RIGA TECHNICAL UNIVERSITY
Faculty of Transport and Mechanical Engineering

Diana SANTALOVA
Candidate for the doctor's degree of the doctoral programme "Transport systems maintenance engineering support"

SEMI-PARAMETRIC REGRESSION MODELS FOR ANALYSIS AND FORECASTING OF FREIGHT AND PASSENGER TRANSPORTATION VOLUMES

SUMMARY OF THE PROMOTIONAL WORK

Scientific supervisor
Dr.hab.sc.ing., professor
A. ANDRONOVS

RTU Publishing House
Riga 2009

UDK [656.025+519.233.5](043.2)
Sa 611 p

Santalova D.
Sa 611 p Semi-parametric regression models for analysis and forecasting of freight and passenger transportation volumes. Summary of the promotional work. – R.: RTU, 2009. – 43 p.

Published according to the decision of the Promotion Council "RTU P-20" 30.09.2009., protocol No 01/2009

This work has been partly supported by the European Social Fund within the National Programme "Support for the carrying out doctoral study program's and post-doctoral researches" project "Support for the development of doctoral studies at Riga Technical University."

This work has been supported by the European Social Fund within the project „Support for the implementation of doctoral studies at Riga Technical University".

ISBN XXX

PROMOTIONAL WORK
IS PRESENTED TO RIGA TECHNICAL UNIVERSITY
TO OBTAIN THE SCIENTIFIC DEGREE OF
DOCTOR OF SCIENCE IN ENGINEERING

The defence of the promotional work will be delivered on December, 7th, 2009, at 14:30, in Riga Technical University, Institute of Transport Vehicle Technology, to the address: 1, Lomonosova street, building V, room 218, Riga, Latvia.

OFFICIAL REVIEWERS:

Professor, *Dr.sc.ing.* Peteris Balckars
Riga Technical University, Latvia

As. professor, *Dr.sc.ing.* Victor Lyumkis
Transport and Telecommunication Institute, Latvia

Professor, *Dr.phys.math.* Nikolai Ushakov
Norwegian University of Science and Technologies, Trondheim, Norway

CONFIRMATION

I confirm that I have independently worked out this promotional work for defence at Riga Technical University for being conferred the degree of Doctor of Science in Engineering. The promotional work has not been presented to any other university to obtaining the scientific degree.

Diana Santalova

Date: October, 8th, 2009.

The promotional work is written in English. The given work consists of 7 chapters. The bibliography includes 127 sources. There are 30 figures, 147 formulas and 60 tables to illustrate the conception of the carried out research. The promotion work contains 162 pages.

ABSTRACT

The promotional work «Semiparametric Regression Models for Analysis and Forecasting of Freight and Passenger Transportation Volumes» has been worked out by Diana Santalova to obtain the scientific degree of Doctor of Science in Engineering. Scientific supervisor of the work is Dr. hab. sc. ing., professor Alexander Andronov.

The objectives of the present work are different kind of transport freight and passenger transportations forecasting applying modern statistical methods. Nowadays is observed not only industrial decentralization all over the world, but also increasing of population mobility. As a consequence is a steady growth of number of passenger and freight transportations by all kind of transport. This fact gives evidence that chosen direction of research is perspective, and presented work is actual.

The first direction of the research is group models development and verification for transportations volumes forecasting for the EU Member States. In this connection four tasks were solved:

- forecasting of total air freight and mail transportations on the basis of parametric multivariate model;
- estimation of correspondence matrix of passenger rail departures applying modified gravity model;
- evaluation and forecasting of rail freight turnover on the basis of semiparametric single-index model;
- analysis and forecasting of total air passenger transportations in conditions of incomplete data using SURE-model.

The second direction is elaboration and estimation of semiparametric single-index models for analysis and forecasting of passenger rail departures for regions of Latvia.

The new methods of models estimation are suggested, criteria for the created models quality verification are developed. Advantages of application of the semiparametric models in comparison with the classical methods of the linear regression are shown.

CONTENTS

1. ACTUALITY OF THE PROBLEM

Present promotional work is devoted to passenger and freight transportations analysis and forecasting for the European Union, placing emphasis on Latvia, on the basis of use of parametric and semiparametric regression models, and to investigating of the efficiency of the elaborated methods of their estimation.

Volumes of transportations are the essential information for drawing up of all plans and forecasts of transport branch functioning and development. In particular, the volume of transportations is used for:

1) making up of vehicle park on prospect,
2) a transport network planning,
3) determination of the tendencies of requirement for capacities and the investment into development of transport and its components, and so on.

Concerning Latvia, its geographical position is one of its main national riches. Huge volumes of freight and passenger transportations pass through its territory. Latvian transport companies work in the conditions of a strong competition with the foreign companies. In this connection the correct economic policy can provide competitiveness and efficiency. Obviously, for getting the qualitative forecasts of transportations volumes the modern efficient mathematical models have to be used.

We accentuate the fact, that our developed semiparametric models and methods have not been applied in Latvia before. In the world the basic researches in the field of nonparametric and semi-parametric regression have begun only in the middle of the twentieth century and are widely spent now. So, the given area of science can be considered of one of the youngest and most perspective.

The received models will allow defining, what factors and in what measure force on directions and intensity of transport flows. The qualitative factors (for example, political conditions in the European Union) are intended to be represented numerically through quantitative factors (for example, national currencies exchange rates or the world prices for oil). It will enable to give recommendation of actions which would provide effective development of transport in Latvia.

The present research can be divided on the following two parts depending on application area:

1) Working out and estimation of models for transportations volumes forecasting for the EU Member States, i.e.:
    a) analysis and forecasting of total air freight and mail transportations on the basis of parametric multivariate model;
    b) estimation of correspondence matrix of passenger rail departures applying modified gravity model;
    c) analysis and forecasting of total air passenger transportations in conditions of incomplete data using SURE-model;

d) evaluation and forecasting of rail freight turnover on the basis of semiparametric single-index model;

2) Elaboration and estimation of semiparametric single-index models for analysis and forecasting of passenger rail departures for regions of Latvia.

2. OBJECTIVES AND TASKS OF RESEARCH

The main objectives of the promotional work are:

1) Developing of the mathematical models for passenger and freight transportations analysis and forecasting.
2) Elaboration of the methods for suggested models estimation.
3) Investigation of the elaborated estimation methods efficiency.

The following tasks are considered:

1) Considering problems of transportations forecasting.
2) Investigation nowadays used models and methods for forecasting.
3) Carrying out the analysis of the factors influencing the volumes of passenger and freight transportations.
4) Creation research information base, making statistical data collections for the Member States of the Europe Union and for regions of Latvia.
5) Developing the models of passenger and freight transportations volumes forecasting on the basis of multiple linear regression models and semiparametric regression models.
6) Developing methods and algorithms for estimation of offered models.
7) Verification and efficiency estimation of developed models.
8) Demonstration of semiparametris model advantage in comparison with multiple linear models.
9) Gravity model modification and method and algorithm developing for the suggested model estimation.
10) Suggested model applying for estimation of passenger departures correspondence matrix.
11) SURE-model formalization and method and algorithm developing for the offered model estimation.
12) Demonstration SURE-models advantages in comparison with multivariate model in condition of data incompleteness.

3. READINESS OF THE THEME

Trends in transport performance follow economic developments. According to *EuroSTAT Yearbook 2008*, while GDP grew at an average yearly rate of 2.4 % from 1995 to 2007, goods transport performance grew at 2.8 % yearly, and passenger transport performance grew at an average yearly rate of 1.7 %. Last one and a half year the economical conditions in the world is unstable. Thereupon now is not reasonable to allege about steady increase in transportations. Now GDP steadily decreases, and number of the unemployed

grows as well. This fact should make against increase in transportations.

According to the press releases, published by Eurostat, is resulted, that in the EU GDP fell up to 2.5% during the first quarter of 2009, compared with the previous quarter. The EU unemployment rate was 9.4% in June 2009, compared with 9.3% in May. Among the Member States, the highest rates in Spain (18.1%), Latvia (17.2%) and Estonia (17.0%). The EU annual inflation was expected to be 1.2% in February 2009 and -0.6% in July 2009.

All stated above is a coherent argument to assume that in such unstable economic conditions is especially important to have authentic forecasts of transportations. In the purpose of getting the qualitative forecasts of transportations volumes, in the present research the following mathematical models have been investigated:

1) Parametric models
   a) linear models
      i) multiple regression models
      ii) multivariate regression models
      iii) SURE-model
   b) Nonlinear models
      i) generalized linear models
      ii) gravity models
2) Nonparametric models
   a) Nadaraya-Watson estimator
   b) single-index model.

In the area of the multivariate statistics and parametric regression is noteworthy to point to such works, as *Applied Regression Analysis* of *N.Draper* and *H.Smith*, *Methods of Multivariate statistics* of *M.Srivastava*, *Theory of Point Estimation* and *Statistical Hypothesis Testing* of *E.L.Lehman*, and *Applied Linear Regression* of *S.Weisberg*.

Author cannot neglect such works of her teachers, as *Probability Theory and Mathematical Statistics* (in Russian) by *A.Andronov*, *E.Kopytov* and *L.Gringlaz* and *Computer Based Methods of Statistical Data Processing* (in Russian) by *I.Jatskiv*.

Multiple regression model

In general the regression model can be described as

$$Y_i = m(x_i) + \varepsilon_i, \tag{1}$$

where Y_i is a dependent variable in the *i*-th observation, $m(\bullet)$ is an unknown regression function, x_i is a d-dimensional vector of independent variables, ε_i is a random term. Furthermore we have a sequence of independent and uncorrelated observations (Y_i, x_i), $i = 1, 2, ..., n$. On that basis the unknown function $m(\bullet)$ should be estimated.

In the simplest case *the multiple linear regression model* is used:

$$Y = X\beta + \varepsilon, \tag{2}$$

where $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_d \end{pmatrix}$ and $X' = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$,

and $x_i = \begin{pmatrix} 1 & x_{i,1} & \ldots & x_{i,d} \end{pmatrix}'$. The main assumptions $E(\varepsilon) = 0$, $Cov(\varepsilon) = \sigma^2 I$ are known as Gauss-Markov conditions. The least squares estimate of β is given by

$$\hat{\beta} = (X'X)^{-1} X'Y. \tag{3}$$

Multivariate regression model

Multivariate regression model can be written in the following way:

$$\mathrm{Y} = \mathrm{XB} + \mathrm{E}, \tag{4}$$

where Y is the $(n \times p)$ matrix of observations on the dependent variables, X is the $(n \times d + 1)$ design matrix, B is the unknown $(d \times p)$ matrix of regression coefficients, E is the $(n \times p)$ matrix of random members, *p* is the quantity of dependent variables, *d* is the quantity of predictors and *n* is the quantity of observations. It is supposed that errors are independent. The least squares estimates for the B coefficients are obtained by:

$$\tilde{\beta}^{\langle k \rangle} = (X'X)^{-1} X'Y^{\langle k \rangle}, \ k = 1, \ldots, p, \tag{5}$$

where $\tilde{\beta}^{\langle k \rangle}$ is the estimate of unknown coefficients for *k*-th dependent variable.

Seemingly Unrelated Regression Equation Model

The *seemingly unrelated regression equation model* (further *SURE-model*) can be considered as continuation of the multivariate model (4) for a case when the part of predictors coincides for all variables of interest, and the part does not coincide. Besides, the number of observations on each variable of interest can be variously. In this field we sign such scientists, as *A.Zellner*, *R.Velu* and *J.Richards*, and *A.Andronov*.

So, we consider a group of *G* objects with numbers *i* = 1, 2, …, *G*. The *i*-th object is examined n_i times, at the time moments $t_{i,1} < t_{i,2} < \ldots < t_{i,n_i}$. At the *j*-th time moment $t_{i,j}$ we register a vector of independent variables $x_{i,j} = (x_{i,j}^{(1)}, x_{i,j}^{(2)}, \ldots, x_{i,j}^{m_i})$, where $m_i < n_i$, and a value of a dependent variable $Y_{i,j}$. It is supposed that the dependent variable $Y_{i,j}$ is formed by the

following linear regression equation

$$Y_{i,j} = \sum_{v=1}^{m_i} \beta_{i,v} x_{i,j}^{(v)} + Z_{i,j}, \qquad (6)$$

where $\beta_{i,v}$ is the coefficient for the *i*–th object and *v*–th independent variable, $Z_{i,j}$ is a normally distributed random term with mean zero and variance σ_i^2.

Further if for two various objects *i* and *i'* the time moments $t_{i,j}$ and $t_{i',j'}$ coincide then the random terms $Z_{i,j}$ and $Z_{i',j'}$ (therefore $Y_{i,j}$ and $Y_{i',j'}$ too) are correlated random variables with the covariance $c_{i,i'}$ whereas for various time moments they are assumed independent ($Z_{i,j}$ and $Z_{i,j'}$ are independent for $j \neq j'$ as well). That is, the disturbances $Z_{i,j}$ are contemporaneously correlated.

As usually it is assumed that for $i = 1, 2, \ldots, G$, $j = 1, 2, \ldots, n_i$ $x_{i,j} = \left(x_{i,j}^{(1)}, x_{i,j}^{(2)}, ..., x_{i,j}^{m_i}\right)$ is a known constant vector, $Y_{i,j}$ is the fixed value. On this base the unknown parameters of the regression model $\{\beta_{i,v}\}$ and unknown covariance $\{c_{i,i'}\}$, where $\sigma_i^2 = c_{i,i}$, should be estimated.

Gravity Models

Gravity model (by *A. Wellington* and *E. Lille*) has the following way:

$$T_{ij} = K \frac{P_i P_j}{D_{ij}^2}, \qquad (7)$$

besides T_{ij} is passenger transportation size between geographical points i and j; P_i and P_j are population living in points i and j, correspondingly; D_{ij} is a distance between points i and j; K is a constant.

The gravity model has received the further developing in forecasting of volumes of air passenger transportations depending on directions. Convincing contribution to air passenger flows forecasting using various modified gravity models has been made by *J.Doganis* in the sixties and by *A.Andronov et al.* in the eighties of the last century. Concerning OD-matix estimation is necessary to underline *Modelling Transport* of *J. de D. Ortuzar* and *L.G.Willumsen* and latest researches of *A.Andronov*.

Overview of Semiparametric and Nonparametric Models

Since parametric models not always give good forecasts, to eliminate this disadvantage *non-* and *semiparametric models* are applied.

In the field of nonparametric regression we need to mention such works, as *Nonparametric and Semiparametric Models* by *W.Härdle, M.Muller, S.Sperlich* and *A.Werwatz, Density Estimation for Statistics and Data Analysis*

by *B.W.Silverman*, *Applied Nonparametric Regression* of *W.Hardle*, *Applied Smoothing Techniques for Data Analysis* by *A.Bowman* and *A.Azzalini*, *Smoothing Methods in Statistics* by *J.Simonoff*, and *Nonparametric Econometrics* by *A.Pagan* and *A.Ullah*.

For general aspects on semiparametric regression we refer to the *Semiparametric Regression for the Applied Econometrician* by *A.Yatchew*, and *Semiparametric Regression* by *D.Ruppert*, *M.P.Wand* and *R.J.Carrol*.

The basis for many semiparametric regression models is the *generalized linear model* (*GLM*) (*Generalized Linear Models* of *P.McCullagh* and *J.A.Nelder*), which is given by

$$m(x_i) = G\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}\right) = G\left(\beta^T x_i\right), \tag{8}$$

where $G(\bullet)$ denotes any *known* continuous function and is called the *link function* of one dimensional variable $\beta^T x_i$, which is called the *index function* or simply the *index*. There are the following kinds of non- and semi-parametric models: additive model (*AM*), partial linear model (*PLM*), generalized additive model (*GAM*), generalized partial linear model (*GPLM*), generalized additive partial linear model (*GAPLM*).

In the given promotional work mainly application of the *single index regression model* (further *SIM*) is investigated. The single index model is one of *GLM* generalizations and it summarizes the effects of the predictors $x = (x_1, x_2, ..., x_d)$ within a single variable called an *index*. The single index model can be expressed by the following formula:

$$E(Y \mid x) = m(x) = g\{v_\beta(x)\}. \tag{9}$$

Here unknown function $m(\bullet)$ is a smooth function. Function $g(\bullet)$ is an *unknown link function* of one-dimensional variable $v_\beta(x)$, which called an *index*. As index function any function can be taken. *W.Härdle* considers two methods for estimation of the *SIM*, i.e. *semiparametric least squares* (*SLS*) and *pseudo maximum likelihood estimation* (*PMLE*).

The ***Nadaraya-Watson Kernel Estimator***, introduced by Nadaraya and Watson in 1964:

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{n^{-1}\sum_{j=1}^{n} K_h(x - X_j)}, \tag{10}$$

is the natural extension of kernel estimator to the problem of estimating an unknown regression function. The bandwidth *h* determines the degree of smoothness of $\hat{m}_h$.

In conclusion, the *individual model* of forecasting considers each object of

forecasting separately. The *group model* is based on simultaneous consideration of some objects in the aggregate.

In spite of SIM is one of the most popular and most investigated semiparametric models, its researches is going on. In this connection is noteworthy to note such scientists, as *J.S.Simonoff* and *C.-L.Tsai*, *M.Hristache*, *A.Juditski*, and *V.Spokoiny*, *L.Xue Y.Xia*, *H.Tong*, *E.Kong*, *W.K.Li*, *L.Zhu*, *H.Wong*, *W.C.Ip* and *R.Zha*.

Most of the latest publications, which are devoted to forecasting of the transportations, contain only simple forecasting models on the basis of the time series methods, as or linear regression models with 2-3 explanatory factors (*A.Baublys*, *J.Butkevičius*, *A.Cokasova*, *N.R.Farnum* and *L.W.Staton*, *Ü.Hunt*, *T.Šliupas*, *E.Spissu*, *N.Taneja*, *S.Wheatcroft* and *G.Lipman*, *S.Makridakis*, *S.Wheelwright* and *R.Hyndman*). At the same time, they do not contain any precise and detailed descriptions both the general technique of forecasting, and criteria of quality of the chosen method of forecasting.

We will keep the following transportations structure in our investigation:

1) By type of transportation:
   a) Passenger;
   b) Freight.
2) By transportation mode:
   a) International;
   b) Internal (domestic).
3) By mode of vehicle:
   a) Rail transport;
   b) Air transport.
4) By object of forecasting:
   a) Countries or regions of country;
   b) OD-pair.
5) By direction:
   a) incoming;
   b) outgoing.
6) By type of forecasted indicator:
   a) Transportations or departures, in thousands of passengers or tonnes;
   b) Turnover, in tonne-km or passenger-km.

Depending on type of described information the factors influencing forecasted indicators are divided on *quantitative* and *qualitative*. The factors which values vary from observation to observation are related to the *quantitative* factors. *Qualitative* factors have the same value for the whole group of observations, and take on some discrete values, according to in advance certain principle. Depending on character of the displayed information all influencing factors can be divided into following basic groups: *economical*, *social* and *structural*.

In the course of the given research two statistical databases have been generated.

The first database is generated entirely on materials of The Statistical

Office of the European Communities (*EuroSTAT*) and contains the statistical data on both passenger and freight transportations on the Member States of the EU from 1995 for 2008. The base also contains the various factors across the Member States of the EU, influencing passenger and freight transportations.

The second database contains the statistical data on rail passenger transportations from regions and cities of Latvia. This base has been generated on the basis of the statistical materials gotten from the Central Statistical Bureau of Latvia (LR CSP) and materials of the annual report of Closed Joint-Stock Company Latvijas Dzelzceļš.

Depending on the spent experiment and on the predicted variable there were used numerous qualitative factors as well.

4. METHODOLOGY AND METHODS OF RESEARCH

The promotion work research is based on:

1) Modern theory of regression analysis, moreover special attention is paid to such kind of generalized regression models, as semiparametric regression model; SURE-model and modified gravity model; nonlinear optimization methods were used as auxiliary means.
2) Statistical data received from "The Statistical Office of the European Communities" (*EuroSTAT*), "Central Statistical Bureau of Latvia" (*LR Centrālā statistikas pārvalde, LR CSP*) and "Annual Report of State Joint-Stock Company Latvijas Dzelzceļš".
3) Scientific literature, press releases and Internet-sources devoted to the investigated problems.
4) Computer based support for necessary calculation and investigation, i.e. Statistica 6.0 package and MathCad 13 environment.

5. SCIENTIFIC NOVELTY

Novelty of the present research consists of:

1) Multivariate regression model for total air freight and mail transportations forecasting.
2) Method and software for estimating parameters of the single index regression models and verification of their adequacy and efficiency for transportations analysis and forecasts.
3) Original non-linear regression model (based on the gravity model) and software for correspondence matrix of transport network.
4) SURE-model and software for total passenger air transportations forecasting.

6. PRACTICAL VALUE, REALIZATION AND APPLICATION OF THE WORK

1) On the basis of the obtained results a part of the course of lectures and practical works on the subject "Mathematical Methods of Traffic Flow Analysis and Forecasting" for the second year foreign student of bachelor's studies programme of the Riga Technical University Mechanical Engineering faculty is prepared.
2) Models and methods developed by the author for forecasting volumes of freight rail transportations for 25 Member Countries of the European Union were used in the scientific project "Mathematical models and their estimation method elaboration for analysis and forecasting of the Baltic Region passenger and Freight flows" which was a component of the II Scientific Project "Zinātniskās darbības attistība augtskolās" and lasted from June, 01st till December, 31st 2006.
3) Models and methods developed by the author for forecasting volumes of passenger rail transportations for regions of Latvia were used in the scientific project "Creation of mathematical models, algorithms and computer programs for Latvia's transport system's analysis, development prognosis and optimization", which was a component of the Scientific Project "Zinātniskās darbības attistība augtskolās" and lasted from June, 01st till December, 31st 2008.
4) The obtained forecasts can be used by transport companies for their work optimal planning, i.e. flights schedule, ticket prices and so on; by transport ministry for optimal distribution of capital expenses in road repairing, railways repairing, terminals reconstruction, building highways etc.

7. STRUCTURE OF THE PROMOTION WORK

Chapter 1. Review of State-of-the-art for Forecasting Problem. The current environment in transportations in the EU is commented. The survey of literature used during performing present Thesis is stated, and the short review of the researches spent in the areas of nonparametric and semiparametric regression is reconciled as well.

Chapter 2. Informational Support of Forecasting Problem. This Chapter is dedicated to analysis of the statistical data used for the present investigation. First of all, types of transportations as predicted indicators are described. The classifications of the factors influencing transportations are given as well. Two statistical data bases generated during the work are resulted. Besides, the corresponding sources of statistical information are characterized, i.e. EuroSTAT, Central Statistical Bureau of Latvia and Annual Report of State Joint-Stock Company "Latvijas Dzelzceļš".

Chapter 3. Mathematical Models for Forecasting and Methods of their Estimation. The theoretical foundation of the used regression models is considered. First, the characteristic of group and individual models is given; is told, for what cases they are applicable. Secondly, it is told about parametrical models. Among them widely known and often applied multiple (either linear or

parametric) model is described. Also there are considered parametrical as well, but more complicated in use multivariate model and SURE-model. A whole Sub-section is dedicated to the gravity model. The special attention is paid to the nonparametric and semiparametric models. The review of various semiparametric models is resulted. As one of accents of the given promotional work is testing of single index model efficiency, the kernel estimators and their properties are described. In particular, the Nadaraya-Watson kernel estimator, applied for estimation of SIM, is considered in details.

Chapter 4. Analysis and Forecasting International Air Freight Transportations for the EU Member States. The results of analysis and forecasting of the total freight air and mail transportations (i.e. internal and external in relation to the EU borders) for the Member States of the European Union are considered. The corresponding multivariate regression model contains main economical factors affected internal and external transportation for each country. It is shown how it can get the total forecast of internal and external transportations for concrete year. The confidence limits for this forecast (at various confidence levels) are calculated as well.

Chapter 5. Evaluation of Rail Passenger Correspondences between Member States of the European Union. Here is talking about estimation of the correspondence matrix of passenger rail departures between the Member States of the European Union on the basis of modified gravity model. The developed efficient algorithm for estimation of unknown parameters of this model is stated. The rail passenger correspondences between 23 Member States of the EU for 2008 are estimated using the suggested approach.

Chapter 6. Application of the Single Index Model for Transportations Values Investigation. This Chapter is devoted to the experimental investigation of single index model efficiency. We have intended to perform such investigation by comparing single index model with linear one. First, the elaborated algorithm of estimation of the unknown parameters of single index model is described. Secondly, the procedures of choosing the most significant single index model and the most significant linear model, developed within the limits of the given Thesis, are resulted. The offered cross-validation approach allows researching the efficiency of considered models not only in case of existing statistical data smoothing, but also in case of forecasting. The results of various experiments, in which obvious preference of single index model in comparison with parametric model has been proved, are stated.

Chapter 7. International Air Passenger Transportations Forecasting for the EU Member States on the Basis of SURE-model. In this Chapter some generalization of the SURE-model is considered. Individually taken observations are supposed not to contain the information about all response variables. An unbiased estimate for a covariance matrix of the model is obtained. The advantage of usage of SURE-model before usual multivariate model in case of the statistical data incompleteness is shown on the example of the total air passenger transportations forecasting for the EU Member States.

8. DESCRIPTION OF THE MAIN RESULTS OF THE WORK

8.1. Analysis and Forecasting International Air Freight Transportations for the EU Member States on the Basis of Multivariate Regression Model

Problem Setting

The freight and mail air international transportations are divided into two parts in relation to the borders of the European Union: internal and external freight and mail air international transportations.

The task is elaboration of regression models for such transportations forecasting. The following problems are considered:

1. *Separately* forecasting of external and internal transportations.
2. *Total* forecasting of external and internal freight and mail transportations.

For the first problem the multiply regression model (2) and for the second the multivariate regression model (4) are used. Our assumption is such: the considered factors should influence internal and external transportations differently. The unknown coefficient and covariance matrix estimation procedures have been represented by package Statistica 6.0 means and software developed in the MathCad 13 environment. The purpose of research is the experimental proof of significance of total forecast.

The main object of consideration, named *object,* is a Member State of the Europe Union, further called *country*. We call as *observation* a data about object for an actual year, i.e. external and internal freight and mail international transportations expressed in thousands of tonnes. The design factors are:

t_1 - time factor (*YEAR*);

t_2 - trade integration of goods (*TI*) as a percentage of GDP;

t_3 - annual average EUR exchange rate versus national currencies (*EURrates*);

t_4 - annual average inflation rate of change in Harmonized Indices of Consumer Prices (*IR*);

t_5 - prices on petroleum products (*PP*), euro per tonne;

t_6 - GDP “per capita” in Purchasing Power Standards (*GDP_PPS*);

t_7 - gradation of countries under value of forecasted parameter (*GradY*), entered into the model with the purpose to allocate the countries with obviously larger specific transportations (such as Belgium, Germany and the United Kingdom) into separate group;

t_8 - gradation of countries under population density (*GradPD*);

t_9 - gradation of countries under area (*GradArea*);

t_{10} - gradation of countries under the duration of membership in EU (*GradMember*);

t_{11} - total population, in thousands of inhabitants (*TP*).

Forecasted values are:

t_{12} - internal freight and mail international transportations by a country for a certain year, in tonnes (FrM_Intra);

t_{13} - external freight and mail international transportations by a country for a certain year, in tonnes (FrM_Extra).

General Structure of Considered Models and Evaluation Procedure

The estimated covariance between two estimates of unknown vectors:

$$Cov^{*}\left(\widetilde{\beta}^{\langle k\rangle},\widetilde{\beta}^{\langle l\rangle}\right)=\widetilde{\sigma}_{kl}\left(X'X\right)^{-1},\ k,l=1,\dots p. \tag{11}$$

Here $\widetilde{\beta}^{\langle k\rangle}$ and $\widetilde{\beta}^{\langle l\rangle}$ are estimated separately using formula (5), and $\widetilde{\sigma}_{kl}$ is a corresponding element from the unbiased estimate of covariance matrix $\widetilde{\Sigma}$. We obtain two several forecasts $\widetilde{\mathrm{Y}}^{\langle 1\rangle}=\mathrm{X}\widetilde{\beta}^{\langle 1\rangle}$ and $\widetilde{\mathrm{Y}}^{\langle 2\rangle}=\mathrm{X}\widetilde{\beta}^{\langle 2\rangle}$. The sum of forecasts is $\widetilde{S}=\widetilde{\mathrm{Y}}^{\langle 1\rangle}+\widetilde{\mathrm{Y}}^{\langle 2\rangle}$. The mean and the variance of this sum are:

$$E\left(\widetilde{S}\right)=E\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle}\right)+E\left(\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right)=\mathrm{X}\left(\widetilde{\beta}^{\langle 1\rangle}+\widetilde{\beta}^{\langle 2\rangle}\right), \tag{12}$$

$$D\left(\widetilde{S}\right)=D\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle}\right)+D\left(\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right)+2Cov\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle},\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right). \tag{13}$$

The first and second terms of (13) are

$$D\left(\widetilde{\mathrm{Y}}^{\langle k\rangle}\right)=D\left(\mathrm{X}\widetilde{\beta}^{\langle k\rangle}\right)=\mathrm{X}\cdot Cov\left(\widetilde{\beta}^{\langle k\rangle}\right)\cdot\mathrm{X}' \tag{14}$$

where $Cov\left(\widetilde{\beta}^{\langle k\rangle}\right)$, $k=1,2$, are covariance matrixes of vectors $\widetilde{\beta}^{\langle 1\rangle}$ and $\widetilde{\beta}^{\langle 2\rangle}$ respectively. The third term of (13) can be defined as follows:

$$Cov\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle},\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right)=Cov\left(\mathrm{X}\widetilde{\beta}^{\langle 1\rangle},\mathrm{X}\widetilde{\beta}^{\langle 2\rangle}\right)=E\left(\left(\mathrm{X}\widetilde{\beta}^{\langle 1\rangle}-\mathrm{X}\beta^{\langle 1\rangle}\right)'\cdot\left(\mathrm{X}\widetilde{\beta}^{\langle 2\rangle}-\mathrm{X}\beta^{\langle 2\rangle}\right)\right)=$$

$$=\mathrm{X}\cdot E\left(\left(\widetilde{\beta}^{\langle 1\rangle}-\beta^{\langle 1\rangle}\right)'\cdot\left(\widetilde{\beta}^{\langle 2\rangle}-\beta^{\langle 2\rangle}\right)\right)\cdot\mathrm{X}. \tag{15}$$

The last member is the joint covariance matrix of two vectors of coefficients (11). The upper confidence limit for total forecast $\overline{E}\left(\widetilde{S}\right)=E(S)$ corresponding to confidence probability γ is $\left(0,\overline{S}_{\gamma}\right)$, and

$$\overline{S}_{\gamma}=E\left(\widetilde{S}\right)+\sqrt{D\left(\widetilde{S}\right)}\cdot\Phi^{-1}(\gamma), \tag{16}$$

where $\Phi^{-1}(\gamma)$ is γ-quantile of standard normal distribution.

Gotten Models for Transportations Forecasting

All the suggested models are *group* models. Both old and new Member States of the EU are chosen for experiments: Austria, Belgium, Czech Republic, Germany, Denmark, Estonia, Spain, Finland, France, Greece, Hungary, Ireland,

Italy, Lithuania, Netherlands, Poland, Portugal, Slovakia and the United Kingdom. The analysed period is from 2001 to 2006. Besides, not all objects of research have observations on all considered years.

The first and second models are multiple linear regression models (2). The dependent variable in the *first model* $Y^{(1)} = t_{12} / t_{11}$ is internal freight and mail international transportations in tonnes, divided by total population in thousands of inhabitants, i.e. *specific* transportations. Explanatory variables are $x_1 = t_1$, $x_2 = t_2$, $x_3 = t_3$, $x_4 = t_4$, $x_5 = t_5$, $x_6 = t_6$, $x_7 = t_7$, $x_8 = t_8$, $x_9 = t_9$, $x_{10} = t_{10}$.

The dependent variable in the *second model* $Y^{(2)} = t_{13} / t_{11}$ is external freight and mail international transportations in tonnes, divided by total population in thousands.

The *third model* is the multivariate one (4) and there are two dependent variables, which are described by the vector $Y^{(3)} = \left(Y_1^{(3)} = t_{12} / t_{11}, Y_2^{(3)} = t_{13} / t_{11} \right)$. The sets of explanatory variables for all the models coincide.

Separate Evaluation of Internal and External Transportations

94 observations have been processed. The estimated models are:

$$\tilde{E}\left(\mathrm{Y}^{\langle 1 \rangle}(x)\right) = -1193 + 0.6x_1 + 0.09x_2 - 0.001x_3 - 0.17x_4 - 0.01x_5 + \\ + 1.99x_7 - 7.1x_8 + 16.47x_9, \qquad (17)$$

R_0 = 2 202, R^2 = 0.82, F^1 = 38, F(10,83) = 1.95, α = 5%.

$$\tilde{E}\left(\mathrm{Y}^{\langle 2 \rangle}(x)\right) = -3394 + 1.7x_1 - 0.001x_3 - 0.02x_5 + 0.07x_6 - 52.5x_7 - \\ - 6.6x_8 + 9.7x_9 + 69.3x_{10}. \qquad (18)$$

R_0 = 21 989, R^2 = 0.95, F = 151, F(10,83) = 1.95, α = 5%.

Evaluation of total transportations

The covariance matrix of errors has been calculated:

$$\tilde{\Sigma} = \begin{bmatrix} 9.12 & 5.24 \\ 5.24 & 22.45 \end{bmatrix}. \qquad (19)$$

Then on the base of (19) we obtain covariance matrices of regression coefficients $\tilde{\beta}^{\langle 1 \rangle}$ and $\tilde{\beta}^{\langle 2 \rangle}$ estimated by (11).

Table 1 contains the total forecasts for 2007 and for 2006 (columns 2007

[1] The calculated value of Fisher criterion

and 2006 respectively), true total transportations for 2006 (column 2006*), and the upper confidence limits (16) with different confidence probabilities γ for total forecasts for 2007. Figure 1 vizualizes presented in Table 1 transportation volumes.

Gotten results show remarkable evidence of proposed approach. However, individual models would be preferable for forecasts for partial countries, which have stable trends in transportations. It is especially important, when there are no any observations on the separate countries. It is not recommended to forecast total transportations on the basis of time series. In this connection it is supposed to improve the present research by means of SURE-model.

Table 1

Total forecasted transportations, in thousands of tonnes

Country	2006*	2006	2007	Upper conf. limits			
				60%	70%	80%	90%
Austria	227.92	174.21	196.60	204.77	213.73	224.09	238.47
Belgium	1124.95	742.95	781.78	788.92	796.75	805.80	818.35
Czech	57.05	90.63	133.22	136.51	140.12	144.30	150.09
Germany	3267.91	3213.93	3444.38	3 498.02	3556.90	3624.96	3719.35
Denmark	6.69	55.71	72.64	73.45	74.35	75.38	76.82
Estonia	10.05	16.08	20.96	24.03	27.40	31.29	36.69
Spain	368.97	673.48	811.22	817.36	824.11	831.91	842.72
Finland	118.43	84.17	101.16	126.16	153.60	185.32	229.31
France	1421.41	1105.82	1307.02	1 340.66	1377.59	1420.27	1479.47
Greece	90.60	154.94	197.92	243.74	294.03	352.17	432.80
Hungary	64.89	41.21	81.90	84.04	86.39	89.10	92.87
Ireland	111.64	86.28	109.44	114.88	120.86	127.77	137.35
Italy	764.05	762.71	949.88	963.28	977.97	994.96	1018.52
Lithuania	12.67	21.94	36.12	40.69	45.70	51.49	59.53
Netherlands	1621.35	1499.07	1545.27	1 577.28	1612.40	1653.00	1709.32
Poland	32.02	104.63	193.87	197.25	200.97	205.26	211.22
Sweden	139.48	66.41	94.66	96.93	99.41	102.29	106.28
Slovakia	5.36	50.44	61.37	68.46	76.25	85.25	97.74
UK	2248.41	2044.67	2248.39	2 289.97	2335.60	2388.36	2461.52

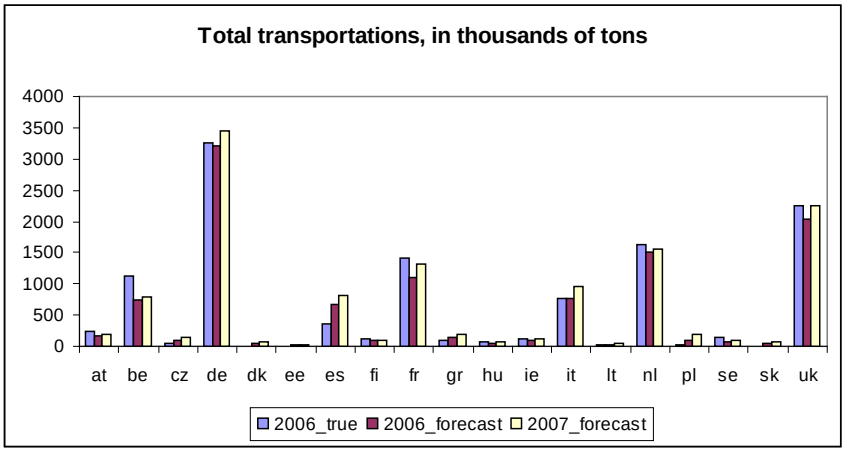

Fig.1. Total observed and forecasted transportations

8.2. EVALUATION OF RAILWAY PASSENGER CORRESPONDENCES BETWEEN THE EUROPEAN UNION MEMBER STATES ON THE BASIS OF MODIFIED GRAVITY MODEL

Problem Setting and Estimation Procedure

We have n corresponding geographical points with numbers $i = 1, 2, \ldots n$. For the point i is known inhabitants number h_i and m regressors $c_{i,j}$, $j = 1, 2, \ldots, m$, those are known constants. For all pairs of the points (i,l) the distance $d_{i,l}$ between them is known as well. In addition, the quantity of the departed passengers Y_i from the point i during considered time interval is known, that is a random variable.

Our aim is to estimate correspondence size $Y_{i,l}$ for all pairs of points (i,l). The matrix of $Y_{i,l}$ is to be called *the correspondence matrix*. Let us denote an estimate of $Y_{i,l}$ by $Y_{i,l}^*$. There are the following requirements: 1) $Y_{i,l}^* > 0$ for $i \neq l$; 2) $Y_{i,i}^* = 0$; 3) $Y_{i,l}^* = Y_{l,i}^*$.

The mathematical model for the particular correspondence (i,l) is:

$$Y_{i,l} = \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}\right), \qquad (20)$$

where a, $\alpha = (\alpha_1 \quad \alpha_2 \quad \ldots \quad \alpha_m)^T$ and $\beta = (\beta_1 \quad \beta_2 \quad \ldots \quad \beta_m)^T$ are unknown regression parameters, τ and θ are unknown form parameters, $c_{(i)} = (c_{i,1} \quad \ldots \quad c_{i,m})$ and $g_{(i,l)} = (c_{i,1}c_{l,1} \quad \ldots \quad c_{i,m}c_{l,m})$ are m–vector-rows, $\{V_{i,l}\}$ are independent identically distributed random variables with zero mean and unknown variance σ^2. It is supposed that $V_{i,l}$ has normal distribution. Then $Z_{i,l} = \exp(V_{i,l})$ has the log-normal distribution. Note that the case $\theta = 1$ and $\tau = 2$ corresponds to the so-called gravity model.

The following presentation for the quantity of the departed passengers from the point i is a corollary of model (20):

$$Y_i = \sum_{\substack{l=1 \\ i \neq l}}^{n} Y_{i,l} = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^\theta}{(d_{i,l})^\tau} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}\right). \qquad (21)$$

The following parameters should be estimated on the basis of fixed values $\{Y_i\}$: θ, τ, a, α, β and σ^2. As to identify parameters a and σ^2 simultaneously is not possible, the united parameter $\tilde{a} = a + \frac{1}{2}\sigma^2$ is introduced. As a criterion of estimates efficiency the WLS criterion is used:

$$R(\gamma, w) = \sum_{i=1}^{n} w_i \left(Y_i - E(Y_i)\right)^2, \tag{22}$$

where $\gamma = \left(\theta \quad \tau \quad \tilde{a} \quad \alpha^T \quad \beta^T\right)^T$ and $w = \left(w_1 \quad w_2 \quad \ldots \quad w_n\right)^T$ is a vector of weights. For a minimization of (22) the gradient method is applied, and

$$\nabla R(\gamma, w) = \left(\frac{\partial}{\partial \theta} R \quad \frac{\partial}{\partial \tau} R \quad \frac{\partial}{\partial a} R \quad \frac{\partial}{\partial \alpha} R \quad \frac{\partial}{\partial \beta} R \right)^T. \tag{23}$$

If w_i = const, the gradient method quickly gives estimates θ^*, τ^*, $\tilde{a}^*$, α^*, β^*. Besides,

$$\sigma^{2*} = \ln\left\{1 + \left(2\sum_{i=1}^{n-1} \sum_{l=i+2}^{n} \left(E\left(Y_{i,l}\right)^*\right)^2 \right)^{-1} \sum_{i=1}^{n} \left(Y_i - E\left(Y_i\right)^*\right)^2 \right\}. \tag{24}$$

Now the estimate of the parameter a is calculated as $a^* = \tilde{a}^* - \frac{1}{2}\sigma^{2*}$.

The statistical data $\{Y_i\}$ and the estimates $\{Y_{i,l}^*\}$ have to be balanced:

$$\sum_{l=1}^{n} Y_{i,l}^* = Y_i \;\; i = 1, \ldots, n. \tag{25}$$

For that the correction coefficient $\delta_i > 0$ for each point i is introduced. Then corrected estimate of $Y_{i,l}$ is

$$\tilde{Y}_{i,l} = \delta_i Y_{i,l}^* \delta_l, \;\; i, l = 1, \ldots, n. \tag{26}$$

Numerical Example

We apply the suggested approach for the passenger rail transportations estimation between 23 Member States of the European Union, further called *countries*. The following characteristics have been taken as regressors:

- c_1 is the average monthly labour cost, EUR;
- c_2 is gradation of countries upon intensity of use of air transport;
- c_3 is gradation of countries upon intensity of use of railway transport;
- c_4 is gradation of countries upon intensity of use of sea transport;
- c_5 is a country gradation upon degree of popularity for tourism;
- c_6 is a country gradation upon the duration of membership in the EU.

All the gradations can take on any positive integer values, only c_6 can take on two values: 0 for the old Member States of the EU, and 1 for the new Member States. The values of gradation factors are determined by means of experts. For example, for Denmark we have $c_1 = 4.34$, $c_2 = 4$, $c_3 = 3$, $c_4 = 4$, $c_5 = 2$, $c_6 = 0$; for Latvia we have $c_1 = 0.68$, $c_2 = c_3 = c_4 = 1$, $c_5 = 0$, $c_6 = 1$. For getting the values of gradations the statistical data for countries about air, sea and rail international transportations for the 2007 year have been analysed.

As distances between points, the distances between capitals of countries have been taken. As a possible measure of distance there could be considered distances between the weighted average coordinates of countries.

We estimate correspondences between origin points (countries of embarkation) and destination points (countries of disembarkation) on the basis of known departures from origin points. Thus, Y_i values are the rail international outgoing transportations in thousands of passengers for countries of embarkation (see Table 2). Thus it is necessary to estimate 15 parameters. The described estimation procedure gives the following values of estimated parameters:

$$\theta^* = 0.788,\ \tau^* = 2.786,\ \tilde{a}^* = -10.01,$$
$$\alpha^* = \begin{pmatrix} 0.074 & 0.061 & 0.077 & 0.063 & 0.078 & 0.33 \end{pmatrix}^T,$$
$$\beta^* = \begin{pmatrix} 0.197 & 0.194 & 0.152 & 0.054 & 0.097 & 0.104 \end{pmatrix}^T.$$

The estimate $\sigma^{2*} = 0.065$. Now using the estimate $\tilde{a}^* = -10.01$ we find

$$a^* = \tilde{a}^* - \frac{1}{2}\sigma^{2*} = -10.01 - \frac{1}{2}0.065 = -10.04.$$

Observed Y_i, estimated Y_i^* departures from each country in thousands of passengers, and correction coefficients δ_i are represented in the Table 2.

Table 2

Estimation results

Country	Y_i	Y_i^*	δ_i	Country	Y_i	Y_i^*	δ_i
EU	44 270	41 210	-	Lithuania	7	0.99	7.062
Belgium	3 187	3 741	0.705	Luxembourg	2 333	905	2.882
Bulgaria	68	23	1.040	Hungary	631	164	3.424
Czech	1015	864	1.212	Netherlands	2 617	2 321	1.211
Denmark	4 974	5 557	0.806	Austria	1 575	2 017	0.362
Germany	5 072	5 279	0.956	Poland	353	202	1.487
Ireland	347	178	1.909	Portugal	98	41	1.629
Greece	18	6	2.050	Romania	197	37	3.888
Spain	329	164	1.554	Slovenia	97	37	2.064
France	5 184	5 240	0.948	Slovakia	1 459	1 081	2.939
Italy	1 600	890	1.654	Sweden	5 023	4 591	1.226
Latvia	2	0.85	2.354	UK	8 082	7 875	1.003

Table 3 contain a part of estimated and true correspondences for some countries. The considerable part of the true data is inaccessible. For the pair “France-Germany” the estimated correspondence is equal to 153 thousands of passengers. The inverse correspondence is 155 thousands of passengers. Unfortunately, we are prevented from seeing true correspondence for this pair. In spite of that fact, obviously the suggested model is able to keep theoretical assumptions stated above and tendencies in transportations.

The estimated correspondence from Germany to Germany is equal to zero. According the stated model (20), such correspondences $Y_{i,i}^{*}$ should be equal to zero. True correspondence here is equal to 5072 thousands of passengers, what contradiscts to that assumption. Analysing Table 2, we can see, that estimated total departure for Germany is very close to the true one. This fact proofs correctness of balancing condition (26).

Table 3

Several estimated correspondences

Country of disembark.	Germany		France		Latvia		Lithuania		Netherlands	
Country of embarkation	$Y_{i,l}$	$Y_{i,l}^{*}$	$Y_{i,l}$	$Y_{i,l}^{*}$	$Y_{i,l}$	$Y_{i,l}^{*}$	$Y_{i,l}$	$Y_{i,l}^{*}$	$Y_{i,l}$	$Y_{i,l}^{*}$
Belgium	:	113	1852	601	:	0	:	0	1280	548
Bulgaria	:	1	:	0.1	:	0	:	0	:	0
Czech	:	581	0	5	:	0	0	0	:	3
Denmark	:	1683	0	37	:	0	:	0	:	51
Germany	5072	0	545	155	:	0.1	0	0.1	807	153
Ireland	:	9	0	15	:	0	:	0	:	6
Greece	:	1	:	0.4	:	0	:	0	:	0
Spain	:	15	315	27	:	0	:	0	:	4
France	:	153	3	0.00	:	0	:	0	520	197
Italy	:	324	777	109	:	0	:	0	:	24
Latvia	:	0	:	0	:	0	2	0.1	:	0
Lithuania	:	1	:	0	2	0.2	0	0	:	0
Luxembourg	:	223	238	572	:	0	:	0	:	113
Hungary	:	39	:	2	:	0	:	0	0	1
Netherlands	:	193	265	251	:	0	:	0	0	0
Austria	:	136	8	6	:	0	0	0	10	2
Poland	:	120	0	1	:	0.1	5	0.5	:	1
Portugal	:	1	0	1	:	0	:	0	:	0
Romania	:	12	0	0	:	0	:	0	:	0
Slovenia	:	7	0	1	:	0	0	0	:	0
Slovakia	:	18	0	0	:	0	0	0	:	0
Sweden	:	1156	0	61	:	0.3	:	0.1	:	49
UK	:	512	1181	3615	:	0	:	0	:	1007

Analyzing the results, can see that better estimates correspond to the old members of the EU. Probably, to improve the estimates for new members, it is necessary to consider ones separately.

So, nonlinear regression model for an estimation of the individually taken correspondence $Y_{i,l}$ between two geographical points has been suggested. The unknown model parameters were estimated using the gradient method. The necessary software has been elaborated in the MathCad 13 environment. Testing suggested approach for passenger railway correspondences estimation between the Member States of the European Union show quite good results.

8.3. APPLICATION OF THE SINGLE INDEX MODEL FOR TRANSPORTATIONS VALUES INVESTIGATION

8.3.1. ESTIMATION OF THE SINGLE INDEX MODEL

Recall that general formula of the single index model is the following:

$$E(Y \mid x) = m(x) = g\{v_\beta(x)\}. \tag{27}$$

Function $g(\bullet)$ is an *unknown link function*. As index function any appropriate smooth function can be taken. In our investigation a linear combination is used:

$$m(x_i) = g(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}) = g(\tau_i), \tag{28}$$

here $\tau_i = \beta^T x_i$ is called an *index*.

The estimation of the single index model consists of two steps. First we estimate the unknown coefficients vector β, and then using the index values for observations we estimate g by ordinary univariate nonparametric regression of Y on $v_{\hat{\beta}}(x)$. We need to perform two important things: to decide by which to replace the unknown link function g and to establish an appropriate object function to estimate the vector of unknown coefficients β. Moreover, there can be variation in choice of method for optimization of chosen object function.

Since there is only one assumption concerning unknown function $m(\bullet)$, i.e. it has to be a smooth function, for the latter the *Nadaraya-Watson kernel estimator* can be applied:

$$\tilde{g}(x) = \frac{1}{\sum_{i=1}^{n} K_h(\tau_i)} \sum_{i=1}^{n} K_h(\tau_i) Y_i, \tag{29}$$

where $\tau_i = (x - x_i)^T \beta$ is a value of index for the i-th observation, Y_i is a value of the dependent variable for i-th observation and $K_h(\bullet)$ is a *kernel function*. In our investigation we use only the Gaussian function as $K_h(\bullet)$:

$$K_h(\tau) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\tau}{h}\right)^2\right), -\infty < \tau < \infty, \tag{30}$$

where h is a *bandwidth*. The unknown coefficients vector β is estimated using the widely known LS criterion:

$$R(\beta)=\sum_{i=1}^{n}\left(Y_i-\tilde{g}(x_i)\right)^2\to\min_{\beta}. \tag{31}$$

For such minimization we will use the *gradient method*. The formula of the corresponding gradient is the following:

$$\nabla R(\beta)=-2\sum_{i=1}^{n}\left(Y_i-\frac{\sum_{i=1}^{n}K_h(\tau_i)Y_i}{\sum_{i=1}^{n}K_h(\tau_i)}\right)\left(\sum_{i=1}^{n}K_h(\tau_i)\right)^{-2}\left(\frac{1}{h}\sum_{i=1}^{n}\frac{\partial}{\partial\tau_i}K_h(\tau_i)\left(Y_i\sum_{i=1}^{n}K_h(\tau_i)-\tilde{\tilde{Y}}\right)x_i\right), \tag{32}$$

where

$$\tilde{\tilde{Y}}=\sum_{i=1}^{n}K_h(\tau_i)Y_i \tag{33}$$

and

$$\frac{\partial}{\partial\tau_i}K_h(\tau_i)=-\frac{\tau_i}{h^2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\tau_i}{h}\right)^2\right) \tag{34}$$

is the first derivative of the Gaussian kernel (30).

For simplification of procedure of the most significant model choice, we suggest comparing parametric and semiparametric models by the *residual sum of squares* R_0, which is calculated in such a manner:

$$R_0=\frac{1}{n-d}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2, \tag{35}$$

where n is a number of observations, d is a number of estimated coefficients, Y_i is an observed value and $\hat{Y}_i$ is an estimated value either by parametric or semiparametric model. So, $\hat{Y}_i$ can be represented by $\tilde{g}(x_i)$ or $\hat{\beta}^T x_i$.

Procedure I of the choice of the most significant single index model.

Begin of Procedure I.

1. To include in model the most significant factors by results of estimation of corresponding linear model (the respective *Procedure II* is described below). Set sequence number of the considered single index model $n = 0$.
2. $n = n + 1$. Choose the start value of bandwidth parameter $h^{(n)}$ on the basis of any tenable arguments. It might be a value chosen on the basis of Mahalanobis distance or a value selected using AMISE formula.
3. Choose the adequate accuracy level ε for the criterion (31) optimal value calculation and the adequate accuracy level η for the gradient (32) optimal value calculation.
4. Let $i = 0$. Choose the start value for unknown coefficients $\beta^{(n,i)}$ on the basis of any reasonable arguments.

5. Calculate value of the gradient $\nabla R(\beta^{(n,i)})$.
6. Calculate the current moving direction $\theta^{(i)}$ on the basis of current gradient $\nabla R(\beta^{(n,i)})$:

$$\theta^{(i)} = \frac{\nabla R(\beta^{n,(i)})}{\sqrt{\nabla R(\beta^{(n,i)})^T \cdot \nabla R(\beta^{(n,i)})}}. \quad (36)$$

7. Calculate value of the current moving step $\nu^{(i)}$ by trial and error method or by some of methods of one-dimensional optimization (dyhotomy method, method of golden section, Fibonacci method, etc.).
8. Let $i = i + 1$. Calculate the current value of unknown coefficients $\beta^{(n,i)}$ using the *steepest descent method* as one of variations of the gradient method:

$$\beta^{(n,i)} = \beta^{(n,i-1)} - \nu^{(i)} \cdot \theta^{(i)}. \quad (37)$$

9. Test for stop. If $\left|R\left(\beta^{(n,i-1)}\right) - R\left(\beta^{(n,i)}\right)\right| > \varepsilon$ then set $n = n + 1$ and go to 6.
10. Test on gradient equality to zero. If $\left|\nabla R\left(\beta^{(n,i)}\right)\right| > \eta$ then set $n = n + 1$ and go to step 5.
11. Declare current $\beta^{(n,i)}$ as optimal value β^*.
12. Construct equation of the best chosen SIM, obtaining the estimates of the forecasted values.
13. Calculate the residual sum of squares $R_0^{(n)}$.
14. If $n = 1$ then go to step 15. Else go to step 16.
15. Current $R_0^{(n)}$ is recognized as optimal value of the residual sum of squares (so called *record* $RE^{(n)}$) and go to step 2.
16. If current $R_0^{(n)}$ is less than last record $RE^{(n)}$ then recognize it as new record, set $RE^{(n)} \leftarrow R_0^{(n)}$.
17. If $h^{(n)}$ still may be changed then go to step 2.
18. Optimal SIM is estimated.

End of Procedure I.

As far the investigation of the SIM efficiency supposes its comparing with linear models (having the similar set of factors), the *Procedure II* for the choice of the most significant linear model is described as well, see full text of promotional work.

We record a fact to be used later, namely that stated above procedures allow arguing efficiency of considered models only in case of existing data smoothing. Thus it is necessary to distinguish criteria of *smoothing* and criteria of actually *forecasting*, i.e. of *cross-validation*.

Procedure of *smoothing* for estimation of coefficients of model is spent on all retrospective data. After that by means of the obtained estimates of regression models coefficients one find settlement values of the dependent variables $\hat{Y}$ corresponding to available observations *Y*. Quality of smoothing is

defined by size of a deviation of settlement estimates $\hat{Y}$ from the observed values of dependent variable *Y*.

In the elementary variant of *cross-validation* the data sample is divided into two parts. The data of the first part is used for obtaining of estimates of considered model coefficients. Then these estimates are applied for forecasting of values of the dependent variable corresponding to observations of second part of the data sample. Comparison of these forecasts to actual values allows arguing quality of forecasting.

We are going to test the efficiency of our suggested models in two cases: existing data smoothing and forecasting. Stated above procedures are able to be easily arranged for the case of cross-validation too.

8.3.2. Analysis and Forecasting of Turnover for International Rail Freight Transport for the EU

Problem Setting

The main *object* of consideration is a Member State of the European Union, named *country*. An *observation* is a data about object for a time moment.

The task of research is to construct regression models, and to choose from them ones given the best forecasts of turnover of rail freight transport.

We consider volumes of turnover, expressed in million tonne-km, for 15 old Member States of the EU. The analyzed time period is from 1996 to 2000. So, for 15 countries and 5 years we have 75 observations. The variable of interest is denoted by t_0. The explanatory variables are:

t_1 – country area, in thousands of km^2 (*SQUARE*);
t_2 – GDP per capita in Purchasing Power Standards (*GDP_PPS*);
t_3 – Comparative Price Level (*CPL*);
t_4 – total length of railways, in thousands of km, (*TOTLEN*);
t_5 – number of locomotives, in thousands, (*LOKOM*);
t_6 – number of goods wagons, in thousands (*WAGONS*).
t_7 – index of country area, (*GradAREA*).

Suggested Models

First two models are linear regression models, denoted *L*1 and *L*2. Other two models are *SIM*, denoted *SIM*1 and *SIM*2. The dependent variable $Y^{(L1)} = t_0$. Explanatory variables are $x_1 = t_2$, $x_2 = t_3$, $x_3 = t_2 / t_3$, $x_4 = t_4$, $x_5 = t_5$, $x_6 = t_6$. The dependent variable $Y^{(L2)} = t_0 / \sqrt{t_1}$. Explanatory factors are the same, in addition the index of the country area is introduced. The dependent variable $Y^{(SIM1)} = t_0 / t_1$. The dependent variable $Y^{(SIM2)} = t_0 / \sqrt{t_1}$. The sets of explanatory factors for the models *SIM*1 and *SIM*2 coincide with the set for model *L*1.

For estimation and comparing of stated four models we use *Procedures I* and *II*. The described abowe cross-validation approach is used as well. Especially for the single index model the series of experiments is performed

with the aim to determine the optimal value of bandwidth *h*.

Estimation of the Linear Models

The estimated models *L*1 and *L*2 are the following:

$$\hat{E}\left(Y^{(L1)}(x)\right) = -3\,713 + 118x_1 + 26x_2 - 11\,769x_3 + 879x_4 + 549x_5 + 158x_6. \quad (38)$$

$R^2 = 0.98$, $F = 384$.

$$\hat{E}\left(Y^{(L2)}(x)\right) = -120.4 - 1.2x_1 + 1.4x_2 + 110.2x_3 + 0.2x_4 + \\ + 5.9x_5 + 0.3x_6 + 29.2x_7. \quad (39)$$

$R^2 = 0.98$, $F = 314$.

Estimation of the Single Index Models

The estimates of coefficients β are obtained by optimal bandwidths $h = 7$ for *SIM*1 and $h = 6$ for *SIM*2. The estimated models can be rewritten in the following form:

$$\hat{E}\left(Y^{(*)}(x)\right) = \frac{\sum_{i=1}^{n} Y_i K_h\left((x - x_i)^T \hat{\beta}\right)}{\sum_{i=1}^{n} K_h\left((x - x_i)^T \hat{\beta}\right)}, \quad (40)$$

where $Y^{(*)}$ could be $Y^{(SIM1)}$ or $Y^{(SIM2)}$ depending on what model is considered at the moment. Vectors of estimated coefficients are $\hat{\beta}^T_{SIM1} = \left(710 \quad -1\times10^3 \quad 18\times10^{-5} \quad 758 \quad 155 \quad -2\times10^3\right)$ and $\hat{\beta}^T_{SIM2} = \left(716 \quad -1\times10^3 \quad -4 \quad 853 \quad 62 \quad -871\right)$.

Cross-validation analysis

We estimate the coefficients β on the basis of the period form 1996 till 1999 and perform the forecast for the 2000. The estimated models *L*1 and *L*2 are the following:

$$\hat{E}\left(Y^{(L1)}(x)\right) = 4154 + 198x_1 + 45x_2 - 20556x_3 + 899x_4 + 532x_5 + 148x_6. \quad (41)$$

$$\hat{E}\left(Y^{(L2)}(x)\right) = -120 + 28x_1 - 1.2x_2 + 1.4x_3 + 110x_4 + \\ + 0.2x_5 + 6x_6 + 0.3x_7. \quad (42)$$

The estimates of coefficients β are obtained by optimal bandwidths $h_0 = 7$ and $h_0 = 8$ for *SIM*1 and $h_0 = 6$ for *SIM*2. Vectors of estimated coefficients are $\hat{\beta}^T_{SIM1} = \left(33160 \quad -22420 \quad 566 \quad 19870 \quad 3996 \quad 56310\right)$ and

$\hat{\beta}^T_{SIM1} = (345 \quad -96 \quad 4 \quad 121 \quad 25 \quad 357)$ and

$\hat{\beta}^T_{SIM2} = (963 \quad 1\times10^3 \quad -2 \quad 604 \quad 49 \quad 690)$.

Table 4 contains the values of R_0 for investigated models. The observed and forecasted values of turnover are displayed on Figures 2 and 3.

Table 4

The values of R_0

Model	*L*1	*L*2	*SIM*1	*SIM*2
Smoothing	11 543 065	4 830 576	894 265	565 407
CV	18 509 464	8 941 875	1 894 237	1 896 287

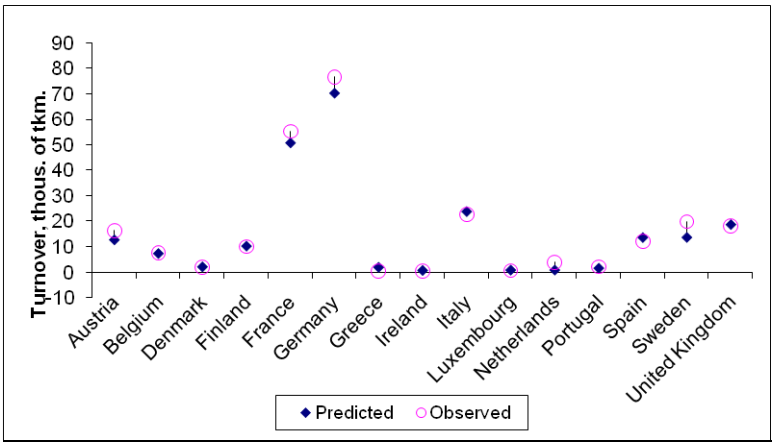

Fig.2. Forecasting by model *L*2

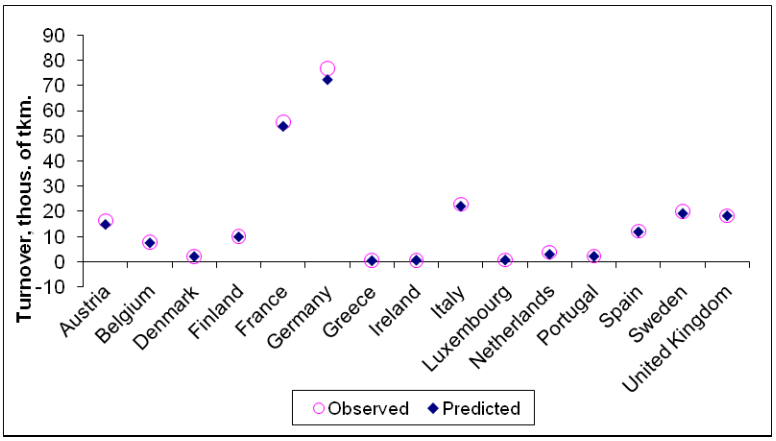

Fig.3. Forecasting by *SIM*2

Four different regression models were constructed and estimated, two of them are linear ones and two others are single index models. The efficiency of these models was investigated by the consideration of turnover for the 15 Member States of the EU. The performed investigations show that the single index regression model gives more precise forecasts than classical methods of linear regression. Moreover, the optimal values of smoothing parameter *h* for considered single index models have been obtained experimentally.

8.3.3. ANALYSING AND FORECASTING OF THE INLAND RAIL PASSENGER TRANSPORTATIONS FROM THE REGIONS OF LATVIA

Principal aim of the research is to construct and improve the group parametric and semi-parametric models equally well describing both large and small rail passenger transportations from the regions of Latvia.

The main problem is in transportations volumes inconsistency. Transportations from the biggest cities exceed transportations from others in times of orders. The reason is that inhabitants use the railway in the different purposes. Inhabitants of big regions use railway transport as usual municipal transport, but of small cities – for long-distance trips.

Many researches have been spent, in which observations have been divided onto two groups depending on transportation size. Model describing only small transportations has been developed. The best results have been achieved in case of outliers excluded according to Mahalanobis distance. It is intended to describe almost all results.

Problem Setting

The main object of consideration named *object* is a region of Latvia performing certain rail passenger transportation, further named *region*. We call as *observation* a data about object for an actual time moment of analyzed period.

The variable of interest, denoted by t_0, is volume of internal rail passenger transportations, expressed in thousands of passengers, for a region of Latvia. All the 33 regions have been selected. The analyzed period is from 2000 to 2003. Total we have 113 observations. The explanatory factors are:

t_1 – the population density, thous. of residents, *TP/1000*SQUARE*;
t_2 – the number of enterprises per a unit of territory, *NE/SQUARE*;
t_3 – the number of enterprises per 1000 residents, *NE/1000*TP*;
t_4 – the density of the unemployed population, in thousands, *UP/SQUARE*;
t_5 – the density of general education institution, *NGEI/SQUARE*;
t_6 – the number of buses per a unit of territory, *NB/SQUARE*;
t_7 – the number of buses per 1000 residents, *NB/1000*TP*;
t_8 – the number of railway stations, *NRS*.

The dependent variables for both models $Y^{(1)}$ and $Y^{(2)}$ are t_0. Explanatory variables in both models are all eight mentioned above.

Each experiment consists of models estimation in smoothing, after that both models are estimated in forecasting using cross-validation approach. The unknown coefficients β are estimated on the base of period from 2000 till 2002. Then using the obtained estimates of β we forecast the transportations for the 2003. It is intended to show only results without surplus comments.

Results of Experiments for Full Data

All 113 observations for all 33 regions have been analyzed. Besides some regions have observations on all the considered period, others do not have.

Smoothing. Equation for the best chosen linear model can be written as follow:

$$\hat{E}(Y(x)) = -75 - 4x_1 + 323x_2 + 168x_3 - 476x_7 + 146x_8. \quad (43)$$

R^2 = 0.94, F = 358.

Estimation of coefficients β for the single index model has been performed with different bandwidths. The best chosen single index model corresponds to h^* = 1 and can be presented in the form (40), where vector of estimated coefficients is $\hat{\beta}^T = (9183 \quad -395 \quad 7375 \quad -3523 \quad 0.1)$.

Cross-validation approach. Equation for the chosen linear model is:

$$\hat{E}(Y_{CV}(x)) = -167 - 3.9x_1 + 311x_2 + 144x_3 - 406x_7 + 140x_8. \quad (44)$$

R^2 = 0.94, F = 245. This model gives negative forecasts for seven cases, i.e. about in 30% of observations.

The best chosen single index model corresponds to h^* = 30 and vector of estimated coefficients is $\hat{\beta}_{CV}^T = (134500 \quad 3975 \quad 8273 \quad -1287 \quad 24270)$.

We have collected the values of R_0 for both investigated models in cases of smoothing and cross-validation in Table 5. Table 6 contains the observed and forecasted transportations obtained by both investigated models for the analyzing period and the relative error δ_i in % for each observation.

Table 5

The values of R_0

	Smoothing	CV
Linear Model	884 141	1 181 086
SIM	133 556	466 016

Results of Experiments for Restricted Data

88 observations according to 22 regions have been analyzed. Only regions on which observations over all 4 considered years present are taken.

Smoothing. Equation for the best chosen linear model can be written as follow:

$$\hat{E}(Y(x)) = -2094 - 19x_1 + 389x_2 + 83x_3 + 75995x_5 - 886x_6 + 155x_8. \quad (45)$$

R^2 = 0.95, F = 306. The best single index model corresponds to h^*= 1 and vector of estimated coefficients is $\hat{\beta}^T = (848 \quad -73 \quad 226 \quad -0.1 \quad -188 \quad 0.1)$.

Cross-validation approach. Equation for the best chosen linear model is:

$$\hat{E}(Y_{CV}(x)) = -2135 - 21x_1 + 380x_2 + 75x_3 + 86963x_5 - 928x_6 + 146x_8. \quad (46)$$

R^2 = 0.96, F = 225. The best SIM corresponds to h^* = 30 and vector of

estimated coefficients is $\hat{\beta}_{CV}^{T} = (21620 \quad 235 \quad 2216 \quad 3 \quad 132 \quad 7522)$. The values of R_0 are collected in Table 7. Unfortunately, linear model gives negative forecasts for eight small transportations, i.e. about in 36% of observations. Single index model predicts the small transportations more efficiently. That fact shows necessity of construction of separate models for forecasting of the large and small transportations.

Table 6

Comparative results in case of cross-validation

	Observed values	Forecasts		Relative error, %	
		Linear Model	SIM	Linear Model	SIM
Riga	23323.05	20067.43	22370.00	13.96	4.09
Daugavpils	259.85	979.57	301.83	-276.98	-16.16
Jelgava	1676.11	1611.31	1412.00	3.87	15.76
Jurmala	6436.22	3360.92	4682.00	47.78	27.26
Rezekne	192.50	1532.21	193.77	-695.95	-0.66
Aizkraukles r.	795.41	427.85	120.98	46.21	84.79
Cesu r.	272.15	332.88	174.91	-22.31	35.73
Daugavpils r.	47.71	-618.34	72.59	1395.97	-52.14
Dobeles r.	43.10	78.08	142.96	-81.17	-231.72
Gulbenes r.	10.25	183.36	94.47	-1689.53	-822.01
Jelgavas r.	267.40	315.14	132.57	-17.85	50.42
Jekabpils r.	251.74	522.49	126.90	-107.55	49.59
Limbazu r.	96.13	-71.44	85.70	174.32	10.85
Ludzas r.	90.78	-152.40	84.19	267.88	7.26
Madonas r.	66.30	-165.41	130.82	349.50	-97.33
Ogres r.	4108.32	2269.00	2149.00	44.77	47.69
Preilu r.	104.37	155.95	100.04	-49.43	4.14
Rezeknes r.	53.34	-119.31	81.32	323.67	-52.46
Rigas r.	6219.64	5660.02	4813.00	9.00	22.62
Saldus r.	9.36	-249.69	115.56	2768.52	-1134.99
Tukuma r.	698.92	332.84	149.75	52.38	78.57
Valkas r.	74.60	-97.24	89.15	230.34	-19.50
Valmieras r.	86.74	62.98	185.40	27.39	-113.74

Table 7

The values of R_0

	Smoothing	CV
Linear Model	823 311	1 335 774
SIM	171 435	909 120

Results of Experiments for Slight Flows

95 observations (only small transportations) on 28 regions are analyzed.

Smoothing. Equation for the best estimated linear model is such:

$$\hat{E}(Y(x)) = -255 - 82x_2 + 16x_3 - 34x_4 + 13033x_5 + 18x_8. \qquad (47)$$

R^2 = 0.48, F = 17. The best chosen single index model corresponds to h^* = 0.5

and vector of estimated coefficients $\hat{\beta}^T = (12 \quad 265 \quad 78 \quad 0.3 \quad 0.1)$.

Cross-validation. Equation for the best estimated linear model is:

$$\hat{E}(Y_{CV}(x)) = -223 - 74x_2 + 14x_3 - 28x_4 + 11258x_5 + 16x_8. \quad (48)$$

R^2 = 0.42, F = 10. The best SIM corresponds to h^* = 2 and vector of estimated coefficients is $\hat{\beta}_{CV}^T = (-38 \quad 2555 \quad 180 \quad 0.1 \quad 1561)$. The values of R_0 are collected in Table 8.

Table 8

The values of R_0

	Smoothing	CV
Linear Model	42 430	79 029
SIM	8 877	40 992

Removal of Outliers Corresponding to Mahalanobis Distance

91 observations have been processed. The following improvements of the present models have been done:

1) introducing the categorized variable into model;
2) removal of outliers according to Mahalanobis distance.

Smoothing. Equation for the best estimated linear model is such:

$$\hat{E}(Y(x)) = -20x_1 + 436x_2 + 163x_3 + 64418x_5 - 734x_6 - 398x_7 + 147x_8. \quad (49)$$

R^2 = 0.96, F = 258.

Now the gradational variable GRAD is introduced. It takes value 1 for regions which have obvious large transportations and value 0 for others. Equation for the linear model after modification has the following form:

$$\hat{E}(Y_G(x)) = -20x_1 + 618x_2 + 81x_3 + 122x_4 + 38525x_5 - 921x_6 + 63x_8 + 2743x_9. \quad (50)$$

R^2 = 0.97, F = 339. The best chosen single index model corresponds to h^* = 1 and vector of estimated coefficients $\hat{\beta}^T = (8959 \quad -386 \quad 7195 \quad 0.1 \quad -3437 \quad 0.1)$.

Regression equation after removal of outliers is such:

$$\hat{E}(Y_M(x)) = 2307x_2 + 59x_3 - 393x_4 - 1207x_6 - 361x_7 + 128x_8. \quad (51)$$

R^2 = 0.91, F = 91. The best chosen single index model corresponds to h^* = 1 and vector of estimated coefficients is $\hat{\beta}_M{}^T = (56 \quad 820 \quad 24 \quad 4 \quad 114 \quad 0.1)$.

Cross-validation. Linear regression equation has the following form:

$$\hat{E}\left(Y_{CV}(x)\right) = 2692 + 3863x_2 - 525x_4 - 117031x_5 - 416x_7 + 110x_8. \qquad (52)$$

$R^2 = 0.91$, $F = 94$. The best chosen single index model corresponds to $h^* = 1$ and vector of estimated coefficients is $\hat{\beta}_{CV}{}^T = (283 \quad -380 \quad 4 \quad 2709 \quad 0.1)$. The values of R_0 are collected in Table 9. The observed and forecasted values of transportations are displayed on Figures 4 and 5.

Table 9

The values of R_0

Model	Smoothing		CV
	before removal of outliers	after	
LM	801 111 / 546 895[2]	259 096	739 151
SIM	165 843	10 053	147 810

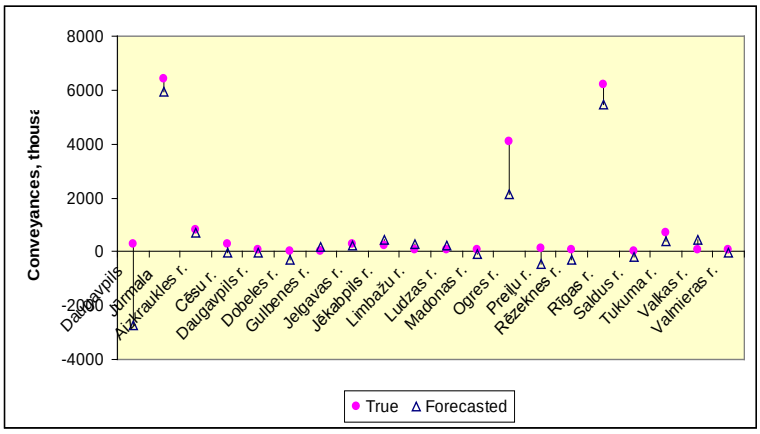

Fig.4. Forecasting by the Linear Model

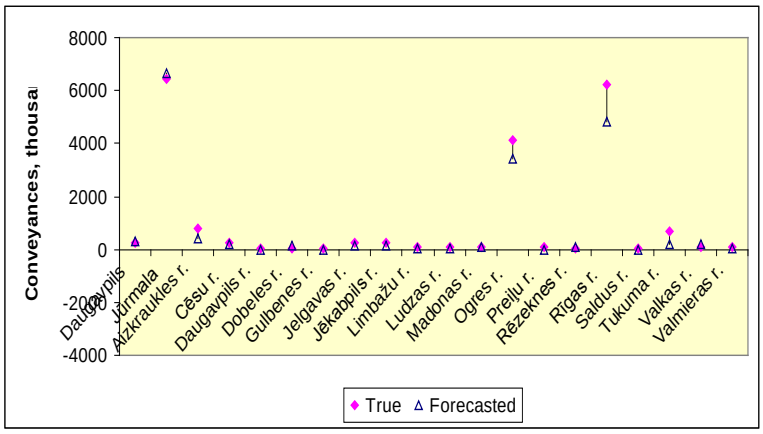

Fig.5. Forecasting by the single index model

Several models of multiple regression, which allow evaluating the influence of the main social-economic factors on the volumes of passenger transportations by the railway transport in the regions of Latvia, have been

[2] including variable GRAD

obtained.As the result two group models were compared: the multiple linear regression model and the single index model. Removal of outliers from data set according Mahalanobis distances has decreased the error of smoothing and, it its turn, has increased precision of forecasting. The results of analysis show the preference of the single index model in cases of smoothing and forecasting.

8.4. INTERNATIONAL AIR PASSENGER TRANSPORTATIONS FORECASTING FOR THE EU MEMBER STATES ON THE BASIS OF SURE-MODEL

SURE-model, proposed by *A.Zellner* in 1962, is appropriate and useful for a wide range of applications in econometrics, logistics and other areas. Some generalization of the SURE-model is considered. Individually taken observations are supposed not to contain the information about all response variables but about a part of them. An unbiased estimate for a covariance matrix is obtained. Evidence of SURE-model usage advantage before multivariate model in case of the data incompleteness on the example of the total air passenger transportations forecasting is stated on the basis of practical results of Master Thesis of *S.Bogdanova*, performed under supervision of the Candidate.

Problem Setting

SURE-model (6) and its important distinctions from multivariate model (4) are described in Chapter 3 of present Summary. So, final aim is to obtain the prognosis of the expectation of the sum

$$W(t) = \sum_{i=1}^{G} Y_{i,j(t)} = \sum_{i=1}^{G} \left(\sum_{v=1}^{m_i} \beta_{i,v} x_{i,j(t)}^{(v)} + Z_{i,j(t)} \right), \tag{53}$$

for future time moment *t*. Unknown coefficients $\{\beta_{i,v}\}$ are estimated using the well known formula

$$\beta_i^* = \left(X_i^T X_i \right)^{-1} X_i^T Y_i, \tag{54}$$

where $\beta_i^* = \left(\beta_{i,1}^*, \beta_{i,2}^*, ..., \beta_{i,m_i}^* \right)^T$ and $Y_i = \left(Y_{i,1}, Y_{i,2}, ..., Y_{i,n_i} \right)^T$ are the vectors of the estimates and the dependent variables. The unbiased estimate for the expectation of the sum $W(t)$ (53):

$$W^*(t) = \sum_{i=1}^{G} Y_{i,j(t)}^* = \sum_{i=1}^{G} x_{i,j(t)} \beta_i^* = \sum_{i=1}^{G} \left(\sum_{v=1}^{m_i} \beta_{i,v}^* x_{i,j(t)}^{(v)} \right). \tag{55}$$

According to (54), the covariance matrix of two coefficient vectors β_i^* and β_l^* is calculated by formula

$$Cov(\beta_i^*, \beta_l^*) = \left(X_i^T X_i\right)^{-1} X_i^T Cov(Y_i, Y_l) X_l \left(X_l^T X_l\right)^{-1}. \quad (56)$$

The covariance $Cov(Y_i, Y_l)$ should be calculated as well. Let $D^{(i,l)}$ be $(n_i \times n_l)$-matrix for which $D_{j,f}^{(i,l)} = 1$ if $t_{i,j} = t_{l,f}$ and $D_{j,f}^{(i,l)} = 0$ otherwise. The covariance of two dependent variables $Y_{i,j}$ and $Y_{l,f}$ for the same time moment $t_{i,j} = t_{l,f}$ is equal to $c_{i,l}$. Therefore $Cov(Y_i, Y_l) = c_{i,l} D^{(i,l)}$. Rewrite the formula (56) in the following form:

$$Cov(\beta_i^*, \beta_l^*) = c_{i,l} \left(X_i^T X_i\right)^{-1} X_i^T D^{(i,l)} X_l \left(X_l^T X_l\right)^{-1}. \quad (57)$$

For unknown covariance $\{c_{i,l}\}$ estimation the LS estimator is used:

$$c_{i,l}^* = \frac{1}{v_{i,l}} \left(Y_i - X_i \beta_i^*\right)^T D^{(i,l)} \left(Y_l - X_l \beta_l^*\right), \quad (58)$$

where $v_{i,l}$ is a constant that is determined by a condition of the unbiased estimator. For constant $v_{i,l}$ defining the expectation of estimate $c_{i,l}^*$ is needed.

Let $R_j(i)^T$ be the *j*-th row of the matrix $\left(I_i - X_i (X_i^T X_i)^{-1} X_i^T\right)$, $f(l, i, j)$ is the observation number of the *l*-th object, for which the time coincides with $t_{i,j}$, and is equal to zero if such number is absent itself:

$$f(l, i, j) = \sum_{v=1}^{n_l} v D_{j,v}^{(i,l)} . \quad (59)$$

Then

$$E(c_{i,l}^*) = \frac{1}{v_{i,l}} \sum_j \sum_f E\left(Z_{i,j} R_j(i)^T D^{(i,l)} R_f(l) Z_{l,f}\right) =$$

$$= \frac{1}{v_{i,l}} \sum_j R_j(i)^T D^{(i,l)} R_{f(l,i,j)}(l) E(Z_{i,j} Z_{l,f(l,i,j)}) = \frac{1}{v_{i,l}} c_{i,l} \sum_j R_j(i)^T D^{(i,l)} R_{f(l,i,j)}(l).$$

With the following value of constant $v_{i,l}$

$$v_{i,l} = \sum_j R_j(i)^T D^{(i,l)} R_{f(l,i,j)}(l). \quad (60)$$

formula (58) gives the unbiased estimate of the covariance $c_{i,l}$. Now the variance of the sum (55) is calculated by usual way:

$$Var(W^*(t)) = Var\left(\sum_{i=1}^{G} x_{i,j(t)}\beta_i^*\right) = \sum_{i=1}^{G} x_{i,j(t)} Cov(\beta_i^*) x_{i,j(t)}^T + \\ + 2\sum_{i=1}^{G-1}\sum_{l=i+1}^{G} x_{i,j(t)} Cov(\beta_i^*, \beta_l^*) x_{l,j(t)}^T. \quad (61)$$

Here $Cov(\beta_i^*, \beta_l^*)$ is calculated by formula (57) and the covariance matrix of the vector β_i^* is calculated by the well known formula

$$Cov(\beta_i^*) = c_{i,i}\left(X_i^T X_i\right)^{-1} = \sigma_i^2\left(X_i^T X_i\right)^{-1}. \quad (62)$$

Forecasting Total Air Passenger Transportations for the EU Member States

Suggested approach is demonstrated on an example of forecasting total air passenger transportations for the EU Member States. There are two forecasted variables: internal and external passenger air international transportations, denoted by Intra-EU and Extra-EU, measured in thousand of passengers. Both variables are forecasted together owing to their strong correlation with each other, i.e. *the total forecast* of two considered dependent variables is obtained.

We intend to obtain forecasts of total transportations under condition of the incompleteness of the statistical data. The corresponding regression model contains main economical, social and structural factors affected internal and external transportation for each country. In this investigation 24 countries have been considered, i.e. all the Member States excluding Bulgaria, Sweden and Finland.

So, we have two objects ($G = 2$) with numbers $i = 1, 2$. Each object has one own dependent variable: for the first object dependent variable Y_1 is the values of the international Intra-EU passenger air transportations, for the second object Y_2 is the values of the international Extra-EU passenger air transportations. The model has 7 independent variables. First three of them are common for both objects:

t_1 – country area *SQUARE*, in thousands of km^2;

t_2 – employment growth, annual percentage change in total employed population, *EGAP*;

t_3 – population change, *PC*.

Next two belongs only to the first object:

t_4 – total population *TP*, in thousands of inhabitants;

t_5 – growth rate of GDP volume, % change on previous year, *GDPrate*.

Last two correspond to the second object only:

t_6 – final energy consumption by transport *FECT*, in 1000 toe;

t_7 – consumption of electricity by industry, transport activities and households/services *CEITH*, GWh.

Analyzed period is from 2000 till 2007. Actually, some data for objects are missing. In this case usually two approaches are used:

- observations for all objects for the corresponding time moment are excluded;
- missing values are estimated by some appropriate way.

Both these approaches deform statistical data and forecasting results. To analyze the efficiency of the stated approach we fix the observations on which data are absent. So, first object has 122 observations, second has 100 observations. It is investigated, how much the number of missing observations influence the variance of sum of interest.

We use two approaches of data processing:

1. some observations are absent and statistical data for all corresponding time moments are extracted for both objects, after that number of observations for each object is 99;
2. some observations are absent and suggested approach is applied.

As efficiency criterion *the estimate of variance of sum of interest* is used. For that aim we process the data for analyzed period and make forecast for 2008 on the basis of estimated coeeficients. We wish to obtain estimates of variance as good as the random variables $Y_{1,j}$ and $Y_{2,j}$ are dependent. For that purpose we are able to use formula (61). In the present example we have two matrixes R (R_1 and R_2), one matrix $D^{(1,2)}$ and one normalizing constant $v_{1,2}$.

Calculations of normalizing constant have been taking a lot of time because of many operations with matrixes of quite high dimension. Estimated regression coefficients and covariance matrix are

by Approach 1:

$$\beta_1^* = \begin{pmatrix} 0.02 \\ 49.59 \\ -21.32 \\ 4.15 \\ -0.66 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} 8.79 \\ -28.35 \\ 28.82 \\ 0.22 \\ 164.29 \end{pmatrix}, \text{ and } \tilde{\tilde{\Sigma}} = \begin{pmatrix} 1.684 \cdot 10^8 & 1.687 \cdot 10^7 \\ 1.687 \cdot 10^7 & 8.234 \cdot 10^6 \end{pmatrix};$$

by Approach 2:

$$\beta_1^* = \begin{pmatrix} -0.19 \\ 49.12 \\ -21.09 \\ 4.14 \\ -0.66 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} 11.93 \\ -30.78 \\ 29.68 \\ 0.21 \\ 102.48 \end{pmatrix}, \text{ and } \tilde{\tilde{\Sigma}} = \begin{pmatrix} 1.668 \cdot 10^8 & 8.538 \cdot 10^6 \\ 8.538 \cdot 10^6 & 8.994 \cdot 10^6 \end{pmatrix}.$$

The normalizing constant is $v_{1,2} = 92.976$.

The results of the total air passenger transportations forecasting are presented in Table 10 and Table 11. The first column of Table 10 contains the observed volumes of total air passenger transportations for 2008, next two columns contain forecasts obtained by both approaches, next two ones contain variances, expressed in thousands, and two last columns represent upper bounds of 95% confidence limits, by both approaches as well. In Table 11 values of R_0

in case of smoothing and forecasting are represented.

Table 10

Forecasting results

Country	2008	2008*, approach No		Variance, thous., approach No		95% upper conf. limits, approach No	
		1	2	1	2	1	2
Belgium	21 982	19 283	19 499	6 197	5 572	35 495	36 061
Czech	13 429	14 548	14 454	2 952	2 602	26 676	26 350
Denmark	24 629	15 217	15 377	10 600	8 764	30 073	30 295
Germany	165 759	129 700	129 710	42 180	36 090	222 560	223 376
Estonia	1 804	3 829	3 692	3 585	3 134	8 958	9 385
Ireland	30 016	18 900	18 916	3 968	3 475	34 053	34 289
Greece	34 790	29 426	29 539	4 111	3 625	51 566	51 584
Spain	161 401	114 130	115 750	30 750	26 100	196 267	198 208
France	122 724	142 710	141 520	18 080	16 200	240 645	239 066
Italy	106 300	105 940	106 630	22 630	20 780	181 543	182 349
Cyprus	7 218	5 209	5 400	4 798	4 197	11 903	12 448
Latvia	3 687	5 717	5 656	3 513	2 999	12 450	12 116
Lithuania	2 552	6 439	6 349	3 262	2 820	13 522	13 166
Luxembourg	1 713	10 221	9 945	3 951	3 642	20 022	19 440
Hungary	8 429	15 116	15 050	3 002	2 649	27 632	27 351
Malta	3 125	2 013	2 203	3 064	2 710	6 001	6 484
Netherlands	50 419	41 197	41 496	5 523	4 843	71 172	71 908
Austria	23 900	14 515	14 624	3 736	3 222	26 748	27 153
Poland	18 727	33 708	33 346	12 030	10 450	60 583	60 376
Portugal	25 047	23 317	23 344	3 034	2 648	40 909	41 141
Romania	8 031	11 260	10 960	6 872	5 842	22 430	22 274
Slovenia	1 649	2 407	2 181	4 189	3 724	7 304	6 742
Slovakia	2 596	2 779	2 684	2 930	2 619	7 364	7 056
UK	213 888	183 880	184 090	48 430	46 640	312 976	313 108

Table 11

The values of R_0

	Approach No	
	1	2
Smoothing	386 427 788	243 593 829
Forecasting	235 066 477	224 884 525

Obvioulsy, forecasts gotten by the suggested approach are closer to the real values of total air passenger transportations than gotten by usual way. It demonstrates the advantage of the stated approach for obtaining the total interest in case of incompleteness of statistical data.

CONCLUSION

1. Promotional work is devoted to development and application of modern statistical methods for the transportations volumes analysis and

forecasting for the EU Member States, accentuating Latvia. As nowadays a steady growth of the number of both passenger and freight transportations all over the world is observed, the chosen direction is perspective, and represented work is actual.

2. The subject mattes are the following kinds of transportations: passenger and freight depending on type of transportation; international and internal – on transportation mode; rail and air – on mode of vehicle; object of forecasting are countries or regions of countries and OD-pairs; forecasted indicator are transportations or departures and turnover. Analysis and forecasting of mentioned above transportations have been performed on the basis of parametric and nonparametric models. Among parametric models there were used both linear (i.e. multiple, multivariate and SURE-models) and nonlinear modified gravity model. Semiparametric single index model has been widely applied as well.
3. The wide range of problems concerning to the tasks of the carried out research is studied: classification of types of transportations, factors influencing the volumes of the transportations; various approaches and methods of forecasting.
4. The review of literature devoted both to parametric and nonparametric methods and models of the transportations volumes forecasting, and to methods of correspondence matrixes estimation.
5. As the initial model for the forecasting of the volumes of the passenger and freight both transportations and turnover, the classical model of the multiple linear regression has been taken. A great number of the linear regression models have been created. In contrast to the linear regression models, presented in majority of considered scientific publications, all our models are group ones, and they include greater number of explanatory factors and their combinations.
6. Let us state considered tasks following to models classification:

- Semiparametric models developing and evaluation for rail freight turnover volumes forecasting for the Member States of the EU and rail passenger departures forecasting from regions of Latvia. In this connection the corresponding statistical data bases have been set, which contain information about factors influencing forecasted indexes. Single index model estimation methodics has been suggested, the procedures for choosing the most singnificant models have been developed. During many experiments the single index model advantage before the linear model has been demonstrated, as in case of data smoothing, and in case of forecasting. In latest the cross-validation approach has been used.
- Nonlinear parametric model working out for correspondence matrix estimation of rail passenger departures between the EU Member States. The suggested model is an original modification of the gravity model. Considered procedure for the model estimation has been developed and

verified. The results of obtained estimates comparative analysis with true correspondences testify the high efficiency of the suggested model.

- Total air freight and total air passenger transportations forecasting for the EU Member States. For this purpose the multivariate regression model and SURE-model have been applied. For the latter the original procedure for variance evaluation of total forecast has been worked out. The advantage of SURE-model use in the case of data incompleteness has been shown by comparing of variances of forecasts, gotten with both approaches.

All the needed calculation have been performed in the Statistica 6.0 package. For procedures suggested during the presented work, the software has been developed in MathCad13 environment.

7. Models and methods developed by the author for forecasting freight rail volumes of transportations for the EU countries were used in the scientific project U7107 "Mathematical models and their estimation method elaboration for analysis and forecasting of the Baltic Region passenger and Freight flows" which was a component of the scientific project II "Zinātniskās darbības attistība augtskolās" and lasted from June, 01st till December, 31st 2006.
8. Models and methods developed by the author for forecasting passenger rail volumes of departures for the regions of Latvia were used in the scientific project U1212 "Creation of mathematical models, algorithms and computer programs for Latvia's transport system's analysis, development prognosis and optimization" which was a component of the scientific project "Zinātniskās darbības attistība augtskolās" and lasted from June, 01st till December, 31st 2008.
9. On the basis of the obtained results a part of the course of lectures on the subject "Mathematical Methods of Traffic Flow Analysis and Forecasting" for the second year foreign students of bachelors studies programme of the RTU Mechanical Engineering faculty has been prepared.
10. The main results of this investigation are published in 9 articles and presented at 11 international scientific conferences held in Latvia, Lithuania, Estonia, Germany, Greece and Switzerland at which the author presented 11 reports on the subject of the promotion work. The list of articles and reports at conferences are given at the end of the summary.

PUBLICATIONS WITH THE AUTHOR'S PARTICIPATION

1. Andronov A., Zhukovskaya C. and Santalova D. On Mathematical Models for Analysis and Forecasting of the Europe Countries Conveyances. RTU Zināniskie Raksti, Datorzinātne. – 5. Sērija, Informācijas tehnoloģija un vadības zinātne, 28. Sējums (2007), pp. 96 – 106.

2. Andronov A., Santalova D. On Nonlinear Regression Model for Correspondence Matrix of Transport Network. The XIIIth International Conference on Applied Stochastic Models and Data Analysis. 2009, June 30 – July 3, Lithuania, Vilnius. – Vilnius: Vilnius University, Applied Stochastic Models and Data Analysis International Society, L.Sakalauskas, C.Skiadas and E.K.Zavadskas (Eds.), 2009, pp. 90-94.
3. Jackiva I, Santalova D. Faktoru analīzes izmantošana transporta nozares attīstības tendenču noteikšanai ES valstīs. RTU zinātniskie raksti, Mašīnzinātne un Transports. – 6. Sērija, Intelektuālas Transporta Sistēmas, 18. Sējums (2005), pp. 139 – 147.
4. Kopytov E, Santalova D. Application of the Single Index Model for Forecasting of the Inland Conveyances. Recent Advances in Stochastic Modelling and Data Analysis. Editor: Christos H. Skiadas. – Singapore: World Scientific Publishing Co Pte Ltd., 2007, pp. 268 – 276.
5. Santalova D. Vairumtirdzniecības noliktavas pārdošanu lieluma regresijas modelis. RTU zinātniskie raksti, Mašīnzinātne un Transports. – 6. Sērija, Intelektuālas Transporta Sistēmas, 18. Sējums (2005), pp. 67 – 73.
6. Santalova D. Regression Model of Sales Volume from Wholesale Warehouse. 13th International Conference on Analytical and Stochastic Modelling Techniques and Applications. 2006, May 28 – 31, Germany, Bonn, Sankt Augustin.– Bonn: Bonn-Rhein-Sieg University of Applied Science, Khalid Al-Begain (Ed.), 2006, pp. 133 – 137.
7. Santalova D. Forecasting of Rail Freight Conveyances in EU Countries on the Base of the Single Index Model. Computer Modelling and New Technologies. – Vol. 11, No. 1. (2007), pp. 73 – 83.
8. Santalova D. The Use of Multivariate Regression Model to Forecast the Freight Air Transportations in the Members Countries of the Europe Union. The International Conference „Modelling of Business, Industrial and Transport Systems – 2008". 2008, May 7 – 10, Latvia, Riga. – Riga: TSI, 2008, pp. 258 – 266.
9. Santalova D. Forecasting of Passenger Conveyances in Latvian Regions Applying Semiparametric Regression Models. RTU zinātniskie raksti, Mašīnzinātne un Transports. – 6. Sērija, Intelektuālas Transporta Sistēmas, 18. Sējums (2008), in appearance.

CONFERENCES IN WHICH THE AUTHOR TOOK PART

1. The 46th Scientific Conference of Riga Technical University. Riga, Latvia, October 13 – 15, 2005. Report: „Vairumtirdzniecības noliktavas pārdošanu lieluma regresijas modelis", author: Santalova D.
2. International Scientific Conference: ECMS 2006, ASMTA 2006. Bonn-Rhein-Sieg University of Applied Science, Bonn, Sankt Augustin, Germany, May 28 – 31, 2006. Report: „Regression Model of Sales Volume from Wholesale Warehouse", author: Santalova D.

3. The 47th Scientific Conference of Riga Technical University. Riga, Latvia, October 12 – 14, 2006. Report: „Aim and problems of the project "Mathematical models and their estimation methods elaboration for analysis and forecasting of the Baltic region passenger and freight flows", author: Santalova D.
4. The 6th International Conference on Reliability and Statistics in Transportation and Communication (RealStat'06). Riga, Latvia, October 25 – 28, 2006. Report: „Forecasting of rail freight conveyances in EU countries on the base of the Single Index Model", author: Santalova D.
5. The XIIth International Conference on Applied Stochastic Models and Data Analysis (ASMDA-2007). Applied Stochastic Models and Data Analysis International Society, Agronomic Institute of Chania – Greece, Crete, Chania, May 28 – June 2, 2007. Report: „Application of the Single Index Model for Forecasting of the Inland Conveyances", authors: Kopytov E., Santalova D.
6. The 8th Tartu Conference on Multivariate Statistics and The 6th Conference on Multivariate Distributions with Fixed Marginals. Tartu, Estonia, June 25 – 31, 2007. Report: „Single Index Model for Railway Passenger Conveyances Forecasting in Regions of Latvia", author: Santalova D.
7. The 48th Scientific Conference of Riga Technical University. Riga, Latvia, October 11 – 12, 2007. Report: „Investigation of classical and semiparametric regression for the forecasting of rail conveyances in Latvia", author: Santalova D.
8. The International Conference "Modelling of Business, Industrial and Transport Systems" (MBITS'08). Riga, Latvia, May 7 – 10, 2008. Report: „The Use of Multivariate Regression Model to Forecast the Freight Air Transportations in the Members Countries of the European Union", author: Santalova D.
9. The 2nd International Workshop on Computational and Financial Econometrics (CFE-2008). University of Neuchatel, Neuchatel, Switzerland, June 19 – 21, 2008. Report: "On Some Generalization of Seemingly Unrelated Regression Equation Models", authors: Andronov A., Santalova D., Svirchenkov A.
10. The 49th Scientific Conference of Riga Technical University. Riga, Latvia, October 13 – 15, 2008. Report: „Forecasting of passenger conveyances in Latvian regions applying semiparametric regression models", author: Santalova D.
11. The XIIIth International Conference on Applied Stochastic Models and Data Analysis (ASMDA-2009). Applied Stochastic Models and Data Analysis International Society, Vilnius University, Vilnius, Lithuania, June 30 – July 3, 2009. Report: „On Nonlinear Regression Model for Correspondence Matrix of Transport Network", authors: Andronov A., Santalova D.