UDK: 3.071.6:328.1:061.1EU 1.01 Original scientific article

An Application of Subgroup Discovery Algorithm on the Case of Decentralization and Quality of Governance in EU

Lan Umek
Faculty of Administration, University of Ljubljana
lan.umek@fu.uni-lj.si

ABSTRACT

This paper analyses a statistical relationship between the decentralization of the EU countries and the quality of their governance. The degree of decentralization is measured from a fiscal and political point of view, and the quality of governance by multiple indicators and citizen opinions. The paper presents a subgroup discovery algorithm which is capable of analysing two sets of several variables, and uses it for the analysis of EU countries. The paper is one of the first to use the data mining methods from the social sciences domain. The used algorithm has discovered some interesting patterns which show a desired relationship. We have discovered that the proportion of public sector employees is one of the most important indicators, which strongly correlates with the degree of trust in the European and national institutions, the government effectiveness and the perception of corruption

Key words: *decentralization, quality of governance, subgroup discovery, data mining, EU countries*

JEL: C38, C12, H11, H50, H77

1 Introduction

Quality of governance is a broad concept that is best expressed through the efficiency and effectiveness of public administration and the quality of its services (Žurga 2001, p. 7). Quality public administration offers services that are consistent with established laws and international standards, and meets the requirements and expectations of its users. If the public administration power is dispersed over several smaller institutions, its performance becomes better and more responsible, and the mutual control is established, which improves the quality of governance (Gerring, Thacker, & Moreno, 2005, p. 567).

The transfer of powers and resources from the centre to the lower levels of governance and political participation is linked with the concept of decentralization (Aristovnik, 2012, p. 6). The paper will discuss in detail two of its aspects, namely the fiscal and the political aspect. Put simply, fiscal decentralization is defined by how much of the expenditure and revenue management

is under the authority of local communities, while political decentralization signifies fragmentation of the country into smaller units, and separation of powers to many political institutions.

In the paper, we use one of the data mining methods, by which we analyse a statistical relationship between the decentralization of a country and its quality of governance, how strong it is and how it is expressed. In the social sciences, researchers use the data mining more and more frequently, but such methods are still quite rare, since established statistical methods (regression, hypothesis testing, factor analysis ...) are preferred. Most commonly, data mining methods (Witten & Frank, 2002) use data tables to generate hypotheses, which are presented to the domain experts in plain language, and in addition to statistical evaluation, allow expert evaluation as well.

In the paper we present an algorithm for subgroup discovery, which is able to analyse two sets of several variables (Umek & Zupan, 2010; Umek, 2011): indicators of decentralization and quality of governance. Unlike the majority of regression analyses, the presented algorithm is able to simultaneously analyse all the indicators of quality of governance, and presents the main results in an interpretable way.

The algorithm is implemented in an open source software package Orange (Demšar et al., 2013), images are plotted with Python library Matplotlib (Hunter, 2007) and with the program Graphviz (Gansner & North, 2000).

In the analysis of the relationship between decentralization and quality of governance, we analysed 27[1] EU countries and 43 indicators. In statistical analysis, the analysis of a small sample with many measured variables represents a challenge, which many existing analyses cannot handle. The discovered patterns are thus even more important.

The principal objective of the paper is the application of a novel approach on a social sciences domain. The effective performance of data mining algorithms on other scientific domains motivates their applications on data sets from other sources (administrative and other social sciences).

The paper is divided into five sections. After the introduction, in the second section we describe the data with which we operated and briefly define the concepts of decentralization and the quality of governance. In the same section we introduce related statistical work and its limitations in analysing such data. In the third section we describe the algorithm for subgroup discovery, and in the fourth we present the results, three interesting subgroups. Conclusion summarizes the paper and presents opportunities for further work.

1 Due to lack of data, we excluded Croatia from the analysis.

2 Description of Data and Related Statistical Approaches

In this section, we will briefly review the most important definitions of quality of governance and decentralization, and describe the variables we used in the analysis. In the paper we used the same 43 indicators as the authors in (Benčina & Mrđa Kovačič, 2013) and (Mrđa Kovačič, 2013): 24 indicators of quality of governance: 15 aspects of trust in several institutions[2] (Samanni et al., 2012) and nine other indicators[3], eight indicators of fiscal[4] and 11 indicators of political decentralization[5].

We will then summarize the related statistical approaches the researchers used for analysing relationships between the decentralization of the country and its quality of governance. Since the main focus of our paper is to present a novel statistical algorithm we will summarize the limitations of the well-established approaches in this area.

2.1 Quality of Governance and Decentralization Indicators

According to the World Bank's definition, quality governance is the »the process by which governments are selected, monitored and replaced; the capacity of the government to effectively formulate and implement sound policies, and the respect of citizens and the state for the institutions that govern economic and social interactions among them« (Dijkstra 2011, p. 1). A slightly different definition is found in Bäck & Hadenius 2008, p. 8. The authors define quality governance as the government's ability to effectively implement its own activities, as well as the absence of corruption. According to Charron (Charron 2009, p. 7) the good governance is characterized by the low level of corruption, the efficiency of bureaucracy and citizens' participation in democratic institutions. According to Vintar the key element of good governance is the level of e-government development (Vintar, 2010).

According to the different definitions our study included 15 indicators of trust in national and European institutions (footnote 2) and nine other indicators, mainly focussed on functioning of government and perception of corruption (footnote 3).

2 Trust in the European Court of Justice, EU Council of Ministers, European Commission, European Central Bank, European Court of Auditors, European Ombudsman, European Parliament, EU Social and Economic Committee, legal system, police, army, political parties, civil service, national government, and in national parliament (presented as the proportion of people who trust in each institution).

3 Functioning of government, Index of objective indicators of good governance, Corruption perceptions index Number of new infringement cases, Transposition of community law, Voter turnout in national and EU parliamentary elections, E-government on-line availability, E-government usage by individuals, Level of citizens' confidence in EU institutions.

4 Tax revenue, Highest marginal tax rate, Expense, Total general government expenditure, Central government expenditure, Local government expenditure, Total receipts from taxes and social contributions (local and central), all measured in (% of GDP).

5 PS (Public sector) Employment, Average district magnitude, Number of districts, Centripetalism, Unitarism, Unitary or federal state, Appointment of regional representatives, Unitary or federal state, Regional authority index, Total fractionalization, Electoral system.

Decentralization is a concept that "includes the transfer of power, formal authority, responsibilities and resources to lower administrative levels of government" (Dubois & Fattore 2009, p. 706). Simply put, it is a transfer of powers from higher to lower levels, initially from the central to local government, and from there to citizens.

According to Schneider (Schneider, 2003, p. 33), there are three aspects of decentralization: fiscal, political and administrative. In our analysis we will focus on the first two. The easiest to measure is the degree of fiscal decentralization, which is concerned with the proportion of transfers of assets (revenue, expenditure) to the local levels, as well as a proportion of fiscal powers which is passed from the centre to the lower administrative levels.

According to Schneider political decentralization covers an area of "the organization, participation and integration of interests in the processes of public administration". In political systems which are decentralized, representatives are operating within the local environment. They represent local interests, by which they participate in the legislative and executive powers.

In our analysis we discussed eight indicators of fiscal decentralization, expressed as a share of GDP (footnote 4) and the selection of 11 indicators of a political decentralization (footnote 5) based on the study (Dubois & Fattore, 2009), and subsequent selection (Benčina & Mrđa Kovačič, 2013).

The majority of the analysed indicators refers to the year 2011 and can be accessible through several sources, the most important are (Samanni et al., 2012) and (Eurostat, 2014).

2.2 Analysis of the Relationship Between Decentralization and Quality of Governance

Several authors have already analysed the impact of decentralization on the quality of governance. Most commonly they have used the regression methods, which have found that fiscal decentralization has a positive impact on the quality of governance (the government's operation), while the political decentralization has a negative one (worse perception of corruption). Similar effects were also confirmed by the studies (Gerring et al., 2005) and (Enikolopov & Zhuravskaya, 2007), which showed that centripetalism has a positive impact on the government performance and economic growth (Kyriacou & Roca-Sagales, 2011).

Several regression studies analysed the impact of decentralization on the level of corruption as one of the aspects of quality of governance. According to (Altunbas & Thornton, 2012; Ivanyna & Shah, 2010), fiscal decentralization has a negative impact on indicators of corruption.

Several authors analysed statistical relations between different aspects (Ahrens & Meurers, 2002; Benčina & Mrđa Kovačič, 2013) and applied

the methods for dimension reduction (Principal component analysis and Factor analysis). On the other hand, by using the structure model, Dreher defined more complex index of corruption (Dreher, Kotsogiannis, & McCorriston, 2007).

All these methods are based on linear combinations of variables, which are often not reasonable or appropriate, and the obtained results are difficult to understand. The main purpose of the paper is to analyse the data with a more general procedure with weaker assumptions (such as linearity) required. The presented approach will be capable of a simultaneous analysis of multiple dependent variables (classical regression analysis can only address one), while the analysed indicators can be completely arbitrary (a mixture of nominal, ordinal and ratio variables). The new approach will present in an understandable way the results which are immediately suitable for further investigation (evaluation from expert's perspective, further testing of generated hypotheses …).

3 Methods

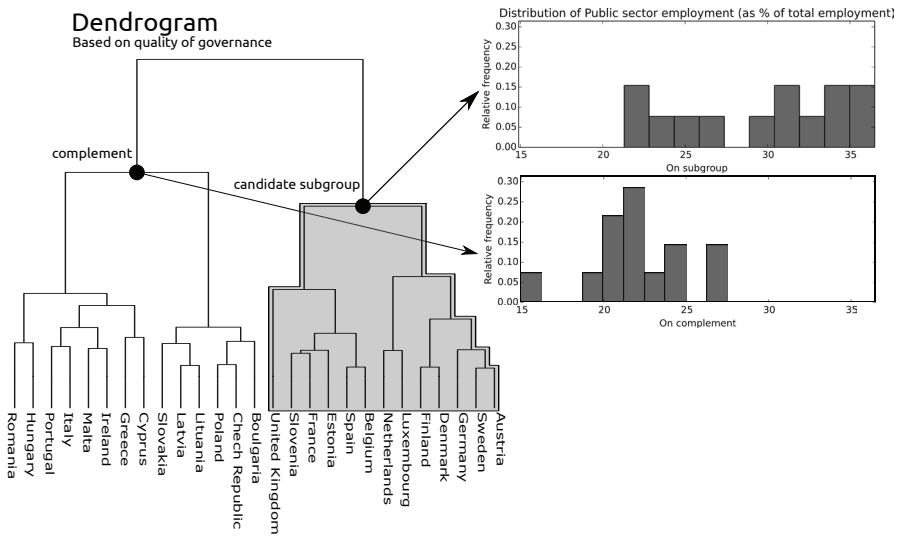
Regarding the objective of the research, we will consider the decentralization indicators presented in Section 2.1, as a set of independent variables (*X*), and the indicators of the quality of governance (section 2.2) as the dependent variables (*Y*). We will analyse the relationships between *X*s and *Y*s. Since we analyse 24 dependent variables, most regression methods cannot be directly applied in our case. Blockeel and Ženko (Blockeel, Raedt, & Ramon, 1998; Ženko, 2008) showed that a separate analysis of each dependent variable weakens the interpretability of the results and extends the computation time. They proposed that in the case of multiple dependent variables, researchers should employ more advanced statistical methods.

In the paper, we applied the algorithm MR-SD (Multiple-Responses Subgroup Discovery), which is primarily aimed at finding interesting subgroups (Klösgen, 1996), and is capable of simultaneously dealing with multiple dependent variables (Umek & Zupan, 2010; Umek, 2011). The MR-SD algorithm tries to find such subgroups of units which are similar in respect to the values of the dependent variables, and then attempts to seek the causes of their similarities in the space of independent variables. In our case, it first searches for the clusters of EU countries that are similar in terms of quality of governance. These clusters (we will call them candidate subgroups) are further evaluated in how well they can be distinguished from other EU member states according to their decentralization. The MR-SD algorithm combines the established methods of clustering and statistical classification, while the results in the form of interesting subgroups reflect the relationship between the two sets of variables.

In a concrete example we show a more detailed description of the algorithm MR-SD. The method first divides EU countries into clusters within which

the members are similar regarding their quality of governance (*Y*) (Ferligoj, 1989). It uses hierarchical clustering, which can be displayed graphically with the so-called clustering tree (dendrogram, left part of Figure 1). The process of hierarchical clustering first treats each country as a separate cluster, and then in each subsequent step merges the two most similar clusters into a new one. The more similar they are, the lower their level of merging is. On the left side of Figure 1, a cluster of 13 countries is highlighted in grey (United Kingdom, Slovenia, ..., Austria). These countries are similar in terms of quality of governance. The aim of further analysis is to determine whether it is possible to find the cause of their similar quality of governance with the indicators of decentralization. We will further evaluate a candidate subgroup with the degree of interestingness "in the space of independent variables" (the right side of Figure 1).

Figure 1: Graphical illustration of the procedure from the paper



Members of the »grey« candidate subgroup are similar in terms of quality of governance (the dendrogram node on the left figure). Double histogram on the right shows that these countries can be clearly distinguished from the others in terms of the indicators of decentralization as well: if the proportion of the PS employees in a country is greater than 30 %, then it is certain one of the countries from the »grey subgroup« (top right).

To illustrate this, let us evaluate the candidate subgroup using a single indicator of decentralization. We illustrate the distribution of the independent variable "public sector employment" separately for the subgroup of the 13 countries and for the complement (right side of Figure 1). The histogram shows that all the units in which the proportion of employees in Public sector (PS) is at least 30 % belong to the subgroup of the 13 countries. When the proportion is less than 20 % it is merely a representative of the complement. At the intermediate interval between the 20 % and 30 %, a subgroup or a comple-

ment representative may be located. Knowing the proportion of employees in PS is therefore a pretty reliable (but not precise) indicator of subgroup's membership.

Let us return to the description of the algorithm MR-SD. Dendrogram nodes which correspond to the clusters of appropriate size, are the candidates for further evaluation. The algorithm translates the analysis of several quality of governance indicators into the analysis of belonging to a subgroup and the problem of analysis of two sets of variables to the statistical classification problem (Hastie, Tibshirani & Friedman, 2009; Witten & Frank, 2002). Several methods have been developed for this task (logistic regression, support vector machine, discriminant analysis, decision trees and rules, ...), which vary regarding the speed, comprehensibility, and the type of data for which they are actually most suitable.

The MR-SD algorithm evaluates the candidates in how well they can be distinguished from the rest of the sample in terms of the values of the independent variables, i.e. decentralization. The preferred evaluation score is therefore the area under the ROC curve (Receiver-Operating-Characteristic), which will be in the paper briefly denoted as the AUC (Area-Under-Curve) (Swets, 1996). The AUC score tells us what is the probability that we are able to distinguish between the members of the subgroup and the complement on the basis of the independent variable values. In our motivational case this means that we choose one of the 13 countries from the "grey" subgroup, and one from the rest of the countries. Based on the indicators of their decentralization[6] we estimate the probability of their belonging to the "grey" subgroup. If this estimate of probability is greater for the subgroup member than for the member of the complement, we have been successful in their separation.

The efficiency of the MR-SD algorithm depends on the suitable selection of the initial parameters. For the clustering in our study we used the weighted Manhattan distance between units (the weights were the reciprocal to variables ranges), and Ward's linkage. We searched for subgroups with at least five countries, for the statistical classification model we used decision trees (Quinlan, 1986) to a depth of three (splitting criteria: information gain). We estimated the AUC score using the leave-one-out method (Geisser, 1993).

The MR-SD algorithm can discover many interesting subgroups, which can be very similar in terms of the belonging countries, reducing the transparency of the obtained results. We have therefore compared the discovered subgroups according to their similarity: if the two subgroups matched in more than 50 % of the countries, we then chose the one with the higher AUC score (Umek, 2011).

6 In the motivational case we consider only the proportion of employees in public sector.

4 Results

The algorithm MR-SD discovered three interesting subgroups with AUC scores greater than 0,6. The subgroups covered all EU countries, which means that each EU country was assigned to at least one interesting subgroup which reflects the relationship between decentralization and quality of governance.

Below we describe the statistical properties of the discovered subgroups. For each subgroup, we report the AUC score and list the belonging countries. In the overview table, we present those indicators of quality of governance, where the arithmetic mean differs significantly between the subgroup and the complement. For the level of statistical significance we chose 0,05, we performed the t-test for independent samples and corrected the calculated p-values using Bonferroni correction (the tables show the original p-values). The tables are arranged according to the ascending p-values (Sig.), which means that more important indicators of quality of governance are presented higher in the table. In addition, we report the mean and standard deviation (stdev) for these indicators, separately for the described subgroup and its complement. If the mean on subgroup is significantly higher than the mean on the complement, we wrote the name of such variable in bold. In one case (subsection 4.1), only the aspect of trust in the European institutions stood out among the most important indicators of quality of governance. In addition to the table, we show and explain the decision tree which successfully distinguishes between the subgroup and the complement, based on the values of decentralization indicators.

4.1 Subgroup of Countries with a Higher Degree of Trust in the European Institutions

The subgroup consists of eight members (29 %) of the EU: Cyprus, Greece, Ireland, Italy, Hungary, Portugal and Romania. The subgroup is well separable from the rest of the EU members, the AUC estimate equals 0,901.

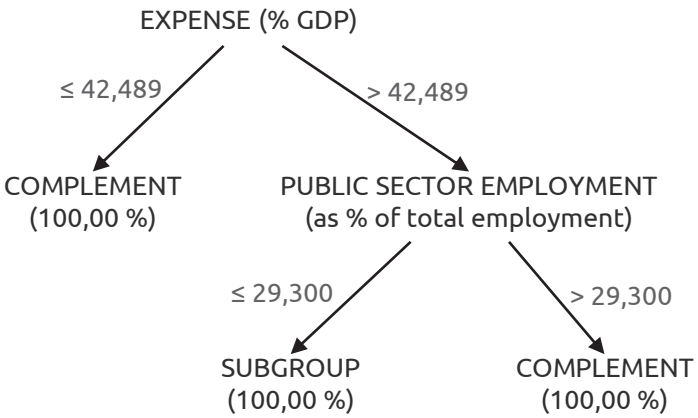
Table 1: Quality of governance indicators which significantly distinguish subgroup from the complement

	subgroup		complement		
variable	mean	stdev	mean	stdev	sig.
Trust and the European Court of Auditors	0,75	0,06	0,60	0,09	3,7E–04
Trust and the EU Council of Ministers	0,77	0,07	0,62	0,09	4,5E–04
Trust in the EU Social and Economic Committee	0,74	0.07	0,58	0,09	4,5E–04
Trust in the European Parliament	0,81	0,07	0,69	0,08	1,3E–03
Trust in the European Central Bank	0,80	0,05	0,69	0,08	1,6E–03
Trust in the European Commission	0,80	0,07	0,66	0,09	1,6E–03

The subgroup is best described by the significantly higher level of trust in the European institutions: the most significant are the differences in trust in the European Court of Auditors, followed by the trust in Council of Ministers, Social and Economic Committee, European Parliament and the Central Bank, while the least significant are the differences in the trust in the European Commission (Table 1). For other indicators, we observed no significant differences, but on average, the trust in national institutions and the index of government functioning were lower.

The decision tree in Figure 2 separates the subgroup from the complement very well, using only two indicators of decentralization. The subgroup is defined by a larger share of expenses as proportion of GDP (> 42,489 %), and a lower proportion of employees in the public sector (≤ 29,3%).

Figure 2: Description of the subgroup with the most characteristic indicators of decentralization



4.2 Subgroup of Countries with Better Governance and a Lower Level of Trust in the EU Institutions

The subgroup consists of 13 members (48 %) of the EU (Austria, Belgium, Denmark, Estonia, Finland, France, Germany, Luxembourg, Netherlands, Slovenia, Spain, Sweden and United Kingdom). The AUC score equals 0,742.

Citizens of the subgroup members use e-government services significantly more, better perceive the level of corruption in the public sector, the government functioning is more successful. This means that the e-government is better accepted among the citizens, the anti-corruption legislation is implemented more consistently, the abuse of power for private gain is rarer. The governments of the subgroup members implement public policies through

the elected representatives more effectively, and are more successful in preventing corruption (Table 2).

Table 2: Quality of governance indicators which significantly distinguish subgroup from the complement

	subgroup		complement		
variable	mean	stdev	mean	stdev	sig.
E-government usage by individuals (%)	43,39	11,95	20,00	6,71	9,7E-06
Corruption Perceptions Index	7,88	1,36	4,78	1,27	5,1E-05
Trust in the EU Social and Economic Committee	0,54	0,08	0,71	0,08	8,5E-05
Trust in the European Commission	0,63	0,08	0,77	0,06	1,0E-04
Trust in the EU Council of Ministers	0,58	0,09	0,74	0,07	1,2E-04
Trust in the European Parliament	0,66	0,08	0,79	0,06	1,2E-04
Functioning of Government	8,65	0,87	6,99	0,91	2,9E-04
Trust in the European Court of Justice	0,70	0,08	0,80	0,06	6,3E-04

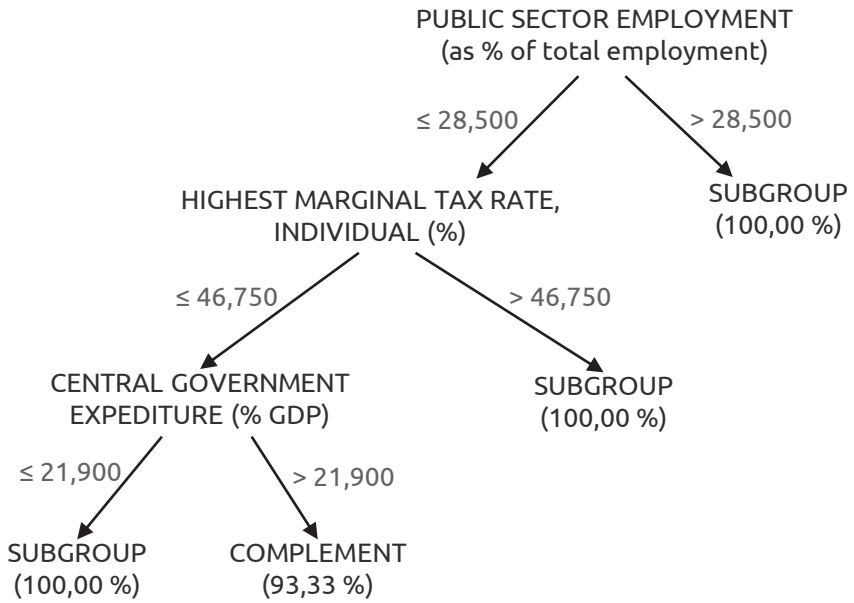
On the other hand, the trust of the citizens from these 13 countries is significantly lower in several European institutions: the most significant are the differences in trust in the European Social and Economic Committee, the European Commission and the Council of Ministers, while the least significant are the differences in trust in the European Parliament and the Court (Table 2).

The decision tree in Figure 3 describes this subgroup with the most important indicators of decentralization. The most discriminating among them is the proportion of employees in the PS: if the proportion is greater than 28,50 %, then certainly one of the 13 countries is concerned. Among the countries in which the proportion of the PS employees is lower than 28,50 %, the subgroup representatives differ most from the other countries by higher marginal tax rate (> 46,75 %). At the minimal depth of a decision tree we find the expenditures of the centre, their effect is less important. The decision tree is not as effective as in the case of subgroups in 4.1. It is clear that the forecast in the leaves is not 100 %, and the AUC score is lower.

The proportion of PS employees again surfaces among the most important indicators of decentralization. Additionally, we calculated the strength of its relationship with the indicators of quality of governance. The calculation of Pearson correlation coefficients (r) showed the strongest relationship with the perception of corruption (r = 0,83), functioning of the government (r = 0,79), and the use of e-government services (r = 0,79); a moderately positive correlation with the level of trust in national institutions (public administration (r = 0,66), police (r = 0,55), political parties (r = 0,52), legal system (r = 0,45), and a moderately negative correlation with the trust in some EU institutions (the European Commission (r = – 0,48), the Parliament (r = – 0,47),

the Social and Economic Committee (r = – 0,46), and the Council of Ministers (r = – 0,42). Most of the above mentioned indicators also appear in Table 2, which helps to explain the discovered subgroup from another point of view.

Figure 3: Description of the subgroup with the most characteristic indicators of decentralization



4.3 The Subgroup of Countries with Low Trust in National Institutions

The subgroup consists of 6 EU members (22 %): Bulgaria, Czech Republic, Latvia, Lithuania, Poland, and Slovakia. The AUC equals 0,631.

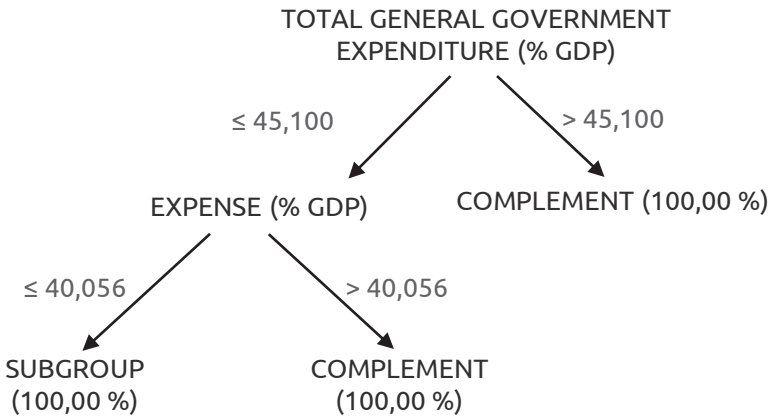
Citizens of the subgroup representatives have significantly lower trust in national parliaments, political parties, judicial system, national government and police, while the corruption perception index and the objective indicators of good governance are significantly lower, and the functioning of the government is typically worse. The subgroup is therefore described by the dissatisfaction of citizens with their own political institutions, and an extremely low (0.03) index of objective indicators of good governance. This suggests that the quality of governance is low: economic conditions are poor, there are too many barriers to entry, the taxes are too high, the contracts are poorly enforced, and there is too much corruption (Table 3).

Table 3: Quality of governance indicators which significantly distinguish subgroup from the complement

	subgroup		complement		
variable	mean	stdev	mean	stdev	sig.
Trust in national parliament	0,24	0,06	0,53	0,12	1,3E–04
Trust in political parties	0,11	0,01	0,25	0,08	1,3E–04
Corruption Perceptions Index	3,98	0,39	6,92	1,83	6,0E–04
Index of Objective Indicators of Good Governance	0,03	0,32	0,59	0,27	9,0E–04
Trust in the legal system	0,30	0,08	0,55	0,14	9,0E–04
Trust in the national government	0,34	0,04	0,52	0,11	9,0E–04
Trust in the police	0,42	0,11	0,66	0,13	1,3E–03
Functioning of Government	6,49	0,56	8,16	1,09	1,5E–03

The decision tree (Figure 4) describes a subgroup of those countries in which the total government spending represents less than 45,1 % of GDP, while expenses account for less than 40,056 % of GDP. The decision tree has otherwise clean leaves, but the AUC score is still relatively low.

Figure 4: Description of the subgroup with the most characteristic indicators of decentralization



5 Conclusion

In the paper we presented a new statistical method that explores the relationship between several independent and dependent variables. We employed the MR-SD algorithm to detect the relationship between the indicators of decentralization (fiscal and political), and quality of governance. The algorithm has proven to be an effective tool and discovered three distinct subgroups within European countries. Significant differences in trust in European

institutions characterized two of them, while significant lower level of trust in national institutions described the third one.

The MR-SD algorithm has shown that the resulting subgroups can be reliably distinguished from the rest of the EU countries by employing a small number of decentralization indicators. The majority of the most significant indicators belonged to the set of fiscal decentralization, the only indicator from the other set was the proportion of employees in PS. Additionally, we separately analysed the relationship of fiscal and political decentralization with the quality of governance. The MR-SD algorithm found some interesting subgroups, the results were better when we used just the political decentralization indicators. The proportion of the PS employees played a key role in this way as the variable with the strongest link to the quality of governance. We do not claim that the proportion has a direct influence on quality of governance – but however, the two concepts are strongly correlated.

The MR-SD algorithm was one of the first subgroup discovery algorithms, which was used on the data about the quality of governance. The promising results of this analysis raise new research questions. In the future work, we will first obtain additional indicators of the quality of governance and decentralization, and verify the stability of the detected subgroups and the principles they express. The most interesting challenge is the change of the results over time. We will first have to adapt the MR-SD algorithm for the time series analysis, and then adequately explain the obtained results.

In the paper we used one of the methods for data mining, which is otherwise rarely used in the social sciences, especially on the data about governance. Since the method has shown that it is capable to generate comprehensible hypotheses and discover certain principles within the data, we hope that in the future, in addition to the existing statistical methods, the data mining methods will be more frequently employed for analysing the data on governance.

***Lan Umek**, PhD, is a Teaching Assistant at the Faculty of Administration for the field of economics of public sector. He holds practical work classes in statistics and quantitative methods. He obtained a BSc in Mathematics in 2005 and PhD in Statistics in 2011 at the University of Ljubljana, Faculty of Mathematics and Physics. His research includes subgroup discovery methods, data mining and other optimization approaches in biological (wine production, genotype--phenotype associations) and administrative sciences (quality of governance, EU countries, questionnaire analysis). Within his research, he developed and applied several algorithms for subgroup discovery in data set with multidimensional responses. He presented the results of his work on several international conferences and journal publications.*

References

Ahrens, J. & Meurers, M. (2002). How Governance Affects the Quality of Policy Reform and Economic Performance: New Evidence for Economies in Transition. *Journal for Institutional Innovation, Development & Transition, 6*, 35–65.

Altunbas, Y. & Thornton, J. (2012). Fiscal Decentralization and Governance. *Public Finance Review, 40*(1), 66–85.

Aristovnik, A. (2012). Fiscal decentralization in Eastern Europe: Trends and selected issues. *Transylvanian Review of Administrative Sciences, 37*, 5–21.

Bäck, H. & Hadenius, A. (2008). Democracy and state capacity: Exploring a J-shaped relationship. *Governance, 21*(1), 1–24.

Benčina, J. & Mrđa Kovačič, A. (2013). The factor model of decentralization and quality of governance in European Union. *Mednarodna revija za javno upravo, 51*(3/4), 57–82.

Blockeel, H., Raedt, L. De, & Ramon, J. (1998). Top-Down Induction of Clustering Trees. In *Proceedings of the Fifteenth International Conference on Machine Learning {ICML} 1998*, (pp. 55–63). Madison, Wisconsin, USA, July 24–27, 1998.

Charron, N. (2009). Government quality and vertical power-sharing in fractionalized states. *Publius, 39*(4), 585–605.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., … Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research, 14,* 2349–2353. Retrieved 29. 9. 2014, from http://jmlr.org/papers/v14/demsar13a.html

Dijkstra, L. (2011). Quality of Government in EU Regions. Retrieved 29. 9. 2014, from http://ec.europa.eu/regional_policy/newsroom/pdf/20110504_shortnote_governance.pdf

Dreher, A., Kotsogiannis, C., & McCorriston, S. (2007). Corruption around the world: Evidence from a structural model. Journal of Comparative Economics, 35(3), 443–466.

Dubois, H. F. W. & Fattore, G. (2009). Definitions and Typologies in Public Administration Research: The Case of Decentralization. *International Journal of Public Administration, 32*(8), 704–727. doi:10.1080/01900690902908760

Enikolopov, R. & Zhuravskaya, E. (2007). Decentralization and political institutions. *Journal of Public Economics, 91*(11–12), 2261–2290.

Ferligoj, A. (1989). *Razvrščanje v skupine.* Metodološki zvezki, 4, p. 182. Ljubljana: Faculty of Sociology, Political Science and Journalism, Research Institute. Retrieved 29. 9. 2014, from http://dk.fdv.uni-lj.si/metodoloskizvezki/Pdfs/Mz_4Ferligoj.pdf

Gansner E. R. & North, S. C. (2000). An open graph visualization system and its applications to software engineering. SOFTWARE - PRACTICE AND EXPERIENCE, 30(11), 1203–1233.

Geisser, S. (1993). *Predictive Inference: an introduction.* New York, London: Chapman and Hall.

Gerring, J., Thacker, S. C., & Moreno, C. (2005). Centripetal Democratic Governance: A Theory and Global Inquiry. *The American Political Science Review, 99*(4), 567–581.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Elements, 1*, 337–387. doi:10.1007/b94608

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering, 9*(3), 90–95.

Ivanyna, M. & Shah, A. (2010). Decentralization (localization) and corruption: new cross-country evidence. *Policy Research Working Paper, 5299*. The World Bank. Retrieved 29. 9. 2014, from http://ideas.repec.org/p/wbk/wbrwps/5299.html

Klösgen, W. (1996). Explora: A Multipattern and Multistrategy Discovery Assistant. In M. Fayyad Usama, G. Piatetsky-Shapiro, & P. Smyth (Eds.) Advances in Knowledge Discovery and Data Mining (pp. 249–271).

Kyriacou, A. P. & Roca-Sagalés, O. (2011). Fiscal decentralization and government quality in the OECD. *Economics Letters, 111* (3), 191–193. doi:10.1016/j.econlet.2011.02.019

Mrđa Kovačič, A. (2013). *Kvantitativen model decentraliziranosti in kakovosti upravljanja*. University of Ljubljana. Retrieved 29. 9. 2014, from http://www.fu.uni-lj.si/diplome/pdfs/magistrska/mrdzakovacicanja.pdf

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Samanni, M., Teorell, J., Kumlin, S., Dahlberg, S., Rothstein, B., Sören, H., & Richard, S. (2012). *The QoG Social Policy Dataset*. University of Gothenburg. Quality of Government Institute. Retrieved 29. 9. 2014, from http://www.qog.pol.gu.se/data/datadownloads/qogsocialpolicydata/

Schneider, A. (2003). Decentralization: Conceptualization and measurement. *Studies in Comparative International Development, 38* (3), 32–56. doi:10.1007/BF02686198

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics : collected papers*. Scientific psychology series.

Umek, L. (2011). *Odkrivanje podskupin v podatkih z več odvisnimi spremenljivkami* (Doctoral dissertation). Ljubljana: University of Ljubljana.

Umek, L. & Zupan, B. (2010). Subgroup Discovery in Data Sets with Multi – Dimensional Responses. *Intelligent Data Analysis, 15* (4), 1–29. doi:10.3233/IDA-2011-0481

Vintar, M. (2010). Elektronsko gradivo za predavanja predmeta e-uprava. Not submitted for publication.

Witten, I. H. & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record, 31*, 76–77. Morgan Kaufmann. Retrieved 29. 9. 2014, from http://portal.acm.org/citation.cfm?id=507338.507355

The World Bank. (2010). World Development Indicators. Retrieved 29. 9. 2014, from http://data.worldbank.org/data-catalog/world-development-indicators

Ženko, B. (2008). Learning Predictive Clustering Rules. Informatica (Slovenia), 32(1), 95–96. Retrieved 29. 9. 2014, from http://www.informatica.si/PDF/32-1/18_Zenko-Learning Predictive Clustering Rules.pdf

Žurga, G. (2001). *Kakovost državne uprave: pristopi in rešitve*. Ljubljana: Faculty of Social Sciences.

POVZETEK

1.01 Izvirni znanstveni članek

Decentralizacija in kakovost upravljanja v EU Uporaba algoritma za odkrivanje podskupin

Ključne besede: decentraliziranost, kakovost upravljanja, odkrivanje podskupin, podatkovno rudarjenje, države EU

V članku predstavimo algoritem za odkrivanje podskupin, s katerim analiziramo, ali pri državah EU obstaja povezanost med njihovo decentraliziranostjo in kakovostjo upravljanja, kako močna je in kako se izraža.

Odkrivanje podskupin v podatkih je področje odkrivanja znanj iz podatkov (angl. knowledge discovery in databases), ki se v naravoslovnih znanostih (biokemija, genomika, medicina …) uporablja čedalje pogosteje, v družboslovnih vedah pa je za zdaj dokaj redko zastopano. Metode odkrivanja znanj iz podatkov iz podatkovnih tabel tvorijo raziskovalne hipoteze, ki področnim strokovnjakom pomagajo pri boljšem razumevanju problema.

V članku predstavimo algoritem MR-SD (Multiple Responses Subgroup Discovery), ki je zmožen obravnave dveh sklopov z več spremenljivimi (neodvisnimi in odvisnimi). Algoritem preizkusimo na dejanskih podatkih o državah EU, pri čemer obravnavamo sklopa indikatorjev decentraliziranosti držav (neodvisne spremenljivke) in kakovosti upravljanja (odvisne spremenljivke). Decentraliziranost merimo z osmimi indikatorji fiskalne in 11 indikatorji politične decentraliziranosti, kakovost upravljanja pa s 24 spremenljivkami, od katerih jih 15 meri stopnjo zaupanja v več državnih in evropskih inštitucij.

Jedro članka je opis algoritma, ki temelji na kombinaciji uveljavljenih metod statističnega uvrščanja in razvrščanja v skupine. V prvi fazi algoritem s hierarhičnim združevanjem razvrsti države EU v skupine, znotraj katerih so si predstavnice podobne glede na kakovost upravljanja. Skupine so predstavljene v drevesni strukturi, ki se lahko grafično ponazori s t. i. drevesom razvrščanja (dendrogramom). Grafično gledano tako vsako njegovo vozlišče predstavlja skupino držav EU s podobno kakovostjo upravljanja. Algoritem nato oceni vsako tako skupino (ki vsebuje ustrezno število članic) zgolj na podlagi decentraliziranosti, tako da ji pripiše stopnjo zanimivosti. Natančneje: algoritem skuša z uveljavljenimi metodami uvrščanja v skupine na podlagi podatkov o decentraliziranosti držav razlikovati med predstavnico analizirane skupine in državo iz preostanka. Stopnjo zanimivosti algoritem MR-SD oceni z mero AUC (Area Under Curve), ki se na področju odkrivanja znanj iz podatkov zelo pogosto uporablja kot ocena kakovosti metod.

Algoritem MR-SD je odvisen od mnogih parametrov: pri razvrščanju v skupine smo uporabili manhattansko razdaljo in Wardovo metodo združevanja, za

statistični model uvrščanja pa smo uporabili odločitvena drevesa do globine 3. Mero AUC smo ocenili z metodo izpusti enega, algoritem pa smo implementirali v odprtokodnem programskem paketu Orange. Algoritem MR-SD smo nato preizkusili na konkretnih podatkih, rezultate v obliki zanimivih podskupin pa smo naknadno nekoliko skrčili, s čimer smo pridobili preglednost, kakovosti pa nismo bistveno zmanjšali.

Postopek se je izkazal za učinkovito orodje in je odkril tri izrazite podskupine evropskih držav. Pri dveh je bila izražena razlika pri zaupanju v evropske institucije, pri tretji pa so se izpostavile značilno nižje stopnje zaupanja v nacionalne institucije.

Najizrazitejši rezultat se je pokazal pri podskupini osmih držav EU (Ciper, Grčija, Irska, Italija, Madžarska, Portugalska in Romunija), ki jo najbolje opiše značilno višja stopnja zaupanja v več evropskih institucij: najbolj izrazita je razlika v zaupanju v Evropsko računsko sodišče, sledijo Svet ministrov, Socialno-ekonomski odbor, Evropski parlament in Centralna banka, najmanj izrazita, a še vedno značilna razlika pa je pri zaupanju v Evropsko komisijo. Podskupina je odločno ločljiva od preostanka držav EU na podlagi indikatorjev decentraliziranosti, saj mera AUC znaša 0,9. Odločitveno drevo je pokazalo, da je ta podskupina zelo dobro ločljiva od komplementa samo z dvema indikatorjema: z večjim deležem odhodkov glede na BDP (> 42,49 %) in manjšim deležem zaposlenih v javnem sektorju (≤ 29,3 %).

Nasprotno pa je za drugo zanimivo podskupino 13 držav (Avstrija, Belgija, Danska, Estonija, Finska, Francija, Luksemburg, Nemčija, Nizozemska, Slovenija, Španija, Švedska in Velika Britanija) značilno nižje zaupanje v več evropskih institucij: najizrazitejša je razlika pri zaupanju v Evropski socialno-ekonomski odbor, sledita Evropska komisija in svet ministrov, najmanj izrazita, a še vedno značilna razlika pa je pri zaupanju v Evropski parlament in Sodišče. V primerjavi z drugimi te države značilno bolje izkoriščajo storitve e-uprave, bolje zaznavajo stopnjo korupcije v javnem sektorju, funkcioniranje vlade je uspešnejše. Podskupina je dobro (AUC = 0,742) ločljiva od preostanka držav EU na podlagi indikatorjev decentraliziranosti: če je delež zaposlenih v JS (javnem sektorju) večji 28,50 %, gre prav gotovo za eno od omenjenih 13 držav. Med državami, pri katerih je delež zaposlenih v JS manjši kot 28,50 %, pa se predstavnice podskupine od drugih držav najbolj ločijo po tem, da je mejna davčna stopnja, tj. razmerje med spremembo celovitega plačila davka in spremembo davčne stopnje, v večini primerov višja od 46,75 %.

Zadnjo podskupino sestavlja šest držav EU (Bolgarija, Češka, Latvija, Litva, Poljska, Slovaška) z nekoliko nižjo mero AUC (0,631). Državljani predstavnic te podskupine značilno manj zaupajo v nacionalne parlamente, politične stranke, pravosodni sistem, nacionalno vlado in policijo, indeksa zaznavanja korupcije in objektivnih indikatorjev dobrega upravljanja sta značilno nižja, funkcioniranje vlade je značilno slabše. Podskupino torej zaznamuje nezadovoljstvo državljanov do lastnih političnih institucij, izrazito nizek pa je indeks objektivnih

indikatorjev dobrega upravljanja. To pomeni, da je kakovost upravljanja izrazito slaba: gospodarski pogoji so slabi, preveč je omejitev vstopa na trg, davki so previsoki, pogodbe se slabo uveljavljajo, preveč je podkupovanja. Odločitveno drevo opiše podskupino kot tiste države, pri katerih celotna potrošnja države predstavlja manj kot 45,1 % BDP, odhodki pa predstavljajo manj kot 40,06 % BDP.

Algoritem MR-SD je pokazal, da je dobljene podskupine mogoče zadovoljivo dobro ločiti od preostanka držav EU z majhnim številom indikatorjev decentraliziranosti. Med najizrazitejšimi je bila večina podskupin iz sklopa fiskalne decentralizacije, iz drugega sklopa se je pojavil le delež zaposlenih v javnem sektorju. Ker je metoda pokazala, da lahko generira razumljive hipoteze in iz podatkov odkrije določene zakonitosti, se nadejamo, da se bodo poleg obstoječih statističnih metod pri analizi podatkov o upravljanju v bodoče pogosteje pojavljale tudi metode odkrivanja podskupin oziroma širše, metode odkrivanja znanj iz podatkov.