INFORMATICA, 2003, Vol. 14, No. 4, 471–486

Data Analysis Strategy for Revealing Multivariate Structures in Social-Economic Data Warehouses

Dale DZEMYDIENE

Department of Software Engineering, Institute of Mathematics and Informatics
Akademijos 4, LT-2021 Vilnius, Lithuania
Department of Legal Informatics, Law University of Lithuania
Ateities 20, LT-2057 Vilnius, Lithuania
e-mail: daledz@ktl.mii.lt

Vitalija RUDZKIENE

Department of Legal Informatics, Law University of Lithuania
Ateities 20, LT-2057 Vilnius, Lithuania
e-mail: vital@ltu.lt

Received: June 2003

Abstract. This research work is aimed at the development of data analysis strategy in a complex, multidimensional, and dynamic domain. Our universe of discourse is concerned with the data mining techniques of data warehouses revealing the importance of multivariate structures of social-economic data which influence criminality. Distinct tasks require different data structures and various data mining exercises in data warehouses. The proposed problem solution strategy allows choosing an appropriate method in recognition processes. The ensembles of diverse and accurate classifiers are constructed on the base of multidimensional classification and clusterisation methods. Factor analysis is introduced into data mining process for revealing influencing impacts of factors. The temporal nature and multidimensionality of the target object is revealed in dynamic model using multidimension regression estimates. The paper describes the strategy of integrating the methods of multiple statistical analysis in cases, where a great set of variables is observed in short time period. The demonstration of the data analysis strategy is performed using real social and economic development of data warehouses in different regions of Lithuania.

Key words: multidimensional statistical methods, data mining, data warehouse, social-economical indicators, criminality.

1. Introduction

Data mining methods make progress from a simple discovery of frequent patterns and regularities toward knowledge-based and interactive decision support processes. In the last decade using multivariate data analysis methods became a topic of interest in the field of data mining. However there is a lack of concrete algorithms how to apply these methods in the cases, where a great set of variables is observed in a short time period. Condition constraints of this kind are frequently met in the analysis and prediction of socio-economic developments. Many social and human-initiated events deal with incomplete or limited by nature information and a complex structure of interdependencies. The criminality is complex phenomenon and depends on different social, political, and economic factors, and their variations in time. Our consideration is aimed at the development of data analysis strategy for revealing important multivariate structures and interdependence spaces which influence criminality in Lithuania. Such problem deals with data mining, patterns recognition techniques in multidimensional data warehouses of social-economic indicators of different regions (Burinskiene *et al.*, 2003; Dzemydiene and Rudzkiene, 2002).

A lot of multidimensional data in data warehouses revealed some problems in the organization of data as well as in supporting data mining techniques for making precise decisions. Relevant patterns are not so easy to be extracted from large volume of data (Hinrichs, 2001; Pedersen and Jensen, 1999). For these reasons some important tasks cannot be sufficiently clear. Some questions, such as “what influences criminality?” become problematic to answer directly from the data space. The integrated methods of data mining and data structuring are required, as well as, the approaches using such methods in recognition processes of complex domain factors.

Such problems deal with treating data as n-dimensional cubes where the data is divided in measures (facts) and dimensions (Gyssen and Lakshmanan, 1997; Parsaye, 1996), and providing structured cube models, which captures dimension hierarchies (Agrawal *et al.*, 1997). But sometimes, the hierarchies and complex interrelationships are not captured explicitly between the parameters by the schema. Some models provide statistical objects where a structures hierarchy is related to explicit aggregation function on a single measure supporting a set of queries (Shoshani, 1996). An extended multidimensional model in which a multidimensional data warehouse is designed as a constellation of facts and dimensions is more efficient (Pedersen and Jensen, 1999; Teste 2001).

There can be several objectives for mining data, including action analysis, efficiency and risk assessment, etc. Distinct tasks require different data structures and separate data mining exercises. The series of organized actions in support of a cause can be analysed separately, because each case will involve patterns with a distinct signature. Mixing the analyses into one data mining exercise will simply dilute the differences between these signatures.

The method under development considers these constraints and evaluate the solution process proposing the algorithm that is considerably different from those algorithms in which a multivariate statistical method of one or another type is applied directly (Gyssen and Lakshmanan, 1997; Tabachnick and Fidell, 1996). Our consideration is attached on exploiting the temporal nature of multidimensional modelling and construction of dynamic model evaluating factor scores. The paper describes an approach of using multiple statistical analysis methods for data mining: time series analysis, cluster analysis, factor analysis, discriminant function, and regression analysis. The time series analysis methods are introduced in a semi-structured analysis session of data mining and allow us to exercise statistical control, forecast the main crime tendencies, and support a decision for crime prevention measures.

2. Architectural View of the Components for Dynamic Structures Revealing Process

To describe the social-economic processes and phenomena, large sets of social-economic indicators are needful. Most of these indicators take the form of time series in data warehouses. This causes some difficulty connected with the establishment of interrelation structure of these indicators. That is why the use of statistical methods for the social-economic process analysis and decision-making is not only justified but also indispensable.

The process of revealing ensembles of classifiers of influence space of criminality consists of interactions of the retrieval process of factors that have the greatest impact on the problem under consideration. The selection mechanism of the appropriate statistical method is related with the testing of the results and consequences. The refinement trends, and patterns within the social and economic data warehouses are the essential conditions for qualified crime detection. A crime rate is defined as the number of crimes per unit of population. A crime is considered registered if it is included into the centralised Register. The regional offices of crime statistics and research collect a great volume of data from different sources that reflect criminal events and their relations (Counties of Lithuania. . . , 2002; Rossmo, 1993; Rudzkiene, 2001). The data are stored in warehouses with multiplex variables and temporal dimensions.

Summarization may be used to reduce data size. But summarization can cause problems, too. Summarization of the same data set by two sampling or summarization methods may result in the same result, and summarization of the same data set by two methods may produce two different results (Parsaye, 1996). Another way could be using a pattern recognition process in a data warehouse (Herden, 2000; Dzemydiene and Rudzkiene, 2002). A pattern warehouse defined by (Hindrichs, 2001) as an information repository that stores relationships between data items and holds historical patterns rather than historical data. In the context of this research patterns are characterized as generic structures won by experience of modelling in the past, they can be abstract (describing the structure of a model) or concrete (describing one special model).

For the purpose of revealing multivariate structures and information patterns which influence criminality, the approach can provide the statistics based on local government areas, which breaks down the reported incidents into categories such as offences against the person, offences against the property, or specific offences like assault incidents (Dzemydiene *et al.*, 2002).

The discovery process is complex, not single-valued and has some different analysis activities. The structured analysis tasks of discovery system are more formal activities in which the influence space of the problem under consideration may be analysed.

The architecture, shown in Fig. 1 represents the components for dynamic structure revealing process. The data mining mechanisms are linked with the reasoning mechanism for retrieving a new problem solving class for pattern discovery. The dynamic model has been constructed by means of multiple regression and correlation analysis. We look for the rate of change and revise the patterns of influence space of the state.

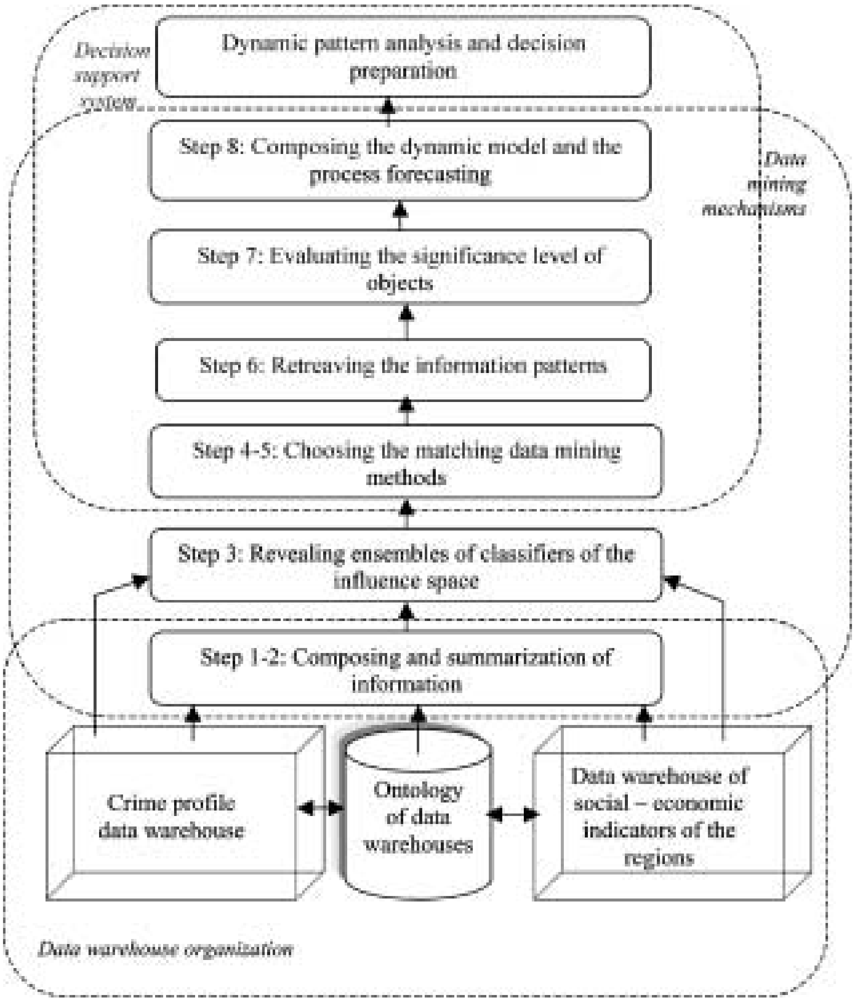
Decision support system
Dynamic pattern analysis and decision preparation
Step 8: Composing the dynamic model and the process forecasting
Data mining mechanisms
Step 7: Evaluating the significance level of objects
Step 6: Retreaving the information patterns
Step 4-5: Choosing the matching data mining methods
Step 3: Revealing ensembles of classifiers of the influence space
Step 1-2: Composing and summarization of information
Crime profile data warehouse
Ontology of data warehouses
Data warehouse of social – economic indicators of the regions
Data warehouse organization

Fig. 1. Architectural view of the components for dynamic structure revealing process. The steps of the algorithm are described in Section 3.

3. An Algorithm for Dynamic Structures Recognition

Multivariate statistical methods from the majority of possible probabilistic-statistical models enable us to make a grounded choice of a model that suits best to the initial statistical data that characterize the real behaviour of the set of objects under consideration and can estimate the present and future situation most exactly as well as to present substantiated and exact conclusions.

In describing the socio-economic situation, a great volume of initial data and indicators are used that characterize the development of a process, therefore it is very important to select the most important ones of them and to consider a small amount of indicators or their groups. Frequently the initial data are transformed so that to ensure the minimal loss of information.

An algorithm of the approach for dynamic structures recognition comprises the following steps:

Step 1: Observation objects of interest (towns, regions, districts, etc.) are selected, i.e., a sample $O = (o_1, o_2, \ldots, o_N)$. The object of a data set is a unit of data whose features are to be investigated. The objects have respective features (or indicators) $X = (x_1, x_2, \ldots, x_n)$ that describe their attributes. These features are measured within particular time intervals (ranges, e.g., a year interval), $\Delta t = (\Delta t_1, \Delta t_2, \ldots, \Delta t_k)$.

Step 2: Compose an $(N \times n \times k)$-dimensional matrix $Q_{ij\Delta t}$ that consists of object features in the time intervals considered, where i is the object considered, j denotes measured features, and Δt is a time interval.

Step 3: Preparing data for a further analysis, we determine homogeneity of the objects observed by investigating their properties. To this end, the methods of cluster or variance analysis should be applied. Cluster analysis belongs to classification algorithms and solves an issue how to organize the observed data into meaningful structures. The general categories of the cluster analysis methods are: joining or tree clustering, two-way joining or block clustering and k-means clustering.

If the clusters are clear heuristically, the methods of variance analysis are usually used. This classification problem can be solved in other ways, too: using heuristics or extremes way. Clusters of objects N are defined by choosing a fixed time interval Δt, and soundness of the clusters formed is verified in other time intervals.

Step 4: When clusters of objects are formed, the structure of features characterizing the clusters is under determination. For this reasons the factor analysis methods are selected for the problem solution. The factor analysis is applied in reducing the number of variables and for detecting a structure in the relationships between the variables. Generally, as a method for data reduction, the principal component analysis is often preferred, and the principal factor analysis is more frequently used in case when the goal of analysis is to detect structure. If the sample of selected objects is small ($N_1 < 50$), a two-dimensional matrix X_{ij} is compiled to solve this problem; here $i = 1, 2, \ldots p$ are objects, considered in time intervals $\Delta t = (\Delta t_1, \Delta t_2, \ldots, \Delta t_p)$, where $p = N_1 \cdot k$.

Step 5: Having verified the data adequacy/suitability to the factor analysis, variables that are not suitable for the analysis are found and eliminated. The adequacy of data (variables) for the factor analysis can be verified by the Kaiser–Meyer–Olkin measure of sampling adequacy KMO (Kim and Mueller, 1978):

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}}{\sum\sum_{i \neq j} r_{ij} + \sum\sum_{i \neq j} \tilde{r}_{ij}}, \tag{1}$$

here r_{ij} is the correlation coefficient, and $\tilde{r}_{ij}$ is the coefficient of partial correlation.

If the KMO value is low, then the indicators considered do not apply to the correlation analysis, since the correlation of these indicators cannot be explained by other indicators.

For making the exploratory data analysis, it is recommended first to analyse principal components (Kline, 1994). The components obtained in this analysis are no correlated and emerge in decreasing order of the amount of variance explained.

Step 6: To obtain a clear pattern of factor loadings, factor rotation strategies should be applied.

The fundamental theorem of factor analysis is invariant within rotations. The results of rotation, however, indicate "the simplest solution among a potentially infinite number of solutions that are equally compatible with the observed correlations" (Krijnen *et al.*, 1996). The simplest case of rotation is an orthogonal rotation. Typical orthogonal rotation strategies are *Varimax*, *Quartimax*, *Equamax*, and *Orthomax*. The *Varimax* rotation method is the most commonly used orthogonal rotation procedure. The overriding criterion of simple structure is that each factor should have a few high loadings with the rest being zero or close to zero. After clearing the patterns of factors, the influence of individual indicators x_n is evaluated and the factor interpretation is performed.

Step 7: Having estimated factor scores, dynamic models are built. Factor scores quantify individual cases on a latent continuum using z-score scale, which ranges from approximately –3.0 to +3.0. The significance (weight) level of the objects investigated is determined by transforming the two-dimensional matrix Z_{ij}, where $i = 1, 2, \ldots, p$ are objects considered in time intervals Δt, and $j = 1, 2, \ldots$ is the number of factors under consideration into three-dimensional matrix $Z^*_{ij\Delta t}$.

Step 8: The interdependence of variables (indicators) composing the factors is evaluated and indicators are predicted by forming a multivariate regression equation for time intervals Δt.

A multiple regression analysis determines the relationship between several independent variables and a dependent variable. The regression function can be estimated, using the least squares estimation or any other loss function (non-linear estimation). After the regression equation has been estimated, the prediction can be computed for a set of independent variables.

The significance of the equation is verified by the criterion F, while the influence of the variables selected is analysed by reciprocally comparing standardized regression coefficients.

4. The Process of Factor Analysis

The target of the research was to explore, estimate, and apply the use of multivariate statistical models in the analysis and prediction of the state situation and tendencies for even distribution of the quality of life in different regions of Lithuania, paying particular attention to the safety of society. In order to estimate the situation and take decisions it is expedient to select the main factors and evaluate their influence in time for social security.

The factor and component analysis methods make it possible to evaluate the multidimensionality of essential data and to explain concisely and simply enough the multivariate structures. They reveal really existing, but directly imperceptible regularities by means of factors or principal components.

The aim of factor analysis is to explain the outcome of p variables in the data matrix X using fewer variables, the so-called factors. These factors are interpreted as latent (unobserved) common characteristics of the observed $x \subset \mathbf{R}^n$. In the factor analysis every observed $x = (x_1, \dots x_n)^T$ can be written as

$$x_j = \sum_{l=1}^{k} a_{jl} f_l + \varepsilon_j, j = 1, \dots n; k \leqslant n. \tag{2}$$

here f_l for $l = 1, \dots, k$ denotes the factors, ε_j is the residual of x_j on the factors. Given the assumption that the residuals are uncorrelated across the observed variables, the correlations among the observed variables are accounted for the factors.

When using the factor analysis we meat a number of problems. Most often they arise in testing data suitability for the factor analysis, in choosing and interpreting factors. According to the logical sequence of problems solved by the factor analysis, the arising problems can be arranged in the following order: the first problem is a robustness, a second one – community, a third one – factors, a fourth one – rotation, a fifth – estimation of factor values, and a sixth one – dynamic models.

Having solved the above five problems, we regard the problem of factor analysis solved.

The sixth problem arises when using the factor analysis for data of several years, by comparing and interpreting the obtained results in time. This problem is related with the principles of building dynamic models. Dynamic models render on opportunity to find out indications the influence of which may decrease in future, or, on the opposite, will increase. Dynamic models enable us to compare or predict the results of factor analysis in a certain segment of time.

In the selection process of observation objects of interest the set of 13 social-economic indicators were collected for the research from 12 Lithuanian towns and 43 regions during time intervals of the period from 1996 until 2001 (Counties of Lithuania ..., 2002). We consider the matrix denotes as $X[n \times N]$. The matrix elements x_{ij} illustrate the value of the jth indicator at the ith research object and have particular values and semantics:

$x_{i},_{1}$ – registered crimes;
$x_{i,2}$ – average annual number of employed;
$x_{i,3}$ – unemployment rate;
$x_{i,4}$ – natural increase;
$x_{i,5}$ – migration;
$x_{i,6}$ – average monthly gross earnings;
$x_{i,7}$ – sales of industrial production;
$x_{i,8}$ – average real estate price;
$x_{i,9}$ – dwelling acquisition;
$x_{i,10}$ – investments in the construction of residential houses;
$x_{i,11}$ – investment in tangible fixed assets;
$x_{i,12}$ – direct foreign investment;
$x_{i,13}$ – turnover of catering,
where $i = 1, 2 \dots N$.

Several important issues are considered preparing data for the factor analysis. First, which variables should be included for the analysis. Second, how many variables should be included. A factor cannot be defined by using a single observed variable. There should be a minimum of three observed variables for each factor expected to emerge (Guttman, 1953). Third, the number of observations should be sufficient to provide reliable estimations of the correlations between the variables. Correlation coefficients tend to be unstable and greatly influenced by the presence of outliers if the sample size is not large. There is a general opinion that it is unwise to conduct the factor analysis on a sample fewer 50 observations. Another recommendation is that the sample size should be twice as large as the amount of variables, that is, there should be at least twice as many cases (N) as variables (n) and the cases should be heterogeneous with respect to the measures studied (Kline, 1994).

While considering the Lithuanian social-economic indices, the sample of objects studied has naturally to be divided into two groups: the first group consists of the largest cities and resort towns, and the second one – of regions. To form the groups, we can use cluster analysis methods, however, in this particular case, group bounds are clear. Substantiation of the division is verified by the hypothesis H_o stating that the average number of registered crimes in towns and regions is equal. This hypothesis is verified by the criterion:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_1^2/N_1 + S_2^2/N_2}}, \tag{3}$$

where $\bar{x}$ is the estimate of mean, and S is the standard deviation.

Arithmetic means of the registered crimes calculated, values of the criterion t, degrees of freedom, and the observed significance level p are presented in Table 1.

The obtained results in Table 1 show that the significance level observed in the years under investigation is lower than 5%. Therefore we have to reject the hypothesis H_o and to consider the criminality in towns and regions separately.

After evaluating the influence of each variable on the KMO measure, we eliminated three variables from the list of indices considered, namely: average monthly gross earnings, dwelling acquisition, and investments in the construction of residential houses. The KMO measure of the rest variables $KMO = 0.75$, so, we conclude that the data are adequate to the factorial analysis.

Table 1

Verification results of the hypothesis that the number of registered crimes in towns and regions is the same

Year	*Average number of crimes in towns*	*Average number of crimes in regions*	*t-value*	*df*	*p*
1996	175.9	132.2	2.8	53	0.006
1997	205.1	156.3	3.1	53	0.003
1998	214.8	153.1	3.7	53	0.001
1999	203.9	155.6	2.6	53	0.011
2000	232.9	189.9	2.0	53	0.049
2001	227.4	176.1	3.0	53	0.004

For making the exploratory data analysis, it is recommended first to analyse principal components. The components obtained in this analysis are non-correlated and emerge in decreasing order of the amount of variance explained.

The number of factors to be extracted can be determined in a scree plot (Fig. 2).

A large first eigenvalue (4.42) and a much smaller second eigenvalue (1.41) suggest the presence of a dominant global factor. The most widely used criterion for finding number of factors is the *Kaiser criterion* (Kaiser, 1953; Kaiser 1960), which recommends to retain only the factors whose eigenvalues are greater than 1. The scree plot (Fig. 2) suggests a maximum of four factors, too. These four factors account for 79.5% of the whole variance.

After evaluating the number of factors to be extracted, the next logical step is to determine the method of rotation. The overriding criterion of simple structure is that each factor should have a few high loadings with the rest being zero or close to zero (Kline, 1994). Applying this criterion the *Varimax* method was selected as providing the simplest structure solution.

When the rotation method is applied, one part of the output from the factor analysis is a matrix of factor loadings (Table 2). A factor loadings or factor structure matrix is a matrix of correlations between the original variables and their factors.

5. Reflection of the Influence Spaces from Social-Economic Data

5.1. *Interpretation of Factors*

The meaning of the rotated factors is inferred from the variables significantly loaded on their factors. A decision needs to be made regarding what constitutes a significant

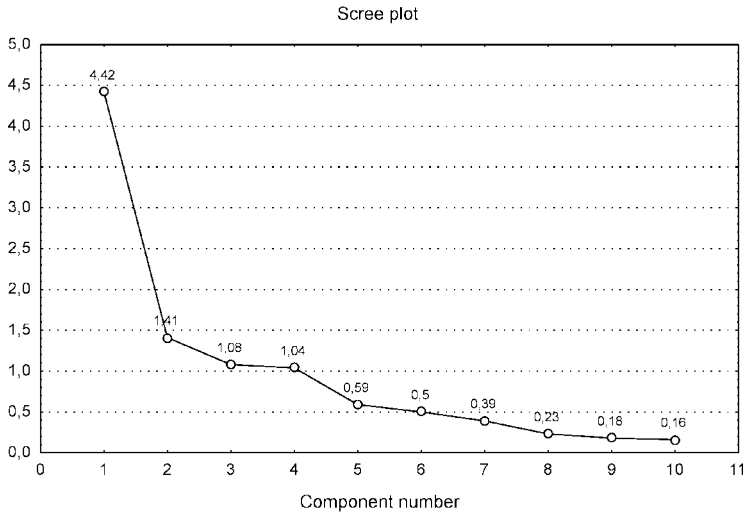

Fig. 2. The rate of change in the magnitude of eigenvalues for the factors.

Table 2

Factor loadings. Clusters of loadings are marked

	Factor 1	*Factor 2*	*Factor 3*	*Factor 4*
Direct foreign investment (x_{12})	**0.773**	0.138	0.111	0.154
Average annual number of employed (x_2)	0.440	**0.721**	0.215	0.018
Unemployment rate (x_3)	0.210	**–0.847**	–0.161	0.179
Natural increase (x_4)	**–0.715**	–0.536	–0.189	0.147
Migration (x_5)	0.127	–0.012	**0.964**	0.032
Sales of industrial production (x_7)	**0.788**	0.435	–0.125	–0.126
Average real estate price (x_8)	**0.742**	0.270	0.158	0.389
Investment in tangible fixed assets (x_{10})	0.110	0.080	0.020	**0.961**
Turnover of catering (x_{13})	**0.784**	0.174	0.138	–0.056
Registered crimes (x_1)	**0.767**	0.195	0.108	0.156

loading. The simplest criterion is that factors loadings greater than 0.30 in absolute value are considered to be significant. As the sample size increases, the criterion may need to be adjusted a bit downward. When the number of factors increases it may be adjusted upward. In general, the larger the absolute size of the factor loading for a variable, the more important the variable is in interpreting the factor.

As we can see from results in Table 2, the most significant variables for the first factor are:

- *Direct foreign investment;*
- *Natural increase;*
- *Sales of industrial production;*
- *Average real estate price;*
- *Turnover of catering;*
- *Registered crimes.*

We may state that this factor reflects the situation that depends on the growth of economy and the improving life conditions of many people. The greatest impact on this factor is made by the variables such as *sales of industrial production* ($L = 0.788$), turnover of catering ($L = 0.784$), that reflect the increasing retail trade, *direct foreign investment* ($L = 0.773$) which indicates the increasing influence of foreign investment, and *an average real estate price* ($L = 0.742$) that is increasing with an increase in the average wages and salary. Naturally, with an improvement of economic situation, the registration of crimes also improves as well as the number of delinquency crimes. We have to acknowledge that this kind of situation is also typical of Lithuania, the influence of the variable r*egistered crimes* ($L = 0.767$) is positive. Demographic laws that, with an increase in economies the birth-rate decreases, are corroborated. The impact of the variable "natural increase" is negative ($L = -0.715$), thus, with an increase in values of other variables, this indicator is inclined to decrease.

The second factor reflects employment and unemployment of population and it consists only of two variables with the opposite effects: *unemployment rate* ($L = -0.847$)

and *the average annual number of employed* ($L = 0.721$).

The third and fourth factors are constituted of individual variables; therefore the factors can be identified by these variables.

5.2. *Estimation of Factor Scores*

A factor score is a numeric value that indicates a relative spacing or standing of object on a latent factor. Factor scores quantify individual cases on a latent continuum using z-score scale, which ranges from approximately –3.0 to +3.0.

There are two general classes of methods for estimating factor scores. The first class of methods yields approximately standardized factor score estimate with different properties. Regression approach produces factor score estimates that maximize determinacy (Bollen, 1989).

$$\mathbf{F} = \mathbf{\Phi} \cdot \mathbf{\Lambda}^{\mathbf{T}}\mathbf{\Sigma}^{\mathbf{-1}}\mathbf{x}. \tag{4}$$

Here **F** are estimated common factors, $\mathbf{\Phi}$ is the covariance matrix of the common factors, $\mathbf{\Lambda}$ is the matrix of loadings, $\mathbf{\Sigma}$ is the model-implied covariance matrix of the measured loadings. Matrices are based on estimated parameters.

Another methods yields factor score estimates that are perfectly orthogonal (uncorrelated) (Krijnen *et al.*, 1996). Each of the refined methods is imperfect. Regression estimates will be correlated even when the factors are orthogonal, and orthogonal estimates will not maximize determinacy.

The second class of scoring procedures estimate the factor score for simply summarizing the responses of subsets of the factored items. Such factor score estimates will invariably be inter-correlated even when the factors are orthogonal, they will not have unit variance.

Choosing among the different factor scoring procedures is thus not a straightforward affair, and the final choice must be dictated by a survey of the extant studies on the differences among the various estimation methods and by the context of particular research program.

Having computed the regression estimates of factor scores, the data were partitioned by separate years and their common diagram drawn (see Fig. 3). Since the data of Birštonas, Neringa, and Visaginas were incomplete, these towns were not included in the factor scores analysis. The scores of the first factor are presented in Fig. 3, and scores of the second factor in Fig. 4.

The results obtained by factor scores analysis revealed the influencing impacts to the first factor. We can conclude that the three large towns (Vilnius, Kaunas, and Klaipėda) make up main share of influence on the first factor and their influence increases every year. The influence of the remaining towns is insignificant, and a change in time is not distinct.

The results obtained by factor scores analysis (Fig. 4) revealed that the Druskininkai have the greatest negative impact on the second factor. The impact of other towns differs a little and with the course of time, it decreases.

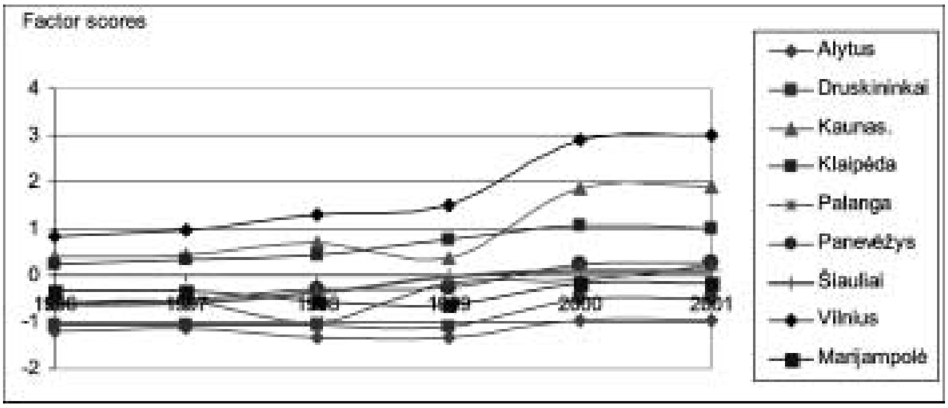

Fig. 3. Actual values (factor scores) of individual cases in 1996–2001 for the first factor.

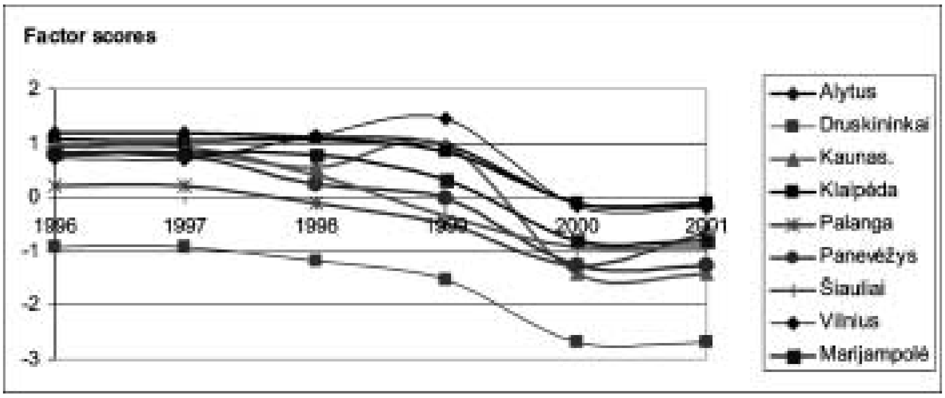

Fig. 4. Factor scores for the 2nd factor during 1996–2001.

6. Dependence of Total Registered Crimes on Well-Fare Indices

After selecting the principle variables that describe the economic development of the state, by the number of these and the registered crimes the system determine their mutual relationship and its change in time. For this we will make use of a linear equation of multiple regression:

$$y_i = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, N, \tag{5}$$

here β_j are unknown parameters, ε_i is a random variable with zero arithmetic average mean and variance δ^2.

In the considered case $k = 5,\ y_i$ denotes total registered crimes, x_{i1} is direct foreign investment, x_{i2} is a natural increase, x_{i3} denotes sales of industrial production, x_{i4} is the average real estate price, and x_{i5} is the turnover of catering.

To verify the significance of the regression equation, the variance analysis criterion F is used. The F-value and the resulting p-value reflect the relationship between the dependent variable and independent variables. Having calculated the value of this criterion for all the data of years analysed, we obtain $F(5,52) = 17.636$ and $p < 0.0001$. The value of multivariate correlation coefficient is $R = 0.79$, and of the determination coefficient $R^2 = 0.63$.

In order to learn which of the selected indicators contributes most to the prediction of total registered crimes, we need to examine the standardised regression coefficients. These coefficients are obtained by standardising the variables to a mean of 0 and the standard deviation of 1.

Thus, the magnitude of these coefficients allows a comparison of a relative contribution of each independent variable in the prediction of dependent variables. The coefficients are obtained separately for 1996–2001 (Fig. 5).

From results obtained by regression analysis (Fig. 5) we can see that the average real estate price acquires more and more weight which has the greatest importance in prediction of criminality. The impact of other indicators is either included to decrease or we can state that the average real estate price is the indicator evaluated most exactly.

No less important it is to verify change of the regression coefficient significance in time, i.e., to verify the hypothesis $H_o\colon \beta = 0$. This hypothesis is verified by the criterion t_j:

$$t_j = \frac{b_j}{[S^2 C_{jj}]^{1/2}}, \tag{6}$$

where C_{jj} is the diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ corresponding to b_j. In the case of the hypotheses $H_0\colon \beta = 0$ being correct, is distributes by Student law with an $n - k - 1$ degree of freedom. The results illustrated in Fig. 6 shows the significance of p-values of each coefficient evaluated by the t_j-criterion during the period 1996–2001.

The results in Fig. 6 demonstrate that the relation between the average real estate price variable and that of total registered crimes remains significant almost during the whole

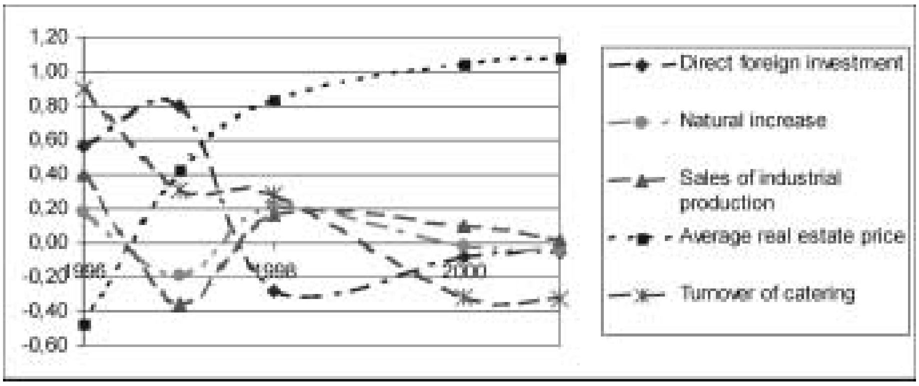

Fig. 5. Standardised regression coefficients in 1996–2001.

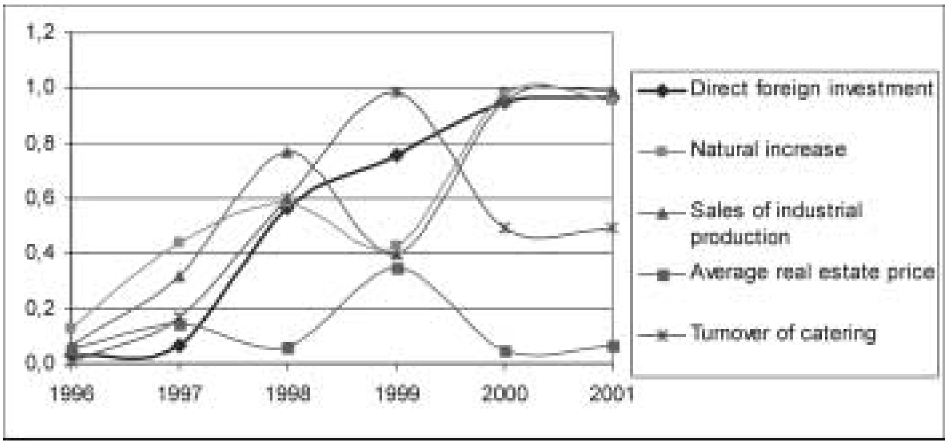

Fig. 6. Significance level of p-values, obtained in verifying the b_j coefficient significance by the t_j-criterion.

period. Meanwhile, other indicators are tending to be less significant or their measuring exactness is insufficient.

Consequently, in order to predict the number of crimes for the other year, first of all we should evaluate the average real estate price.

Conclusions

When analysing the socio-economic situation, we have to use many interrelated initial data and indicators that characterize the development of the process. In employing multivariate statistical methods from many possible probabilistic-statistical models there was obtained the model such that describes the real behaviour of the explored set of objects best and that provides substantiated and exact conclusions.

In order to evaluate the situation and make decisions the described data analysis strategy enables the evaluation of main factors and selection the influencing space of the social security in towns and regions of Lithuania. To this purposes the framework of integration of factorial and component analysis methods have been created. These models allowed the estimation of essential data multidimensionality and a concise and simplified explanation of multivariate structures of data warehouses. By means of factors and principal components they displayed existing in reality but directly imperceptible regularities.

After analysing the factor scores it has been established the influence of individual cases on the factor using the z-score scale. This allows us to evaluate a dynamic model of the situation considered. After choosing the principle variables that describe the economic development of the state, a dynamic model has been constructed that defined the interdependence change of these indicators in time. To this end, the multiple regression models have been used.

Results of factor analysis show that the three largest towns (Vilnius, Kaunas, and Klaipėda) make up main share of influence not only on the growth of economy and people well being, but also on the criminality increasing during 1996–2001. The increasing of criminal level of smaller towns is not such remarkable, and changes in time are not such distinct. The results obtained by regression analysis show that among well-being indicators the average real estate price has the greatest importance in prediction of criminality. The relation between the average real estate price variable and total registered crimes remains significant during the whole investigated period.

Acknowledgements

The authors wish to thank the Department of Statistics of Lithuania for the facilities to use the databases of social-economic indicators of Lithuania, the Centre for Crime Prevention in Lithuania for opportunity to use the criminological databases, the Chief of the Department of Urban Engineering at Vilnius Gediminas Technical University prof. M. Burinskienė for helpful comments that greatly helped to improve the interpretation of economic indicators. The authors are grateful to anonymous referees for their helpful remarks that helped to improve this paper.

References

Agrawal, R., A. Gupta, A. Sarawagi (1997). Modeling multidimensional databases. In *Proceedings of ICDE'97*, IBM Research Report.

Bollen, K.A. (1989). *Structural Equation with Latent Variables*. Wiley-Interscience.

Burinskiene, M., D. Dzemydiene and V. Rudzkiene (2003). An approach for recognition of significant factors for sustainable development strategies. In *Modelling and Simulation Business Systems. International Conference Proc.*, Vilnius. pp. 90–96.

Counties of Lithuania: Economic and Social Development (2002). Department of Statistics of Lithuania, Vilnius.

Dzemydiene, D., V. Rudzkiene. (2002). Multiple regression analysis of crime pattern warehouse for decision support. *Lectute Note in Computer Science*, Vol. 2453. Database and Expert systems Applications. A. Hameurlain, R. Cicchetti, R. Traunmuller (Eds.), Springer. pp. 249–258.

Dzemydiene, D., E. Kazemikaitiene, R. Petrauskas (2002). Knowledge representation in advisory informations system of crime investigation domain. In H.-M.Haav, A. Kalja (Eds.), *Databases and Information Systems*, II. Kluwer Academic Publishers, Netherlands. pp. 135–147.

Herden, O. (2000). Parameterised patterns for conceptual modelling of data warehouses. In *Proc. of the 4th IEEE International Baltic Workshop on Databases and Information Systems*. Technika, Vilnius. pp. 152–163.

Hinrichs, H. (2001.) Statistical quality control of warehouse data. In J. Barsdinš, A. Caplinskas (Eds.), *Databases and Information Systems*. Kluwer Academic Publishers. pp. 69–84.

Gyssen, M., L.V.S. Lakshmanan (1997). A foundation for multi-dimensional databases. In *VLDB'97*. Athens.

Guttman, L. (1953). Image theory for the structure of quatitative variables. *Psychmetrica*, **18**, 277–296.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Pyrometrical*, **23**, 187–200.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, **20**, 141–151.

Kim, J.O., C.V. Mueller (1978). *Introduction to Factor Analysis*. Beverly Hills, Sage Publications.

Kline, P. (1994) *An Easy Guide to Factor Analysis*. London, Routledge.

Krijnen, W.P., T. Wansbeek T., J.M.F. Ten Berge (1996). Best linear predictors for factor scores. *Communications in Statistics: Theory and Methods*, **25**, 3013–3025.

Parsaye, K. (1996). *Rotational Schemas: Multi-Focus Data Structures for Data Mining*. Information Discovery Inc.

Pedersen, T.B., C.S. Jensen (1999). Multidimensional data modeling for complex data. In *Proc. ICDE'99*. San Diego, California, USA.

Rossmo, D.K. (1993). Target patterns of serial murders: a methodological model. *American Journal of Criminal Justice*, **17**(2), 1–21.

Rudzkiene, V. (2001). Mathematical modelling of criminality in Lithuania in a context of the East European Countries. *Liet. Matem. Rink.*, **41**, 548–551 (in Lithuanian).

Shoshani, A. (1996). Statistical databases and OLAP: similarities and differences. Invited talk. In *CIKM'96*. Rockville, USA.

Tabachnick, G.B., L.S. Fidell (1996). *Using Multivariate Statistics*. New York, HarperCollins.

Teste, O. (2001). Towards conceptual multidimensional design in decision support systems. In A. Caplinskas, J.Eder (Eds) *Proc. ADBIS'2001*. Research Communications. Vilnius. pp. 77–87.

D. Dzemydienė holds a diploma with honour in applied mathematics and software engineering (Kaunas University of Technology), and ScD in mathematics and informatics. Dr. Dzemydienė is an associate professor of the Department of Legal Informatics of the Law University of Lithuania and a senior researcher at the Department of Software Engineering at the Institute of Mathematics and Informatics. Her research interests include information system engineering, artifitial intelligence methods, knowledge representation, and decision-support systems.

V. Rudzkienė is a doctor of technological sciences, an associate professor of the Department of Legal Informatics at the Law University of Lithuania. Dr. Rudzkienė graduated from Department of Informatics of Kaunas Technological University, Lithuania and received the ScD degree in management of technological systems. Her research interests include an investigation of social-economic dynamic systems, statistical technologies in law and management, multivariable modelling of sustainable development.

Socialinių-ekonominių duomenų saugyklų daugiamačių struktūrų analizės strategija

D. Dzemydienė, V. Rudzkienė

Straipsnyje pateikiama daugiamačių statistinių metodų integravimo strategija, skirta socialinių-ekonominių duomenų struktūrų ir ryšių atskleidimui, kai stebima didelė kintamųjų aibė trumpais laiko intervalais. Nagrinėjami faktorinės ir klasterinės duomenų analizės metodų integruoto taikymo aspektai. Pasiūlytas algoritmas leidžia atskleisti ir įvertinti sudėtingas duomenų struktūras, informacijos kitimo šablonus ir jų tarpusavio ryšius socialinių-ekonominių duomenų saugyklose. Šių metodų pagrindu konstruojami dinaminiai modeliai, atspindintys socialinės gerovės rodiklių įtaką nusikalstamumo lygio kitimui. Pateikiami šios metodikos realizavimo algoritmo taikymo pavyzdžiai ir gauti rezultatai analizuojant situaciją Lietuvos miestuose.