ISSN 1822-8011 (print)
ISSN 1822-8038 (online)
INTELEKTINĖ EKONOMIKA
INTELLECTUAL ECONOMICS
2011, Vol. 5, No. 2(10), p. 224–243

MINING SOCIAL SCIENCE DATA: A STUDY OF VOTING OF THE MEMBERS OF THE SEIMAS OF LITHUANIA BY USING MULTIDIMENSIONAL SCALING AND HOMEGENEITY ANALYSIS

Tomas KRILAVIČIUS[1], Vaidas MORKEVIČIUS[2]

[1] Informatics Faculty, Vytautas Magnus University and Baltic Institute of Advanced Technology, e-mail: t.krilavicius@gmail.com
[2] Policy and Public Administration Institute, Kaunas University of Technology, e-mail: vaidas.morkevicius@gmail.com

Abstract. Multidimensional scaling (MDS) is a well known statistical and data mining technique. It is applicable for an exploratory data analysis and visualization in many different areas, such as economics, especially marketing, credit risk analysis, psychology and computer science. However, it suffers from some serious drawbacks, i.e. it depends on several subjective parameters: choice of data coding, similarity measures and modeling type. We demonstrate these drawbacks in a novel application of MDS analyzing a roll-call voting of the members of Lithuanian Parliament (MPs). We propose using a different technique allowing to escape from the mentioned problems in social science data mining, a homogeneity analysis. We briefly discuss it, illustrate its application on the same data and demonstrate its advantages over MDS. In the paper we concentrate on the technical and methodological aspects of the both methods, therefore, it can be easily reapplied to analyze various economic data, such as customers churn in telecommunications or customers groups in marketing. We discuss all the used tools, coding of votes, similarity measures, division (or non-division) of roll calls into the substantive periods, dimensionality of the solutions of MDS and homogeneity analysis as well as diverse visualization techniques. We compare different visualization techniques of the results of homogeneity analysis where most of the objects in the produced plots represent MPs: 2D and 3D *object plots*; *span plots*, where for each class of objects (in our case a faction) a minimal spanning tree is drawn; 2D and 3D *star plots*, where each object is connected with its class centroid. We conclude with recommendations for researchers modeling social science data and present our future plans regarding voting analysis.

JEL classification: C10, C15, C46, C65.
Keywords: Multidimensional scaling, Homogeneity analysis, Data mining, Data visualization, Similarity measures, Roll-call analysis, Parliamentary voting.
Reikšminiai žodžiai: daugiamačių skalių metodas, homogeniškumo analizė, duomenų gavyba, duomenų vaizdavimas, panašumo matai, balsavimų analizė, balsavimai parlamente.

1. Introduction

Proliferation of information technologies changes all aspects of life, not excluding research methodologies and techniques in technical and social sciences, including economics and political science (King 2009). Data mining provides powerful techniques for analysis of huge amounts of data. Often these techniques are applied in the area of marketing and sales, where companies' posses huge amount of precisely recorded data, that can potentially provide information about customer's loyalty, their behavior (Witten, Frank 2005). It can be applied for credit risk analysis by financial institutions (Žliobaitė, Krilavičius 2009). One of the techniques, often used for exploratory analysis and visualization of data is *Multidimensional Scaling* (MDS) (Shepard 1962a,b; Kruskal 1964a,b). While MDS is a well known classical technique, it suffers from a certain serious drawbacks; namely, it depends on several rigid and subjective parameters: all data should consist of variables measured only in quantitative, ordinal or binary scales, similarity measure and a type of MDS should be chosen by a researcher. The latter two, usually very subjective, choices considerably influence outcomes of analysis. We propose to use homogeneity analysis (de Leeuw, Mair 2009) that treats all data as categorical and makes choices of similarity measures and coding obsolete. This is very attractive when modeling social science data, since many variables in the analyzed data sets are categorical (for example, country, region, municipality, nationality, occupation).

As an illustration and a case study, in this paper we analyze voting behavior of the members of Lithuanian Parliament (MPs). We apply Shepard-Kruskal formulation of non-metric multidimensional scaling (Shepard 1962a,b; Kruskal 1964a,b) and homogeneity analysis (de Leeuw, Mair 2009). Aspects, important for a political science, such as the split of National Revival Party (TPP) faction in the middle of 2009, were discussed elsewhere (see Krilavičius, Morkevičius 2010a,b), and this paper concentrates on the technical and methodological aspects, outlining all the steps of analysis and in such a way making it easy to apply them when analyzing marketing and sales data, or other social science data. We shortly discuss all the used tools, propose and investigate several ways of coding votes and compare MDS solutions obtained using Euclidean and Manhattan similarity measures. We examine the effect of dividing roll calls into substantive periods as well. Further, we explore another very important aspect of the analysis—dimensionality of obtained solutions after performing MDS and homogeneity analysis (Krilavičius, Žilinskas 2008; Hix, Noury, Roland 2006; Hix, Noury 2008; Poole 2005). As a final touch, we experiment with diverse visualization techniques, such as 2D and 3D *object plots*, where each object represents an MP; *span plots*, where for each class of objects (in this case a parliamentary faction) a minimal spanning tree is drawn; 2D and 3D *star plots*, where each object is connected with its class centroid. Such techniques can be applied to visualize results of MDS or other data as well.

Data on MPs voting from the 2008-2012 term of the Seimas of the Republic of Lithuania (LRS) is analyzed, starting from the first plenary session (November 17,

2009) till the beginning of the fifth plenary session (September 9, 2010), overall 4951 roll calls. Analysis demonstrates that results of MDS are dependent upon the chosen similarity measures and coding of the data. In order to overcome instability of solutions due to subjective decisions made by an analyst we propose using homogeneity analysis for voting analysis. This statistical method is suitable for the categorical data which is very common in social science. It makes selection of similarity measures and coding of votes unnecessary. As is expected, dividing roll calls into substantive periods not only makes results more interpretable, but also reveals political dynamics of the Parliament—changes in voting behavior of the political groups and separate MPs over time. We illustrate our results using several different visualization techniques and conclude with recommendation for working with social science data and our future plans.

2. Data

2.1. Data description

In this paper we analyze data on Lithuanian MPs voting from the 2008-2012 term of the LRS. Data for this investigation was collected from the Lithuanian Parliament web page (Seimas 2010) by atviras-seimas.info project (Krilavičius, Cimmperman, Žalandauskas 2010). Only *roll call votes*[1] are considered for a period from 2008 11 17 to 2010 09 09, i.e. from the first plenary session to the beginning of the fifth plenary session. In the analysis we divide the data into three periods reflecting the changes in political composition of the Seimas (splits and establishments of factions, which sometimes results in MPs' moving between the position and opposition)[2]:

1. TPP1, the first plenary session—split of the TPP faction: 2008-11-17–2009-07-15;
2. TPP2, split of TPP faction—establishment of the Christian party (KP) faction: 2009-07-16–2010-02-10;
3. KP, establishment of the Christian party faction – the fourth plenary session: 2010-02-11–2010-09-09.

We summarize some statistics describing the periods in Table 1. We have removed from the analysis MPs with more than 10 missing votes during the analyzed period. The amount of votes was chosen intuitively, i. e. we did not want to remove MPs that have started a bit later because of specifics of the elections in Lithuania (Seimas 2010), but we did not want to include MPs who were not present in the Parliament for a long period as well, because such data would create outliers in our analysis.

[1] A roll call vote is such a vote, when the name of the MP with her voting position is recorded.

[2] The whole period is called ALL and covers data for 2008-11-17—2010-09-09.

Table 1. Voting in Lithuanian Parliament, 2008 – 2012: Descriptive statistics

Title	Period	Total Votes	Number of MPs	Number of MPs with MPs with >10 missing votes removed
All	2008 11 17 2010 09 09	4951	146	132
The first plenary session – TPP split (TPP1)	2008 11 17 2009 07 15	2153	143	134
TPP split – creation of Christian party faction (TPP2)	2009 07 16 2010 02 10	1415	142	138
Creation of Christian party faction – the fifth plenary session (KP)	2010 02 11 2010 09 09	1383	142	141

Periods were selected based on the major changes in the political composition of the Parliament. Therefore, periods reflect different composition of factions and coalitions. We illustrate it in the Tables 2 and 3.

Table 2. Factions in Lithuanian Parliament, 2008-2012

Abbreviation	Faction
AF-TPPF	"Oak" faction, later renamed back to National revival party faction
DPF	Labour party faction
FVL-KPF	Faction "One Lithuania", later renamed into Christian party faction
FTT	Faction "Order and Justice"
KPF	Christian party faction
LSDPF	Lithuanian social democratic party faction
LCSF	Liberal and center union faction, later merged with National revival party faction
LSF	Liberal movement faction
LVLS	Representatives of the Lithuanian peasants' people party
TPPF	National revival party faction
TS-LKDF	Homeland union—Lithuanian Christian democrats faction
Kiti	Other MPs

Table 3. Governmental status of factions in Lithuanian Parliament, 2008-2012

Governmental status	Factions		
	The first plenary session – TPP split	**TPP split – establishment of KP**	**Establishment of KP – the fifth plenary session**
Position	LCSF	LCSF	LCSF
	LSF	LSF	LSF
	TPPF	AF-TPPF	TPPF
	TS-LKDF	TS-LKDF	TS-LKDF
Opposition		FVL-KPF	KPF
	DPF	DPF	DPF
	FTT	FTT	FTT
	LSDPF	LSDPF	LSDPF
	Kiti	Kiti	Kiti

It is not surprising, that Lithuanian Parliament is quite dynamic with respect to its political composition. This is a common feature of all the parliaments in Eastern and Central Europe, the so-called "new democracies" (see Crawford 1996; Pridham, Ágh 2001; White, Batt, Lewis 2007; Hix, Noury 2008). We hypothesize that such dynamics should be reflected in the voting of MPs.

2.2. Coding of votes

Voting behavior of MPs is not just voting for (aye) or against (no) a bill. The possible outcomes when MPs are voting in the Seimas are the following:

1. Aye—voted for the bill or proposal;
2. No—voted against the bill or proposal;
3. Abstain—abstained during the voting;
4. No-vote—registered for the voting, but did not vote;
5. No-partic. —did not participate in the plenary sitting when the voting was taking place.

It is quite obvious that the scale of measurement for voting variable is nominal (Stevens 1946). This poses difficulties in most of the standard analytical data reduction techniques since as an input they require either binary or at least ordinal data. Therefore, in order to perform MDS original data must be mapped into numerical values. This step is insufficiently discussed in the voting data analysis literature. Voting data are usually coded in binary format: 1—Aye, and 0—all the rest (Poole, Rosenthal 1997; Poole 2005; Hix, Noury, Roland 2006; Hix, Noury 2008). This coding scheme is based on the empirical analysis of the voting data in the US Congress, where it was shown that Congressmen's abstentions are rather low and turnout is usually high (Poole, Rosenthal 1997). However, it appears that these results were very culture-specific as in most of the other parliaments around the world (including the Lithuanian Seimas) MPs' abstentions and absenteeism is a rather common practice (Krilavičius, Morkevičius 2010b).

Moreover, this mapping scheme is not adequate in other contexts of applications where data is nominal and non binary.

Therefore, in this section we briefly state how the original data were recoded (into ordinal scales) for MDS (for an extended overview of different mappings see Krilavičius 2007; Krilavičius, Žilinskas 2008; Morkevičius 2008; Morkevičius, Krilavičius 2009; Krilavičius, Morkevičius 2010a,b) and compare an effect of the different codings to the results of MDS. The summary of some of the voting data (re)coding schemes is presented in Table 4.

Table 4. Schemes of coding of votes

Vote	Mapping		
	PBDB Most Popular	**PBDC "Better" Alternative**	**PBDA "Optimal" Alternative**
Aye	1	1	1
No-vote	0	0	0
No-partic.	0	0	0
Abstain	0	-1	0
No	0	-2	-1

Different mappings correspond to the different theoretical interpretations of the MPs behavior. For example, in the most popular analytical strategy DW-NOMINATE votes are recorded as "aye" (1) and "all the others" (0) on the assumption that other voting outcomes are rare (see the discussion above). This assumption might be true for the US Congress or Senate, but it clearly does not correspond to the reality in most of other parliaments around the world including Lithuanian Seimas (Hix, Noury 2008; Krilavičius, Morkevičius 2010a,b). Therefore, numerical transformations of the original data must be clearly specified along the lines of empirical evidence in order to perform MDS or any other similar statistical analysis (cluster analysis, factor analysis, unfolding analysis etc., see Krzanowski, 2007). We use three coding schemes defined in table 4 for coding voting data for MDS. As an optimal alternative for the Lithuanian Parliament we propose an ordinal level of measurement with the "aye" (1) on one end of the scale and "no" (-1) at the other. All other outcomes we interpret as "something in between the two end points", "neither clearly aye, neither clearly no" and code accordingly (0). Clearly, this scale (this also applies to the other two coding schemes) allows only non-metric MDS to be performed on the data. All these considerations become unnecessary when performing *homogeneity analysis*, (see section 4 for the details) a technique that allows using the original nominal scale data in the analysis.

2.3. Tools and data representation

We use free open source tools for extracting the data for the analysis. Data of voting of MPs are obtained from the LRS web site (www.lrs.lt) and stored in MySQL database

(MySQL Oracle 2010) that is provided by atviras-seimas.lt (Krilavičius, Cimmperman, Žalandauskas 2010). Data transformation and analysis are performed using statistical analysis tool R (R Team 2010) and its libraries MASS (Venables, Ripley 2002) and HOMALS (de Leeuw, Mair 2009), for MDS and homogeneity analysis, respectively.

We store voting data in *voting matrix* $V(K,S)$, where K is a number of MPs and I is a total number of roll calls. Each $V(k,i)$ represents a voting result of kth MP at ith roll call. For MDS, original roll call data is transformed into numerical scales presented in Table 4, while for HOMALS we use numerical representation just to speed-up the tools.

2.4. Dissimilarity Measures

We believe that the political structure of the Parliament may be reconstructed from MPs' voting data by performing MDS analysis employing different proximity/dissimilarity measures. To uncover the structure of the Parliament based on the voting patterns of MPs we treat them as points (objects) in S-dimensional vector space, and dissimilarity among MPs is defined by Minkowski (p-norm) distance (1).

$$d_p(k,m)=\left(\sum_{i=1}^{S}|V(k,i)-V(m,i)|^p\right)^{1/p} \tag{1}$$

where V is the voting matrix and $p \geq 1$ is a norm type.

In the analysis we use two most common dissimilarity measures:

- Manhattan (also known as city block) or 1-norm distance (2)

$$d_1(k,m)=\sum_{i=1}^{S}|V(k,i)-V(m,i)| \tag{2}$$

- Euclidean or 2-norm distance (3)

$$d_2(k,m)=\sqrt{\sum_{i=1}^{S}|V(k,i)-V(m,i)|^2} \tag{3}$$

See Žilinskas and Žilinskas (2007) for the details on measures impact on the MDS results.

3. Multidimensional Scaling

Multidimensional scaling is a well known technique (Torgerson 1958; Shepard 1962a,b; Kruskal 1964a,b) for dimensionality reduction of a matrix. It is based on the pair-wise similarity of the objects. In our case the objects (MPs) are represented by the points in a multidimensional vector space where dissimilarity is measured by Manhattan or Euclidean distance between them. Therefore, voting data matrix is transformed to a dissimilarity matrix of the form $D_p = (d_p(k,m))$, where $d_p(k,m)$ is dissimilarity between

*k*th and *l*th MPs measured by Manhattan ($p = 1$) and Euclidean ($p = 2$) distances.

MDS maps the set of considered objects to the low dimensional space of images where objects are represented by points and distances among points describe their dissimilarity. Usually, objects are mapped onto a two- or three-dimensional space.

Let dissimilarity between pairs of K objects is given by the matrix $D = (d(k.m)), k, m = 1,\ldots,K$ and $d(k.m) = d(m.k)$. Points $z_i \in R^r, i = 1,\ldots,K$ in r-dimensional space of images should be found such that their inter-point distances fit the given dissimilarities. Different accuracy measures can be chosen to define different images of analyzed objects. If the objects are points in a high dimensional space, the images can be interpreted as non-linear projections of the set of points from the higher to the lower dimensional space. Images construction problem usually is reduced to minimization of badness of fit criterion, e.g. commonly used least squares function STRESS (4)

$$\text{STRESS}(Z) = \sum_{i<j} w_{ij} \left(\delta_{ij}(Z) - d(i,j)\right)^2 \tag{4}$$

where $Z = (z_1, \ldots, z_{K1}, z_2, \ldots, z_K)'$ is a vector of coordinates of images, $\delta_{ij}(X)$ is the distance between *i*th (z_i) and *j*th (z_i) points in image space, and $w_{ij}, i, j = 1,\ldots,K$ are positive weights. Given STRESS formula defines a class of minimization criteria. It allows selecting different distances, in our case, Manhattan and Euclidean distances. See Žilinskas and Žilinskas (2007) for the details on distances impact on the MDS results.

We use a less ambitious approach, namely, Shepard-Kruskal formulation of non-metric multidimensional scaling that attempts only to approximate the ranks of the dissimilarities, i.e. inter-point distances between images approximate a monotonic transformation of the original dissimilarities. Stress function (5) is reformulated in the following way

$$\text{STRESS}(Z) = \sqrt{\frac{\sum_{i,j} \left(\delta_{ij}(Z) - d(i,j)\right)^2}{\sum_{i,j} \delta_{ij}(Z)}} \tag{5}$$

In our investigation we apply a standard R routine *isoMDS* (Venables, Ripley 2002) based on the Shepard-Kruskal formulation of non-metric MDS.

4. Homogeneity analysis

Homogeneity analysis (de Leeuw, Mair 2009) is a variant of *multiple correspondence analysis* (MCA) technique that minimizes the divergence from the homogeneity defined by a loss function almost identical to STRESS function in MDS, but adapted for categorical features.

It can be defined in a following manner (see de Leeuw, Mair 2009). For $k = 1,\ldots,K$ objects, data on I categorical features is collected where each of the $i = 1,\ldots,S$ features takes on l_i different values (categories). It is coded using $K \times l_i$ binary indicator (dummy) matrices G_i, that are collected to a block matrix $G = [G_1 \vdots G_2 \vdots \cdots \vdots G_I]$. Missing values

are coded as zero sum rows, i.e. if object k is missing for feature i, then the row sum of G_j is 0, otherwise the sum is 1, because the category entries are disjoint. All row sums are used to construct a diagonal matrix M_i for each feature, where each diagonal element $M_i(k,k)$ is 0, if object k has a missing value on feature i, and 1 otherwise. Let $M_\bullet$ be the average of M_i.

The goal is the same, as in MDS case, i.e. objects should be mapped from R^S to R^r. Let $X(K,r)$ is a matrix containing objects mapped to *r*-dimensional space and Y_i is the $l_i \times r$ matrix containing the category quantifications. The following loss function is used (6):

$$\sigma(X;Y_1,\ldots,Y_S)=\frac{1}{S}\sum_{i=1}^{S} tr\left(X-G_iY_i\right)' M_i\left(X-G_iY_i\right) \qquad (6)$$

with normalization conditions $\mathrm{u}' M_\bullet X = 0$ and $X'M_\bullet X = 0$ to avert trivial zero values solutions.

Alternating least squares (ALS) algorithm is applied. The idea is following: (a) at the zero iteration we begin with an initial solution $X^{(0)}$ and update category quantifications $Y_i^{(1)}$; in the following step we update $X^{(1)}$ and normalize them; then we continue with updating category quantifications and object scores until loss function difference between two iterations becomes lower the a specified threshold ε.[3]

Homogeneity analysis might be performed with different statistical packages (SPSS, R). Here we present analysis performed with an open source package "homals" (de Leeuw, Mair 2009) implemented in the statistical analysis platform R (R Team 2010). This package has the advantages over other implementations since it includes multiple visualization options, which can further be extended by the analysts themselves.

5. Results

5.1 MDS

First, we present results of MDS application. We implemented Shepard-Kruskal formulation of non-metric MDS on differently coded data (3 coding schemes, see section 2.2 and Table 4) using different dissimilarity measures (Euclidean and Manhattan, see section 2.4) for the different periods of voting in Lithuanian Parliament (4 periods, see Table 1). In total, we obtained 24 solutions, which were further analyzed and interpreted. Our main goals in this section are to evaluate 1) dimensionality of the solutions, 2) impact of the different dissimilarity measures on the configuration of objects in the solutions, and 3) influence of the different codings of data on the configuration of objects in the solutions, and an interpretation of some of the results for the different periods of MPs' voting under study.

In order to evaluate the dimensionality of the solutions we present a *scree plot*, where STRESS values of different solutions are plotted against the number of dimensions (see Figure 1). The plot shows that quite unexpectedly STRESS values for the

[3] We refer an interested reader to de Leeuw and Mair (2009) for the details.

last period under study are higher than for the others (see section 2.1 for definitions of periods). This goes contrary to our preliminary hypothesis that the highest values of STRESS should be expected for the whole period under study since many of the MPs were moving from one faction to the other and some factions moved between position and opposition. Therefore, the data for the whole period should have contained more contradictory ("noisy") entries (votes) than data for the substantive periods divided according to the major political events in the Seimas. However, this is only true for the two dimensional solution. Notwithstanding the above, starting from the third dimension STRESS values converge to very similar ones for all the solutions.

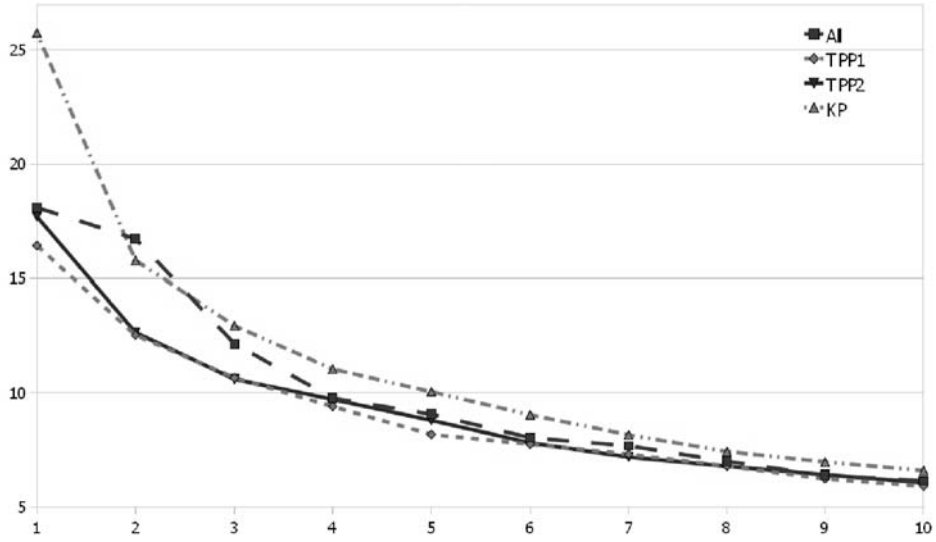
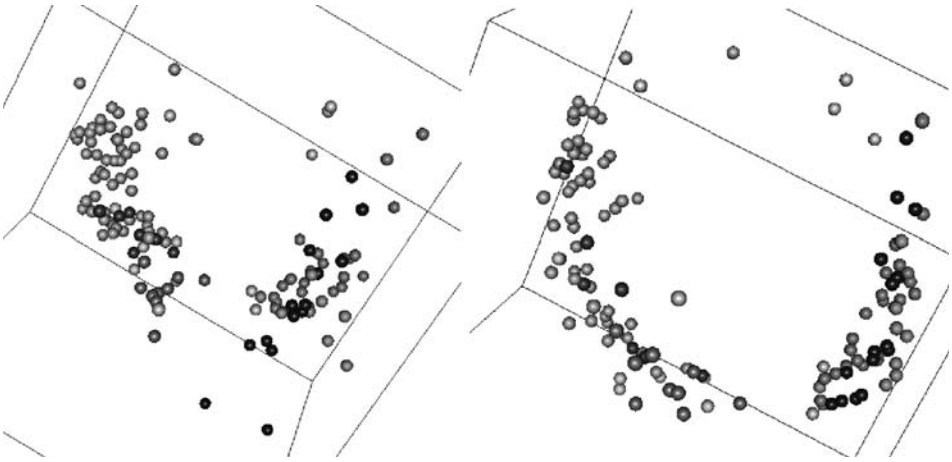


Figure 1. Scree plot for dimensionality of voting data from the Lithuanian Parliament, 2008-2012 (Euclidean dissimilarity measure, data coded as PBD_B, see Table 4)

STRESS values stop descending steeply at dimension 3 in all the solutions, which we interpret as an optimal number of dimensions for all the data. This result is similar to other results obtained in other countries similar to Lithuania (see Hix, Noury 2008).

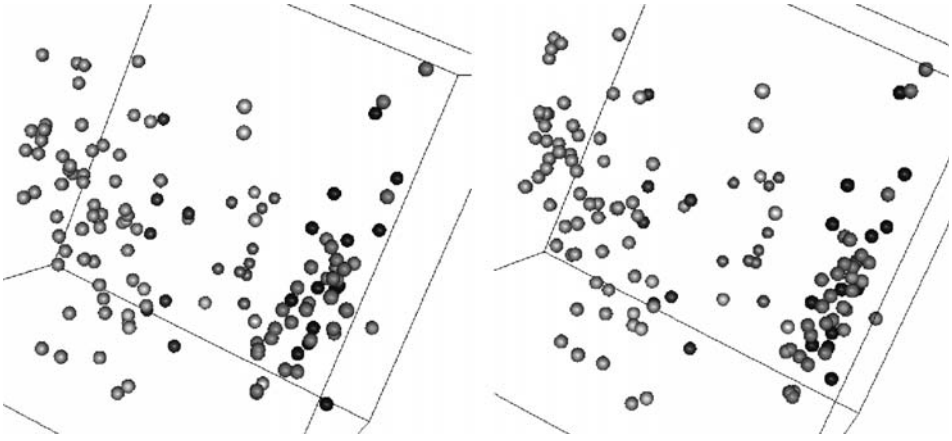
For evaluating the impact of different dissimilarity measures we tried to compare configuration of objects (MPs) in results obtained using Euclidean and Manhattan dissimilarity measures. The results indicate that different dissimilarity measures have certain influence on the spatial positions of objects (see figures 2-3). Even though general trends seem to be the same, objects are more dispersed using Manhattan dissimilarity measure. Moreover, configurations of points are somewhat different.



1) 2)

Figure 2. Object (MPs) plots of voting data from the Lithuanian Parliament, 2008-2012 (data for the whole period, Euclidean dissimilarity measure, data coded as 1) PBD_A and 2) PBD_B, see Table 4, colors indicate different factions)

For evaluation of the influence of the different codings of data on the configurations of objects in the solutions we again compared configurations of objects (MPs) in the results obtained from the data coded using different schemes (see Table 4). The results show that mapping of the original data has certain influence on the configuration of objects in space (see figures 2-3). Again, although the configurations of objects are similar between the solutions, one can spot some, seemingly random, movements of some of the MPs between the configurations.



1) 2)

Figure 3. Object (MPs) plots of voting data from the Lithuanian Parliament, 2008-2012 (data for the whole period, Manhattan dissimilarity measure, data coded as 1) PBD_A and 2) PBD_B, see Table 4, colors indicate different factions)

Finally, we present results obtained from the voting data for different substantive periods looking for meaningful configurations of objects in the solutions. The results reveal several major trends (see figure 4):

1) Objects in the configuration can be clearly divided into two opposing blocks – position (left) and opposition (right), with nonaligned MPs placed in between the two sides. Interestingly, MPs from the National revival party faction (TPPF) and (the later established) Christian party faction (KPF) are also placed between the two poles.

2) The second dividing line seems to be between the conservatives (upper part) and liberals (lower part). However, this interpretation needs to be confirmed in more elaborate further studies, where issues of voting are (more) clearly identified (for a possible strategy see Morkevičius 2009).

3) The only clear dynamic aspect of the data seems to be movement of the splinter MPs from TPPF (later members of KPF) into the side of opposition.

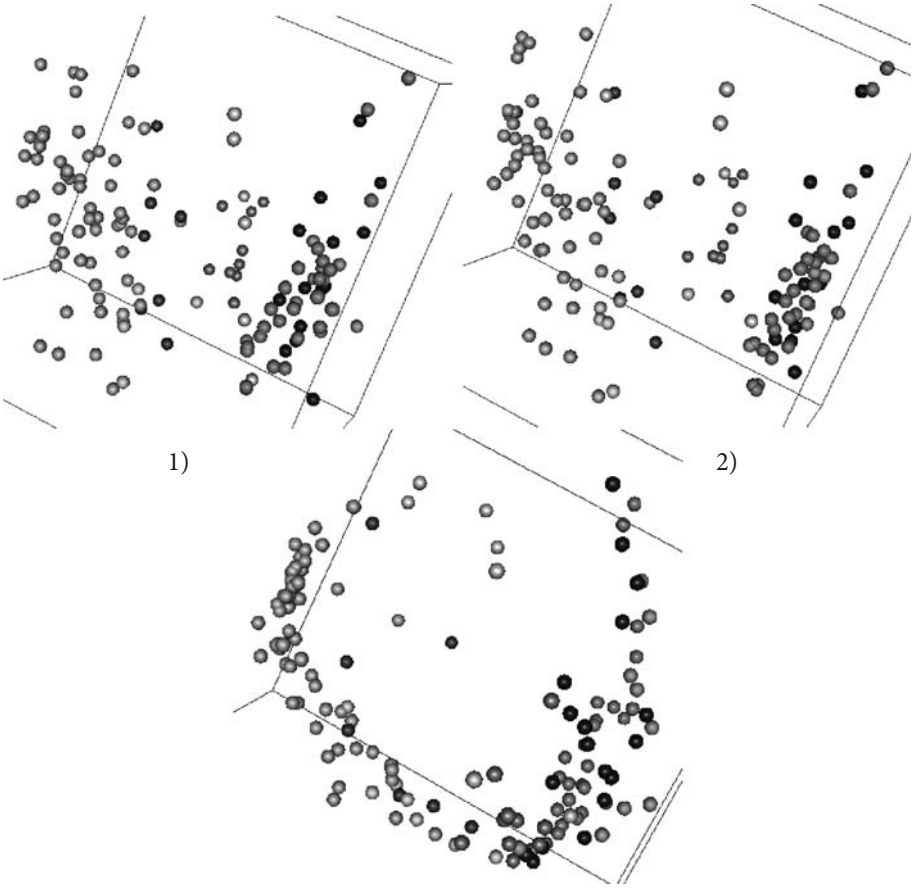


Figure 4. Object (MPs) plots of voting data from the Lithuanian Parliament, 2008-2012 (data divided into substantive periods: 1) TPP1, 2) TPP2, 3) KP, see Tables 1 and 3, Manhattan dissimilarity measure, data coded as PBD_B, see Table 4, colors indicate different factions)

Based on the MDS analysis results we conclude with two theses:

1) Different variants of data coding and dissimilarity measures strongly influence configurations of points in the solutions of MDS. Such variance of results makes their substantive interpretation very subjective, dependent on the data analytic methods chosen. Moreover, usually, criteria for choosing the appropriate parameters of the models are difficult to substantiate empirically or theoretically. Consequently, choosing the best method or strategy for analysis becomes complicated as well.

2) Dividing data into politically meaningful periods does not impact results of MDS considerably, compared to influence of data coding and dissimilarity measures, which is a counter-intuitive result.

Therefore, we do not recommend using MDS for the exploratory data mining in social sciences since researchers need to make several important decisions during the analysis that have considerable influence on the results. Our recommended strategy is using homogeneity analysis (see Section 5.2), since in most cases it allows avoiding introduction of unwanted subjectivity and instability into the analysis.

5.2 Homogeneity analysis

In this section we present results of homogeneity analysis. As it was mentioned in sections 2 and 4, this method does not require much preparatory work in order to be able to implement the analysis. Moreover, the algorithm itself does not require specifying any important parameters. This is due to the fact that homogeneity analysis can be performed on categorical data (for example, on voting data). Our main goals in this section are twofold: (1) to evaluate dimensionality of the solutions, and (2) to present an interpretation of results of dimensional reduction for the different periods of MPs' voting under study.

In order to evaluate the dimensionality of the solutions we present a scree plot where eigenvalues for data from the different substantive periods are plotted against the number of dimensions (see figure 5). The plot shows only marginal differences between the periods. Again, eigenvalues stop descending steeply at dimension 3 in all the homogeneity analysis solutions which we interpret as an optimal number of dimensions for all the data (see also Hix, Noury 2008).

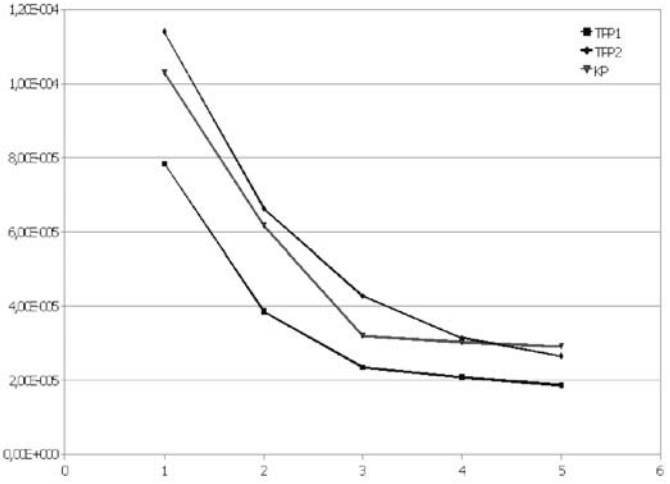
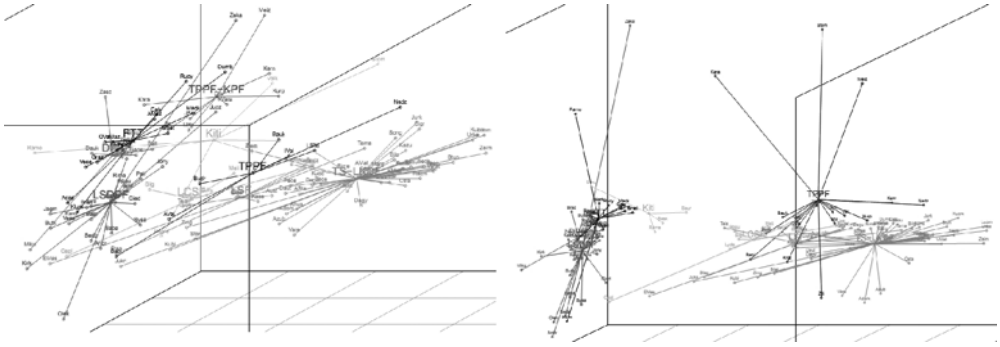


Figure 5. Eigenvalues of solutions with different number of dimensions for MPs' voting data in the Lithuanian Parliament, 2008-2012

Further, we present substantive results obtained for the different politically meaningful periods trying to interpret configurations of objects in the solutions. The results indicate several major trends (see figures 6-7):

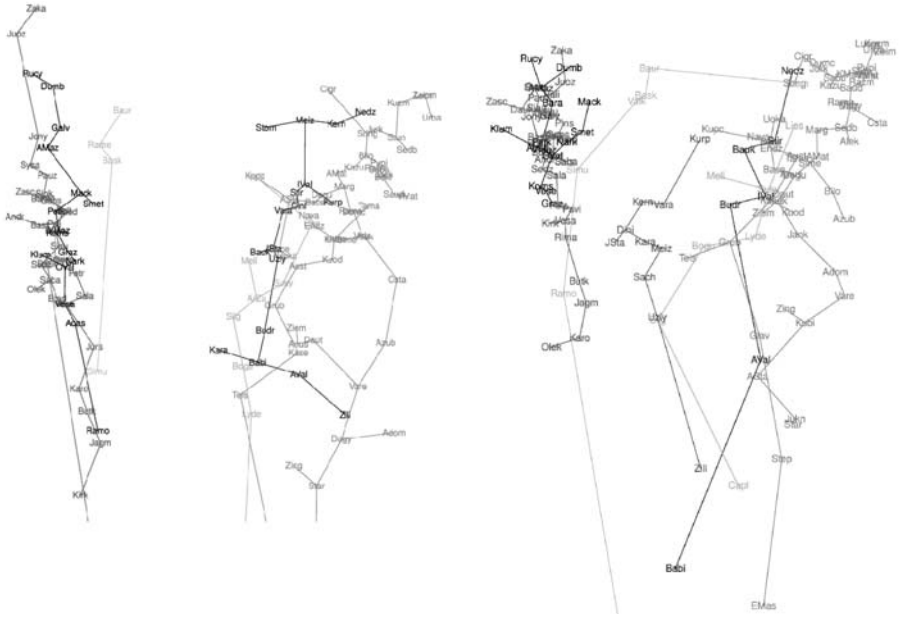
1) Star plots show that data for the whole period under study is more "noisy" and dispersed, whereas objects (MPs) in the star plots for different substantive periods are more tightly attached to their centroids. Therefore, division of data into meaningful periods makes results more interpretable.

2) Similar to MDS results, objects (MPs) in the configuration can be clearly divided into two opposing blocks—position (right) and opposition (left)—with nonaligned MPs in between the two sides. Again, one of the dimensions seems to reflect liberal-conservative divide, however, further studies are needed (see Morkevičius 2009).

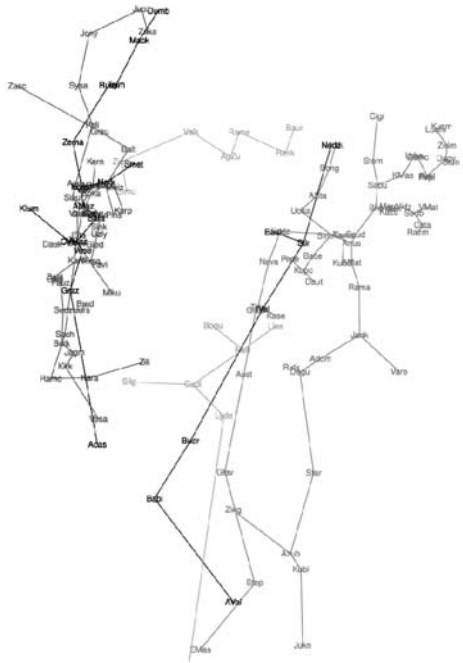


1) 2)

Figure 6. Star plots of MPs' voting data in the Lithuanian Parliament, 2008-2012 (data for the whole period (1) and for the first substantive period TPP1 (2), see Tables 1 and 3 for definitions)



1) 2)



3)

Figure 7. Span plots of MPs' voting data in the Lithuanian Parliament, 2008-2012 (1st and 2nd dimensions, data for substantive periods: TPP1 (1), TPP2 (2) and KP (3), see Tables 1 and 3 for definitions, colors indicate different factions)

3) Looking at the span plots we see that they rather "neatly" portray political dynamics within the Parliament: MPs that splintered from TPPF moves towards the opposition side in the second period (TPP2) and appeared among the opposition in the third (KP). Similarly, MPs from the Lithuanian peasants' people party remain between the position and opposition even after signing an agreement with the ruling coalition. Since the agreement was only partial, these representatives were free to vote on the majority of bills and initiatives (some of which reflect support for the position, and others—support for the opposition).

4) Shifting political preferences of some of MPs were clearly captured by the analysis. For example, Saulius Stoma even when he was a member of different factions, he was voting in the same way as the majority of conservative MPs and finally joined this faction. Similarly, voting analysis shows that Žilvinas Šilgalis was steadily "moving" towards the center of the political spectrum even when he was a member of Liberal and center union faction (LCSF) and, consequently, left the faction.

All in all, the results of homogeneity analysis, especially well illustrated with more advanced and refined visualization techniques, are much more consistent and much better represent political dynamics of the Parliament. This allows us to conclude that homogeneity analysis is a much better suited method for voting analysis in the Lithuanian parliament. Moreover, this method seems to be better suited for social science researchers performing exploratory data mining than MDS: it eliminates much of the subjectivity from the analysis.

6. Conclusions, Recommendations and Future Plans

After performing MDS and homogeneity analysis with voting data of the Lithuanian Parliament we conclude by recommending using homogeneity analysis in the similar studies (see also Desposato 1997). First, researchers employing this method escape difficult decisions of choosing an appropriate data coding (or recoding) scheme and dissimilarity measure. Second, configurations of objects are more meaningful and interpretable when performing homogeneity analysis. Homogeneity analysis (and its counterpart, multiple correspondence analysis) could be useful not only for political science data analysis, but for marketing and sales, financial, economic or other social sciences data. It provides excellent means for working with mixed categorical, ordinal and numerical data, which can also be hierarchically structured (Escofier, Pagès 1998; Pagès 2002; Le Dien, Pagès 2003; Pagès 2004) and, therefore, could be very useful for credit risk analysis (Žliobaitė, Krilavičius 2009) and other economic and social data analysis tasks.

Moreover, we advise using appropriate visualization techniques, e.g. star plots and span plots, which reveal various important aspects of the results by including additional information about the objects.

In the future research we will extend our study in several important directions. First of all, interpretation of dimensions should be performed more systematically. In this paper we only hypothetically discussed possible meaning of the dimensions. Further study requires performing experiments with a sample of votes representing

different periods and clearly defined issues (ideological positions). Preliminary analysis was already attempted by one of the authors (see Morkevičius 2009).

Additionally, we intend to perform an analysis "inside" the different factions searching for patterns of voting and ideological positions among the members of like-minded groups. It would be interesting to further explore cohesion of factions and "dividing lines" within them.

Finally, development of analytical tools and ever increasing computer power will allow in the future performing a real-time analysis of parliamentary voting. This undertaking would allow closely follow political dynamics of the Seimas and detect important changes (as they are occurring) of political behavior of MPs, on the basis of which we could predict their future "movement" between the factions. This would also allow monitoring political behavior of the whole factions and predicting changes of their position regarding participation in the governing coalitions.

References

1. Crawford, K. 1996. East Central European Politics Today: From Chaos to Stability? Manchester and New York: Manchester University Press.
2. Burt, C. 1950. The Factorial Analysis of Qualitative Data. British Journal of Psychology, Statistical Section, 3: 166-185.
3. de Leeuw, J. and Mair, P. 2009. Gifi Methods for Optimal Scaling in R: The Package homals. Journal of Statistical Software, 31(4): 1-20. <http://www.jstatsoft.org/v31/i04>.
4. Desposato, S. W. 1997. Estimating Legislators' Ideal Points Using HOMALS. Poster presentation for the 14th Annual Political methodology summer conference, Ohio State University, July 23-27, 1997.
5. Escofier, B. and Pagès, J. (1998). Analyses factorielles simples et multiples, 3e edition [Simple and multiple factorial analysis, 3rd ed.]. Dunod. Paris.
6. Hix, S. and Noury, A. 2008. Government-Opposition or Left-Right? The Institutional Determinants of Voting in Fourteen Parliaments. Working paper, version 1, May 2008 <http://personal.lse.ac.uk/hix/Working_Papers/Hix-Noury-GOorLR-01May08.pdf>.
7. Hix, S., Noury, A. and Roland, G. 2006. Dimensions of Politics in the European Parliament. American Journal of Political Science, 50(2): 494-520.
8. Young, F. W. 1981. Quantitative Analysis of Qualitative Data. Psychometrika, 46(4): 357-388.
9. Young, F. W., de Leeuw, J. and Takane, Y. 1980. Quantifying Qualitative Data, in Similarity and Choice, eds. E. D. Lanterman & H. Feger. Bern: Hans Huber, 150-179.
10. King, G. 2009. The Changing Evidence Basis of Social Science Research, in The Future of Political Science: 100 Perspectives, eds. G. King, K. L. Schlozman, and N. Nie. New York: Routledge, 91-93.
11. Krilavičius, T. and Morkevičius, V. 2010a. Kaip balsuoja naujasis Seimas? Erdvinių konfigūracijų sudarymas ir interpretacija [How the New Seimas is Voting? Constructing and Interpreting Spacial Configurations], in Theses of the Lithuanian Social Science Forum 2010, ed. A. Ramonaitė. Vilnius, 14-15.
12. Krilavičius, T. and Morkevičius, V. 2010b. Lietuvos parlamentarų ideologinės pozicijos ir jų raida 2008-2012 m. kadencijos Seime: statistinė Seimo narių balsavimų analizė [Ideological positions of Lithuanian MPs and their dynamics in the term 2008-2012:Statistical Analysis

of MPs Voting]. Paper presented at the annual conference of the Lithuanian Political Science Association, 2010 11 20, Vilnius.
13. Krilavičius, T. and Žilinskas, A. 2008. On Structural Analysis of Parliamentarian Voting Data. Informatica 19, 3 (August 2008), 377-390. <http://www.mii.lt/informatica/pdf/INFO727.pdf>.
14. Krilavičius, T., Cimmperman, P. and Žalandauskas, T. 2010. Duomenų užteks visiems [Plenty of data for everyone]. Poster at the annual conference of the Lithuanian Political Science Association, 2010 11 20, Vilnius. < http://www.atviras-seimas.info>.
15. Kruskal, J. B. 1964a. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika, 29(1): 1-27.
16, Kruskal, J. B. 1964b. Nonmetric Multidimensional Scaling: A Numerical Method. Psychometrika, 29(2): 115-129.
17. Krzanowski, W. J. 2007. Statistical Principles and Techniques in Scientific and Social Investigations. Oxford, New York: Oxford University Press.
18. Le Dien, S. and Pagès, J. 2003. Analyse factorielle multiple hiérarchique [Hierarchical multiple factorial analysis]. Revue de Statistique Appliquée, 51(2): 47-73.
19. MySQL Oracle. 2010. MySQL.com. < http://www.mysql.com>.
20. Morkevičius, V. 2009. Neideologinis Seimas? Statistinė svarbių 2004-2008 m. kadencijos Lietuvos Seimo balsavimų analizė [Non-Ideological Seimas? Statistical Analysis of Voting on Important Bills during the 2004-2008 Term of the Seimas of Lithuania], in Partinės demokratijos pabaiga? Politinis atstovavimas ir ideologijos [The End of Party Democracy? Political Representation and Ideology], ed. A. Ramonaitė. Vilnius: Versus aureus, 53-87.
21. Morkevičius, V. and Krilavičius, T. 2009. Kaip analizuoti Seimo narių balsavimus? Metodologinės pastabos [How to Analyze Voting of Members of the Seimas?
22. Methodological Remarks], in Theses of the Lithuanian Social Science Forum 2009, ed. A. Ramonaitė. Vilnius, 16-17.
23. Pagès, J. 2002. Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes [Multiple factorial analysis for qualitative variables and mixed data].
24. Revue de Statistique Appliquée, 50(4): 5-37. Pagès, J. 2004. Analyse factorielle de données mixtes [Factorial analysis of mixed variables]. Revue de Statistique Appliquée, 52(4): 93-111.
25. Poole, K. T. & Rosenthal, H. 1997. Congress: A Political-Economic History of Roll Call Voting, New York: Cambridge University Press.
26. Poole, K. T. 2005. Spatial Models of Parliamentary Voting, New York: Cambridge University Press.
27. Pridham, G. and Ágh, A. eds. 2001. Prospects for Democratic Consolidation in East-Central Europe, Manchester and New York: Manchester University Press.
28. R Team. 2010. The R Project for Statistical Computing. <http://www.r-project.org>.
29. Seimas 2010. Seimas of the Republic of Lithuania. URL: http://www.lrs.lt.
30. Shepard, R. N. 1962a. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. Psychometrika, 27(2): 125–140.
31. Shepard, R.N. 1962b. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II. Psychometrika, 27(3): 219–246.
32. Stevens, S. S. 1946. On the Theory of Scales of Measurement. Science, 103(2684): 677-680. Tenenhaus, M. and Young, F. W. 1985. An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data. Psychometrika, 50(1): 91-119.

33. Torgerson, W. S. 1958. Theory and Methods of Scaling, New York: John Wiley & Sons.
34. Venables, W. N. and Ripley B. D. 2002. Modern Applied Statistics with S, 4th ed., New York: Springer.<http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/isoMDS.html>.
35. White, S., Batt, J. and Lewis, P.G. eds. 2007. Developments in Central and East European Politics 4, Houndmills, Basingstoke and New York: Palgrave Macmillan.
36. Witten, I.H and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.
37. Žilinskas, A. and Žilinskas, J. 2007. Two Level Minimization in Multidimensional Scaling. Journal Of Global Optimization, 38(4): 581-596.
38. Žliobaitė, I. and **Krilavičius T. 2009**, CLAN: Clustering for Credit Risk Assessment. An entry to PAKDD 2009 Data Mining Competition, 4 p. (unreviewed).

SOCIALINIŲ MOKSLŲ DUOMENŲ GAVYBA: LIETUVOS RESPUBLIKOS SEIMO NARIŲ BALSAVIMO ANALIZĖ NAUDOJANT DAUGIAMAČIŲ SKALIŲ METODĄ IR HOMOGENIŠKUMO ANALIZĘ

Tomas KRILAVIČIUS
Informatikos fakultetas, Vytauto Didžiojo universitetas ir Baltijos pažangių technologijų institutas
Vaidas MORKEVIČIUS
Politikos ir viešojo administravimo institutas, Kauno technologijos universitetas

Santrauka. Daugiamačių skalių metodas (MDS) yra gerai žinomas statistikoje ir duomenų gavyboje. Jis gali būti taikomas tiriamajai duomenų analizei ir rezultatų vaizdavimui daugelyje sričių, pvz. ekonomikoje, ypač marketinge, kredito rizikos analizėje, psichologijoje ir informatikoje. Deja, šis metodas turi ir tam tikrų trūkumų – jis priklauso nuo keleto subjektyviai parenkamų parametrų: duomenų kodavimo būdų, panašumo matų ir modeliavimo tipų. Šiame straipsnyje mes atskleidžiame MDS trūkumus, pritaikydami jį naujame kontekste, t. y. analizuodami Lietuvos Respublikos Seimo (LRS) narių balsavimus. Taip pat siūlome duomenų analizės metodą, leidžiantį išvengti minėtų problemų – homogeniškumo analizę. Straipsnyje trumpai apžvelgiamas šis metodas ir pademonstruojamas jo efektyvumas taikant jį tiems pat duomenims.

Straipsnyje taip pat gana detaliai aptariami techniniai ir metodologiniai darbo aspektai, kad mūsų pademonstruotus metodus būtų galima lengvai pritaikyti kitose srityse, pvz. analizuojant ekonominius duomenis – klientų kaitą ryšio paslaugų bendrovėse ar klientų grupavimą marketinge. Aprašomi ir visi darbo etapai: naudoti įrankiai, balsavimų kodavimas, panašumo įvertinimo matai, balsavimų (ne)skaidymas į prasmingus periodus, MDS ir homogeniškumo analizės sprendinių dimensijų skaičiaus analizė bei įvairūs gautų rezultatų vaizdavimo būdai. Taip pat aptariami bei lyginami skirtingi homogeniškumo analizės rezultatų vaizdavimo metodai: objektų[1] atvaizdavimas 2-matėje ir 3-matėje erdvėse (angl. *object plot*), minimalaus jungimo medis objektams (angl. *span plot*), objektų centroidų jungtys su objektais 2-matėje ir 3-matėje erdvėje (angl. *star plot*), Voronojaus mozaikos (angl. *Voronoi plot*) ir kiti.

Straipsnis baigiamas rekomendacijomis darbams su socialinių mokslų duomenimis bei tolimesniais tyrimo planais.

Tomas Krilavičius is an associate professor at Vytautas Magnus University and a senior research fellow at Baltic Institute of Advanced Technologies. His main research interests are formal methods, information retrieval and applications of quantitative techniques in social sciences. He defended his PhD thesis Hybrid Techniques for Hybrid Systems in 2006, Twente University, The Netherlands.

Vaidas Morkevičius is a research fellow at Policy and Public Administration Institute, Kaunas University of Technology. His main research interests are text analytics, survey methodology and data analysis, qualitative comparative analysis and parliamentary analysis. He was analyzing content of Lithuanian parliamentary debates in his doctoral thesis defended in 2006 at Kaunas University of Technology. Currently, he is working on the project which aims to establish Lithuanian data archive for social sciences and humanities (LiDA) where he is responsible for coordination of major international survey research projects (European Social Survey and others).

Tomas Krilavičius dirba docentu Vytauto Didžiojo universitete ir vyresniuoju mokslo darbuotoju Baltijos pažangių technologijų institute. Pagrindinės jo tyrimų kryptys yra formalūs metodai, informacijos paieška ir kiekybinių metodų taikymas socialiniuose moksluose. Jis apgynė daktaro disertaciją „Hibridinės technikos hibridinėms sistemoms“ 2006 metais, Twente universitete, Nyderlanduose.

Vaidas Morkevičius yra Kauno technologijos universiteto Politikos ir viešojo administravimo instituto mokslo darbuotojas. Pagrindinės jo tyrimų kryptys yra teksto analitika, socialinių apklausų metodologija ir duomenų analizė, kokybinė lyginamoji analizė bei parlamento analizė. Savo daktaro disertacijoje, apgintoje 2006 m. Kauno technologijos universitete, jis analizavo Lietuvos Seimo debatų turinį. Šiuo metu jis dirba projekte, kurio tikslas – sukurti Lietuvos humanitarinių is socialinių mokslų duomenų archyvą. Projekte jis atsakingas už Europos socialinio tyrimo įgyvendinimo Lietuvoje koordinavimą.