CONTENTS

Metodološki zvezki, Vol. 11, No. 2, 2014

Metodološki zvezki, Vol. 11, 2014

Reviewers for Volume Eleven

Rok Blagus
Andrej Blejec
Lea Bregar
Matevž Bren
Germa Coenders
Lluís Coromina
Patrick Doreian
Anuška Ferligoj
Herwig Friedl
Georg Heinze
Nataša Kejžar
Katarina Košmelj
Irena Krizman
Nada Lavrač
Giovanni Millo
Irena Ograjenšek
Jože Rovan
Janez Stare
Damjan Škulj
Aleš Toman
Vasja Vehovar
Gaj Vidmar
Anja Žnidaršič
Aleš Žiberna

Metodološki zvezki, Vol. 11, No. 2, 2014, 79-92

Confidence Interval for the Process Capability Index C_p Based on the Bootstrap-*t* Confidence Interval for the Standard Deviation

Wararit Panichkitkosolkul[1]

Abstract

This paper proposes a confidence interval for the process capability index based on the bootstrap-*t* confidence interval for the standard deviation. A Monte Carlo simulation study was conducted to compare the performance of the proposed confidence interval with the existing confidence interval based on the confidence interval for the standard deviation. Simulation results show that the proposed confidence interval performs well in terms of coverage probability in case of more skewed distributions. On the other hand, the existing confidence interval has a coverage probability close to the nominal level for symmetrical or less skewed distributions. The code to estimate the confidence interval in R language is provided.

1 Introduction

Statistical process quality control has been widely applied in many industries. One of the quality measurement tools used for improvement of quality and productivity is the process capability index (PCI). Process capability indices are practical tools for establishing the relationship between the actual process performance and the manufacturing specifications. Although there are many process capability indices, the most commonly used index is C_p (Kane, 1986; Zhang, 2010). In this paper, we focus on the process capability index C_p, defined by Kane (1986) as:

$$C_p = \frac{USL - LSL}{6\sigma}, \tag{1}$$

where USL is the upper specification limit, LSL is the lower specification limit, and σ is the process standard deviation. The numerator of C_p gives the size of the

[1] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand; wararit@mathstat.sci.tu.ac.th

range over which the process measurements can vary. The denominator gives the size of the range over which the process actually varies (Kotz and Lovelace, 1998). Due to the fact that the process standard deviation is unknown, it must be estimated from the sample data $\{X_1,...,X_n\}$. The sample standard deviation S; $S=\left((n-1)^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^{1/2}$ is used to estimate the unknown parameter σ in Equation (1). The estimator of the process capability index C_p is therefore

$$\hat{C}_p=\frac{USL-LSL}{6S}. \tag{2}$$

Although the point estimator of the capability index C_p shown in Equation (2) can be a useful measure, the confidence interval is more useful. A confidence interval provides much more information about the population characteristic of interest than does a point estimate (e.g., Smitson, 2001; Thompson, 2002; Steiger, 2004). The confidence interval for the capability index C_p is constructed by using a pivotal quantity $Q=(n-1)S^2/\sigma^2\sim\chi^2_{(n-1)}$. Therefore, the $(1-\alpha)100\%$ confidence interval for the capability index C_p is

$$\left(\hat{C}_p\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}},\hat{C}_p\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\right), \tag{3}$$

where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the $(\alpha/2)100^{th}$ and $(1-\alpha/2)100^{th}$ percentiles of the central chi-square distribution with $n-1$ degrees of freedom.

The confidence interval for the process capability index C_p shown in Equation (3) is to be used for data that are normal. The coverage probability of this confidence interval is close to a nominal value of $1-\alpha$ when the data are normally distributed. However, the underlying process distributions are non-normal in many industrial processes. (e.g., Chen and Pearn, 1997; Bittanti et al., 1998; Wu et al., 1999; Chang et al., 2002; Ding, 2004). In these cases, the coverage probability of the confidence interval can be appreciably below $1-\alpha$. Cojbasic and Tomovic (2007) presented a nonparametric confidence interval for the population variance based on ordinary t-statistics combined with the bootstrap method for a skewed distribution. In this paper, we propose a new confidence interval for the process capability index C_p based on the bootstrap-*t* confidence interval proposed by Cojbasic and Tomovic (2007).

The paper is organized as follows. In Section 2, the theoretical background of the existing confidence interval for the C_p is discussed. In Section 3, we provide an analytical formula for the confidence interval for the C_p based on the bootstrap-*t* confidence interval for the standard deviation. In Section 4, the performance of the confidence intervals for the C_p are investigated through a Monte Carlo simulation study. Conclusions are provided in the final section.

2 Existing confidence interval for the process capability index

Suppose $X_i \sim N\left(\mu,\sigma^2\right), i=1,2,...,n,$ a well-known $(1-\alpha)100\%$ confidence interval for the population variance σ^2, using a pivotal quantity $Q=(n-1)S^2/\sigma^2$, is (Cojbasic and Loncar 2011)

$$\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \tag{4}$$

where $S^2=(n-1)^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2,$ and $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$ are the $(\alpha/2)100^{th}$ and $(1-\alpha/2)100^{th}$ percentiles of the central chi-square distribution with $n-1$ degrees of freedom, respectively. From Equation (4), we have

$$\begin{aligned}
&\mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right) &&= 1-\alpha\\
&\mathrm{P}\left(\frac{\chi^2_{\alpha/2,n-1}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{1-\alpha/2,n-1}}{(n-1)S^2}\right) &&= 1-\alpha\\
&\mathrm{P}\left(\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{(n-1)S^2}} < \frac{1}{\sigma} < \sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{(n-1)S^2}}\right) &&= 1-\alpha\\
&\mathrm{P}\left(\frac{(USL-LSL)}{6}\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{(n-1)S^2}} < \frac{(USL-LSL)}{6\sigma} < \frac{(USL-LSL)}{6}\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{(n-1)S^2}}\right) &&= 1-\alpha\\
&\mathrm{P}\left(\frac{(USL-LSL)}{6S}\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}} < C_p < \frac{(USL-LSL)}{6S}\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\right) &&= 1-\alpha.
\end{aligned}$$

We obtain a $(1-\alpha)100\%$ confidence interval for the C_p based on the confidence interval for the standard deviation which is

$$CI_1 = \left(\hat{C}_p\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}}, \hat{C}_p\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}\right). \tag{5}$$

3 Proposed confidence interval for the process capability index

The bootstrap introduced by Efron (1979) is a computer-based and resampling method for assigning measures of accuracy to statistical estimates (Efron and Tibshirani, 1993). For a sequence of independent and identically distributed (i.i.d.) random variables, the bootstrap procedure can be defined as follows (Tosasukul et al., 2009). Let $X_1, X_2, ..., X_n$ be independently and identically distributed random

variables from some distribution with mean μ and variance σ^2. Let the random variables $\{X_j^*, 1 \le j \le m\}$ be the result of sampling m times with replacement from the n observations $X_1, X_2, ..., X_n$. The random variables $\{X_j^*, 1 \le j \le m\}$ are called the bootstrap samples from original data $X_1, X_2, ..., X_n$. A confidence interval for the population variance can be constructed using the aforementioned pivotal quantity $Q = (n-1)S^2/\sigma^2$. For large sample sizes, a central chi-square distribution with $n-1$ degrees of freedom can be approximated by a normal distribution with mean $n-1$ and variance $2(n-1)$ (Cojbasic and Tomovic, 2007). Therefore, the distribution of the standardized variable

$$Z = \frac{\frac{(n-1)S^2}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} = \frac{S^2 - \sigma^2}{\sqrt{\text{var}(S^2)}}$$

converges to a standardized normal distribution as n increases to infinity. The bootstrap confidence interval for the σ^2 is calculated based on the statistic

$$T = \frac{S^2 - \sigma^2}{\sqrt{\widehat{\text{var}}(S^2)}},$$

where $\widehat{\text{var}}(S^2)$ is a consistent estimator of the variance of S^2. Casella and Berger (2001) have shown the estimator of $\text{var}(S^2)$ for a non-normal distribution such that

$$\widehat{\text{var}}(S^2) = \frac{1}{n}\left(\hat{\mu}_4 - \frac{n-3}{n-1}S^4\right) \text{ and } \hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4.$$

After re-sampling B bootstrap samples, in each bootstrap sample we compute the value of the following statistic

$$T^* = \frac{S^{*2} - S^2}{\sqrt{\widehat{\text{var}}(S^{*2})}}, \tag{6}$$

where S^{*2} is a bootstrap replication of statistic S^2, $\widehat{\text{var}}(S^{*2}) = \frac{1}{n}\left(\hat{\mu}_4^* - \frac{n-3}{n-1}S^{*4}\right)$ and $\hat{\mu}_4^* = \frac{1}{m}\sum_{i=1}^{m}(X_i^* - \bar{X}^*)^4$. The $(1-\alpha)100\%$ bootstrap-t confidence intervals for the σ^2 is

$$\left(\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2} + \sqrt{2(n-1)}}, \frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2} + \sqrt{2(n-1)}} \right),$$

where $\hat{t}^*_{\alpha/2}$ and $\hat{t}^*_{1-\alpha/2}$ are the $(\alpha/2)100^{th}$ and $(1-\alpha/2)100^{th}$ percentiles of T^* shown in Equation (6). Additionally, the $(1-\alpha)100\%$ confidence interval for the standard deviation σ is

$$\left(\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2} + \sqrt{2(n-1)}}\right]^{1/2}, \left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2} + \sqrt{2(n-1)}}\right]^{1/2} \right). \tag{7}$$

Then, from Equation (7), we construct the confidence interval for the C_p based on the bootstrap-t confidence interval for the standard deviation which is

$$\mathrm{P}\left(\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2}+\sqrt{2(n-1)}} \right]^{1/2} < \sigma < \left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2}+\sqrt{2(n-1)}} \right]^{1/2} \right) = 1-\alpha$$

$$\mathrm{P}\left(\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} < \frac{1}{\sigma} < \left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} \right) = 1-\alpha$$

$$\mathrm{P}\left(\frac{USL-LSL}{6}\cdot\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} < \frac{USL-LSL}{6\sigma} < \frac{USL-LSL}{6}\cdot\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} \right) = 1-\alpha$$

$$\mathrm{P}\left(\frac{USL-LSL}{6}\cdot\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} < C_p < \frac{USL-LSL}{6}\cdot\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} \right) = 1-\alpha.$$

Therefore, the confidence interval for the C_p based on the bootstrap-*t* confidence interval for the standard deviation is given by

$$CI_2 = \left(\frac{USL-LSL}{6}\cdot\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2}, \frac{USL-LSL}{6}\cdot\left[\frac{S^2\sqrt{2(n-1)}}{2\hat{t}^*_{1-\alpha/2}+\sqrt{2(n-1)}} \right]^{-1/2} \right). \quad (8)$$

All confidence intervals were implemented using the open source statistical package R (Ihaka and Gentleman, 1996); source code is available in Appendix.

4 Simulation study

To assess the performance of the proposed confidence interval, we conducted a Monte Carlo simulation study to estimate the coverage probabilities and expected lengths of the proposed confidence interval under different situations and compare them with the existing confidence intervals. The estimated coverage probability and the expected length (based on M replicates) are given by

$$\widehat{1-\alpha} = \frac{\#(L \le C_p \le U)}{M},$$

and

$$\widehat{Length} = \frac{\sum_{j=1}^{M}(U_j - L_j)}{M},$$

where $\#(L \le C_p \le U)$ denotes the number of simulation runs for which the true process capability index C_p lies within the confidence interval. The right-skewed data were generated with the population mean μ = 50 and the population standard deviation σ = 1 given in the Table 1.

Table 1: Probability distributions generated and the coefficient of skewness for Monte Carlo simulation.

Probability Distributions	Coefficient of Skewness
$N(50,1)$	0.000
$Uniform(48.268, 51.732)$	0.000
$10 \times Beta(4.4375, 13.3125) + 47.5$	0.506
$Gamma(9,3) + 47$	0.667
$Gamma(4,2) + 48$	1.000
$Gamma(2.25,1.5) + 48.5$	1.333
$Gamma(1,1) + 49$	2.000
$Gamma(0.75, 0.867) + 49.1340$	2.309
$Gamma(0.5, 0.707) + 49.2929$	2.828
$Gamma(0.4, 0.6325) + 49.3675$	3.163
$Gamma(0.3, 0.5477) + 49.4523$	3.651
$Gamma(0.25, 0.5) + 49.5$	4.000

The true values of the process capability index C_p, LSL and USL are set in the Table 2.

Table 2: True values of C_p, LSL and USL used for Monte Carlo simulation.

True Values of C_p	LSL	USL
1.00	47.00	53.00
1.33	46.01	53.99
1.50	45.50	54.50
1.67	44.99	55.01
2.00	44.00	56.00

The sample sizes simulated were 10, 25, 50 and 100 and the number of simulation trials was set to 10,000. The number of bootstrap samples is 1,000. The nominal confidence level was fixed at 0.95. All simulations were performed using programs written in the open source statistical package R (Ihaka and Gentleman, 1996).

The simulation results are presented for four cases. As can be seen from Figures 1 and 2, the existing confidence interval (CI_1) provides more estimated coverage probabilities than the proposed confidence interval (CI_2) when the data were generated from symmetrical and less skewed distributions (coefficient of skewness between 0 and 2) for all sample sizes. Namely, CI_1 provides estimated coverage probabilities close to the nominal level 0.95, which is more than those of the CI_2 for the normal distribution. In addition, the expected lengths of CI_2 were shorter than those of CI_1 for all sample sizes (see Figures 5 and 6).

On the other hand, for more skewed distributions (coefficient of skewness between 2.309 and 4), the estimated coverage probabilities of CI_2 were greater than those of CI_1 for almost all sample sizes as shown in Figures 3 and 4. Figures 7 and 8 present the results on the expected lengths of CI_1 and CI_2 in case of right-skewed distributions. We found that the expected lengths of CI_1 were shorter than those of CI_2 for all sample sizes.

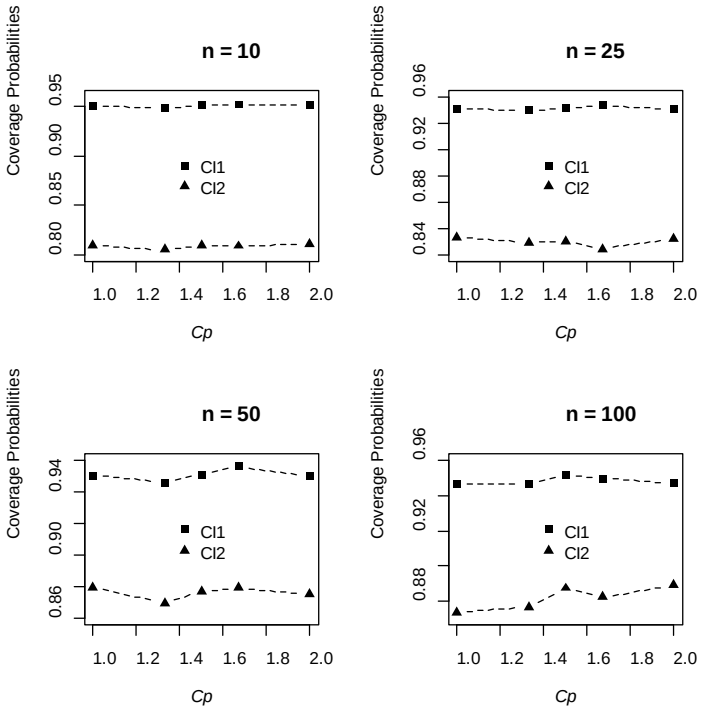

Figure 1: The estimated coverage probabilities of CI_1 and CI_2 for C_p in case of $N(50,1)$

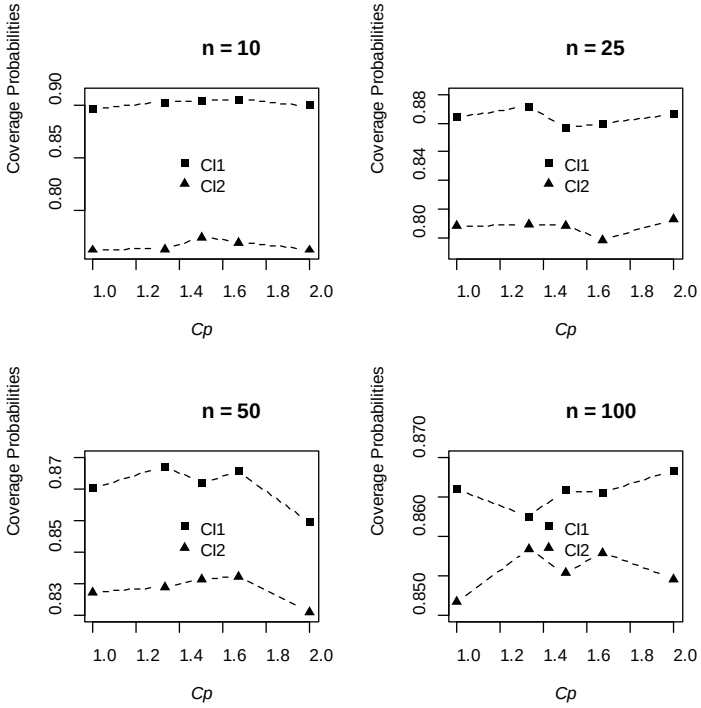

Figure 2: The estimated coverage probabilities of CI_1 and CI_2 for C_p in case of $Gamma(4,2)+48$

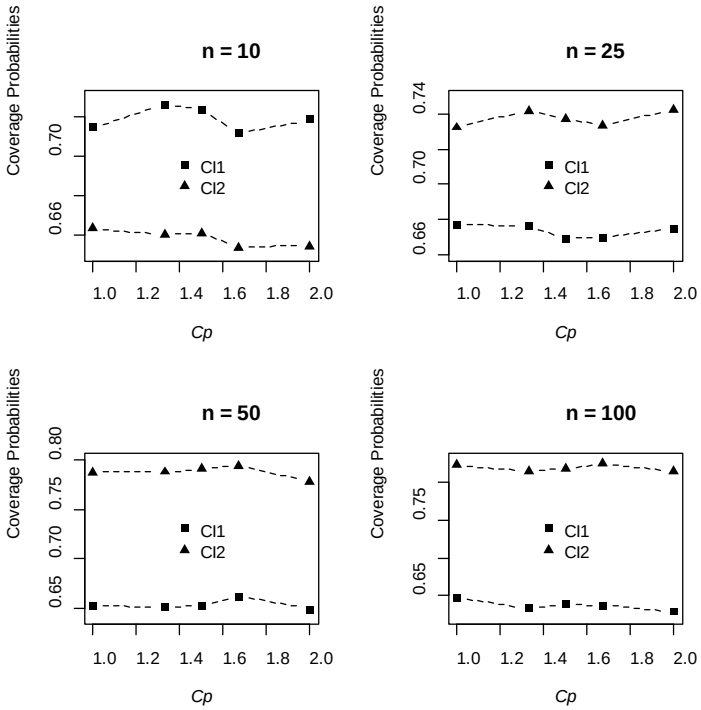

Figure 3: The estimated coverage probabilities of CI_1 and CI_2 for C_p in case of $Gamma(0.75,0.867)+49.1340$

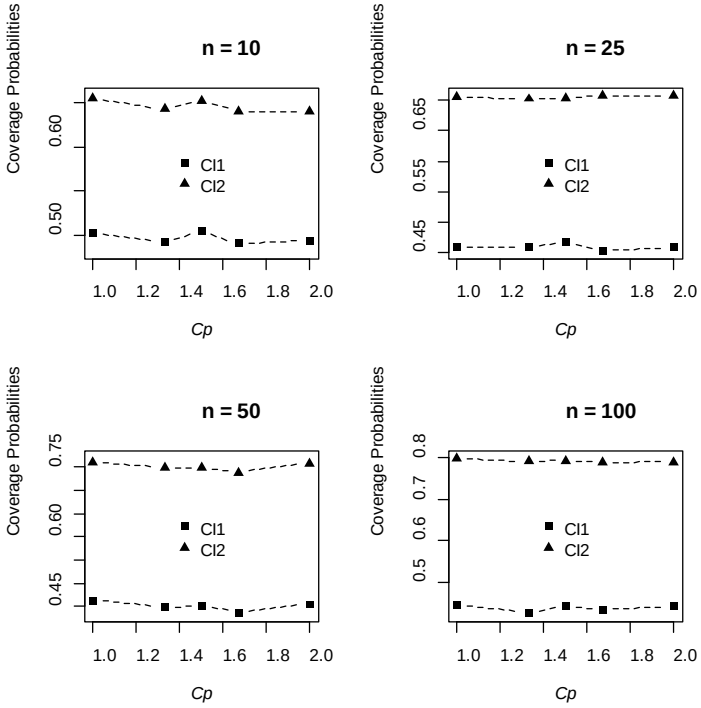

Figure 4: The estimated coverage probabilities of CI_1 and CI_2 for C_p in case of $Gamma(0.25, 0.5) + 49.5$

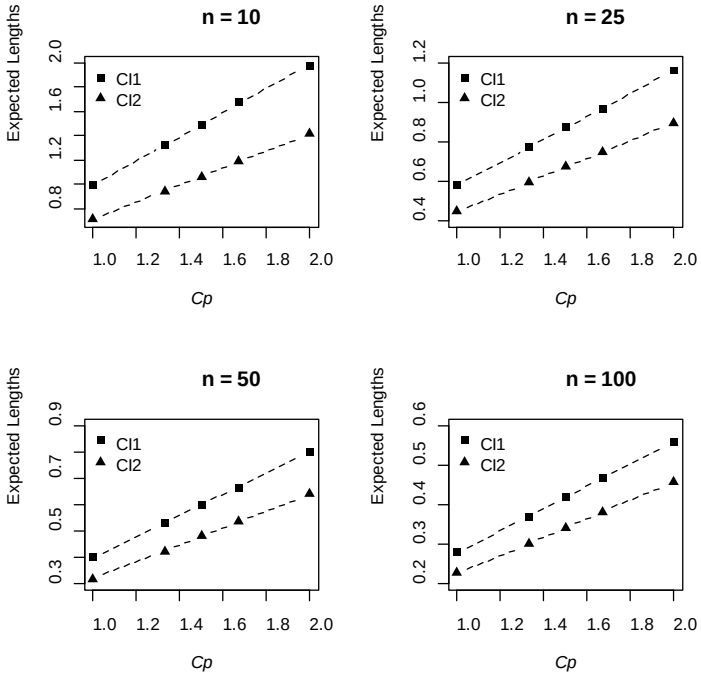

Figure 5: The expected lengths of CI_1 and CI_2 for C_p in case of $N(50,1)$

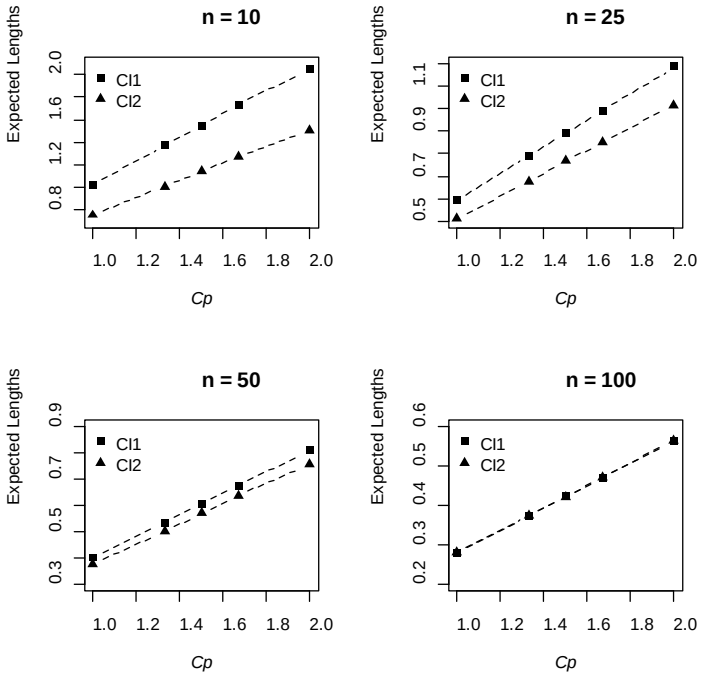

Figure 6: The expected lengths of CI_1 and CI_2 for C_p in case of $Gamma(4,2)+48$

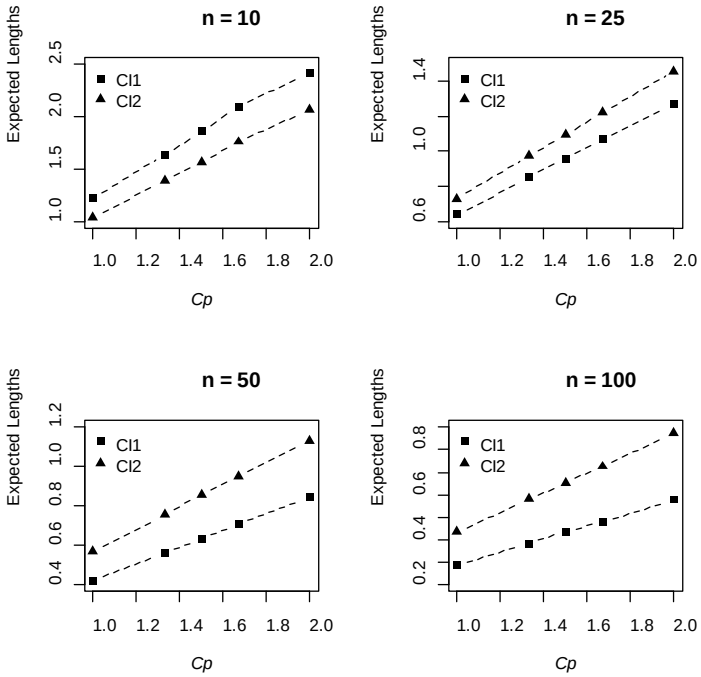

Figure 7: The expected lengths of CI_1 and CI_2 for C_p in case of $Gamma(0.75,0.867)+49.1340$

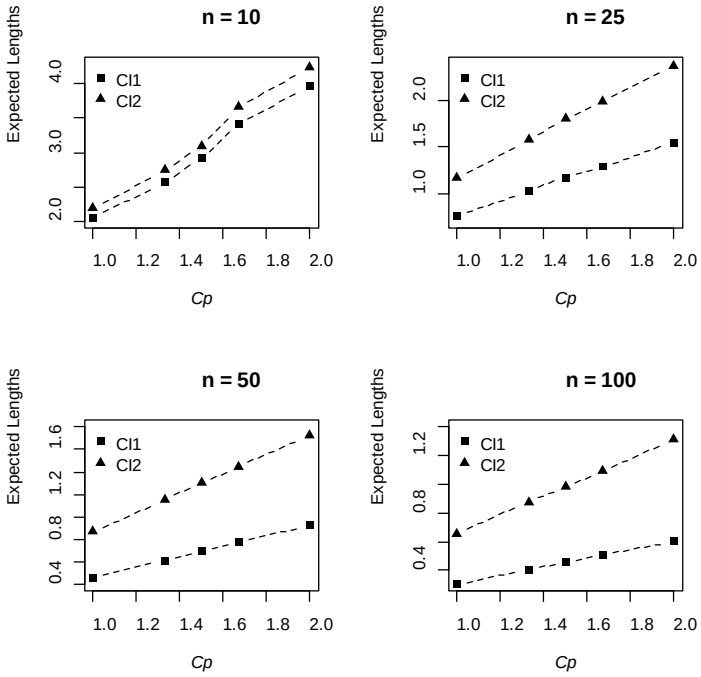

Figure 8: The expected lengths of CI_1 and CI_2 for C_p in case of $Gamma(0.25, 0.5) + 49.5$

5 Conclusions

The existing confidence interval for the capability index C_p based on the confidence interval for the standard deviation was based on a normal distribution. However, the underlying distribution may be non-normal or skewed in some circumstances. A confidence interval for the capability index C_p based on the bootstrap-*t* confidence interval for the standard deviation was developed. The proposed confidence intervals were compared with the existing confidence interval through a Monte Carlo simulation study. The proposed confidence interval proved to be better than the existing confidence interval in terms of the coverage probability when the data have a coefficient of skewness > 2. On the other hand, when the data are symmetrical or have a coefficient of skewness ≤ 2, the estimated coverage probability of the existing confidence interval can be close to the nominal level.

Appendix: Source R code for all confidence intervals

```
CI1 <- function (x,LSL,USL,alpha)
{
        n <- length(x)
        S <- sd(x)
        chisq1 <- qchisq(alpha/2,df=n-1)
        chisq2 <- qchisq(1-alpha/2,df=n-1)
        K <- (USL-LSL)/(6*S)
        ci.low <- K*sqrt(chisq1/(n-1))
        ci.up  <- K*sqrt(chisq2/(n-1))
        out <- cbind(ci.low,ci.up)
        return(out)
}

CI2 <- function (x,LSL,USL,alpha)
{
        n  <- length(x)
        s2 <- var(x)
        percentile.T.S <- percentile.T.star(x,alpha)
        T1 <- percentile.T.S[1]
        T2 <- percentile.T.S[2]
        K1 <- (USL-LSL)/6
        K2 <- s2*sqrt(2*(n-1))
        ci.low <- K1*(K2/(2*T1+sqrt(2*(n-1))))^(-1/2)
        ci.up  <- K1*(K2/(2*T2+sqrt(2*(n-1))))^(-1/2)
        out <- cbind(ci.low,ci.up)
        return(out)
}

percentile.T.star <- function (x,alpha)
{
        B  <- 1000
        n  <- length(x)
        S2 <- var(x)
        T.star <- numeric(B)
                for (i in 1:B){
                        xs <- sample(x,n,replace=TRUE)
                        s2.star <- var(xs)
                        T.star[i] <- sqrt((n-1)/2)*((s2.star/S2)-1)
                }
        T1 <- quantile(T.star,probs=alpha)
        T2 <- quantile(T.star,probs=1-alpha)
        out <- cbind(T1,T2)
        return(out)
}
```

Acknowledgements

The author would like to thank the anonymous referees for their helpful comments, which resulted in an improved paper. The author is also thankful for the support in the form of the research funds awarded by Thammasat University.

References

[1] Bittanti, S., Lovera, M. and Moiraghi, L. (1998): Application of non-normal process capability indices to semiconductor quality control. *IEEE Transactions on Semiconductor Manufacturing*, **11**, 296-303.

[2] Casella, G. and Berger, R.L. (2001): *Statistical Inference*. Duxbury Press: Pacific Grove.

[3] Chen, K.S. and Pearn, W.L. (1997): An application of non-normal process capability indices. *Quality and Reliability Engineering International*, **13**, 335-360.

[4] Cojbasic, V. and Tomovic, A. (2007): Nonparametric confidence intervals for population variances of one sample and the difference of variances of two samples. *Computational Statistics & Data Analysis*, **51**, 5562-5578.

[5] Ding, J. (2004): A model of estimating process capability index from the first four moments of non-normal data. *Quality and Reliability Engineering International*, **20**, 787-805.

[6] Efron, B. (1979): Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

[7] Efron, B. and Tibshirani, R.J. (1993): *An Introduction to the Bootstrap*. Chapman & Hall: New York.

[8] Ihaka, R. and Gentleman, R. (1996): R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.

[9] Kane, V.E. (1986): Process Capability Indices. *Journal of Quality Technology*, **18**, 41-52.

[10] Kotz, S. and Johnson, N.L. (1993): *Process Capability Indices*. London: Chapman & Hall.

[11] Kotz, S. and Lovelace, C.R. (1998): *Process Capability Indices in Theory and Practice*. Arnold: London.

[12] Pearn, W.L. and Kotz, S. (2006): *Encyclopedia and Handbook of Process Capability Indices: A Comprehensive Exposition of Quality Control Measures*. Singapore: World Scientific.

[13] Smithson, M. (2001): Correct confidence intervals for various regression effect sizes and parameters: the importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, **61**, 605-632.

[14] Steiger, J.H. (2004): Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, **9**, 164-182.

[15] Thompson, B. (2002): What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher*, **31**, 25-32.

[16] Tosasukul, J., Budsaba, K. and Volodin, A. (2009): Dependent bootstrap confidence intervals for a population mean. *Thailand Statistician*, **7**, 43-51.

[17] Wu, H.-H., Swain, J.J., Farrington, P.A., and Messimer, S.L. (1999): A weighted variance capability index for general non-normal processes. *Quality and Reliability Engineering International*, **15**, 397-402.

[18] Zhang, J. (2010): *Conditional confidence intervals of process capability indices following rejection of preliminary tests*. Ph.D. Thesis, The University of Texas at Arlington, USA.

Metodološki zvezki, Vol. 11, No. 2, 2014, 93-106

Investments in Renewable Energy Sources in OPEC Members: a Dynamic Panel Approach

Antonio A. Romano[1] and Giuseppe Scandurra[2]

Abstract

In this paper we analyze the key factors promoting the investments in renewable energy sources in a panel dataset of Petroleum Exporting Countries (OPEC) members. To address these issues, a dynamic panel analysis of renewable investments in the sample of OPEC with distinct economic and social structures, in the years between 1980 and 2009, is proposed. Results confirm that key factors promoting investments in renewable energy sources are similar to other studies which include more developed countries. However, lack of grants and/or incentives to promote the installations of new renewable power plants is a limit for the future and sustainable development of these countries.

1 Introduction

Renewable Energy Sources (RES) are becoming increasingly important in the energy mix of countries, because of their ability to limit the environmental impact of energy production and counter the gradual appreciation of the raw materials used in the process of traditional generation based on gas and / or oil power plants.

The centrality represented by investments in renewable sources is confirmed by the attention by the international scientific community in recent years. Sadorsky (2009) studied the relationship between renewable energy sources (wind, solar and geothermal power, wood and wastes) and economic growth in a panel framework of 18 emerging economies for the period 1994-2003 and found that increases in real GDP had a positive and statistically significant effect on renewable energy consumption per capita. Wolde-Rufael (2012) analyzes the causal nexus between

[1] Department of Management Studies and Quantitative Methods, University of Naples "Parthenope", via Generale Parisi, 13, 80132 Napoli, Italy; antonio.romano@uniparthenope.it.
[2] Department of Management Studies and Quantitative Methods, University of Naples "Parthenope", via Generale Parisi, 13, 80132 Napoli, Italy; giuseppe.scandurra@uniparthenope.it.

nuclear consumption and GDP. Yuksel (2010) and Baris and Kucukali (2012) analyze RES deployment in Turkey and find that, thanks to the potential for renewable use, Turkey is working towards a clean and sustainable energy development. Menz and Vachon (2006) and Carley (2009) study the renewable investments in the USA, the former with a regression into countries and the latter using a panel regression. Marques et al. (2010) analyze the drivers promoting renewable energy in European countries and finds that lobbies of traditional energy source and CO_2 emission restrain renewable deployment. Evidently, the need for economic growth suggests an investment that supports, but does not replace, the before installed capacity. Romano and Scandurra (forthcoming-a) investigate the drivers of investments in Renewable sources in panel of OECD countries and including some development countries and the divergences in countries that produce electricity using or not using nuclear power plants while the same authors (forthcoming-b), in a forthcoming paper, explore the drivers promoting the investments in renewable energy sources and the divergences on the basis of development stage of the countries employing a large sample of 60 countries split into 3 different sub-samples, following the classification proposed by World Bank (low income and lower middle income; upper middle income; high income). Gan and Smith (2011) identify key factors that may have driven the differences in the shares of renewable energy in total primary energy supply among OECD countries for renewable energy in general and bioenergy in particular. Masini and Menichetti (2012) propose and test a conceptual model in order to analyze factors affecting the investor decisions and the relationship between the investments in RES and the portfolio performances.

The need to meet the demand for energy and environmental sensitivity leads policy makers to plan further investments in generation plants based on renewable sources. However, despite the exponential growth in the production of energy from renewable sources in recent years, yet most of the energy demand is met through the use of fossil fuels (IEA, 2012).

Currently there is great interest in development of RES due to the prospect of the all available of reserves of fossil fuel getting depleted and the environment pollution caused by burning of fossil fuel. However there are some disadvantages of using renewable energy. These are described below.

- Availability of fuel obtained from plants that can be used as economical energy practically is limited. Though lot of research and development activities is going on around to world to develop plants that could provide suitable fuels economically and in sufficient quantities.
- The total potential of renewable energy sources as wind power and tidal power is limited and/or intermittent.
- The current capital cost for equipment to convert renewable energy such as solar, wind and tide is very high.

- Plant for generating power from wind, and tides can be located only in places where suitable conditions of tide or wind exist.
- The plant for generating energy from sun light, wind and solar energy have to be spread around large areas.
- Solar power is dependent on availability of sunlight. Thus the availability of power fluctuates from zero to maximum every day.
- There have been some allegations that large scale use of wind power can interfere pattern of wind flow and disturb the set weather pattern. Use of hydro power is already known change the pattern of silting in rivers.

With this in mind, we analyze the drivers of investment in renewable energy sources in Petroleum Exporting Countries (OPEC). OPEC is a permanent, intergovernmental Organization, created on 1960 by Iran, Iraq, Kuwait, Saudi Arabia and Venezuela. The organization now has 12 members having since been joined by Algeria, Angola, Ecuador, Libya, Nigeria, Qatar and the United Arab Emirates. The objective is to co-ordinate and unifies petroleum policies among Member Countries in order to secure fair and stable prices for petroleum producers; an efficient, economic and regular supply of petroleum consuming nations; and a fair return on capital to those investing industry.

In this paper we analyze the determinants of investments in renewable sources (hydroelectric and other renewable sources) and the divergences in the composition of the energy mix of countries. In practice, we test the impact of key factors in renewables, highlighting the progressive adaptation to the changing energy needs. This paper addresses these issues by means of a dynamic panel analysis of the renewable investment in a sample of OPEC countries with distinct economic and social structures as well as different levels of economic development. The data are the annual time series from 1980 to 2009.

In the model proposed we include the main policy, environmental, socio – economic and generation factors. We use a dynamic specification of the equation that takes into account past investments in renewable energy sources. A widely used methodology for dynamic panel modeling applies Generalized Method of Moments (GMM) estimators proposed by Arellano and Bond (1991). In particular, we try to understand if RES significantly contribute to climate change and if OPEC characterized by a large availability of fossil fuel invests in RES.

The organization of the paper is as follows: Section 2 describes data; Section 3 we briefly explain the method proposed. Section 4 reports the model, the empirical results and discusses the policy implications. Section 5 concludes.

2 Data

The data used in this paper are from U.S. Energy Information Administration (EIA) and International Energy Agency (IEA) databases.
Following the literature (e.g. Carley, 2009; Marques and Fuinhas, 2011), the explanatory variables try to capture main socioeconomic, political and environmental factors from which investment decisions originate.
For the environmental factors we consider the per capita Carbon Dioxide Emissions (CO_2) from the Consumption of Energy. CO_2 emission is one of the main factors of the greenhouse gas (GHG) effects and it could be considered as a proxy of environmental degradation and not the only responsible. The expected results are estimates with a significant positive effect. The presence of a negative effect emphasizes the persistence of an economy tied to fossil fuels, which are still unable to replace the traditional energy sources. The last class of factors (Socioeconomic) includes per capita GDP, per capita Consumption of Energy and a proxy for the energy security of supply. The GDP is directly related to energy consumption (Sadorsky, 2009). The per capita Consumption of Electricity is considered a proxy for economic development of the country (e.g. Toklu, 2011) but it also represents the evolution of energy demand. The need to meet the energy demand can lead to the creation of new power plants based on RES, increasing investment. However, if the increasing demand is met through traditional power plants based on fossil fuel, then the effect on investment will be negative. A similar argument applies to energy security, approximated by the degree of dependence on foreign supplies of electricity. The need to increase their share of production (reducing the energy bill) and to reduce dependence could increase investment in RES. Considering the main production of the countries, we include also the annual oil extraction. The expected result is an estimate with a significant positive effect. The increasing in oil extraction can suggest to countries to increase the investment in RES.
Various forms of incentives are currently adopted and many of those directly affected by the wealth of countries, of which we have detailed information[3]. However, there is a lack of information about the availability of grant to promote the renewable in the OPEC countries. In particular, seems that these countries, at the best of our knowledge, do not provide any incentives for renewable investments. For this reason we do not include a policy variable. In order to reduce variability, GDP, EI, electricity consumption, oil supply and CO_2 are expressed through natural logarithm.
The analysis of data on generation sources (see Table 1) in the dataset considered (OPEC) highlight different patterns in the countries:

- Some countries do not have generation based on RES (Kuwait; Libya; Qatar, Saudi Arabia).
- Angola, Ecuador and Venezuela, generate most of their electricity from RES.

[3] For example, the European Commission with the Directive 2001/77/EC aim to promote the electricity produced from renewable energy sources.

- Iran and Nigeria generate an appreciable share of electricity from RES.

The United Arab Emirates have a small share of generation from RES, since 2009, when the first solar power plants were put into operation.
In the entire sample we observe, however, that the generation from RES is obtained almost entirely from hydroelectric plants.
Given the great availability of fossil fuels for the production of electrical energy, these countries have little considered the possibility of generation sources based on renewable.
Considering the generation share from RES in the countries included in our dataset, we reduce its sectional dimension, analyzing only countries that generate electricity from RES. In addition, Iraq has not been included due to missing data in the GDP series. The countries we have included in the final sample are: Algeria, Angola, Ecuador, Iran, Nigeria and Venezuela.

Table 1: Mean Electricity generation by sources and countries (1980 – 2009).

Countries	Share of total renewable power generation (%)	Share of renewable – not based on hydroelectric power plants (%)	Share of thermal power generation (%)
Algeria	1.88	0	98.22
Angola	65.60	0	34.40
Ecuador	64.70	0.54	35.30
Iran	11.86	0.01	88.14
Iraq	5.00	0	95.00
Kuwait	0	0	1
Libya	0	0	1
Nigeria	34.56	0	65.44
Qatar	0	0	1
Saudi Arabia	0	0	1
United Arab Emirates	0.99	0.01	99.00
Venezuela	64.39	0	35.61

Different ways to evaluate the development of RES are proposed in literature. Bird et al. (2005) measure the total amount of renewable energy produced while Marques et al. (2010) use the contribution of renewable to energy supply. Following Romano and Scandurra (forthcoming-a) we explain the investment in RES (ShRen) as the ratio between Renewable Generation and Total Net Electricity Generation. The share

of Renewable Electricity Net Generation can be considered a proxy of investments in RES.

3 Method

Dynamic panel data (DPD) models contain one or more lagged dependent variables, allowing for the modeling of a partial adjustment mechanism, i.e.:

$$y_{i,t} = \delta y_{i,t-1} + \mathbf{x}'_{\mathbf{it}}\beta + u_{i,t} \tag{3.1}$$

where for country i ($i=1,\ldots,N$) at time t ($t=1,\ldots,T$), δ is a scalar, $y_{i,t}$ is the outcome variable, $y_{i,t-1}$ is the lagged dependent variable, $\mathbf{x}'_{\mathbf{it}}$ is the vector of independent variables while the error term

$$u_{i,t} = \alpha_i + \tau_{i,t} \tag{3.2}$$

follows a one - way error component model where α_i denote a country – specific effect, $\tau_{i,t}$ denotes a observation – specific effect and $\alpha_i \sim IID(0, \sigma^2_\alpha)$ and $\tau_{i,t} \sim IID(0, \sigma^2_\tau)$.

The dynamic panel data regression described in (3.1) and (3.2) is characterized by two sources of persistence over time: autocorrelation due to the presence of a lagged dependent variable among the regressors and individual effects characterizing the heterogeneity among the individuals.

Several econometric problems may arise from estimating the parameters in eq. (3.1) (cf. Hsiao, 2003): *i*) the variables in x_{it} are assumed to be endogenous; *ii*) time-invariant country characteristics (fixed effects) may be correlated with the explanatory variables; *iii*) the presence of the lagged dependent variable $y_{i,t-1}$ gives rise to autocorrelation. With these assumptions, the estimations with fixed effects (OLS) or random effects (GLS) would not be appropriate since the obtained estimates would be biased.

Since $y_{i,t}$ is a function of α_i, it immediately follows that $y_{i,t-1}$ is also a function of α_i. Therefore, $y_{i,t-1}$, a right-hand regressor in (3.1), is correlated with the error term. This renders the OLS estimator biased and inconsistent even if $\tau_{i,t}$ are not serially correlated.

One way to solve this problem is to estimate a dynamic panel data model based on the Generalized Method of Moments (GMM) estimator proposed by Arellano and Bond (1991). The GMM procedure is more efficient than the Anderson and Hsiao (1982) estimator, while Ahn and Schmidt (1995) derived additional nonlinear moment restrictions not exploited by the Arellano and Bond (1991) GMM estimator. Arellano and Bond argue that the Anderson–Hsiao estimator, while consistent, fails to take all of the potential orthogonality conditions into account. A key aspect of the

method proposed by Arellano and Bond is the assumption that the necessary instruments are 'internal': that is, based on lagged values of the instrumented variable(s) (Baltagi, 2005). The estimators allow the inclusion of external instruments as well. For instance, let us consider a simple autoregressive model with no regressors:

$$y_{\mathrm{i,t}} = \delta y_{i,t-1} + u_{i,t} \tag{3.3}$$

where $u_{i,t} = \alpha_i + \tau_{i,t}$ with $\alpha_i \sim IID(0, \sigma^2{}_\alpha)$ and $\tau_{i,t} \sim IID(0, \sigma^2{}_\tau)$, independent of each other and among themselves.

In order to get a consistent estimate of δ as $N \to \infty$ with T fixed, we first difference (3.3) to eliminate the individual effects

$$\begin{aligned} \Delta y_{i,t} &= y_{i,t} - y_{i,t-1} = \delta\left(y_{i,t-1} - y_{i,t-2}\right) + \left(\tau_{i,t} - \tau_{i,t-1}\right) = \\ &= \delta \Delta y_{i,t-1} + \Delta \tau_{i,t} \qquad t = 3,\ldots, T \end{aligned} \tag{3.4}$$

and note that $\left(\tau_{i,t} - \tau_{i,t-1}\right)$ is MA(1) with unit root.

Equation (3.4) is equivalent to a system of simultaneous equations with (T-2) equations with N observations, or:

$$\begin{cases} \Delta y_{i3} = \delta \Delta y_{i2} + \Delta \tau_{i3} & instruments: y_{i1} \\ \Delta y_{i4} = \delta \Delta y_{i3} + \Delta \tau_{i4} & instruments: y_{i1}; y_{i2} \\ \vdots & \\ \Delta y_{i,T} = \delta \Delta y_{,iT-1} + \Delta \tau_{i,T} & instruments: y_{i1}; y_{i2}; \ldots; y_{i,T-2} \end{cases}$$

where the instruments are uncorrelated with the error terms.

The variance\covariance of the error term can be expressed in the following matrix:

$$V = E\left(\Delta\tau_i \Delta\tau_i'\right) = \sigma_\tau^2 \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{bmatrix}$$

is *(T-2)×(T-2)*, since $\left(\tau_{i,t} - \tau_{i,t-1}\right)$ is MA(1) with unit root. Define the $(T-2) \times C$ matrix,

$$Z_i = \begin{bmatrix} y_{i1} & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \cdots & 0 \\ 0 & y_{i1} & y_{i2} & 0 & 0 & 0 & \dots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & y_{i1} & y_{i2} & y_{i3} & \dots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & y_{i1} & y_{i2} & \dots & y_{iT-2} \end{bmatrix},$$

where $C = \sum_{j=1}^{T-2} j$ and lines contain the instruments.
Then, the $N(T-2) \times C$ matrix of instruments is $Z = [Z_1', \dots, Z_N']'$ and the moment equations described above are given by $E(Z_i \Delta\tau_{i3}) = 0$. Premultiplying the differenced equation (3.4) in vector form by Z', one gets

$$Z'\Delta y = Z'(\Delta y_{-1})\delta + Z'\Delta\tau \tag{3.5}$$

Performing GLS on (3.5) one gets the Arellano and Bond (1991) preliminary one-step consistent estimator:

$$\widehat{\delta_1} = [(\Delta y_{-1})'Z(Z'(I_n \otimes V)Z)^{-1}Z'(\Delta y_{-1})]^{-1}[(\Delta y_{-1})'Z(Z'(I_n \otimes V)Z)^{-1}Z'(\Delta y)] \tag{3.6}$$

One can gets the two-step Arellano and Bond (1991) GMM estimator by replacing the matrix of the second population moments with that of the corresponding second sample moments. For a more detailed discussion see e.g. Baltagi (2005).

4 Model and discussion

In this paper we employ a panel dataset including 6 OPEC countries from 1980 to 2009[4]. There are three main issues that can be solved using a panel dataset. In fact, a panel dataset allows us to have more degrees of freedom than with time-series or cross-sectional data, and to control for omitted variable bias and reduce the problem of multi-collinearity, hence improving the accuracy of parameter estimates (Hsiao, 2003), having more informative data. Furthermore, annual data avoids the seasonality problems. Since static regression models can suffer from a number of problems, including structural instability and spurious regression, we employ a

[4] Arellano and Bond's (1991) GMM estimator is consistent for large *N* (number of countries) with *T* fixed. In our empirical research, Initially, the current sample was broader and included all of the OPEC members. Considering that some of them do not have sources of generation based on renewable energy, or SHRen = 0 in he analysed years, we employ a subset of countries. The sectional component of the error remains in the variables and must thus refer to the wholeness of the sample. Furthermore, we tries to use only the most recent instruments (but also simple OLS estimation) but without sensible variations in the significance.

dynamic analysis that allows for slow adjustment. The dynamic model captures the "persistence effect" on investment in RES[5]. The assumed model is as follows:

$$ShRen_{i,t} = c + (1+\gamma)ShRen_{i,t-1} + \sum_{k=0}^{K} \varphi_{1k}\Delta lnGDP_{i,t-k} + \sum_{k=0}^{K} \varphi_{2k} lnOil_{i,t-k} + \sum_{k=0}^{K} \varphi_{3k} lnCO_{2;i,t-k} + \sum_{k=0}^{K} \varphi_{4k} lnEI_{i,t-k} + \sum_{k=0}^{K} \varphi_{5k} lnConsumption_{i,t-k} + u_{i,t}, \quad (4.1)$$

where for country i ($i = 1,\ldots, N=6$) at time t ($t = 1,\ldots, T=30$), $ShRen_{i,t}$ are the renewable investments, $\Delta\ln GDP_{i,t}$ is the first differences of natural logarithm of GDP per capita (growth of GDP per capita), $\ln EI_{i,t}$ is the natural logarithm of Energy intensity, $\ln Consumption_{i,t}$ is the natural logarithm of per capita electricity consumption, $\ln Oil_{i,t}$ is the natural logarithm of oil supply while u_{it} is the error component. We include also the natural logarithm of per capita carbon dioxide emission $lnCO_{2;i,t.}$ It is considered predetermined, or:

$$E(u_{i,s}|CO_{2;i,t}) \neq 0 \text{ where } s < t.$$

In fact, variation in carbon dioxide emissions are uncorrelated with past (and potentially current) investments, but will be correlated with future investments. Here, $lnCO_2$ is predetermined but not strictly exogenous.

The consistency of the estimation depends on whether lagged values of the endogenous and exogenous variables are valid instruments in our regression[6]. Also, this methodology assumes that there is no second-order autocorrelation in the errors, therefore a test for the previous hypotheses is needed.

In this model we take into account the full electricity generation mix. In fact, the remaining part, not included in the model, is all ascribable to fossil fuel. We employ the robust one-step GMM estimator.

The consistency of the estimations is assessed applying a set of tests (Table 2). The Wald test fails to accept the null hypothesis that all the coefficients except the constant are zero. In order to obtain consistent GMM estimates the assumption of no serial correlation in the residual in levels is essential. The presence of first order autocorrelation in the difference residuals does not imply the estimates are inconsistent, but the presence of second order autocorrelation would imply that the

[5] In the growth of investments, persistence may reflect the existence of a long term relationship as conduits of knowledge helping countries to continuously upgrade and maintain their generation capacity.

[6] We estimate two version of the model, obtaining similar standard errors. In the former, we include all the instruments while in the latter we consider only the most recent.

estimates are inconsistent (Arellano and Bond, 1991- pp. 281-282). The test statistic satisfies the specification requirements. In eq. (4.1) we assume that there is a first order autocorrelation present for the observed responses. Moreover, we fail to reject the null hypothesis of no second order autocorrelation in all specifications. Having annual data, we also report AR(3) and AR(4) autocorrelation test. Both tests accept the null hypothesis[7].

Table 2: Parameter estimates and test statistics

Variable	Estimates
$ShRen_{(-1)}$	0.77***
$lnCO_2$	-0.1***
$lnCO_{2(-1)}$	0.10***
$\Delta lnGDP$	0.17***
$lnEI$	0.06***
$lnConsumption$	-0.06***
$lnOil$	-0.02***
$Constant$	-0.63***
Test Statistics	
Wald test	675.31***
1st order autocorrelation	-2.05**
2nd order autocorrelation	-0.79
3rd order autocorrelation	1.12
4th order autocorrelation	-0.61

Significance levels: ***: 1%; **: 5%

The estimation results for eq. (4.1) are in Table 2.

The result of the estimations shows that GDP, energy efficiency, per capita electricity consumption and oil supply are significant. Almost all coefficients also show the expected signs. Only the CO_2 emission, which is traditionally seen as directly linked to investments in renewable energy, and the electricity consumption have a negative sign. Furthermore, the share of renewable presents a significant and positive coefficient. Obviously, the investments made over the years are to increase the share of energy produced from renewable sources.

The GDP growth is significant in the sample, and it has a positive sign. This expected result, suggests the progressive increasing of the living condition of the population give to these countries the opportunity to increase the investments in RES.

[7] Sargan test for the validity of the instruments is not reported in Table 2 because we employ a one-step GMM robust estimator. Arellano and Bond (1991) recommend using the one step results for inference on the coefficients and using two – step Sargan test for inference on model specification. In our model, two-step Sargan test supports the assumption that model is correctly specified (χ^2=129.56).

Evidently, GDP grew at a faster average rate than investments in renewable energy sources. This result is also encouraged by the consistency with energy efficiency.

The per capita electricity consumption depresses investment in RES. This result is unexpected. In fact, main idea suggest that need to meet the increasing electricity consumption is to invest in new power plants based on renewable sources. This is supported by the cost of raw materials for thermic power plants which in the recent years have increased. However, considering the nature of the countries, we observe that the dynamics of production and the energy demand has led the system to find an equilibrium using more traditional sources and with a little attention to energy efficiency. The high availability of fossil fuel suggests to satisfy the increasing consumption with thermic power plants.

The CO_2 emissions are significant and show a negative sign in level and a positive sign at lag 1. The combined effect is still negative (-0.015). An increasing in carbon emissions depresses the investments in RES. This is partly unexpected even if this phenomenon has been repeatedly highlighted in the literature (e.g. Marques et al, 2010; Romano and Scandurra, forthcoming-a), especially when rich countries are analyzed. It portends an energy production system more advanced but still tied to traditional sources that compress the dynamics of development of RES. We remember, however, that these countries have no CO_2 emission targets.

The coefficient for the oil supply is also significant and presents a positive sign. Increasing in oil extraction encourages the investment in renewable energy, and the positive effect prevails.

The amount of energy required for the production of a unit of GDP is in line with the expected results. This result confirms that the technological progress increase the investment in RES. Energy efficiency offers a powerful and cost-effective tool for achieving a sustainable energy future. Improvements in energy efficiency can reduce the need for investment in energy infrastructure, cut fuel costs, increase competitiveness and improve consumer welfare. Environmental benefits can also be achieved by the reduction of GHG emissions.

There are many similarities among the key factors in investments in OPEC countries and other countries. Comparing the results with other studies we observe that the decisions depend by the diversification of the energy mix.

The environmental aspect is primary aspect and the estimates have revealed as CO_2 emissions depress investments. This aspect is robust with most of the literature, where the effect is often negative because of the mix of generation based mainly on fossil fuels (e.g. Marques et al, 2010; Romano and Scandurra, forthcoming-a). The breakdown by source of generation allows, however, assessing the impact of emissions on investment and ensuring that it depends directly from the sources themselves.

Stable with the literature (e.g. Romano and Scandurra, forthcoming-a) is the sign of the GDP. Basic idea is that larger income allows countries to handle the costs of developing the RES. The positive effect of income in the investments in RES, yet

verified by Menz and Vachon (2006) and Marques et al. (2010) is confirmed also for OPEC.

5 Summary and conclusions

This paper analyzes the driving of investment in RES in a sample of OPEC members. In the model proposed we include environmental and socio – economic determinants identified by literature (Carley, 2009; Marques et al, 2010; Romano and Scandurra, forthcoming-a), through a dynamic panel regression that takes into account past investments in renewable energy sources.

Results suggest that these countries invest in renewable sources but their use is conditioned by the orography of territory. In general, these countries have invested in RES only in the recent years and, at this moment, their use is limited and the investments are not relevant. Furthermore, there are not policies promoted by Government in order to stimulate the investments in RES and this could be a point that depresses their use. As previously demonstrated, policies to support investment in renewable energy sources have positive and significant coefficients and promote the growth in generation capacity. In fact, renewable power generation policies remain the most common type of support policy. The Feed-in-tariffs (FITs) and/or renewable portfolio standards (RPS) are the most commonly used policies in this sector and many countries adopt this policies in order to promote the investments in RES. Probably, OPEC members have to adopt some grants to ensure a rapid development of generation based on renewable power plant. Lack of policy grants and/or incentives in order to promote the investments in RES is a criticism for the future. It does not stimulate the renewable power generation and could be a limit for a sustainable future.

There has been little linking of energy efficiency and renewable energy in the policy arena to date, but countries are beginning to wake up to the importance of tapping their potential synergies. We think that enhanced scientific and engineering knowledge should lead to performance improvements and cost reductions in RE technologies. Knowledge about RE and its climate change mitigation potential continues to advance. The existing scientific knowledge is significant and can facilitate the decision-making process. Under most conditions, increasing the share of renewable sources in the energy mix will require policies to stimulate changes in the energy system.

Acknowledgements

The authors wish to thank the editor and two anonymous reviewers for detailed comments and suggestions. The usual disclaimer applies.

References

[1] Ahn, S.C., Schmidt, P. (1995): Efficient estimation of models for dynamic panel data, *Journal of Econometrics*, **68**, 5–27.

[2] Anderson, T.W., Hsiao, C. (1982): Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, **18**, 47–82.

[3] Arellano, M., Bond, S. (1991): Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, **58**, 277-297.

[4] Baltagi, B. H. (2005): *Econometric Analysis of Panel Data 3rd Ed.*. Chichester: John Wiley & Sons Ltd.

[5] Baris, K., Kucukali, S. (2012): Availability of renewable energy sources in Turkey: Current situation, potential, government policies and the EU perspective. *Energy Policy*, **42**, 377–391.

[6] Bird, L., Bolinger, M., Gagliano, T., Wiser, R., Brown, M., Parsons, B. (2005): Policies and market factors driving wind power development in the United States. *Energy Policy*, **33**, 1397 – 1407.

[7] Carley, S. (2009): State renewable energy electricity policies: an empirical evaluation of effectiveness. *Energy Policy*, **37**, 3071 – 3081.

[8] Gan, J., Smith, C.T. (2011): Drivers for renewable energy: A comparison among OECD countries. *Biomass and Bioenergy*, **35**, 4497 – 4503..

[9] Hsiao, C. (2003): *Analysis of Panel Data, 2nd edition.* Cambridge*:* Cambridge University Press.

[10] IEA, 2012. *World Energy Outlook, Executive Summary*, IEA.

[11] Marques, A.C., Fuinhas, J.A. (2011): Do energy efficiency measures promote the use of renewable sources? *Environmental sciences & policy*, **14**, 471 – 481.

[12] Marques, A.C., Fuinhas, J.A., Pires Manso, J. R. (2010): Motivations driving renewable energy in European countries: a panel data approach. *Energy Policy*, **38**, 6877 – 6885.

[13] Masini, A., Menichetti, E. (2012): The impact of behavioural factors in the renewable energy investment decision making process Conceptual framework and empirical findings. *Energy Policy*, **40**, 28 – 38.

[14] Menz, F., Vachon, S. (2006): The role of social, political and economic interests in promoting state green electricity policies. *Environmental Science and Policy*, **9**, 652-662.

[15] Romano, A.A., and Scandurra, G. (forthcoming-a): "Nuclear" And "Non Nuclear" Countries: Divergences on Investment Decisions in Renewable Energy Sources. *Energy Sources, Part B*: *Economics, Planning, and Policy*. *Doi: 10.1080/15567249.2012.714843*

[16] Romano, A.A., and Scandurra, G. (forthcoming-b): Investments in Renewable

Energy Sources in Countries Grouped by Income Level. *Energy Sources, Part B*: *Economics, Planning, and Policy. Doi: 10.1080/15567249.2013.834006*

[17] Sadorsky, P. (2009): Renewable energy consumption and income in emerging economies. *Energy Policy*, **37**, 4021-4028.

[18] Toklu, E., Guney, M.S., Isik, M., Comakh, O., Kaygusuz, K. (2010): Energy Production, consumption, policies and recent developments in Turkey. *Renewable and Sustainable Energy Reviews,* **14**, 1172 – 1186.

[19] Wolde-Rufael, Y. (2012): Nuclear Energy consumption in Taiwan. *Energy Sources, Part B:* : *Economics, Planning, and Policy*, **7**, 21 – 27.

[20] Yuksel, I. (2010): As a renewable energy hydropower for sustainable development in Turkey. *Renewable and Sustainable Energy Reviews*, **14**, 3213**–** 3219.

Appendix

All of the data analyses were done using *xtabond* procedure implemented in Stata ver. 11. Data employed are freely available from U.S. Energy Information Administration (http://www.eia.gov) and International Energy Agency (http://www.iea.org).

Metodološki zvezki, Vol. 11, No. 2, 2014, 107-120

A Distance Based Measure of Data Quality

Pavol Kráľ[1], Lukáš Sobíšek[2], Mária Stachová[3]

Abstract

Data quality can be seen as a very important factor for the validity of information extracted from data sets using statistical or data mining procedures. In the paper we propose a description of data quality allowing us to characterize data quality of the whole data set, as well as data quality of particular variables and individual cases. On the basis of the proposed description, we define a distance based measure of data quality for individual cases as a distance of the cases from the ideal one. Such a measure can be used as additional information for preparation of a training data set, fitting models, decision making based on results of analyses etc. It can be utilized in different ways ranging from a simple weighting function to belief functions.

1 Introduction

According to Cox (Cox 1972) "issues of data quality and relevance, while underemphasized in the theoretical statistical and econometric literature, are certainly of great concern in much statistical work". Nevertheless, data quality issues are mostly discussed in connection to data collection, data storage and data extraction and preparation processes, not statistical and data mining procedures themselves. In the presented paper we focus on data quality as a possible input for further data analysis and/or decision making based on results of this analysis. The main goal is to propose a simple and easily applicable measure for data quality. In our opinion, such a measure should aggregate various aspects of data quality, for example completeness, uncertainty, imprecision etc. (Berti-Equille 2007, Parsons 1996). Assuming that each aspect of data quality for a particular data entry can be assessed by a single number from the unit interval, data quality of a particular variable can be expressed by a corresponding n-tuple of mappings where each mapping maps values of a variable recorded in a data set into the unit interval. Data quality of a particular case then aggregates data quality of all corresponding variables in the form of a family of n-tuples. If we use the above mentioned data quality description, it allows us to represent data quality of a case as a distance from the ideal case, i.e. the case without any data imperfection. In the rest of our paper we call such a distance Data Quality Index and denote it DQI. The DQI can be used as prior information for further modelling (classification,

[1] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica; pavol.kral@umb.sk

[2] University of Economics, Prague W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic; lukas.sobisek@vse.cz

[3] Department of Quantitative Methods and Information Systems, Faculty of Economics, Matej Bel University, Tajovského 10, 975 90 Banská Bystrica; maria.stachova@umb.sk

clustering etc.), e.g. in the form of weights for particular cases. Instead of using DQI as a direct input for our analyses, we can use it as a source for measuring data quality of the whole data set. Data quality (reliability, validity) of the whole data set can be then used as a supplement to decision making based on results of statistical analysis.

The paper is organized as follows. In Section 2 we review data quality issues discussed in literature. Section 3 forms the main part of the paper: first we propose data quality description of individual variables and statistical units, then, on the basis of this representation, we construct DQI, a simple real valued measure of data quality for statistical units. Finally, in Section 4 we apply the proposed distance based data quality measure to a real data set.

2 Data quality

Data quality is a term with very broad meaning. In (Berti-Equille 2007) the author presents the following main data quality issues: duplicate and redundant data, imperfect data with low accuracy, missing values and incomplete databases and stale, i.e. non-fresh data. It is obvious that importance of these particular aspects of data quality depends on the problem we are trying to solve, what are our goals, what methods we intend to use, whether a particular data issue can be solved etc. For example, duplicate and redundant data can be effectively handled by fusion or deletion of records in the process of data extraction from a warehouse and many authors described how to deal with missing values and incomplete data in the past (see Imielinski and Lipski 1984, Grahne 2002 and Naumann, Leser, Freytag 1999). Contrary, in the case of two remaining families of data quality issues, data freshness and data accuracy, it is quite impossible to deal with them prior to the assumed analysis. Therefore we focus on them in the rest of our paper.

2.1 Data Freshness

Segev and Fang in (Segev, Fang 1990); Theodoratos and Bouzeghoub in (Theodoratos, Bouzeghoub 1999) use the traditional freshness definition called currency. It takes into account the difference between Query Time[1] and Extraction Time[2]. Another notion of freshness, called timeliness, describes the ageing of data. It describes how often data changes, it means it takes into account the difference between Query Time and Last Update Time[3] (Naumann, Freytag, Leser 2004).

The freshness factors and their corresponding metrics, summarized in (Peralta 2006), are listed in Table 1.

The relevance of data freshness factors and metrics from the point of view of statistical analysis depends on goals of analysis. For example, it is more relevant for frequent basic reporting than for supervised learning.

[1] Query Time is the instant time, when users retrieve data.
[2] Extraction Time refers to the starting time, when extracted data is used.
[3] Last Update Time corresponds to the time, when data was last updated.

Table 1: Summary of freshness factors and metrics.

Factor	Metric	Description
Currency	Currency	The time elapsed since data was extracted from the source (the difference between the delivery time and extraction time).
	Obsolescence	The number of updates operations to a source since the extraction time.
	Freshness ratio	The percentage of tuples in the view that are up-to-date.
Timeliness	Timeliness	The time elapsed from the last update to a source (the difference between the delivery time and last update time).

2.2 Data accuracy

Data accuracy plays a key role in data quality studies. Data with low accuracy can be defined as imperfect data. This is a very broad term further characterized by Parsons (Parsons 1996). Parsons compiles earlier works of Bonnissone and Tong (Bonnissone, Tong 1985), Bosc and Prade (Bosc, Prade 1993), and splits imperfect data into five separate parts, namely incomplete information, uncertainty, imprecision, vagueness and inconsistency. Moreover, Parsons specifies the above mentioned terms, describes their sources and offers solutions how to deal with these issues.
In our analysis we focus on uncertainty in data. Motro (Motro 1993) claims "Uncertainty permeates our understanding of the real world. The purpose of information systems is to model the real world. Hence information systems must be able to deal with uncertainty." If a system provides poor data to data users (analysts, researchers), they must incorporate uncertainty into their modelling strategies.

It is obvious that this factor cannot be easily exactly defined. It is strictly context dependent and has to be evaluated with respect to the analyzed problem. It can include expert information and intuitive approach based on users' (analysts) expectations and combine them with exact statistical techniques (e.g. clustering, classification, regression,...).

3 Data quality description and a distance based data quality measure

As it was mentioned in the previous chapter, data quality attributes are often context dependent. In our opinion, regardless the problem we are trying to solve, data quality can be viewed from the three different perspectives:

1. data quality of variables,
2. data quality of particular cases (statistical units),
3. data quality of a data set.

We can evaluate different data quality attributes (uncertainty, freshness, missingness etc.) from a local or global point of view. The global view means that we are able to decide whether the examined variable or statistical unit is appropriate for our analysis, e.g. we can remove variables and statistical units with high missingness or penalize variables with high uncertainty. The local view means that we are interested in data quality of a variable (a statistical unit) for a particular statistical unit (a particular variable), e.g. data entries for a particular case were made just a moment before a data set extraction, therefore freshness of this variable for that particular case is very good. The local view can be used to decide if a statistical unit would be used for our analysis unchanged, penalized or boosted. Obviously, if we aggregate local data quality of a variable for all available statistical units, we get global data quality of this variable. Analogously, if we aggregate local data quality of all variables for a statistical unit, we get global data quality of this statistical unit. On the other hand, we are often able to assess global data quality of a variable (a statistical unit) without aggregating its local data quality for statistical units (variables).

In the rest of our paper we assume for simplicity that we work with data sets already prepared for analysis, i.e. variables with the high number of missing values were already removed, duplicity in data entries was resolved etc. Moreover, we do not deal with variables with obvious 100 % data quality, i.e. variables without uncertainty, irrelevant freshness etc. Gender is an example of such a variable. It means that we focus primarily on the local view of data quality.

3.1 Data quality description

The basic element of our data quality description is formed by the definition of data quality of a variable with respect to a chosen data quality attribute (freshness, uncertainty, etc.) or a set of attributes, and a particular data set.

Definition 1. Let X denotes an observed variable, A denotes an attribute of data quality and $\mathcal{C}$ denotes a set of statistical units. Then the data quality of X with respect to A and $\mathcal{C}$ is a mapping $D_{X,A,\mathcal{C}} : \mathrm{ran}(X) \times \mathcal{C} \to [0,1]$, where $\mathrm{ran}(X)$ denotes the range of the variable X. If $\mathrm{ran}(D_{X,A,\mathcal{C}}) = \{1\}$, the variable X has 100 % data quality with respect to the attribute A and the set $\mathcal{C}$. If $\mathrm{range}(D_{X,A,\mathcal{C}}) = \{0\}$, the variable X has 0 % data quality with respect to the attribute A and the set of statistical units $\mathcal{C}$.

Clearly, even if the variable X takes the same value for two statistical units $c, c' \in \mathcal{C}$, the mapping $D_{X,A,\mathcal{C}}$ can take completely different values. Definition 1 can be generalized to a set of attributes in the following way.

Definition 2. Let X be an observed variable, $\mathcal{A} = \{A_1, A_2, \ldots, A_p\}$ be a set of data quality attributes and $\mathcal{C}$ be a set of statistical units. Then the data quality of X with respect to $\mathcal{A}$ and $\mathcal{C}$ is a p-tuple

$$\left(D_{X,A_1,\mathcal{C}}, D_{X,A_2,\mathcal{C}}, \ldots, D_{X,A_p,\mathcal{C}}\right), \tag{3.1}$$

where $D_{X,A_i,\mathcal{C}}$ denotes the data quality of X with respect to the attribute A_i and the set of statistical units $\mathcal{C}$.

Using the data description of variables from Definition 1 and 2 we can characterize data quality of a particular case (a statistical unit) $c \in \mathcal{C}$ with respect to a variable X and a set of attributes $\mathcal{A}$.

Definition 3. Let $\mathcal{C}$ be a set of statistical units, X be a variable and $\mathcal{A} = \{A_1, A_2, \ldots, A_p\}$ be a set of data quality attributes. Then the data quality of a statistical unit $c \in \mathcal{C}$ with respect to X, $\mathcal{A}$ and $\mathcal{C}$ is a p-tuple defined as follows

$$\left(D_{X,A_1,\mathcal{C}}(x,c), D_{X,A_2,\mathcal{C}}(x,c), \ldots, D_{X,A_p,\mathcal{C}}(x,c)\right), \tag{3.2}$$

where x is a value of X measured on c.

For simplicity, we denote the p-tuple (3.2) by $D_{X,\mathcal{A},\mathcal{C}}$.

The previous definition can be straightforwardly extended to a set of variables as follows.

Definition 4. Let $\mathcal{C}$ be a set of statistical units, $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ be a set of variables and $\mathcal{A} = \{A_1, A_2, \ldots, A_p\}$ be a set of data quality attributes. Then the data quality of a statistical unit $c \in \mathcal{C}$ with respect to $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{C}$ is an m-tuple

$$\left(D_{X_1,\mathcal{A},\mathcal{C}}, D_{X_2,\mathcal{A},\mathcal{C}}, \ldots, D_{X_m,\mathcal{A},\mathcal{C}}\right). \tag{3.3}$$

In the following example we illustrate Definitions 1-4.

Example 1. Let us assume $\mathcal{C} = \{c_1, c_2, c_3\}$, $\mathcal{X} = \{X_1, X_2\}$ and $\mathcal{A} = \{A_1, A_2\}$. Moreover, let $X_1(c_1) = x_{11}$, $X_1(c_2) = x_{11}$, $X_1(c_3) = x_{13}$, $X_2(c_1) = X_2(c_2) = X_2(c_3) = x_{21}$. Then, according to Definition 1, the mappings

$$D_{X_1,A_1,\mathcal{C}} = \begin{cases} 0.5 & \text{for } (x_{11}, c_1), \\ 0.4 & \text{for } (x_{11}, c_2), \\ 0.7 & \text{for } (x_{13}, c_3), \\ 0 & \text{elsewhere,} \end{cases}, \quad D_{X_1,A_2,\mathcal{C}} = \begin{cases} 0.2 & \text{for } (x_{11}, c_1), \\ 0.8 & \text{for } (x_{11}, c_2), \\ 0.6 & \text{for } (x_{13}, c_3), \\ 0 & \text{elsewhere,} \end{cases}$$

$$D_{X_2,A_1,\mathcal{C}} = \begin{cases} 0.3 & \text{for } (x_{21}, c_1), \\ 0.5 & \text{for } (x_{21}, c_2), \\ 0.4 & \text{for } (x_{21}, c_3), \\ 0 & \text{elsewhere,} \end{cases}, \quad D_{X_2,A_2,\mathcal{C}} = \begin{cases} 0.1 & \text{for } (x_{21}, c_1), \\ 0.2 & \text{for } (x_{21}, c_2), \\ 0.9 & \text{for } (x_{21}, c_3), \\ 0 & \text{elsewhere,} \end{cases}$$

are examples of data quality of X_1 and X_2 with respect to A_1, C and A_2, C. Applying $D_{X_1,A_1,\mathcal{C}}$, $D_{X_1,A_2,\mathcal{C}}$, $D_{X_2,A_1,\mathcal{C}}$, $D_{X_2,A_2,\mathcal{C}}$ we get the following data quality of X_1 and X_2 with respect to $\mathcal{A}$ and $\mathcal{C}$ (see Definition 2):

$$\left(D_{X_1,A_1,\mathcal{C}}, D_{X_1,A_2,\mathcal{C}}\right) \text{ and } \left(D_{X_2,A_1,\mathcal{C}}, D_{X_2,A_2,\mathcal{C}}\right).$$

Then, following Definition 3, for data quality of the statistical unit c_1 it holds, with respect to X_1, $\mathcal{A}$ and $\mathcal{C}$,

$$\left(D_{X_1,A_1,\mathcal{C}}(x_{11}, c_1), D_{X_1,A_2,\mathcal{C}}(x_{11}, c_1)\right) = (0.5, 0.2)$$

and, with respect to X_2, $\mathcal{A}$ and $\mathcal{C}$,

$$(D_{X_2,A_1,\mathcal{C}}(x_{21}, c_1), D_{X_2,A_2,\mathcal{C}}(x_{21}, c_1)) = (0.3, 0.1)$$

Analogously, we get, with respect to X_1, $(0.4, 0.8)$ for c_2 and $(0.7, 0.6)$ for c_3. With respect to X_2, we get $(0.5, 0.2)$ for c_2 and $(0.4, 0.9)$ for c_3. Finally, data quality of c_1, c_2 and c_3 is the following, with respect to $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{C}$,

$$((0.5, 0.2), (0.3, 0.1)), ((0.4, 0.8), (0.5, 0.2)) \text{ and } ((0.7, 0.6), (0.4, 0.9)), \text{ respectively.}$$

Data quality description of variables and statistical units can be a basis for data quality description of the whole data set. Let the dimension of the whole data set be $n \times m$. Then data quality of the whole data set can be characterized either as an n-tuple, where each element represents data quality of a particular case (statistical unit), or as an m-tuple, where each element represents data quality of a particular variable. The above mentioned data quality description is exhaustive, incorporates data quality of all variables and statistical units with respect to any set of attributes. Unfortunately, from the practical point of view our description is not easily applicable (large dimensions, complex interpretation etc.). Therefore we construct on its basis a simple data quality measure aggregating the complete description of data quality of each statistical unit or statistical variable into a single real number.

3.2 A distance based data quality measure – Data Quality Index

There are many possibilities how to use our data quality description as a basis for further analysis or as additional information which supplements results of our analysis. Because we do not assume that attributes of data quality are independent we restrict ourselves to the distance based data quality measure, DQI. It means that data quality of a statistical unit is defined as a distance of its m-tuple from an m-tuple describing the ideal statistical unit, i.e. the statistical unit without any data quality issues. In our paper the term distance coincides with the term metric, i.e. we require its non-negativity, identity of indiscernible, symmetry and triangle inequality.

Definition 5. Let $\mathcal{C}$ be a set of statistical units, let data quality of each statistical unit $c \in \mathcal{C}$ be described by (3.3), let d be a distance function. Then a mapping $\mathrm{DQI}\colon \mathcal{C} \to [0, 1]$ is defined as follows

$$\mathrm{DQI}(c) = d((D_{X_1,\mathcal{A},\mathcal{C}}, D_{X_2,\mathcal{A},\mathcal{C}}, \ldots, D_{X_m,\mathcal{A},\mathcal{C}}), \mathbf{1}), \tag{3.4}$$

where $\mathbf{1}$ denotes the m-tuple $\left(\underbrace{(1, 1, \ldots, 1)}_{p}, \ldots, \underbrace{(1, 1, \ldots, 1)}_{p}\right)$.

DQI can take values from the unit interval, where 0 means that a statistical unit c has not any data issues with respect to $\mathcal{A}$ and 1 means that a statistical unit c has 0 % data quality.

Data quality of the whole data set we can characterize as a sum or an appropriate measure of central tendency, e.g. mean or median, of all $\mathrm{DQI}(c)$, where $c \in \mathcal{C}$.

We have many possibilities how to choose an appropriate distance used in formula (3.4). It is similar to selecting an appropriate distance in the case of clustering. It is obvious that our data quality description of a statistical unit is mathematically equivalent to so called hesitant fuzzy sets (Zeshui, Meimei 2011), although its interpretation is completely different. Therefore, in the rest of our paper, we restrict ourselves to two distances similar to those used in the case of hesitant fuzzy sets, the normalized Hamming like distance

$$d_{NHD}(c,c') = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{p}\sum_{j=1}^{p}|D_{X_i,A_j,\mathcal{C}}(x_i,c_i) - D_{X_i,A_j,\mathcal{C}}(x_i',c_i')|\right], \tag{3.5}$$

and the normalized Euclidean like distance

$$d_{NED}(c,c') = \left(\frac{1}{m}\sum_{i=1}^{m}\left[\frac{1}{p}\sum_{j=1}^{p}\left(D_{X_i,A_j,\mathcal{C}}(x_i,c_i) - D_{X_i,A_j,\mathcal{C}}(x_i',c_i')\right)^2\right]\right)^{\frac{1}{2}}, \tag{3.6}$$

where $c, c' \in \mathcal{C}$.

Remark Similarly, we can introduce DQI for variables as a mapping $\mathrm{DQI_v}\colon \mathcal{X} \to [0,1]$:

$$\mathrm{DQI}(X) = d((D_{X,\mathcal{A},c_1}, D_{X,\mathcal{A},c_2}, \ldots, D_{X,\mathcal{A},c_n}), \mathbf{1}),$$

where $\mathbf{1}$ denotes the n-tuple $\left(\underbrace{(1,1,\ldots,1)}_{p},\ldots,\underbrace{(1,1,\ldots,1)}_{p}\right)$, $D_{X,\mathcal{A},c_i}$ denotes a p - tuple $\left(D_{X,\mathcal{A}_1,\mathcal{C}}(X(c_i),c_i), D_{X,\mathcal{A}_2,\mathcal{C}}(X(c_i),c_i),\ldots,D_{X,\mathcal{A}_p,\mathcal{C}}(X(c_i),c_i)\right)$ and d denotes a distance function. Therefore we can apply our data quality measuring algorithm to statistical units as well as to statistical variables.

Our approach is roughly inspired by the TOPSIS method (Hwang, Yoon 1981) but our algorithm assumes the best alternative independently of existing statistical units and does not assume the worst alternative. Moreover, in the case of TOPSIS method we are interested in ranking of alternatives in order to choose the best alternative, in our approach we are interested in an absolute measure of data quality of a particular statistical unit allowing us to decide whether and how this statistical unit can be used in our further analyses. On the other hand, similarly to TOPSIS, our method allows a trade-off between data quality of attributes, where one attribute can be compensated by another one. The level of compensation depends on the number of attributes and variables.

4 Application of data quality analysis to insurance data

The application of data quality analysis depends strongly on the studied research problem. Nevertheless, we can pose some recommendations how to incorporate data quality analysis into a data analysis process. In order to illustrate such possible inclusion, we present here partial data quality analysis in the context of statistical analysis we performed on a real data set. The data set comes from a Czech insurance company and consists of 677,284

real customer contracts (units, i.e. rows) with 9 characteristics (variables, i.e. columns). One of these characteristics (the dependent variable) represents classification of the contract and other variables are described in Table 2. 261,402 cases belong to the customers with a lapsed insurance policy and 415,882 cases to the customers with a policy in force. Although we fully support the idea of reproducible research, the insurance company did not give us permission to share data in any form due to its confidentiality policy.

Table 2: Description of used variables and their notation.

Variable	Type
type of product	dichotomous
payment frequency (within one year)	categorical with 5 levels
region	categorical with 14 levels
gender	dichotomous
the age of policyholder at the time of conclusion of the contract (in years)	numeric
number of policyholder migrations	categorical with 11 levels
freshness (in years)	numeric
policy duration (in years)	numeric

Variables listed in Table 2 can be classified into 2 types: policyholder's characteristics (age, gender, region and number of migrations) and contract's characteristics (product type, payment frequency, freshness and policy duration). The first step of our analysis consists of elements of exploratory data analysis.

4.1 Elements of exploratory data analysis for insurance data

In order to better understand a relationship between the independent variables and lapse, we did exploratory visualization (mosaic plots, density plots,...) and found out that there is no relationship between gender and lapse in our data. Data also indicates, that there is a difference between lapsed policies of two different types of insurance (a Unit Linked Life insurance and a Traditional Life insurance). This may be caused by the fact, that the unit linked life insurance product is more expensive and less easy understandable than the traditional insurance product. It also seems that policies with quarterly payment have the highest lapse rate. On the other hand, the policies paid with one single payment and policies with monthly payments have the lowest lapse rate. The policies of customers who migrated five times have the largest lapse rate. Generally, lapse contracts have shorter duration, hence they have lower number of migrations.

From our analyses it followed that the policies of customers who are at the age between 20 and 35 at the time of conclusion of the contract are more likely to lapse than the policies of older customers. Moreover, the shorter time elapsed since contract information was updated the lower risk of contract lapse occurs. There is a very similar dependency between lapses and policy duration. The lapse rate is higher for policies with shorter duration.

The region and the number of migrations are relevant behavioural characteristic for lapse prediction only if assumption, that lapse rate is higher in the poorer regions, is cor-

rect. In order to validate this assumption we examine the relationship between lapse rate and selected macro economical aggregates by regions. We have chosen the net disposable income of households per capita, GDP per capita and unemployment rate. Values of income and GDP come from the Czech Statistical Office and the source of values of unemployment rate is the Czech Ministry of Labor and Social Affairs. In Table 3, we summarize selected indicators and add the proportion of people with overdue liabilities to the total population at the age of 18 and over in each region (source: `www.solus.cz`) and proportion of lapse contracts to total contracts per region (source: the insurance company). The highest lapse rate occurs in regions with the highest unemployment rate (Ústecký region 42%) and the lowest disposable income (Liberecký 41%, Karlovarský 40%, Olomoucký 40%). Also the highest rate of people having problem with paying off their debts occurs in the poorer regions (Ústecký 14%, Karlovarský 13%, Liberecký 11%).

Table 3: Selected macro economical aggregates, payment behaviour and lapse rate by region.

Region	Income	GDP	Unempl.	Liab.	Lapse
Capital city Prague	250,121	768,173	0.04	0.06	0.36
Středočeský region	206,669	325,797	0.07	0.08	0.37
Jihomoravský region	184,823	341,024	0.10	0.07	0.37
Královéhradecký region	179,715	315,307	0.08	0.07	0.38
Pardubický region	177,064	297,755	0.08	0.07	0.39
Region Vysočina	180,102	303,263	0.09	0.05	0.39
Zlínský region	178,580	308,642	0.09	0.05	0.39
Moravskoslezský region	176,135	317,835	0.11	0.10	0.39
Jihočeský region	181,215	306,576	0.08	0.07	0.4
Plzeňský region	187,924	326,513	0.07	0.08	0.4
Karlovarský region	171,785	260,083	0.10	0.13	0.4
Olomoucký region	172,415	281,540	0.11	0.07	0.4
Liberecký region	178,750	279,733	0.10	0.11	0.41
Ústecký region	170,925	289,851	0.13	0.14	0.42

Income = Net disposable income of households per capita (in CZK, year 2011),
GDP = GDP per capita (in CZK, year 2011),
Unempl. = Unemployment rate (%, value to date 31.12.2011),
Liab. = Proportion of people with overdue liabilities to the total population (% value to date 31.3.2012),
Lapse = Proportion of lapsed contracts to total contracts per region (%).

Table 4 shows correlation coefficients among indicators. The lapse rate negatively correlates with disposable income (-0.73), i.e. the lower income, the higher lapse rate. We can observe a positive correlation between the lapse rate and two indicators: the unemployment rate (0.67) and payment behaviour (0.58). Consequently, we may assume that the poorer regions of Czech Republic might have a higher risk of lapse.

Our exploratory analysis indicates that gender can be omitted from lapse prediction modelling. This fact can be used also for decision that gender does not need to be collected by the insurance company.

Table 4: Pearson's Correlation Coefficients.

	Income	GDP	Unempl.	Liab.	Lapse
Income	1.00	-	-	-	-
GDP	0.94	1.00	-	-	-
Unempl.	-0.81	-0.69	1.00	-	-
Liab.	-0.34	-0.31	0.58	1.00	-
Lapse	-0.72	-0.63	0.68	0.60	1.00

4.2 Data quality analysis of insurance data

In effort to increase the reliability of our analyses, e.g. lapse prediction modelling, we can choose a suitable set of variables for our basic model not only on the basis of performed exploratory analysis, but also on the basis of data quality. Moreover, if we describe data quality of all available statistical units, we can use this information for data set preparation and for better understanding of a resulting model. If we would like to assess data quality of a contract, it is necessary to start with data quality of individual variables. For simplicity, and in coherence with our statements in the previous sections, in further analysis we restrict ourselves to the three elements of data - currency, timeliness and uncertainty. It is obvious that variables gender, age of a client and type of product are constant values at the time of conclusion of a contract, therefore unimportant for the intendet data quality analysis. We omit them from the rest of our data quality analysis assuming that there are no data issues for these three variables, i.e. the data quality with respect to uncertainty is 1, timeliness and currency are irrelevant for these variables.

Currency and timeliness were already defined in Section 2. But for our purposes it is necessary to transform them to the unit interval in order to get 1 as the best possible option and 0 as the worst one. We use a very simple transformation $\frac{a}{(a+x)}$, where x represents currency or timeliness, respectively, and a represents our sensitivity to changes in data quality with respect to currency and timeliness. For simplicity, presented results were computed for $a = 1$. The extraction date was October 1, 2013 and the delivery date was November 10, 2013. In our case, currency is the same for all variables and also for all statistical units. Timeliness was computed using the same formula for all variables but, in general, it is different for different statistical units.

Uncertainty represents our doubts about data quality of an individual variable. In our opinion, contrary to timeliness and currency, evaluation of uncertainty cannot be entirely based on a particular value of the variable corresponding to the selected contract, but we should primarily evaluate the whole variable. In the case of the presented data set, the variable region is validated by a financial intermediary (an agent, a broker). The insurance company records all characteristics of the contract proposal received from the intermediary into its primary production information system. After that the client receives his or her contract and confirms correctness of information by the act of acceptance, hence all data can be considered reliable at the time of inception. Contract's characteristics are under the insurer's control. The contact address region could be invalid if the client is not motivated to update it after migrating to a different place. Therefore in our example uncertainty is interesting only for the variables region and the number of migrations, data

quality with respect to uncertainty equals to 1 for the rest of variables .

We decided to compute uncertainty of the number of migrations as follows. We suppose that the contract with a positive number of policyholder migrations is correct because the client is rigorous and updates his or her personal data. Similarly, we suppose the correct data for contracts within the three-month period of the confirmation process. For these contracts the uncertainty degree (ud) is 1 $(\mathrm{ud}(\mathrm{contract}_k) = 1, k = 1, 2, \ldots, n_1)$, where n_1 is the number of contracts with a positive number of policyholder migrations or contracts shorter than three months.

The uncertainty degrees for the rest of contracts were determined using the following procedure. The average of the variable number of policyholder migrations was computed for each region. For each remaining contract, the uncertainty degree was computed according to the formula

$$\mathrm{ud}(\mathrm{contract}_k) = \mathrm{P}(0 \text{ migrations in a region}_l), k = 1, 2, \ldots, n_2; l = 1, 2, \ldots s, \quad (4.1)$$

where n_2 is the number of contracts older than three months or with zero policyholder migrations, s is the number of regions and P is the probability mass function of the Poisson probability distribution with the mean λ_l estimated by the mean number of policyholder migrations in the region l. Formula (4.1) is coherent with intuition that migrations are more likely for regions with the higher average number of policyholders migrations. The uncertainty degrees of contracts with respect to the number of policyholder migrations were used also for regions.

Using the above mentioned procedure, we assign to each variable (except variables gender and age) in each contract a triplet (currency, timeliness, uncertainty). Therefore each contract is characterized by a family of triplets, one for each variable, i.e. it is modelled as a hesitant fuzzy set. Then DQI for the contract can be computed as a distance between the contract itself and the ideal contract. In the paper we compute the distance using two basic distances, the normalized Hamming distance (NHD) (3.5) and the normalized Euclidean distance (NED) (3.6). Values of DQI are visualized in Figure 1 in the form of empirical density plots. From Figure 1 it is obvious that, regardless of the metric, the number of contracts with low data quality of selected variables with respect to currency, timeliness and uncertainty is quite high. Using computed DQI we can conclude that due to the low data quality with respect to currency, timeliness and uncertainty, we can expect the low predictive power of resulting models. Moreover, in addition to results of exploratory data analysis on the basis of low data quality, we can exclude variables region and number of migrations from the set of variables assumed as predictors for our lapse models.

As it was already mentioned before, we can use DQI also to compute weights for individual cases. Despite the fact that in our example weighting of cases would not decrease error rates of the resulting lapse prediction models, we prefer to include DQI into the model fitting process because it could boost our confidence in results we got.

5 Discussion and conclusions

The main result of the presented paper is a measure of data quality, so called Data Quality Index (DQI). It allows us to evaluate data quality of individual contracts by a single

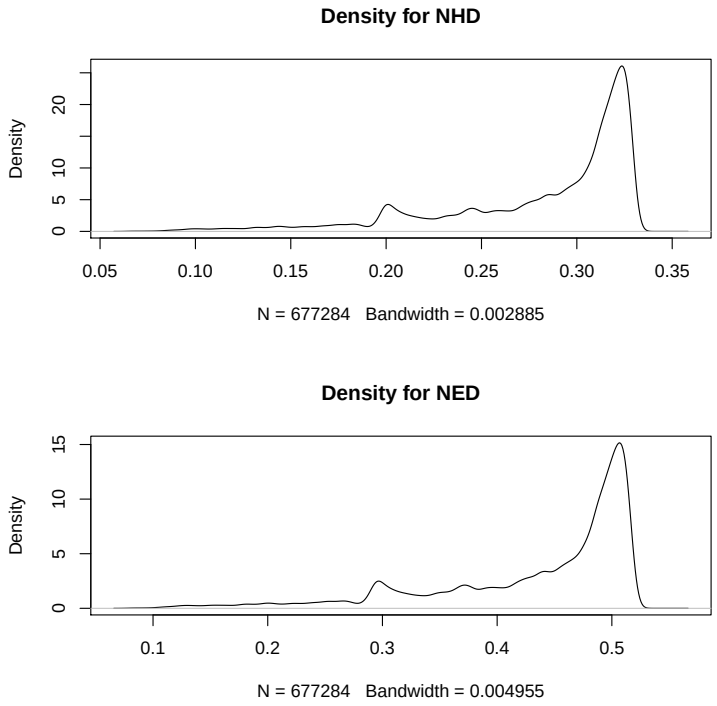

Figure 1: Density plots for DQI computed NHD (upper plot) and NED (lower plot)

number from the unit interval. The starting point of the whole procedure is based on evaluation of data quality of individual variables. The proposed measure is illustrated using the real insurance data set and some possibilities how to incorporate data quality analysis into complex data analysis procedures are pointed out. Although in the paper we restrict ourselves only to currency, timeliness and uncertainty, and we use very simple distances to assess each of them, it is obvious that our analysis can be further extended using more components and more sophisticated mappings for these components. Alternatively, using the same mathematical representation of data quality for individual contracts, we can use an appropriate aggregation function instead of a distance based measure to evaluate data quality. These possible extensions, as well as other behavioural factors (occupation, education) and distances between individual contracts, will be further investigated. Moreover, because all key elements of the presented data quality analysis, such as values of DQI, their interpretation, appropriate data quality components and distances etc., are strictly context dependent, we will focus on DQI in particular contexts in our future research, e.g. on verification of possibility to use DQI as a prior to adjust lapse probability models constructed using some well established classification methods (logistic regression, random forests etc.).

The codes for all the examples given above are written in R (R core team 2013) and are included in supplementary materials of the paper. In order to further simplify possible adoption of the proposed methodology we also included a small artificial data set mimicking some properties of the original one.

Acknowledgment

This work was supported by projects Mobility - enhancing research, science and education at Matej Bel University, ITMS code: 26110230082, under the Operational Program Education co-financed by the European Social Fund, VEGA 1/0647/14 and IGA VSE F4/17/2013.

We would like to thank prof. Hana Řezanková for her valuable comments and suggestions.

References

[1] Berti-Equille, L. (2007): Quality Awareness for Data Management and Mining. *Habilitation a Diriger des Recherches*, Universit'e de Rennes 1, France, [available online]

[2] Bonnissone, P.P. and Tong, M. (1985): Editorial: Reasoning with Uncertainty in Expert Systems, *Int'l J. Marl Machine Studies*, **22**, 241–250.

[3] Bosc, P. and Prade, H. (1993): An Introduction to Fuzzy Set and Possibility Theory Based Approaches to the Treatment of Uncertainty and Imprecision in Database Management Systems,*Proc. Second Workshop Uncertainty Management in Information Systems: From Needs to Solutions*, 44–70, Catalina, Calif.

[4] Cox, D.R. (1972): Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

[5] Grahne, G. (2002): Information Integration and Incomplete Information, *IEEE Data Eng. Bull.*, **25(3)**, 46–52.

[6] Hwang, C. L. and Yoon, K. (1981): *Multiple Attribute Decision Making: Methods and Applications*. New York: Springer-Verlag.

[7] Imielinski, T., Lipski, W.JR. (1984): Incomplete Information in Relational Databases.*J. ACM*, **31(4)**, 76–791.

[8] Motro, A. (1993): Sources of Uncertainty in Information Systems, *Proc. Second Workshop Uncertainty Management and lnformation Systems: From Needs to Solutions*, 9–26, Catalina, Calif.

[9] Naumann, F., Freytag, J.-Ch., Leser, U. (2004): Completeness of Integrated Information Sources. *Inf. Syst.*, **29(7)**, 58–615.

[10] Naumann, F., Leser, U., Freytag, J.-Ch. (1999): Quality Driven Integration of Heterogenous Information Systems, *Proceedings of the 25th International Conference on Very Large Data Bases*, 447–458, Edinburgh, Scotland, UK.

[11] Parsons, S. (1996): Current approaches to handling imperfect information in data and knowledge bases. *IEEE Transactions on Knowledge and Data Engineering*, **8**, 353–372.

[12] Peralta, V. (2006): Data Quality Evaluation in Data Integration Systems. *Ph.D. thesis*, Université de Versailles, France and Universidad de la República, Uruguay.

[13] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[14] Segev, A. and Fang, W. (1990): Currency-Based Updates to Distributed Materialized Views. *Proceedings of the 6th International Conference on Data Engineering*, ICDE 1090, 51–520, Los Angeles, CA, USA.

[15] Theodoratos, D. and Bouzeghoub, M. (1999): Data Currency Quality Factors in Data Warehouse Design, *Proceedings of the International Workshop on Design and Management of Data Warehouses*, DMDW'99, Heidelberg, 15.1–15.16, Germany.

[16] Zeshui Xu and Meimei Xia (2011): Distance and similarity measures for hesitant fuzzy sets, *Information Sciences*, **181**, 2128–2138.

INSTRUCTIONS TO AUTHORS

Language: *Metodološki zvezki – Advances in Methodology and Statistics* is published in English.

Submission of papers: Authors are requested to submit their articles (complete in all respects) to the Editor by e-mail (MZ@stat-d.si). Contributions are accepted on the understanding that the authors have obtained the necessary authority for publication. Submission of a paper will be held to imply that it contains original unpublished work and is not being submitted for publication elsewhere. Articles must be prepared in LaTeX or Word. Appropriate styles and example files can be downloaded from the Journal's web page (http://www.stat-d.si/mz/).

Review procedure: Manuscripts are reviewed by two referees. The editor reserves the right to reject any unsuitable manuscript without requesting an external review.

Preparation of manuscripts

Tables and figures: Tables and figures must appear in the text (not at the end of the text). They are numbered in the following way: Table 1, Table 2,…, Figure 1, Figure 2,…

References within the text: The basic reference format is (Smith, 1999). To cite a specific page or pages use (Smith, 1999: 10-12). Use "et al." when citing a work by more than three authors (Smith et al., 1999). The letters a, b, c etc. should be used to distinguish different citations by the same author(s) in the same year (Smith, 1999a; Smith, 1999b).

Notes: Essential notes, or citations of unusual sources, should be indicated by superscript number in the text and corresponding text under line at the bottom of the same page.

Equations: Equations should be centered and labeled with two numbers separated by a dot enclosed by parentheses. The first number is the current section number and the second a sequential equation number within the section, e.g., (2.1)

Author notes and acknowledgements: Author notes identify authors by complete name, affiliation and his/her e-mail address. Acknowledgements may include information about financial support and other assistance in preparing the manuscript.

Reference list: All references cited in the text should be listed alphabetically and in full after the notes at the end of the article.

References to books, part of books or proceedings:

[1] Smith, J.B. (1999): *Title of the Book.* Place: Publisher.

[2] Smith, J.B. and White A.B. (2000): *Title of the Book.* Place: Publisher.

[3] Smith, J. (2001): Title of the chapter. In A.B. White (Ed): *Title of the Proceedings*, 14-39. Place: Publisher.

Reference to journals:

[4] Smith, J.B. (2002): Title of the article**.** *Name of Journal***, 2**, 46-76.

Metodološki zvezki
Advances in Methodology and Statistics

Published by		Izdajatelj
Faculty of Social Sciences University of Ljubljana, for Statistical Society of Slovenia		**Fakulteta za družbene vede Univerze v Ljubljani za Statistično društvo Slovenije**
Editors	**Valentina Hlebec** **Lara Lusa**	Urednika
Founding Editors	**Anuška Ferligoj** **Andrej Mrvar**	Prva urednika
Cover Design	**Bojan Senjur** **Gregor Petrič**	Oblikovanje naslovnice
Typesetting	**Lara Lusa**	Računalniški prelom
Printing	**Littera picta d.o.o. Ljubljana, Slovenia**	Tisk
	MZ	
is indexed and abstracted in	**SCOPUS** **EBSCO** **ECONIS** **STMA-Z** **ProQuest**	je indeksirana in abstrahirana v
Home page URL	http://www.stat-d.si/mz/	Spletna stran

ISSN 1854 - 0023