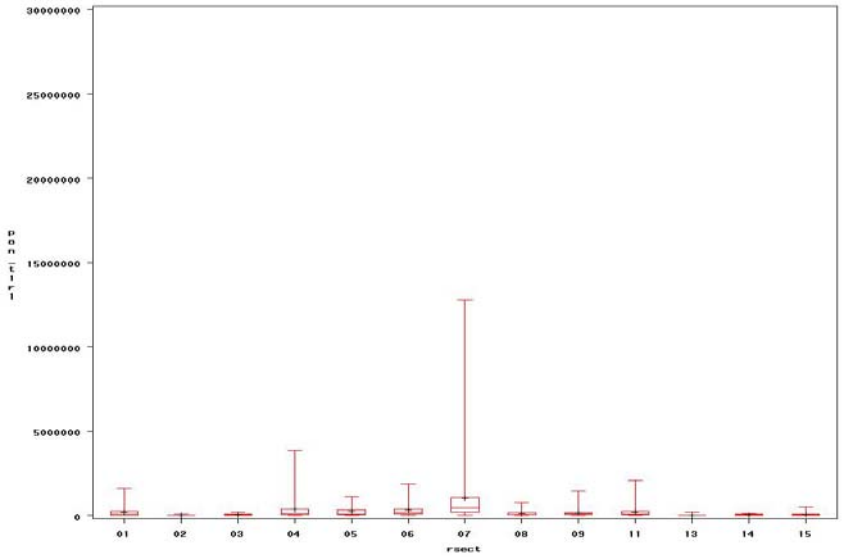

pon_t1r1 – weighted income, rsect - section

***Boxplot5.*** *Dataset after excluding outliers detected setting* $c_u = 25$



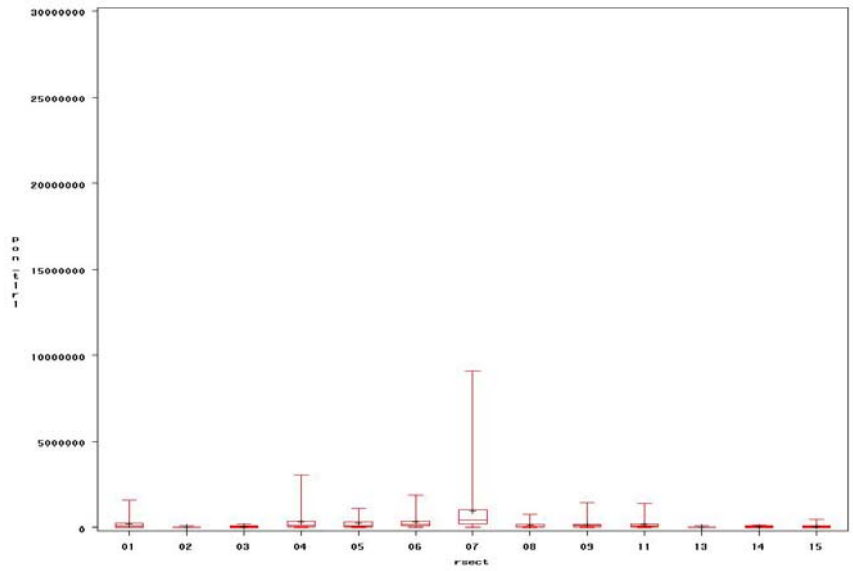

pon_t1r1 – weighted income, rsect - section

***Boxplot6.*** *Dataset after excluding outliers detected setting* $c_u = 20$



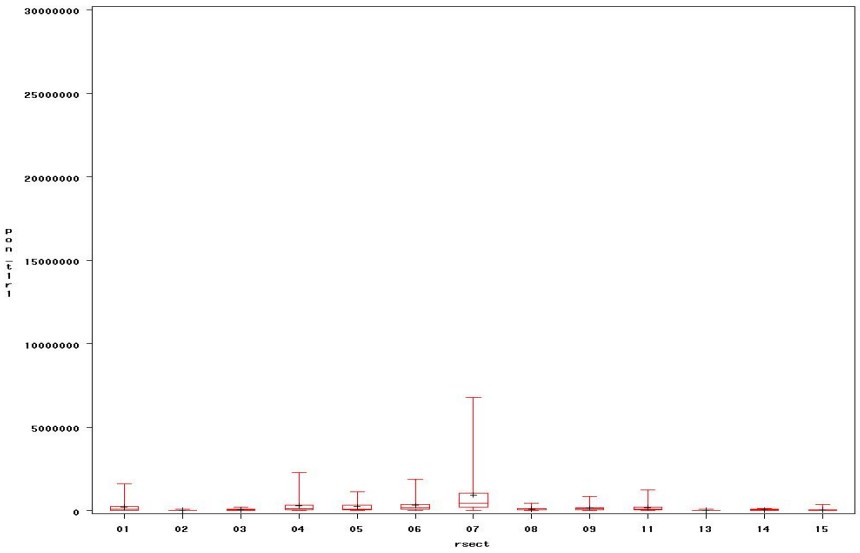

pon_t1r1 – weighted income, rsect - section

***Boxplot7.*** *Dataset after excluding outliers detected setting* $c_u = 15$




pon_t1r1 – weighted income, rsect - section

***Boxplot8.*** *Dataset after excluding outliers detected setting* $c_u = 10$




pon_t1r1 – weighted income, rsect - section

*Nataša Cvetković*

# DETEKCIJA EKSTREMNIH VREDNOSTI SA PRIMEROM ISTRAŽIVANJA O STRUKTURNIM BIZNIS STATISTIKAMA U SRBIJI

## REZIME

*Kvartalno istraživanje o strukturnim biznis statistikama (SBS) uvedeno je u Republičkom zavodu za statistiku Srbije 2007. godine. U ovom istraživanju prikupljaju se podaci o poslovnim prihodima, poslovnim rashodima, zalihama, broju zaposlenih i investicijama. Neki od metoda za detekciju ekstremnih vrednosti (outliers) su već bili primenjeni i neke ekstremne vrednosti detektovane. U ovom radu diskutuje se o primeni nekih novih metoda, i daju rezultati poređenja različitih metoda.*

*Marija Panovic, Statistical Office of the Republic of Serbia*
*e-mail: mpanovic@stat.gov.rs*

# THE IMPORTANCE OF INITIAL DATA ANALYSIS IN THE PHASE OF FRAME CONSTRUCTION

## ABSTRACT

*Initial examination of data involves scrutinizing and summarizing data, assessing the quality and structure of data. It is a valuable stage in the phase of frame construction, because other steps, stratification and allocation, depend on the quality of auxiliary variables. It is particularly important to find out how the data were collected. The possible presence of errors, outliers and missing auxiliary values should be investigated. The aim of this paper is to present one example, the quarterly catering survey, which shows that initial examination is necessary or desirable.*

## 1 INTRODUCTION

Frame construction should usually begin by scrutinizing, summarizing and exploring a set of data. The aim is to clarify the general structure of data, investigate the possible presence of errors, outliers and missing auxiliary values. This paper will present the phase of frame construction for the quarterly catering survey as illustration of importance of initial data analyses. It will show that although computer can easily be programmed to check data, human eye is also very efficient at spotting suspect values when the number of observations is not very large.

## 2 THE QUARTERLY CATERING SURVEY

The quarterly catering survey was conducted for the first time in 2007, as a pilot survey and in 2008 it became a regular survey. The object of this survey is to obtain:

- Data on number of units by territory;
- Total turnover, turnover by type of service ( food, drink, overnight stay) and by month;
- Total value added tax, value added tax by type of service and by month;
- Number of employees

The main goal is to estimate catering turnover indices, last quarter with respect to the previous quarter.

### 2.1 Population and frame

Population of interest consists of entrepreneurs dealing with Catering activities (division 55 of NACE 1) who are obliged to pay value added tax. The frame is constructed using Statistical Business Register (SBR), state of March in year *t*, where *t* is the reference year of the survey. SBR contains information on address, NACE activity and turnover. The data on turnover is from balance sheet or from tax form for the year *t-2*. If entrepreneur didn't submit the balance sheet for the year *t-2* than the data on turnover is taken from the tax form for the year *t-2*.

The preliminary frame consists of all active entrepreneurs from SBR that fulfill the given conditions. The final frame was constructed by sorting preliminary frame by turnover and excluding 5% of the smallest entrepreneurs according to the subsume of the turnover. The main reasons for construction of the frame with cut-off are financial and personal constrains that limit the size of the sample. The remaining units in the frame are more homogenous and it is also expected that there is a higher chance of obtaining data from them in the field. The final frame, apart from cut-off units, does not cover entrepreneurs that started their activity in the year *t-1* and have considerable turnover.

### 2.2 Sampling plan and sampling allocation

Business surveys are typically characterized with a few large units that account for a good portion of the total for the variable of interest and many small units. To obtain higher precision in estimation, usually "big" units are separated in special stratum called "take-all stratum" and considered as census units."Take-some stratum" is the stratum of all remaining units that are sampled.

Hidiroglou algorithm was used to divide the frame units into census and the remaining ones and for the sample allocation. Based on information on an auxiliary variable (correlated with the study variable) an optimum cut-off value is determined that minimizes the sample size for a given precision for an estimate of the total of this variable.

Hidiriglou algorithm was applied to strata of units defined by cross-classifying activity classes (five digits level) and territory classes (Belgrade, Central Serbia without Belgrade and Vojvodina). The data on turnover was used as auxiliary variable and 0.085 was given level of CV for the estimated total. The final stratification included take-all and take some units within strata already defined by Nace activity and territory classes.

Simple random sample was selected from each stratum using permanent random numbers (PRN). One desirable feature of this sampling system is possibility to sample redraw with maximize overlap in the next survey occasion.

### 2.4 Estimation

The estimates were obtained in a standard way for stratified simple random sampling. Units in the sample which filled in the questionnaire but are not dealing with catering activity or not obliged to pay value added tax and units with some specific reasons for non-

response (enterprise, temporary inactive) are also considered as responding units. In the phase of estimation they were assigned zero for turnover value. In our opinion this is one way to reduce effects of frame imperfection, due to overcoverage.

## 3 THE IMPORTANCE OF INITIAL DATA ANALYSIS IN THE PHASE OF FRAME CONSTRUCTION

Initial examination of data involves data description and various data analytic techniques. In the phase of frame construction it can be very useful in detecting errors in auxiliary variables. If these errors are not detected that can have significant impact on the estimates of parameters derived from a sample that is selected from such a frame. This will be illustrated in the next example of the quarterly catering survey in 2007 and 2008.

For the survey in 2008, using the updated SBR, the same procedure for construction of the frame, stratification, allocation and selection of the sample was repeated. A large overlap of frames and samples in 2007 and 2008 was expected, changes being introduced by deaths and births of the entrepreneurs and in a few cases significant change in auxiliary variable (turnover).

***Table 1.***

| | number of units | turnover | STDturnover | MINturnover | MAXturnover | MEANturnover | Cvturnover |
|---|---|---|---|---|---|---|---|
| Frame 2007 | 2668 | 7483732 | 4717.5 | 54 | 79273 | 2805 | 168.2 |
| Frame 2008 with error | 1763 | 10278432 | 64228.3 | 0 | 2666406 | 5917 | 1085.4 |
| Frame 2008 | 1763 | 7614692 | 6755.3 | 0 | 81884 | 4383 | 154.1 |

From row 1 and row 2 we can see that the number of units in the frame for 2008 is smaller than number of units in the frame for 2007, but sum of turnover is significant bigger. Data on maximal turnover indicated possible presence of error. The value was in RSD, not thousands of RSD, as it is for all others units.

The frame is constructed with cut-off, therefore some units that were included in the frame for 2007 may not belong in the frame for 2008 anymore. Some units that were „take-all“ units in 2007 were not in „take-all strata“ in 2008. About 40 „take-all“ units from 2007 were not in the sample for 2008. Since these units were „take-all“, meaning significantly „big“, it was interesting to see data on their economic activities in 2008 and therefore they were intentionally included in sample and treated as census units. The planned sample in 2007 was about 390 units and in 2008 about 370 units. For illustration, sample was drawn from the original frame and from the frame with error.

***Table 2***. *Frame 2007*

| Strat | Nh | pnh | weight |
|---|---|---|---|
| 0553001 | 782 | 44 | 17.8 |
| 0553002 | 318 | 20 | 15.9 |
| 0553003 | 300 | 17 | 17.6 |
| 0554001 | 474 | 40 | 11.9 |
| 0554002 | 182 | 16 | 11.4 |
| 0554003 | 385 | 31 | 12.4 |
| census units | 227 | 227 | 1 |

***Table 3.*** *Frame 2008 with error*

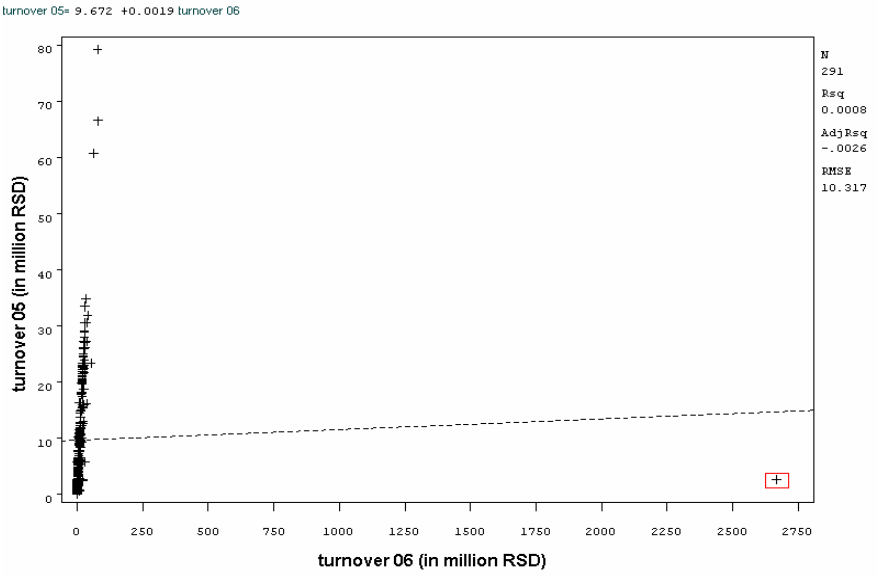
| Strat | Nh | pnh | weight |
|---|---|---|---|
| 0553001 | 575 | 40 | 14.4 |
| 0553002 | 207 | 38 | 5.4 |
| 0553003 | 217 | 21 | 10.3 |
| 0554001 | 240 | 27 | 8.9 |
| 0554002 | 111 | 17 | 6.5 |
| 0554003 | 214 | 19 | 11.3 |
| census units | 199 | 199 | 1 |

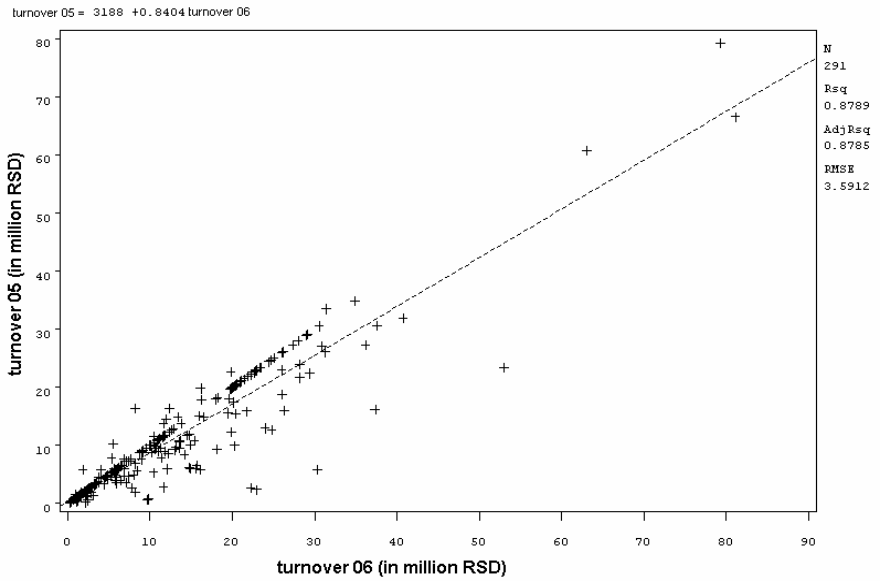
***Table4.*** *Frame 2008*

| Strat | Nh | pnh | weight |
|---|---|---|---|
| 0553001 | 567 | 32 | 17.7 |
| 0553002 | 176 | 13 | 13.5 |
| 0553003 | 212 | 16 | 13.3 |
| 0554001 | 229 | 20 | 11.5 |
| 0554002 | 104 | 10 | 10.4 |
| 0554003 | 210 | 16 | 13.1 |
| census units | 265 | 265 | 1 |

In case of small sets of data, inaccuracy of data may become evident using appropriate graphical presentations, such as scatter plots of data. . For instance, scatter plots of sample units that are in both samples, for the year 2007 and 2008, are presented on the Graph 1. Existence of a unit with extremely large, error value, of turnover can easily be detected.

***Graph 1**. Scatter plot of mutual units for sample in 2007 and planned sample for 2008 before and after repairing the error:*



## 4 CONCLUSION

Initial examination of data is a valuable stage in the phase of the frame construction involving data description and various data analytic techniques. Detecting errors in auxiliary variables, outliers and missing auxiliary values are very important benefits of initial analysis of data. In repeated surveys comparisons should be made with previous frame and sample information, on the number of units and sums of auxiliary variable values. Small difference in number of units and significant difference in sum of auxiliary variable values can indicate possible presence of error, like in an example presented in this paper. If these errors are not detected they can have significant impact on the estimates of parameters derived from a sample that is selected from such a frame. Example in this paper is a very simple one and error can be easily detected arranging the data in rank order or by graphical presentation. In some cases, in order to check for inaccuracies, a more complex analysis of data is needed.

## REFERENCES

[1] Chatfield C. (1985) *The Initial Examination of Data,* J.R.Statist.Soc A.148, 214-53
[2] Melovski Trpinac O. (1995), *Inicijalna analiza* podataka, Zbornik radova Analiza grupisanja I
[3] Hidiroglou, M.A. (1986) *The construction of Self-Representing Strata of Large Units in Survey Design*, The American Statistician, Vol. 40, No. 1

*Marija Panović*

# ZNAČAJ INICIJALNE ANALIZE PODATAKA PRI KONSTRUKCIJI OKVIRA ZA IZBOR UZORKA

## REZIME

*Pod inicijalnom analizom se podrazumeva pažljivo ispitivanje podataka sa ciljem da se utvrdi kvalitet i struktura podataka, da se podaci sažmu i sagledaju njihova ključna svojstva. Inicijalna analiza je važan korak pri konstrukciji okvira za izbor uzorka, jer druge faze, stratifikacija i alokacija uzorka, zavise od kvaliteta pomoćnih promenljivih. Veoma je važno saznati i na koji način su podaci prikupljeni. Neophodno je ispitati moguće prisustvo grešaka, ekstremnih vrednosti i nedostajućih vrednosti za pomoćne promenljive. Inicijalna analiza podataka je neophodna ili poželjna što pokazuje i primer kvartalnog istraživanja u ugostiteljstvu, čija metodologija je prikazana u ovom radu.*

*Jelena Kovijanic*
*Statistical office of the Republic of Serbia, Serbia*
*e-mail: jkovijanic@ stat.gov.rs*

# FRAME IMPERFECTIONS IN THE SURVEY ON ROAD FREIGHT TRANSPORT

## ABSTRACT

*The access to a good sampling frame is of vital importance when planning for a survey. However, sampling frame is rarely a perfect match to the target population. This paper aims to study approaches to dealing with frame imperfections on the example of the Survey on road freight transport. The amount and type of imperfection in the exemplary frame will be discussed.*

## 1. INTRODUCTION

This paper examines frame imperfections in the pilot (survey) on Road freight transport conducted during two weeks in 2008, March 23-30 and March 31- April 6 by the Statistical Office of the Republic of Serbia. The goal of this pilot research was to test the methodology and organization of the survey which is designed according to European Commission Regulation 1172/98 on statistical returns in respect of the carriage of goods by road.

The problem of imperfect frame is one of the most important issues which affect the results of this research. In this paper we will discuss frame imperfections which were familiar in the phase of the survey design and those which were observed after the first analysis of the response. Also we will discuss the methods of estimation which could be used in order to make the allowance for frame imperfections.

### 1.1. Introduction to the survey

The purpose of this survey is to collect data on road performances of road freight vehicles in and outside the transport sector of the national economy. These data provide information such as origin and destination of transport, the type of goods carried, type of packaging, the category of dangerous goods, degree of utilization of vehicles, axle configuration, age of vehicles etc.

Statistical unit in this survey is a tractive vehicle. Reporting unit is the owner of the vehicle which provides the data. Time unit is one week.

### 1.2. Population and frame

Population consists of road motor vehicles which are divided into two main sub groups: Lorries and road-tractors (road motor vehicles designed exclusively or primarily to haul other road vehicles which are not power-driven, mainly semi-trailers).

Register of vehicles is under the responsibility of the Ministry of Interior Affairs of the Republic of Serbia (MUP). Register contains data on vehicle registration, name of the owner, identifier of the owner, NACE rev. 1.1 economic activity code of the owner, city/town, address of the owner, municipality code, date of the last registration, vehicle make, country of manufacture, year of the production, engine power, engine volume, load capacity, vehicle use code, vehicle body type code and age etc. It is updated continuously.

There is an agreement between the Statistical Office of the Republic Serbia and The Ministry of Interior Affairs permitting Statistical Office to obtain data on vehicles from the Register. Data were delivered to Statistical Office in two separate files, one for vehicles owned by legal persons (state of Register at the end of 2006) and other for vehicles owned natural persons (Register data from March 2008). The later does not contain necessary information about the owner and his address. That is the reason why only the data base of legal persons was used to construct the sampling frame. The frame was defined by excluding vehicles with laden weight of less then 3, 5 tons and lories with missing laden weight.

In order to get more reliable data, the frame data base was linked to the data base of active enterprises SBR1(state March 2008) by identity number of the legal person to which vehicle belongs. All records could not be linked to SBR because some vehicle records did not have correct identity number of the legal person and some did not have it at all, in all 7, 2 % were not matched. For economic activity we had two sources: Register of vehicles and SBR. In this pilot survey, we decided to use economic activity of the enterprise from the SBR where available, otherwise from the Register of vehicles. About 20% of the records could be classified differently depending of the source of the economic activity in use. Although information on address exists in both the Register of vehicles and the SBR, this information has been used from the Register of vehicles because that is where the vehicle was registered and where there was a higher chance of finding the vehicle. SBR information has been used only as additional control data.

In Table 1 we can see the following information:

Column (1) shows the number of road freight vehicles and road tractors, owned by legal persons, by classes of permissible laden weight, at the end of 2006 (data that Statistics obtained from MUP).

Column (2) shows the number of all vehicles in the frame.

Column (3) shows the number of vehicles in the frame that were linked to SBR by identity code of the legal person.

[1] Statistical office of Serbia has a Statistical business register (SBR) that is still in developing phase. This Register contains data for enterprises that correspond, for now, in 1-1 relation to legal entities. SBR for KAU units of the enterprise is not yet in a usable form.

***Table 1*** *Number of vehicles by laden weight in the register and frame*

| | Register data of vehicles owned by legal persons | Frame | |
|---|---|---|---|
| | | All vehicles | Vehicles linked to SBR |
| | (1) | (2) | (3) |
| Total | 24993 | 19831 | 18405 |
| <3.5 tons | 3652 | | |
| 3.5 – 10 tons | 12206 | 12206 | 11166 |
| >= 10 tons | 5063 | 5063 | 4806 |
| Missing laden weight | 1510 | | |
| Road Tractors | 2562 | 2562 | 2433 |

### 1.3. Sampling design

For the pilot survey stratified simple random sampling was employed. Strata were defined by cross-classifying size classes, activity classes and territory classes. There were three size classes according to the weight capacity, three territory classes and six activity classes.

For each of the two weeks, sample of 350 units was allocated to strata proportionally to 4th root of the number of vehicles in strata. The reason for such allocation was to allocate more units to smaller strata and to spread the sample more uniformly across strata. Smaller strata in most cases are with vehicles of higher permissible laden weight and/or of less frequent activity. We wanted to give more chance to them than in a proportional allocation.

The sample was drawn from the part of the frame which was matched with SBR.

## 2. FRAME IMPERFECTION

### 2.1 Imperfections in the frame for the Survey on road freight transport observed during the construction of the frame

*Missing values*

As we can see from the column 1 of Table 1, there were 1510 vehicles with missing laden weight, which therefore could not be classified by weight.

As mentioned earlier, the data on road vehicles owned by natural persons (entrepreneurs, small enterprises) was given in a separate data base which does not contain information on the owner, his address and economic activity, due to MUP privacy regulations, that are necessary for collection of data. Therefore this data base could not be used for sample selection.

*Differences in database update*

The register of vehicles owned by enterprises was outdated and showed the state of registered vehicles at the end of 2006. On the other hand the register of vehicles owned by natural persons was up to date (state March 2008).

*Different classifications*

The database of vehicles owned by legal persons consists of data collected by MUP Serbia and MUP Vojvodina. Different classifications of vehicle body type were used by these two institutions. Recoding was necessary in order to standardize the codes which additionally complicated merging of the databases of these two institutions. The same goes for the database of vehicles owned by natural persons.

*Problems with linking to SBR*

All records could not be linked to SBR because some vehicle records did not have correct identity number of the legal person and some did not have it at all.

Additional imperfections were found after the analysis of response.

## 2.2 Response rates

***Table 2*** *Response rates*

| Total | Response code | Week 1 (num. of vehicles) | Week 2 (num. of vehicles) | Week 1 (%) | Week 2 (%) |
|---|---|---|---|---|---|
| No reply | 10 | 20 | 37 | 5,7 | 10.6 |
| Unknown address of the enterprise | 11 | 12 | 6 | 3,4 | 1.7 |
| A Refusal | 12 | 13 | 16 | 3,7 | 4.6 |
| Vehicle working in the observed period | 21 | 82 | 85 | 23,4 | 24.3 |
| -item non-response | 21 | 10 | 4 | 2,9 | 1.1 |
| -substitute | 21 | 16 | 6 | 4,6 | 1.7 |
| Vehicle was not working in the observed period | 22 | 48 | 41 | 13,7 | 11.7 |
| Scraped vehicle | 23 | 1 | 1 | 0,3 | 0.3 |
| Sold vehicle | 24 | 14 | 11 | 4,0 | 3.1 |
| Leased vehicle | 25 | 1 | 5 | 0,3 | 1.4 |
| Other reasons why the vehicle did not work | 27 | 8 | 9 | 2.3 | 2.6 |
| Out of scope vehicles | 28 | 16 | 16 | 4,6 | 4.6 |
| Load capacity less then 3,5 tons | 29 | 0 | 1 | 0 | 0.3 |
| Questionnaire missing | | 109 | 112 | 31,1 | 32,0 |
| **Total** | | **350** | **350** | **100** | **100** |

Table 2 shows response rates from the pilot Survey on road freight transport. Here is an additional explanation of some categories:

**No reply** - no response is received from the unit despite all attempts made to contact the unit. This makes the biggest part of non-response in both weeks (5, 7% and 10, 6% respectively).

**Out of scope vehicle** – vehicle was working inside the enterprise (construction sight, factory), vehicle is not designed for the carriage of goods, vehicles were carrying passengers or freight which is out of the scope of this survey, agricultural vehicles, military vehicles, and public administration and public service vehicles.

**Questionnaire missing** - is a category which was added after the analysis of response, due to the lack of feedback from our regional offices Belgrade, Novi Sad and Kraljevo. There is still a dilemma on how to treat these responses since they could be classified both as no reply or unknown address of the enterprise. One possible solution is to classify them as no reply because the majority of missing questionnaires in all regions fall to this category. Another solution is to classify them proportionally to the responses in other regions.

It is obvious that the main reason for these response rates is an outdated vehicle register. This has caused both undercoverage and overcoverage errors.

It can be concluded that the lag between reference time point for the target population and reference time point for the frame population should be as short as possible. According to the Road freight transport methodology made by Eurostat, in most European countries frequency of access to draw a sample is quarterly while the register is updated continuously. The registry of Ministry of Interior Affairs of the Republic of Serbia is updated continuously. However, the accessibility of this data is still limited therefore it is planed to draw a sample on yearly basis in the next realization of this survey.

### 2.3 Frame imperfections found after response analysis

There have been errors in the construction of the frame that caused the overcoverage which became evident after reviewing the answers.

The road vehicles which are not designed for the carriage of goods were left in the frame (such as garbage vehicles or snow removal vehicles) as well as vehicles of the public administration. After excluding these vehicles from the frame by dismissing vehicles with certain body type and vehicles owned by public administration and defense and compulsory social security (section 75 in the NACE rev.1 classification of economic activities) we cut down the number of vehicles in register from 24993 (Table 1) to 20308.

In the sample these vehicles were classified in the category “Out of scope vehicles” and they added significantly to its share (see Table 2).

### 2.4. Estimation

Estimation is as for stratified simple random sampling but with taking into account the specific characteristics of the survey:

1. Time component. Total park of vehicles in that strata should be multiplied by the number of relevant survey periods (if we are estimating two weeks with 2, a quarter with 13, or if we want an estimate of a year with 52);
2. Characteristics of the frame.
3. Implicit assumption is that non-respondents have the same characteristics as respondents.
4. The database of vehicles owned by natural persons will be used in the estimation phase. We have no information about the owners of vehicles in this database and therefore no information about their activity class. In order to allocate the vehicles into strata some assumptions have to be made.

   One possibility is to assume that the number of vehicles owned by natural persons, corresponding to particular territory and permissible leaden weight, in particular activity class is proportional to the number of vehicles in that class of legal persons. However, analysis showed that the distribution of natural persons in SBR over activity classes differs from the distribution of legal persons in the sampling frame over activity classes.

   The second possibility, which we will most likely accept, is to allocate the vehicles into different activity classes according to the distribution of natural persons in SBR. This means that we will assume that every natural person owns only one vehicle.

   We will also assume that natural persons have the same characteristics as respondents.
5. Weighting factors are calculated for each strata. It is unnecessary to calculate them for each time period (a week) of a survey. Weighting factors will be based upon data for both weeks in the pilot survey.

Weighting factor to be applied to the vehicle and journey records for the sampled vehicles in each stratum is calculated in the following way:

$$K \cdot \frac{N}{S + S'}$$

where: $K$ is the time period for which the estimates are needed, $K$=2 (two weeks).

$N$ is the total number of vehicles in stratum. The number $N$ is not the number used when the sample was drawn. This total number includes also: the vehicles that were not successfully linked to SBR and vehicles owned by natural persons.

$S$ is the number of usable questionnaires (working and not working during the reference week);

*S'* is the number of scrapped and out-of scope vehicles (for example good vehicles under 3.5 tone, specialist vehicles not adapted for carrying goods such as cranes) in the relevant stratum.

Frame may contain vehicles that have been scrapped or sold or outside the scope of the survey and as a consequences tones and tones-kilometers may be overestimated if all questionnaires where there is positive information about the status of the vehicle are not included. Scrapped and out of scope vehicle constitute additional useable questionnaires *S'*. It is assumed that the proportion of those vehicles found in the survey is representative of their proportion in the frame. Where the vehicle has been sold or leased, business closed or questionnaire cannot be delivered, it is not possible to use this information because there is no positive information about whether the vehicle still exists, is being used or not.

Taking into consideration the frame imperfections that were found after the response analysis estimation can be done according to the principles given in the section and by exclusion of the vehicles which are not designed for the carriage of goods, and vehicles of public administration from both the frame and the sample. That also means that they are excluded from *S'*.

***Table 3*** *Number of vehicles in updated register and frame, and register of vehicles owned by natural persons*

| | Updated register data of vehicles owned by legal persons | Updated frame | | Register of vehicles owned by natural persons |
|---|---|---|---|---|
| | | All vehicles | Vehicles linked to SBR | |
| | (1) | (2) | (3) | (4) |
| Total | 20308 | 18864 | 17960 | 39418 |
| <3.5 tons | 290 | | | 15726 |
| 3.5 – 10 tons | 11404 | 11404 | 10756 | 17344 |
| >= 10 tons | 4899 | 4899 | 4745 | 4399 |
| Missing laden weight | 1154 | | | 823 |
| Road Tractors | 2561 | 2561 | 2459 | 1126 |

Column (1) shows the updated register data of vehicles owned by enterprises where the vehicles which are not designed for the carriage of goods, and vehicles of public administration were excluded. Column (2) shows the number of all vehicles in the updated frame. Column (3) shows the number of vehicles in the updated frame that were linked to SBR by identity code of the legal person. Column (4) shows the number of vehicles owned by natural persons where the vehicles which are not designed for the carriage of goods, and vehicles of public administration were excluded.

If we would get the updated Register of vehicles (state March 2008) we could use calibration in order to correct HT estimates an in that way reduce undercoverage errors.

## 3. CONCLUSION

Suitable re-weighting methods should be used in order to make the allowance for the inaccuracies of the vehicle register. The response analysis has enabled us to observe the level of inaccuracies in frame which can be fixed in future and help adjust estimation in this pilot research.

Frame imperfections which are the result of the imprecise definition of the frame can be corrected in future. Some frame imperfections can't be corrected because of the quality and accessibility of the database. An improved register of MUP will be introduced at the beginning of 2009. We expect that it will give us the possibility to construct a much better sampling frame.

## REFERENCES

[1] Road freight transport methodology, Volume 1, Reference manual for the implementation of Council Regulation N° 1172/98 on statistics on the carriage of goods by road, Eurostat

*Jelena Kovijanic*

# NEDOSTACI OKVIRA U ISTRAŽIVANJU O PREVOZU ROBE DRUMOM

## REZIME

*Pristup dobrom okviru za uzorak je veoma bitan kod planiranja istraživanja. Ipak, okvir za uzorak se retko kada idealno podudara sa ciljnom populacijom. Cilj ovog rada je da razmotri pristupe u rešavanju problema nesavršenosti okvira na primeru istraživanja o prevozu robe u drumskom saobraćaju. Razmotrićemo broj i tip nedostataka u proučavanom okviru.*

*Author: Mrs. Elizabeta Dumik , State Statistical Office of the Republic of Macedonia*
*e-mail: elizabeta.dumik@stat.gov.mk*

# DATA ANALYSIS AND DATA MODELLING IMPLEMENTED IN THE PROCESS OF CREATION OF THE BUSINESS REGISTER SOLUTION

## ABSTRACT

*This paper treats activities done in the phase of establishing the Statistical business register (SBR) in The Republic of Macedonia. For that purpose administrative and other sources were examined and data frame has been defined. A conceptual and a physical data model have been created in order to define the SBR database concept. Procedures for initially loading of the database and update on monthly /annual bases were programmed. All the procedures for updating the SBR statistical Units, on monthly and annual basis, are documented and the documentation is updated regularly, as any change appears.*

## INTRODUCTION

Until 31.12.2005, the State Statistical Office (SSO) was responsible for an administrative business register and was involved in administrative authorisation of Legal Units through maintaining the evidence of business entities. It was mainly responsible for attachment of identification (ID) numbers and determination of the main activity and had data on all Legal Units at disposal.

The situation changed with the adoption of the Law on One-stop-shop System, in 2006, which enabled the registration of Legal Units at one place in the Central Register (CR). Following this act, the SSO's competencies (ID attachment, determination of activity), as well as its previous register were transfered to the Central Register. That transfer of competencies caused the SSO's activities to be focused on establishing a Statistical Business Register (SBR), as a main goal. Presently, the database of the SBR is established and is in the process of permanent improvement.

Activities taken in the initial phase of establishing the SBR, which resulted in the current situation, were:

- a draft of the national methodology has been prepared in accordance with EU regulations and standards, the national legal base and the national conditions;
- variables(identification, demographic and stratification) have been defined;

- at the moment, only Legal Units, Enterprises and Local Units could be (and have been) included in the SBR;
- data from the main administrative sources (the Central Register, the Pension and Disability Insurance Fund (PF), and some data from the Public Revenue Office, without on-line connection) have been assessed, examined, arranged, investigated, categorized and used;
- as an activity, which will serve for the purposes of the SBR is the Census of Business Entities that is planed to be conducted in 2008 by the SSO;

For development of the SBR, methodological and practical knowledge was necessary, with a specific accent on knowledge of data modelling.

The basic SBR methodological rules were implemented in the phase of data modelling in order to define mechanisms, methods and procedures for establishing, maintenance and updating a sustainable Statistical Business Register.

Data from different sources were thoroughly assessed, examined, arranged, investigated, categorized, codified and re-codified, in order to:

- compare available data and
- define business rules for establishing SBR

**Sources** that were used for establishing the Statistical Business Register database are:

1) Trade register data bases

- Confirmed (database composed of business entities whose data are verified by the Central Register)
- Unconfirmed (database composed of business entities whose data are not verified by the Central Register)

2) Register of Other Legal Entities database;

3) Pension and Disability Insurance Fund database;

4) Central Register financial (annual) accounts database;

5) Central Register Legal Entities and Local Units database (the Central Register delivers changes of business entities on a monthly basis in .xml data format)

- Foundations (new founded business entities within the reference month)
- Changes (changes in data on business entities within the referenced month, such as change of Name, Address, Owner, Local Units, Director, Activity etc.)
- Ceases (termination of business entities within the reference month)

**Procedures** were written/programmed for quantity comparison of all the data sources mentioned above.

Quantitative comparison has been made between:

- The Pension and Disability Insurance Fund data base and the Central Register-annual financial accounts;
- The Trade Register- confirmed database and the Trade Register- unconfirmed database
- The Register of Other Legal Entities database and CR-annual financial accounts
- Trade Register database and confirmed Pension and Disability Insurance Fund database
- The CR-annual financial accounts database and the Trade Register and the Register of Other Legal Entities

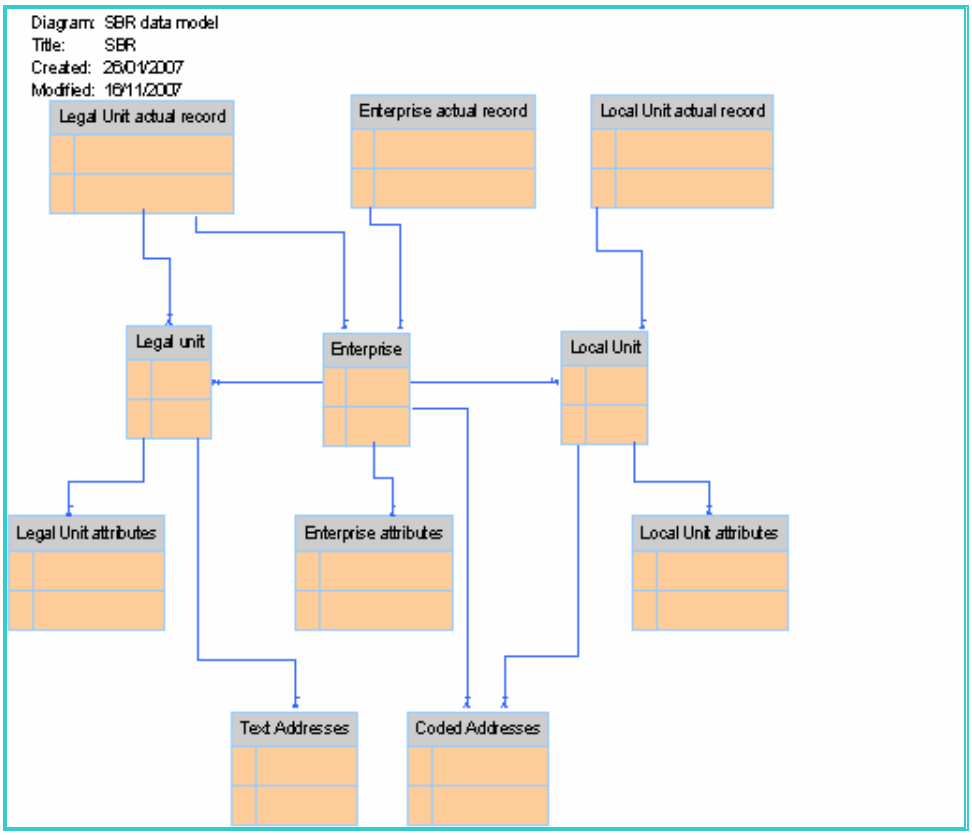
**Implementation** of business rules in the phase of designing the SBR concept is still underway and susceptible to changes. However, up-to-date, we faced some problems such as:

- Adjustment to the very new concept of database received from the CR;
- Different quantity of variables used in both concepts;
- Different approach in the definition of some of the variables (ex. changes in Local Units did not reflect on Legal Entity changes level);
- Codes of conversion from our system to CR system (organization type, activity, change type);
- Resolving unsuitable codes for some of the variables (organization type, activity);
- Resolving missing address codes (lack of street and settlement codes, only given by names);
- Dealing with duplicate data from various data sources;
- Unsuitable/undefined street codes (new ones related to unknown settlements) received from some of the sources;
- Statutory changes, defined as a string instead of a code variable (still a problem);
- Lack of legal acts for obligatory registering of seat address and other contact changes, ownership and capital origin;
- We are forced to keep parallel address variables (statistical ones, from various - statistical surveys);
- We were forced to write procedures to derive ownership and capital origin,
- in accordance with rules for obtaining those categories. The CR did not provide those data in their first phase of registration of the business entities;
- Lack of software tool for "easy" import of nested .xml files that were provided from the CR on a monthly basis;
    - That have led us to a design of an application that can read those files and then reorganize them in a "useful" and "user friendly" form as a temporary solution. That had to be done for a very short period of time (it was time consuming job despite the high expectations of the employees).

Technical solution for importing nested .xml files and distributing them to the users involved in the creation of the BR has been made in order to continue the process of updating the BR data base and to create a final solution.

**Implementation** of business rules in the SBR concept design include the activities - mentioned bellow:

**A conceptual model** of the BR was established using Power Designer tool.




The conceptual model eased the discussion between system developers and users, eased the understanding of the business and gave a structure over the things that were interesting to keep information about.

It also clears out the relationships between entities, identify and describe Object, entities and relations, identify and describe Properties and Identity, and provide good ground to build a Physical model

**Physical data model** was created from the conceptual model**.**

The purpose of Physical data modeling was to make a model that shows how the data should be stored physically.
Modern design tool supports a direct conversion from a class model to a Physical one. That was aimed with the Power designer tool (knowledge of it was gained through the trainings organized in our office by SIDA).

This type of model enables the designer to remodel the database concept as many times as needed for it contains a script that recreate or change the database. It also contains the current documentation of the database.

The optimal solution for a database managing system is not always possible to be applied for every system, but some demands have to be fulfilled.
Therefore, modeling can be compared to a balance, where the settings to increase the performance are weighed against other parameters like safety, user friendliness etc.

Every database table should have unique index - primary key, which mainly protects the database from accepting duplicate records.
Further, the job of the primary key is to locate, as fast as possible, a particular data record in a table. This operation must be carried out whenever data from several tables are assembled - in short, very often indeed.

Redundant storages could be controlled or uncontrolled. Controlled redundancy is usually used in output databases.

Uncontrolled redundant storage can sometimes be useful in an input database but is most likely a case of bad modeling.

To increase the performance and eased the process of data retrieval, de-normalization and controlled redundancy of the database could be needed. This means that redundancy or double storage of some data increases maintenance demand, but makes usage easier. That has been done in our model, but only on output level and only for easier data retrieval and data analysis purpose.

## TECHNICAL ASPECTS OF SBR IMPLEMENTATION

This, so-called, temporary solution, is planed to be part of the major SBR solution, which, at this phase, has fulfilled these steps:

- Database model - a relational normalized database, made with Power Designer database modeling software, has been created.
- Stored procedures, with Transact SQL syntax, have been written in order to load the relational database. This has been done in accordance to the given business rules from the Methodological Department. Data have been inserted from the defined data frame (Trade Register, Register of Other Legal Entities, PF database, CR-annual financial account, Founded Legal Units from the monthly CR database).
- Stored procedures for updating data from the CR database on monthly bases are functional, but still susceptible to improvement.
- Stored procedures for updating data from statistical sources are still in process of adaptation (for the time being, only procedures for insertions of statistical addresses are functional)
- User interface was designed and installed in each of the subject matter departments for update of addresses (update of activity and number of employed

persons is planed to be included, too) on annual, monthly or quarterly basis (depending on the survey periodicity)
  - The results of this activity are considered as a first step in the process of updating the statistical Units from the statistical sources
  - The results are implemented into the SBR

As mentioned above, the SBR is composed of data from administrative sources. But some variables such as Addresses, Activity, Number of employed persons etc., should be updated from the statistical sources. In that way, the SBR should keep track of parallel data for those variables from administrative sources, as well as from statistical sources. Thus, our introductory steps in gathering data from statistical sources, in order to update the SBR with statistical data, are done via user interface for update of addresses (gathering statistical data on activity, number of employed persons and some other variables should be included later in the same user interface). The gathered data on statistical addresses are implemented into the SBR.

All the procedures for updating the SBR statistical Units, on monthly and annual - basis, are documented and the documentation is updated regularly, as any change appears.

## CONCLUSIONS

Finally, we must highlight the importance of keeping the register systematically updated and highlight the need for permanent flow of information for this purpose.

Still, designing, implementing and maintaining a sustainable SBR is complex and takes time to mature.

International experience also shows that constructing and implementing SBR due to its complexity, takes a long time.

However, after implementation of all the conclusions and experts' recommendations in the solution, we hope that SBR shall successfully serve to all purposes such as:

- Creation of a sample frame for other departments need;
- Coordinating of samples;
- Grossing up;
- Imputation and
- Disseminating data on Business Population.

*Elizabeta Dumik*

# ANALIZA I MODELIRANJE PODATAKA U PROCESU IZRADE REŠENJA POSLOVNOG REGISTRA

## REZIME

*Ovaj rad razmatra aktivnosti izvedene tokom uspostavljanja Statističkog poslovnog registra (SPR) u Republici Makedoniji. Za te potrebe ispitani su admistrativni i drugi izvori i definisan je okvir podataka. Konceptualni i fizički model podataka je urađen sa ciljem da se definiše koncept baze podataka SPR. Programirane su procedure za početno punjenje baze podataka i mesečno / godišnje ažuriranje. Dokumentovane su sve procedure za ažuriranje statističkih Jedinica SPR, mesečno i godišnje, a dokumentacija se redovno ažurira, čim se pojave promene.*

*Arijana Amina Muhić, Federal Office of Statistics, Bosnia and Herzegovina*
*E-mail: arijanam@fzs.ba*

# DEALING WITH FRAME IMPERFECTION – CASE OF THE STRUCTURAL BUSINESS STATISTICS (SBS) SURVEY 2008 IN BOSNIA AND HERZEGOVINA (BIH)

## ABSTRACT

*This document is explaining a way of dealing with imperfections in frame which was made for the Structural Business Statistics (SBS) sample survey in Bosnia and Herzegovina (BiH). Results of few business sample surveys, which were conducted in past years in BiH, showed that the non-response rate was mostly high due to the non-updated data, such as address, activity status, activity code… of the enterprises, that are coming out of existing administrative registers. Due to the lack of existence of a business register and the non-updated data in the existing administrative registers, there was a need to simulate some small business registers using data from existing and active business surveys for making a frame for SBS. Each entity statistical institution in our country had to prepare the most updated frame of active enterprises for its entity.*

### *List of Acronyms*

| | |
|---|---|
| *AFR* | *Annual Financial Report (Annex)* |
| *BD* | *Brcko District* |
| *BiH* | *Bosnia and Herzegovina* |
| *FBiH* | *Federation of Bosnia and Herzegovina* |
| *FOS* | *Federal Office of Statistics* |
| *HIFRS* | *Health Insurance Fund of Republic Srpska* |
| *PDIFRS* | *Pension and Disability Insurance Fund of Republic Srpska* |
| *RS* | *Republic Srpska* |
| *RSIS* | *Republic Srpska Institute for Statistics* |
| *SBS* | *Structural Business Statistics* |
| *TARS* | *Tax Administration of Republic Srpska* |
| *UIN* | *Unique Identification Number* |

## 1 INTRODUCTION

The purpose of sample surveys is to gather information about a certain finite population by estimating finite population parameters such as means, totals or fractions. In sampling theory, observations obtained from the sampling units are regarded as fixed. The population to be sampled (the sampled population) should coincide with the population about which information is wanted (the target population). Before selecting the sample, the population must be divided into parts that are called sampling units. In principle, these units must cover the whole population and they must not overlap, in the sense that each element

in the population belongs to exactly one sampling unit. The construction of a list of sampling units, called a sampling frame, is often one of the major practical problems. Sampling frames are often found to be incomplete, or partly eligible, or contain an unknown amount of duplication.

The noncoverage errors may be caused by the use of faulty frames of sampling units. If the frames are not updated or old frames are used as a device to save time or money, it may lead to serious bias. Frame imperfections can bias the estimates in the following ways: If units are not represented in the frame but should have been part of the frame, this results in zero probability of selection for those units omitted from the frame. On the other hand, if some units are duplicated, these result in overcoverage with such units having larger probabilities of selection. Thus noncoverage refers to the negative errors resulting from failure to include elements that would, under normal circumstances, belong to the sample. Positive errors of overcoverage also occur due to inclusion in the sample of elements that do not belong there.

Noncoverage errors differ from nonresponse. Corrections and weighting for noncoverage are much more difficult than for nonresponses. The extent of nonresponse can be measured from the sample results by comparing the selected sample with the achieved sample. By contrast the extent of noncoverage can only be estimated by some kind of check external to the survey operation. This strictly refers to losses and distortions within the sampling frame.

'Miscoverage' by delays in recording real-life events changes the eligibility of population units. For example, new units generally appear on the frame some time after they came into existence and units that have ceased to exist are not removed from the frame immediately. In some countries statistical offices attempt to predict the births and deaths of businesses by developing and using some special methodologies and models for it. The most effective way to reduce coverage error is to improve the frame by excluding erroneous units and duplicates and updating the frame through filed work in order to identify units missing from the frame.

## 2 STRUCTURAL BUSINESS STATISTICS

SBS are based on legal obligation. New Council Regulation No 295/2008 concerning SBS covers all market activities of the business economy with the exception of agricultural activities, which are covered by other statistics. SBS provide important indicators concerning the structure, activities, employment, investment activities and performance of enterprises and entrepreneurs at the national and regional levels in the breakdown by economic branches in accordance with NACE Rev.1.1.

In the year 2008, for the first time this survey was conducted in Bosnia and Herzegovina (BiH). The specificity of it was that for the first time in BiH a structural survey concerning small economical units, the entrepreneurs, was conducted.

## 3 BUILDING UP THE FRAMES

In BiH very often some businesses can be operating for years without appropriate registration because it has changed its activity or name or any other important data, or it has just merged with some other business or split into more than one business due to some expecting benefits, but these changes did not reach the final users because of the complicated way of registration and the delay of update. So the possibility to meet un-updated registers of business units in BiH is very high. This was also shown through some industrial and trade statistics surveys conducted before the preparation of the SBS survey. The undercoverage is the most important type of frame imperfection in BiH.

Due to the lack of existence of the business register and the non-updated data in the existing administrative registers and for the purpose of planning and conducting the SBS sample survey, a simulation of a business register was done, using data from existing and active business surveys in all the three statistical offices of both entities (Federation of BiH (FBiH) and Republic Srpska (RS)) and the Brcko District (BD). Methodologies for building up the frames were developed separately by the statistical offices depending on availability of sources of the requested information.

### 3. 1 Frame for FBiH

A simulation of a business register for FBiH was done on the base of the existing administrative register of business units kept by Federal Office of Statistics (FOS) of BiH, and from the statistical reports: Annex of Annual Financial Report (AFR) for 2006, monthly reports for 2007 concerning employment and salaries and quarterly reports for 2007 concerning trade.

For enterprises (category 02), the business units for which the activity status was A (active), D (moved into canton[1)] or O (moved away from the canton) were selected out of the administrative register. The business units for which the status was O (moved away) were considered as inactive in the canton from which they moved away, but were taken into account for the frame because there had been only few of them and they had completed Annex for 2006, too.

All business units selected by previously described criteria were matched with Annex 2006 and only those which had completed Annex were taken over. In the same way, selection and matching with RAD-1 survey (monthly statistical report on employees and salaries) from August 2007 were performed and, subsequently, with TRG-1 survey (quarterly statistical report on trade statistics) from August 2007.

Matched units were integrated in the following way: firstly, the business units matched with Annex were taken, then business units matched with RAD (which were not included in matching with Annex), and, finally, business units matched with TRG (which were not found during matching with Annex and RAD) were added. It means that all units were present only once.

---

[1] territorial regional unit in FBiH

Apart from identification number for each unit, the following data were also taken over: entity, canton, municipality, status of organization, form of ownership, NACE activity and number of employees. All these data were taken over from the register of survey from which the unit was selected after the matching was carried out. In case that some data was missing in the register of survey from which the business unit was selected, then an adequate data was assigned – if it existed – from the register of the next survey with which matching was carried out. If the data was missing for all surveys with which matching was done, then that data was taken over from the administrative register of the business units.

Partially, the same procedure was used for entrepreneurs (category 03). The business units for which the activity status was A (active) or D (moved in) were selected, but business units for which the activity status was O (moved away) were excluded as it was not for sure that entrepreneurs who moved away from one canton, would continue their business activity in another canton.

As there were no adequate surveys with which matching could be performed, only administrative register of business units from which all data were taken over was used as a source of data. Since the number of employees has not been updated in this register, and most frequently, that number was 0, 1 or 2, meaning that different strata would not be created for entrepreneur's sample, this number has not been presented on the list. For the purpose of creating the list, all enterprises and entrepreneurs, for whom the number of employees was 0, were treated as active enterprises with 1 employee. Later on, after the insight in one of the new National Account's surveys for the years 2006 and 2007, whose results were available after carrying out the SBS survey, it has been realised that the number of entrepreneurs in one division (division 55, group 55.1) was significantly enlarged. This number was changed in frame and also in the sample of the survey afterwards, so that the sample size for FBiH was considerably decreased.

For both categories of business units (enterprises and entrepreneurs), new unique identification numbers were made on the basis of the existing unique numbers.

### 3. 2 Frame for RS

Starting lists were produced by using all available sources of information.

For creating a list of enterprises, the data from the existing administrative register of business units kept by Republic Srpska Institute for Statistics (RSIS) of BiH, and from the statistical reports: Annex of Annual Financial Report (AFR) for 2006, reports for 2007 concerning labour force, transport statistics, trade and tourism statistics.

A starting list of legal entities from the register contained information on enterprises, which have reported changes in the administrative registers since 1997 up to considerable date. Enterprises established in 2007 were deleted from that list as well as enterprises which initiated bankruptcy procedures. That list was compared with information obtained from statistical surveys, with the purpose of possible adding to the list the enterprises, which were active according to those surveys. Afterwards, the enterprises, which were active according to the surveys and did not report any changes in administrative register over the past ten years, were also added to the list.

The code of activity was taken over from the administrative register, while the number of employees was taken from the labour force survey, from the profit and loss accounts and from the administrative register, whereas the data sources were listed according to the given priority.

The units for which there was no information concerning the number of employees were deleted from the list.

List of entrepreneurs was created by using the administrative data from the following institutions:

- Health Insurance Fund of Republic Srpska (HIFRS)
- Pension and Disability Insurance Fund of Republic Srpska (PDIFRS)
- The Tax Administration of Republic Srpska (TARS).

As a starting point for forming the list of entrepreneurs, it was used the HIFRS list of entrepreneurs integrated with the list from PDIFRS (from which the units that were not entrepreneurs were deleted). That data also contained PDIFRS identification number and TARS Unique Identification Number (UIN) for the taxpayer. The reason for integrating it with the data received by the PDIFRS was the fact that the UINs missing in the HIFRS data were updated in the data concerning entrepreneurs. UIN was necessary for integration with the TARS data where observed unit could be connected with the basic unit. For the sake of registration with the PDIFRS, each unit in the system obtained a new registration number and there was no possibility to be linked to the basic unit.

The data on number of employees and code of activity were taken over from the HIFRS data.

### 3. 3 Frame for BD

For the purpose of creating a list of enterprises, the data from the administrative register of business units (which is jointly kept and maintained by Tax Authority for direct taxation and BD Branch for Statistics), Annex of Annual Financial Report (AFR) for 2006, statistical reports on employees and salaries and industry, construction and trade were used.

During the first phase of creating lists, an estimation of activity status of enterprises contained in the register was performed on the base of the data from AFR for the enterprises which have already submitted those reports. The data on the value of turnover and number of employees taken from AFR were integrated in the existing lists of the register. Afterwards, an additional analysis of the data on number of employees was performed by comparing the data on the number of employees taken over from the AFR with the same data from the RAD-1G (an annual statistical report on employees and salaries). For the purpose of creating the list, all enterprises, which reported certain turnover in AFR and for which number of employees was 0, were treated as active enterprises with 1 employee. Additional updating of the lists of active enterprises was carried out on the base of data taken over from the statistics of industry, construction and trade.

For the purpose of creating the entrepreneurs lists, the data were also used from the above mentioned register.

It is important to point out that mandatory revised registration of all business units (enterprises and entrepreneurs) in BD was carried out in 2002. The form for registration of business unit encompassed two data on number of employees: 1) number of employees at the time of registering and 2) planned number of employees. This means that the business units with revised registration in the register (which were already established in 2002) have data on employees from 2002. For the business units registered after 2002, the data included in the register referred to the number of employees in the year of their registration. The list of entrepreneurs integrated data on the number of employees at the time of its registration. So the created list contained data on the number of employees referring to different years (from 2002 to 2006). There was no possibility for checking and updating the data from the statistical sources, because the entrepreneurs have not been monitored through statistical surveys in sectoral statistics. No other (administrative) sources for entrepreneurs have data on the number of employees, their business activities and business addresses.

Additional problem (due to the specific status of BD), faced during lists' creation and taking over the data from AFR, was actual use of two entity systems of accounting in BD. The business units with FBiH accounting system had to be separated on the list from the business units using RS accounting system. The forms with links to the corresponding chart of accounts have been delivered to the reporting units.

## CONCLUSION

The frame imperfection is a problem that can have huge influence on many issues, i.e. collecting enough and sufficient data, final estimation… For the situation in BiH the most proper way to solve this problem, at least temporary, was the simulation of an updated business register. In all the three statistical offices statisticians worked separately, but they used almost identical way to obtain as good as possible frame for the first SBS survey in BiH. For the first time those preparations of the frame were made in the statistical offices of BiH, and this will be continued doing if the real business register does not come to the use in the coming time. Of course, the data from the SBS register will be also used as one of the sources for updates of the business register, when that one comes in use.

## REFERENCES

[1] BHAS, FOS, RSIS and ISTAT, *The BiH Structural Business Survey 2008 , Draft of the Methodological Report* (2008), Sarajevo

[2] Biemer P.P. and Lyberg, L. E., Introduction *to Survey Quality*: Wiley Series in Survey Methodology, (2003), Wiley, Hoboken

[3] Renssen R.H, *A Course in Sampling Theory* (*1998*), Heerlen

*Arijana Amina Muhić*

# REŠAVANJE NESAVRŠENOSTI OKVIRA: ISTRAŽIVANJE STRUKTURNE EKONOMSKE STATISTIKE (SBS) 2008 U BOSNI I HERCEGOVINI (BIH)

## REZIME

*Ovaj rad objašnjava način rešavanja nesavršenosti okvira koji je urađen za istraživanje strukturne ekonomske statistike (SBS) na uzorku u Bosni i Hercegovini (BiH). Rezultati nekoliko istraživanja o poslovanju na uzorku, koja su sprovedena poslednjih godina u BiH, pokazuju da je stopa odbijanja ankete bila uglavnom visoka zbog neažuriranih podataka, kao što su adresa, status aktivnosti, šifra aktivnosti... preduzeća, koji potiču iz postojećih administrativnih registra. Zbog nepostojanja poslovnog registra i neažuriranih podataka u postojećim administrativnim registrima, ukazala se potreba za simuliranjem malih poslovnih registara, koji koriste podatke iz postojećih i aktivnih istraživanja o poslovanju, za izradu okvira za istraživanja strukturne ekonomske statistike (SBS). Statističke ustanove u entitetima u našoj zemlji treba da pripreme najažurniji okvir aktivnih preduzeća za svoj entitet.*

# PRIKAZI KNJIGA

*Slobodan Nićin , magistrant Poljoprivrednog fakulteta u Novom Sadu, smer Kvantitativna analiza*

# PRIKAZ KNJIGE *"APPLIED MULTIVARIATE ANALYSIS"*

Knjiga "Applied Multivariate Analysis" autora Neil H. Timm-a izašla je u izdanju izdavačke kuće Springer, u ediciji Springer Texts in Statistics 2002 godine. Napisana je na 693 strane.

Iza detaljnog sadržaja dati su spiskovi tabela i dijagrama koje se u knjizi nalaze.

Ova knjiga sastoji se iz deset poglavlja.

Prvo poglavlje nosi naziv *"Introduction"* i u njemu je dat pregled multivarijacionih modela i metoda, kao i cilj i namena ove knjige.

Drugo poglavlje *("Vectors and Matrices")* sadrži uvod u vektorsku i matričnu algebru, gde su objašnjeni vektori, vektorski prostori i podprostori. Dalje, objašnjene su vektorske norme i algebra vektorskih prostora. U jednoj delu ovog poglavlja date su osnovne operacije sa matricama: jednakost, sabiranje i množenje matrica, transponovanje matrica, trag matrice, Euklidska matrična norma, Kronecker- ovi i Hadamard- ovi proizvodi, direktne sume. Slede pojmovi ranga matrice, inverzne matrice, determinante, koji su detaljno prikazani i potkrepljeni sa odgovarajućim primerima.

Kasnije su dati sistemi jednačina, transformacije i kvadratne forme: linearne transformacije, projekcione transformacije, svojstvene vrednosti i svojstveni vektori, matrične norme, kvadratne forme i ekstremi.

Treće poglavlje *("Multivariate Distributions and the Linear Model")* sadrži multivarijacione distribucije i linearni model. U ovom poglavlju detaljno su opisani slučajni vektori i matrice, multivarijaciona normalna distribucija (osobine, ocene i matrična normalna distribucija), $\chi^2$ – distribucija, Wishart-ova distribucija, kao i druge multivarijacione distribucije - univarijantna t i F distribucija, Hotelingova ( Hotelling) $T^2$ – distribucija , $\beta$ – distribucija, kao i multivarijacione $t, F$ i $\chi^2$ -distribucija. U delu ovog poglavlja pod naslovom opšti linearni model dati su regresioni model, model analize varijanse i analize kovarijanse, zatim multivarijaciona regresija, multivarijaciona analiza varijanse i multivarijaciona analiza kovarijanse i opšti multivarijacioni model analize varijanse. U narednim delovima ovog poglavlja daju se evaluacija ( vrednovanje) normalnosti, testovi kovarijacionih matrica ( jednakost kovarijacionih matrica, testiranje specifične kovarijacione matrice, test sferičnosti, test nezavisnosti, testovi za linearnu strukturu, testovi lokacije) i power-kalkulacije.

Poglavlje 4 *("Multivariate Regression Models")* sadrži podpoglavlja multivarijaciona regresija sa urađenim primerom, jednosmerna multivarijaciona analiza varijanse i multivarijaciona analiza kovarijanse potkrepljene sa odgovarajućim primerima, kao i simultane

test procedure za multivarijacionu analizu varijanse i multivarijacionu analizu kovarijanse, dvodimenzionalnu multivarijacionu analizu varijanse i multivarijacionu analizu kovarijanse, dvodimenzionalnu multivarijacionu analizu varijanse sa interakcijom, aditivnu dvosmernu multivarijacionu analizu varijanse , testove neaditivnosti. Tu su nabrojani i objašnjeni i neortogonalni planovi dvodimenzionalne multivarijacione analize,fiksni efekti višeg reda, „nested“ i ostali planovi ogleda ( designs). Takođe, dati su i planovi sa ponovljenim merenjima, analiza robustnosti i power- analiza multiregresionog modela, kao i testiranja razlike između aritmetičkih sredina sa nejednakim kovarijacionim matricama.

Poglavlje broj 5 *("Seemingly unrelated regression models")* sadrži SUR (seemingly unrelated regression) model ( ocena i testiranje hipoteza, predviđanje), što je potkrepljeno jednim primerom. Tu su i CG MANOVA model sa ilustracijom, GMANOVA model sa primenom, test neaditivnosti sa odgovarajućim primerom, testovi sa nemogućnošću fitovanja ( prilagođavanja), kao i multivarijacioni SUR (MSUR) model.

U 6. poglavlju *("Multivariate Random and Mixed Models*") dati su regresioni modeli slučajnih koeficijenata, zatim univarijantni opšti linearni mešoviti modeli, nekoliko primera iz mešovitih modela, zatim mešoviti multivarijantni modeli, primeri za balansirane mešovite multivarijantne modele, dvostruki multivarijantni model sa primerom, multivarijantni hijerarhijski linearni model i testiranje sredina sa nejednakim kovarijacionim matricama.

Sedmo poglavlje *("Discriminant and Classification Analysis")* odnosi se na diskriminacionu i klasifikacionu analizu, gde su date Fišerova linearna diskriminaciona funkcija, testiranje koeficijenata diskriminacione funkcije, pravila klasifikovanja, primer za diskriminacionu analizu dve grupe, diskriminaciona analiza za više grupa ( testiranje značajnosti diskriminacione funkcije, izbor promenljivih) i primer za diskriminacionu analizu za više grupa.

U osmom poglavlju *("Principal Component, Canonical Correlation, and Exploratory Factor Analysis")* dat je detaljan opis metoda glavnih komponenata, kanoničke korelacije i objašnjavajuće faktorske analize.

Što se glavnih komponenata tiče, dat je populacioni model za metod glavnih komponenata, broj i struktura komponenata, glavne komponente sa kovarijatima, uzorački metod glavnih komponenata, otkrivanje autlajera i dr.

Statistički testovi kod metoda glavnih komponenata na bazi kovarijacione i korelacione matrice takođe su dati u ovom poglavlju. Analiza glavnih komponenata praćena je brojnim primerima iz različitih naučnih oblasti. Dalje, opisana je primena regresije na glavne komponente sa primerom.

Jedan deo ovog poglavlja nosi naziv kanonička korelaciona analiza, gde su objašnjeni populacioni model, uzoračka kanonička korelaciona analiza, testovi značajnosti asocijacija, parcijalna kanonička korelacija, predviđanje u multivarijacionoj regresiji primenom kanoničke korelacione analize, izbor promenljivih u kanoničkoj korelacionoj analizi. Za kanoničku korelacionu analizu data su i dva primera njene primene.

Ovde je detaljno opisana objašnjavajuća faktorska analiza (populacioni model, ocena parametara modela, određivanje prilagođenosti modela, rotacija faktora, ocena faktorskih skorova) i dato je nekoliko primera za objašnjavajuću faktorsku analizu.

Deveto poglavlje *("Cluster Analysis and Multidimensional Scaling")* odnosi se na klaster analizu i višedimenzionalno skaliranje ( mere različitosti, sličnosti i promenljive za klasterizaciju). Dat je detaljan opis klaster analize (aglomerativni hijerarhijski modeli klasterizacije, nehijerarhijski metod klasterizacije, broj klastera), što je i potkrepljeno brojnim primerima iz raznih oblasti. U jednom od delova ovog poglavlja opisan je metod višedimenzionalnog skaliranja i data su tri primera koji ilustruju ovaj metod.

Modeli strukturnih jednačina dati su u desetom poglavlju *("Structural Equation Models")*, gde podpoglavlja nose nazive potvrđujuća faktorska analiza ( sa primerima), „PATH"-analiza ( sa primerima), strukturne jednačine za manifestne i latentne promenljive ( sa primerom), longitudinalna analiza sa latentnim promenljivim i egzogenost u modelima strukturnih jednačina.

Osim ovih deset poglavlja, knjiga sadrži i *Appendix, References, Author Index* i *Subject Index*.

U delu pod nazivom *Appendix* (dodatak) date su tablice sa kritičnim vrednostima za pojedine teorijske distribucije.

Deo *References* ( pregled literature*)* sadrži 642 literarna navoda.

*Author Index* ( indeks autora) dat je na 8 strana, a *Subject Index* ( indeks pojmova) dat je na 19 strana.

Ovu knjigu mogu koristiti svi oni koji se bave primenom metoda multivarijacione analize u istraživanjima. Takođe, ona se može preporučiti i kao literatura u nastavi primenjene multivarijacione analize, kako i sam autor navodi u predgovoru.

*Predrag Čanović, Republički zavod za statistiku*

# DVA VEKA RAZVOJA SRBIJE – STATISTIČKI PREGLED

Republički zavod za statistiku je 24. oktobra 2008. godine, na Sajmu knjiga održanom u Beogradu, predstavio knjigu "Dva veka Srbije – statistički pregled".

U njoj su prikazani konstituisanje i razvoj Srbije od Prvog srpskog ustanka do danas.

Istovremeno, stučnoj i široj javnosti predstavljeni su razvoj, delovanje i rezultati državne statistike Srbije, koja je doprinela afirmaciji statistike kako u ondašnjem srpskom društvu, tako i van granica zemlje. Briga o ugledu i ozbiljnost nastupa državne statistike pred međunarodnim auditorijumom ogleda se i u činjenici da su njene publikacije u periodu od 1862. do 1939. godine štampane dvojezično: na srpskom i na francuskom jeziku. Ovom prilikom, tekst na srpskom jeziku prati prevod na engleski.

Na 302 stranice, u dvadeset tematskih poglavlja, prikazane su najvažnije oblasti društveno-ekonomskog života u tom periodu. Obuhvaćeni su svi segmenti o kojima je državna statistika tokom protekla dva veka zabeležila ponešto, a korišćeni su i drugi verodostojni izvori iz raspoloživih arhiva i publikacija.

Svako poglavlje počinje tekstom u kom su opisani prilike i uslovi, navedeni izvori podataka i data metodološka objašnjenja kao pomoć čitaocu u boljem razumevanju podataka.

Tabele koje slede sadrže duge serije podataka o najznačajnijim pojavama u posmatranom periodu. Predstavu upotpunjuju grafički prikazi, a izvori podataka su dati u fusnoti. Problemi uporedivosti podataka koji su proizašli iz promene teritorije, metodologije, pa i same pojave, delom su rešeni kompozicijom poglavlja.

Podaci su prikazani po sledećim periodima:
1. Pre Prvog svetskog rata,
2. Između dva svetska rata i
3. Posle drugog svetskog rata.

Teritorija Srbije se u protekla dva veka više puta menjala, pa je prvo poglavlje posvećeno teritorijalnim promenama i sadrži karte Srbije nakon svake od promena granica.

Izdvojili smo jednu tabelu u kojoj je uspešno rešen problem uporedivosti podataka, s obzirom na promenu teritorije.

## Становништво на историјским областима Србије
## Population in historical regions of Serbia

хиљ./*thous.*

| | Београдски пашалук *Belgrade Pašaluk* | Нахије припојене 1833. *Annexed countries (nahias), 1833* | Окрузи припојени 1878. *Annexed districts 1878* | Области припојене 1912. *Annexed regions 1912* | Војводина и друге области припојене 1918. *Vojvodina and other annexed regions* | Припојене општине у Крајинском и Зајечарском срезу *Annexed municipalities in the districts of Krajina and Zaječar* |
|---|---|---|---|---|---|---|
| 1833 | 475 | 203 | … | … | … | … |
| 1884 | 1 082 | 464 | 356 | … | … | … |
| 1910 | 1 645 | 705 | 415 | … | … | … |
| 1921 | 1 445 | 619 | 364 | 767 | 1 537 | 48 |
| 1931 | 1 830 | 784 | 461 | 971 | 1 624 | 57 |
| 1948 | 2 111 | 878 | 819 | 956 | 1 719 | 45 |
| 1953 | 2 285 | 934 | 859 | 1 063 | 1 792 | 44 |
| 1961 | 2 510 | 977 | 877 | 1 235 | 2 004 | 39 |
| 1971 | 2 747 | 1 010 | 906 | 1 537 | 2 211 | 36 |
| 1981 | 2 993 | 1 043 | 931 | 1 898 | 2 417 | 31 |
| 1991[1)] | 3 060 | 1 027 | 931 | 2 288 | 2 465 | 26 |
| 2002[2)] | 3 049 | 997 | 895 | 353 | 2 576 | 23 |

[1)] Подаци за Косово и Метохију и општине Бујановац, Медвеђа и Прешево су процењени.
[2)] Попис 2002. године није спроведен на Косову и Метохији.

[1)] Data for Kosovo and Metohia and municipalities of Bujanovac, Medveđa and Preševo are estimated.
[2)] The census in 2002 was not carried out in Kosovo and Metohia.

Nažalost, za veći deo građe nije se mogao obezbediti sličan pregled.

Oskudni izvori iz prve polovine XIX veka ne obezbeđuju duge serije podataka, posebno kada je reč o ekonomskim statistikama. Ipak, veći problem je predstavljala podela Kraljevine Jugoslavije na banovine, pošto je teritorija Srbije bila uključena u pet banovina. Za taj period mogle su se rekonstruisati samo serije podataka koji su objavljivani na nivou srezova, na osnovu čega se izdvajao deo teritorije koji je sačinjavao Srbiju. Najbogatiji su prikazi posle Drugog svetskog rata, pošto je statistička služba u tom periodu znatno ojačala i bila u mogućnosti da obezbedi obilje statističke građe.

Deo koji se odnosi na stanovništvo je najobimniji. To je i razumljivo s obzirom na dugu istoriju popisa stanovništva, koji je i utemeljen kao instrument za planiranje poreske politike. Popisni podaci dobro su izbalansirani sa podacima vitalne statistike. Upečatljivi su uticaji promene teritorije, teške posledice brojnih ratova, kao i migracionih faktora.

Već u delu koji se odnosi na školstvo počinju problemi sa uporedivošću, na šta je čitalac upozoren. Naime, osnovna škola iz prve polovine XIX veka nije isto što i osnovna škola danas. Problemi u oblasti zdravstvene zaštite prisutni danas takođe su veoma različiti od tadašnjih. O problemima koje je trebalo prevazići kada su u pitanju ekonomska kretanje

ne treba trošiti reči: promena monete, promena strukture proizvodnje, pa čak i promena metodologije propisane za praćenje pojedinih pojava otežavali su uporedivost podataka u ovako dugom razdoblju.

Čini nam se da je ipak svako od autora uložio dosta napora i da su problemi rešeni na najbolji mogući način. Nemerljiv doprinos ovom poduhvatu dao je dr Miodrag Nikolić, koji je zahvaljujući svom bogatom iskustvu postavio okvire ove publikacije i usmeravao autore u nalaženju optimalnih rešenja.

Obilje podataka, duge serije i raznovrsnost obuhvaćenih oblasti dovoljna su preporuka za ovu knjigu, koja može biti od velike koristi naučnim radnicima, ekonomistima i istoričarima. Istovremeno, knjiga će zainteresovati svakoga ko želi da se suoči sa objektivnom slikom Srbije, koja je u procesu raspada Jugoslavije bila izložena neargumentovanim, tendencioznim, pa i zlonamernim uticajima.

Delovi knige posvećeni izborima, pravosuđu i školstvu svedoče o demokratskom i sekularnom karakteru države, koja je često prednjačila na tom polju u odnosu na mnoge države u Evropi. Stoga je ova knjiga, u borbi protiv predrasuda o Srbiji, korisna za svakog ko želi da sazna istinu o našoj zemlji.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

311

**STATISTIČKA revija /** glavni i odgovorni urednik Miladin Kovačević. - God. 1, br. 1 (1951) - . - Beograd (Milana Rakića 5) : Statističko društvo Srbije, 1951 - (Beograd : Republički zavod za statistiku Srbije). - 25 cm

Tromesečno
ISSN 0039-0534 = Statistička revija
COBISS.SR-ID 46175239

UDK 31 ISSN 0039-05-34

# STATISTICAL REVIEW

YEAR LVII Number 1-2 FOR 2008

## CONTENS



ISSUED BY STATISTICAL ASSOCIATION OF SERBIA - BELGRADE