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

Promotional work

Scientific supervisor
Dr. hab. sc. ing., professor
Alexander ANDRONOV

Riga 2009

ANOTĀCIJA

Promocijas darbu «Pusparametriskie regresijas modeļi krāvu un pasažieru pārvadājumu apjomu analīzei un prognozēšanai» izstrādājusi Diāna Santalova inženierzinātņu doktora zinātniskā grāda iegūšanai. Darba zinātniskais vadītājs ir Dr. hab. sc. ing., profesors Aleksandrs Andronovs.

Darba mērķis ir dažādu transporta veidu pasažieru un kravu pārvadājumu apjomu prognozēšana ar mūsdienu statistisko metožu pielietošanu. Patlaban visā pasaulē novērojama kā ražošanas dekoncentrēšana, tā arī iedzīvotāju mobilitātes pieaugums. Kā secinājums ir visu transporta veidiem pasažieru un kravu pārvadājumu apjomu nepārtraukts pieaugums. Tas liecina par izvēlētā darba virziena perspektīvumu, tātad darbs ir aktuāls.

Pirmais pētījuma virziens ir grupu modeļu izstrādāšana un verifikācija pārvadājumu apjomu prognozēšanai ES valstīm. Šajā ziņā bija atrisināti četri uzdevumi:

- summāro kravu aviopārvadājumu prognozēšana uz parametriskā daudzdimensiju modeļa bāzes;
- pasažieru dzelzceļa nosūtījumu korespondenču matricas novērtēšana ar modificētā parametriskā gravitācijas modeļa palīdzību;
- dzelzceļa kravu apgrozījuma novērtēšana un prognozēšana uz pusparametriskā viena indeksa modeļa bāzes;
- summāro pasažieru aviopārvadājumu prognozēšana datu nepilnības nosacījumā ar parametriskā SURE-modeļa izmantošanu.

Otrais virziens ir grupu pusparametrisko modeļu izstrādāšana un izvērtēšana pasažieru dzelzceļa nosūtījumu analīzei un prognozēšanai Latvijas reģioniem.

Tika piedāvātas jaunas modeļu novērtēšanas metodes, izstrādāti veidojamo modeļu kvalitātes verifikācijas kritēriji. Parādītas viena indeksa modeļu pielietošanas priekšrocības salīdzinājumā ar lineārās regresijas modeļiem.

Promocijas darbs ir uzrakstīts angļu valodā, satur ievadu, 7 nodaļas, nobeigumu, literatūras sarakstu, 2 pielikumus un glosāriju. Darbā ietilpst 147 formulas, 60 tabulas, 30 zīmējumi, kopā 164 teksta lappuses. Literatūras sarakstā ir 128 nosaukumi.

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 are 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.

The promotion work is written in English. The given work consists of 7 chapters. The bibliography includes 128 sources. There are 30 figures, 147 formulas and 60 tables to illustrate the conception of the carried out research. The promotion work contains 164 pages.

АННОТАЦИЯ

Промоционная работа «Полупараметрические регрессионные модели для анализа и прогнозирования объемов грузовых и пассажирских перевозок» разработана Дианой Санталовой на получение научной степени доктора инженерных наук. Научный руководитель хабилитированный доктор инженерных наук, профессор Александр Андронов.

Целями работы является прогнозирование объемов грузовых и пассажирских перевозок различными видами транспорта с применением современных статистических методов. В настоящее время в мире наблюдается как рассредоточение производства по разным странам, так и увеличение подвижности населения. Следствием этого является неуклонный рост объемов грузовых и пассажирских перевозок всеми видами транспорта. Это свидетельствует о перспективности выбранного направления исследования, так что данная работа представляется актуальной.

Первым направлением исследования является разработка и оценивание групповых моделей для прогнозирования объемов перевозок для стран ЕС. В этой связи были решены четыре задачи:

- прогнозирование суммарных грузовых авиационных перевозок на базе параметрической многомерной модели;
- оценивание матрицы корреспонденций пассажирских железнодорожных отправок с помощью модифицированной параметрической гравитационной модели;
- оценивание и прогнозирование железнодорожного грузооборота на базе полупараметрической одноиндексной модели;
- прогнозирование суммарных пассажирских авиационных перевозок при условии неполноты данных с использованием параметрической SURE-модели.

Вторым направлением является разработка и оценивание групповых полупараметрических моделей для анализа и прогнозирования пассажирских железнодорожных отправок для регионов Латвии.

Предложены новые методы оценивания моделей, разработаны критерии верификации их качества. Показано преимущество применения полупараметрических моделей по сравнению с линейной регрессионной моделью.

Промоционная работа написана на английском языке и состоит из 7 глав. Библиография включает в себя 128 источников. В работе 30 рисунка, 147 формул и 60 таблиц, которые поясняют концепцию проведенного исследования. Промоционная работа состоит из 164 страниц.

CONTENTS

"Good roads, canals and navigable rivers, by diminishing the expense of carriage, put the remote parts of the country more nearly on a level with those in the neighbourhood of the town. They are upon that account the greatest of all improvements."

Adam Smith, The Wealth of Nations, 1776

ABBREVIATIONS

pdf	probability density function
AMISE	asymptotic MISE
CV	cross-validation
EU	European Union
GAM	generalized additive model
GAPLM	generalized additive partial linear model
GDP	Gross Domestic Product
GLM	generalized linear model
GPLM	generalized partial linear model
IRLS	iteratively reweighted least squares
LS	least squares
MISE	mean integrated squared error
MLE	maximum likelihood estimator
MSE	mean squared error
OLS	ordinary least squares
PLM	partial linear model
PMLE	pseudo maximum likelihood estimator
RSS	residual sum of squares
SIM	single index model
SLS	semiparametric least squares
SURE-model	Seemingly Unrelated Resgression Equation model
U	unity element
WLS	weighted least squares

INTRODUCTION

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.

Actuality of the problem

Statistically, the transport sector is an important part of the economies of the Member States of the European Union, accounting for a substantial proportion of private spending, employment and government expenditure. Changes within the sector potentially have serious and widespread effects on welfare, and so it is important that the economic understanding of the sector is comprehensive. It is only with a sound base of understanding that policy-makers can most effectively develop the sector and preempt, and minimise, the effects of any adverse changes that inevitably occur. Establishing this understanding is the role of the transport economist [81].

Transport is vital to the economy and the way we live. Taking a long-term view of the likely trends in the key metrics (traffic, congestion and environmental impacts) is important to be able to make the right policy decisions early enough to have an impact. Furthermore, there is uncertainty regarding the development of many of the key drivers of transport demand – such as the development of the economy, and changes in fuel prices – and it is important to consider the range of transport outcomes to which this uncertainty could give rise [80].

Transport is the complicated system which working capacity substantially defines rates, rhythm and productivity of social and economic development of the country and its regions.

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

- making up of vehicle park on prospect,
- a transport network planning,
- determination of the tendencies of requirement for capacities and the investment into development of transport and its components,
- distribution of freight and passenger transportations volumes between modes of transport,

- planning of development of a network of transport objects,
- placements of vehicle parks and development of a network of maintenance depots.

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 only the correct economic policy in this sphere can provide competitiveness and efficiency. Thus, application of modern mathematical models and methods is necessary for correct decisions developing and for right actions planning. Obviously, for getting the qualitative forecasts of transportations volumes the modern efficient mathematical models have to be used.

In this connection in the present research the following mathematical models have been investigated:

1. Parametric models
    1.1. Linear models
        1.1.1. Multiple regression models [95], [120], [127]
        1.1.2. Multivariate regression models [95], [120]
        1.1.3. SURE-model [13], [115], [116]
    1.2. Nonlinear models
        1.2.1. Generalized linear models [75]
        1.2.2. Gravity models [12], [28], [121]
2. Nonparametric models
    2.1. Nadaraya-Watson estimator [53], [76], [104]
    2.2. Single index model [53], [63].

We would like to accentuate the fact, that similar models and methods have not been applied in Latvia before – neither for forecasting of passenger transportations, nor for forecasting of freight transportations. 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 as one of the youngest and most perspective.

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

Objectives and tasks of the 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.

Concerning to the stated division and objectives of the present research the following *tasks* are considered:

1. Considering problems of passenger and freight transportations forecasting.
2. Investigation nowadays used statistical models and methods for forecasting.
3. Analyzing models of modern regression theory, considering opportunities of their application for forecasting.
4. Performing the analysis of the factors influencing the volumes of passenger and freight transportations.
5. Creation research information base, making statistical data collections for the Member States of the Europe Union and for regions of Latvia.
6. Developing the models of passenger and freight transportations volumes forecasting

on the basis of multiple linear regression models and semiparametric regression models.

7. Developing methods and algorithms for estimation of offered regression models.
8. Verification and efficiency estimation of developed models.
9. Demonstration of semiparametris model advantage in comparison with multiple linear models.
10. Gravity model modification and method and algorithm developing for the suggested model estimation.
11. Suggested model applying for estimation of passenger departures correspondence matrix.
12. SURE-model formalization and method and algorithm developing for the offered model estimation.
13. Demonstration SURE-models advantages in comparison with multivariate model in condition of data incompleteness.

Methodology and methods of the 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 [4], [46].
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*". [5], [7], [30], [31, [32].
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 [123], [125], [126].

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.

Practical importance and applying

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 of new highways etc.

Author' publication

The main results of the present investigation are published in 9 articles [3], [6], [12], [14], [63], [84], [85], [86], [87] and presented at 11 international scientific conferences held in Latvia, Lithuania, Estonia, Germany, Greece and Switzerland at which the author has presented 11 reports according to the promotion work subject. The list of reports on the conferences are given at the Appendix 1.

Structure of the promotional work

Chapter 1. Review of State-of-the-art for Considered 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. The structures of 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 is 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.

1. REVIEW OF STATE-OF-THE-ART FOR FORECASTING PROBLEM

1.1. THE IMPORTANCE OF TRANSPORTATIONS FORECASTING

As it is noted in Introducion, transport is one of the major branches of any country economy developing. Roads, railway lines and inland waterways, as well as seaports, airports and railway stations, form the basic transport infrastructure in the European Union. A modern transport infrastructure of a high standard is the basis for the mobility of goods and passengers and thus essential both for regional economic development and for the creation of an internal European market. In keeping with the high importance of inland transport infrastructure for the economic development of the European regions, investments in road and rail infrastructure account for a major share of the Community's regional budgets [30].

Trends in transport performance, especially those of goods transport, follow economic developments. According to Panorama of Transport [36] and EuroSTAT Yearbook 2008 [31], while gross domestic product (GDP) grew at an average yearly rate of 2.4 % from 1995 to 2007, goods transport performance, measured in tonne-kilometres, grew at 2.8 % yearly. Over the period, passenger transport performance, measured in passenger-kilometres, grew at an average yearly rate of 1.7 %. Besides, according to forecasts based upon a comprehensive survey of the airline industry performed by International Air Transport Association (IATA), international air passenger traffic was supposed to increase at an average annual growth rate of 4.8% between 2006 and 2010 [58].

Changes in the transport sector, and especially private transport, clearly have wide-reaching effects across economies. The transport sector also accounts for a substantial proportion of employment.

Till the middle of 2008, i.e. before the world financial crisis began, that tendency was kept. Last one and a half year the economical conditions in the world, in the EU in particular, and especially in so-called new Member States of EU, is unstable enough. Thereupon now is not reasonable to allege about steady increase in transportations from the end of 2007 till nowadays. Unfortunately, the statistical data of transportations volumes for the period is not available on the EuroSTAT at the moment. However, in our disposal there is an proof of that GDP steadily decreases, and number of the unemployed grows as well. This fact should make against increase in transportations.

So, in the press release 100/2009, published by Eurostat [38], is resulted, that in the Euro area[1] (EA16) GDP fell by 2.5% and EU27 GDP fell by 2.4% during the first quarter of 2009, compared with the previous quarter, according to second estimates from Eurostat. In the fourth quarter of 2008, growth rates were -1.8% in both zones. In comparison with the same quarter of the previous year, seasonally adjusted GDP declined in the first quarter of 2009 by 4.9% in the Euro area and by 4.7% in the EU27, after -1.7% and -1.6% respectively in the previous quarter.

In the first quarter of 2009, all Member States for which seasonally adjusted GDP data are available registered a negative growth rate compared with the previous quarter, except Poland (+0.4%) and Cyprus (0.0%).

Moreover, in the press releases 112/2009 and 25/2009 [40], [41] is told about the fact, that Euro area (EA16) seasonally-adjusted unemployment rate was 9.4% in June 2009, compared with 9.3% in May. It was 7.5% in June 2008. The EU27 unemployment rate was 8.9% in June 2009, compared with 8.8% in May. It was 6.9% in June 2008. For the Euro area this is the highest rate since June 1999 and for the EU27 since June 2005.

Eurostat estimates that 21.526 million men and women in the EU27, of which 14.896 million were in the Euro area, were unemployed in June 2009. Compared with May, the number of persons unemployed increased by 246 000 in the EU27 and by 158 000 in the Euro area. Compared with June 2008, unemployment went up by 5.024 million in the EU27 and by 3.170 million in the Euro area.

Among the Member States, the lowest unemployment rates were recorded in the Netherlands (3.3%) and Austria (4.4%), and the highest rates in Spain (18.1%), Latvia (17.2%) and Estonia (17.0%). Compared with a year ago, all Member States recorded an increase in their unemployment rate. The smallest increases were observed in Germany (7.3% to 7.7%), Romania (5.7% to 6.2% between the first quarters of 2008 and 2009) and the Netherlands (2.7% to 3.3%). The highest increases were registered in Estonia (4.6% to 17.0%), Latvia (6.4% to 17.2%) and Lithuania (5.1% to 15.8%). Figure 1.1 demonstrates unemployment rate increasing compared to January 2009.

According to the press release 111/2009, Euro area annual inflation was expected to be 1.2% in February 2009 and -0.6% in July 2009 according to a flash estimate issued by

[1]The euro area (EA16) consists of Belgium, Germany, Ireland, Greece, Spain, France, Italy, Cyprus, Luxembourg, Malta, the Netherlands, Austria, Portugal, Slovenia, Slovakia and Finland. The EU27 includes Belgium, Bulgaria, the Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, the Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden and the United Kingdom.

Eurostat, the Statistical Office of the European Communities. It was 1.1% in January and -0.1% in June 2009 [39].

It is logical to assume that in such unstable political and economic conditions is especially important to have authentic forecasts of transportations for competent state budget redistribution. For authentic forecasts obtaining the adequate mathematical models should be used. Besides not only widely known and used parametrical ones, but semiparametric and nonparametric as well. Latest are the models, allowing to consider large sets of the factors influencing transportations now and in the future. In the present Thesis both parametric and semiparametric models are investigated, a lot of attention has been paid to the single index model. Late is more flexible, than parametrical one, but has disadvantages as well.

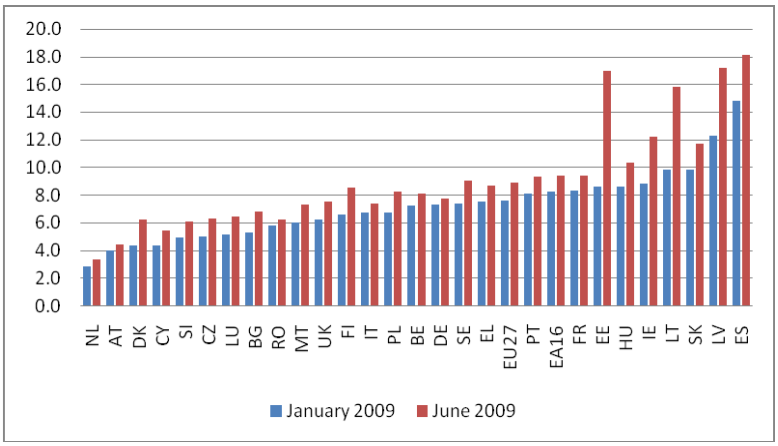

Fig.1.1. Unemployment rates in the EU

Thereupon further the short review of the researches spent in the areas of nonparametric and semiparametric regression is reconciled.

1.2. THE SURVEY OF LITERATURE

First of all, semiparametric and nonparametric models have been widely investigated by *W.Hardle*, *M.Muller*, *S.Sperlich* and *A.Werwatz* in [53]. This book divides itself into two parts. First of them is dedicated to nonparametric models and covers the methodological aspects of nonparametric function estimation for cross-sectional data, in particular kernel smoothing methods. A closely related topic to consider is nonparametric density estimation. Since many techniques and concepts for the estimation of probability density functions are also relevant for regression function estimation, histograms and kernel density estimates are

considered in more details. Several methods of nonparametrically estimating regression functions are introduced as well. Mainly it is kernel regression, but other approaches such as splines, orthogonal series and nearest neighbor methods are also covered.

The second part is devoted to semiparametric regression models, in particular extensions of the parametric generalized linear models (*GLM*). In fact, widely used in the framework of the present Thesis single index model (*SIM*) is a nonparametric extension of *GLM*. Single index models focus on the nonparametric error distribution in an underlying latent variable model. In addition to the methodological aspects, the second part also covers computational algorithms for the considere models.

In the field of nonparametric regression we need to mention such works, as *Density Estimation for Statistics and Data Analysis* by *B.W.Silverman* [88], *Applied Nonparametric Regression* of *W.Hardle* [51], *Applied Smoothing Techniques for Data Analysis* by *A.Bowman* and *A.Azzalini* [18], *Smoothing Methods in Statistics* by *J.Simonoff* [89], and *Nonparametric Econometrics* by *A.Pagan* and *A.Ullah* [78].

For general aspects on semiparametric regression we refer to the textbooks of *A.Pagan* and *A.Ullah* [78], *Semiparametric Regression for the Applied Econometrician* by *A.Yatchew* [114], and *Semiparametric Regression* by *D.Ruppert*, *M.P.Wand* and *R.J.Carrol* [83].

Comprehensive presentations of the generalized linear model can be found in such fundamental investigation, as *Generalized Linear Models* of *P.McCullagh* and *J.A.Nelder* [75], *An introduction to Generalized Linear Models* by *A.J.Dobson* [27], *Nonparametric Regression and Generalized Linear Models* by *P.J.Green* and *B.W.Silverman* [47], and *Generalized Linear Models and Extensions* by *J.Hardin* and *J.Hilbe* [50].

The idea of using parametric objective function and substituting unknown components by nonparametric estimates has first ben proposed by *S.Cosslett* in 1983 [22]. The maximum score estimator introduced by *C.Manski* in 1985 and the maximum rank correlation estimator stated by *A.Han* in 1987 are of the same type [71], [49]. Their resulting estimates are still very close to the parametric estimates. For that reason the semiparametric least squares (further *SLS*) method of *H.Ichimura* (1993) may outperform them when the parametric models are misspecified [59]. The pseudo maximum likelihood version of *SLS* was found independently by *S.Weisberg* and *A.H.Welsh* in 1994. They present it as a straightforward generalization of the *GLM* algorithm and discuss numerical details [107]. A different idea for adding nonparametric components to the maximum likelihood function is given by *A.Gallant* and *D.Nychka* in 1987. They use a Hermite series to expand the densities of the objective function

[45]. The different methods for *SIM* estimation have been compared in a simulation study by *M.Bonneu*, *M.Delecroix* and *E.Malin* in 1993. They also include a study of *M.Bonneu* and *M.Delecroix* in 1992 for a slightly modified pseudo likelihood estimator [17], [118]. An alternative average derivative estimation (further *ADE*) method (without weight function) was proposed by *W.Hardle* and *T.M.Stoker* in 1989. This estimator shares the asymptotic properties of the weighted ADE, but requires for practical computation a trimming factor to guarantee that the estimated density is bounded away from zero [54].

Partial linear models were first considered by *P.J.Green* and *B.S.Yandell* in 1985, *L.Denby* in 1986, *P.E.Speckman* in 1988 and *P.M.Robinson* in 1988. See respectively [48], [26], [93], [82].

Additive models were first considered for economics and econometrics by *W.Leontief* in 1947 [65], [66]. Intensive discussion of their application to economics can be found in *A.Deaton* and *J.Muellbauer* (1980) and *M.Fuss*, *D.McFadden* and *Y.Mundlak* (1978), see respectively [25] and [44]. *W.Wecker* and *C.Ansley* introduced especially the backfitting method in economics in 1983 [105]. The marginal integration estimator was first presented by *D.Tjøstheim* and *B.Auestad* in 1994 [99] and *O.Linton* and *J.P.Nielsen* in 1995 [68]. *E.Masry* and *D.Tjøstheim* in 1995 and 1997 use marginal integration and prove its consistency in the context of time series analysis, see [73] and [74].

The semiparametric approach to partial linear models can already be found in *P.J.Green* and *B.S.Yandell* in [48]. The way they present it was developed by *P.E.Speckman* in 1988 [93] and *P.M.Robinson* in 1988 [82]. A variety of generalized models can be found in the monograph *Generalized Additive Models* of *T.J.Hastie* and *R.J.Tibshirani*. This concerns in particular backfitting algorithms [55]. The literature on marginal integration for generalized models is very rare. Particularly interesting is the combination of marginal integration and backfitting to yield efficient estimators as discussed in 2000 by *O.Linton* [67].

Despite fitting of single index models does not require distributional assumptions on the error term, the properties of the estimates depend on such assumptions. In this connection, *Simonoff J.S.* and *Tsai C.-L.* in 2000 described a score tests for three potential misspecifications of the single index model: heteroscedasticity in the errors, autocorrelation in the errors, and the omission of an important variable in the linear index. These tests have a similar structure to corresponding tests for nonlinear regression models [90].

In spite of *SIM* is one of the most popular and most investigated semiparametric models, its researches is going on. *M.Hristache*, *A.Juditski*, and *V.Spokoiny* suggested in 1998

a new method of estimating the index coefficients in a single index model which is based on iterative improvement of the average derivative estimator. The resulting estimate is $\sqrt{n}$-consistent under mild assumption on the model [56]. *Y.Xia*, *H.Tong*, *W.K.Li* and *L.Zhu* proposed in 2002 an adaptive estimation method of dimension reduction space [110].

In 2006 *Y.Xia* derived asymptotic distributions of two mentioned above estimators of *SIM*. Efficiency comparisons between these estimation methods are made as well [112]. *E.Kong* and *Y.Xia* show in their paper [61] how the delete-m-out cross-validation method (CV(m)) can be used for variable selection in *SIM*. Applying the method to the Swiss banknotes data, a *SIM* with selected variables indeed has much better prediction ability than a *SIM* with all the variables.

H.Wong, *W.C.Ip* and *R.Zhang* proposed the varying-coefficient single-index model (*VCSIM*) in 2008. It can be seen as a generalization of the semivarying-coefficient model by changing its constant coefficient part to a nonparametric component, or a generalization of the partially linear single-index model by replacing the constant coefficients of its linear part with varying coefficients. They obtained based on the local linear method, average method and backfitting technique. The estimates of the unknown parameters and the unknown functions of the *VCSIM* and their asymptotic distributions are derived. Both simulated and real data examples are given to illustrate the model and the proposed estimation methodology [109].

L.Xue and *L.Zhu* considered a semiparametric regression model for longitudinal data. They developed two calibated empirical likelihood approaches for the regression coefficients and the baseline function estimation. Compared with methods based on normal approximations, the empirical likelihood does not require consistent estimators for the asymptotic variance and bias [113].

L.J.Keele in [60] demonstrates the potential of semiparametric techniques using detailed empirical examples drawn from the social and political sciences. This book includes also software for implementing the methods in *S-Plus* and *R*. *M.R.Fengler* in [43] gives recent advances in the theory of implied volatility and refined semiparametric estimation strategies and dimension reduction methods for functional surfaces.

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* [29], *Methods of Multivariate statistics* of *M.Srivastava* [95], *Applied Multivariate Statistics for the Social Sciences* of *J.Stevens* [97], *Theory of Point Estimation* and *Statistical Hypothesis Testing* of *E.L.Lehman* [64], [124], *Applied Linear Regression* of *S.Weisberg* [106], *The Analysis of*

Time Series: An Introduction of *C.Chatfield* [21], and *Discrete Choice Analysis: Theory and Applications to Travel Demand* of *M.Ben-Akiva* and *S.Lerman* [16].

Author cannot neglect such works of her teachers, as *Probability Theory and Mathematical Statistics* (in Russian) by *A.Andronov*, *E.Kopytov* and *L.Gringlaz* [120] and *Computer Based Methods of Statistical Data Processing* (in Russian) by *I.Jatskiv* [127].

In the field of pdf estimation we would sign researches of *V.Lyumkis* [69] and *N.Ushakov* [102].

Concerning OD-matix estimation is necessary to underline such works as *Modelling Transport* of *J. de D. Ortuzar* and *L.G.Willumsen* [77] and latest researches of *A.Andronov* [9], [10], [12].

In the field of such improved parametric model investigation, as SURE-model, we sign such authors, as *A.Zellner* [115], [116], *R.Velu* and *J.Richards* [103], and *A.Andronov* [13].

Convincing contribution to air passenger flows forecasting using various modified gravity models has been made by *A.Andronov et al.*; see [121], [12] and by *J. Doganis* in [28].

In the field of passenger and freight transportations analysis and forecasting by means of modern methods of multivariate statistics we can underline such researchers, as *A.Baublys* [15], *J.Butkevičius et al.* [19], [20], *A.Cokasova* [23], [24], *N.R.Farnum* and *L.W.Staton* [42], *Ü.Hunt* [57], *T.Šliupas* [92], *E.Spissu* [94], *N.Taneja* [98], *S.Wheatcroft* and *G.Lipman* [108], *S.Makridakis*, *S.Wheelwright* and *R.Hyndman* [70] and promotional works of *H.Afanasyeva* [122] and *V.Demidovs* [2]. The results of wide research concerning National Transport Model spent by Dept. for Transport (UK) can be found in [81].

CONCLUSION

In the given Chapter adverse changes of such main economic indicators of each country, as gross domestic product, inflation and unemployment rates, are shown. It means that the situation, recent developed in the world, most likely will affect volumes of transportations not in the best way. Thereupon for authentic forecasts obtaining authors recommend to use more flexible semiparametric models. The review of the researches spent to areas of semiparametric and nonparametric estimation, testifies that the given area of research is new and perspective.

2. INFORMATIONAL SUPPORT OF FORECASTING PROBLEM

In the previous Chapter it has been told about importance of forecasting of volumes passenger and freight transportations for more effective and comprehensive planning of the basic economic indicators of any country. As it is known, one of pledges of the qualitative forecast obtaining is presence of representative statistical date. Therefore in the given Chapter it will be a question on the statistics used for carrying out of the present investigation. First of all, kinds of transportations as predicted indicators are described. Further classifications of the factors influencing transportations are given. Two bases of the statistical data gathered during work are described. The statistics from these databases is used for carrying out of the basic experiments. Also the characteristic of sources of this statistical data is given.

2.1. TYPES OF TRANSPORTATIONS AND INFLUENCING FACTORS

On the accepted international classification distinguish two indicators concerning the passengers' transport [80]:

- *Passenger transportation* (*or departure*) is the volume of passengers carried from one place in another for the certain period of time. The unit of measure is thousands of transported passengers.
- *Passenger turnover* is the volume of transport work for passengers' transportation. The unit of measurement of turnover is passenger-kilometres. It is determined as a sum produced by multiplying the number of passengers for each position of transportation by the length of distance covered.

This principle is applicable for freight transport as well:

- *Freight transportation* (*or departure*) is the volume of goods carried from one place in another for the certain period of time. The unit of measure is tonnes of transported goods.
- *Freight turnover* is the volume of transport work for goods' transportation. The unit of measurement of turnover is tonne-kilometres. It is determined as a sum produced by multiplying the goods in tonnes for each position of transportation by the length of distance covered.

By the mode of vehicle the indicators can be divided on *air*, *rail*, *road*, *sea*, *pipeline* and *total*. By transportation mode the indicators can be divided on *international*, *internal* (*domestic*), *long-distance*, *suburban* and *intracity*. Depending on frequency of registering the

indicators can be divided on *annual*, *quarter*, *monthly*, *week*, *daily* and *hour*. As objects of forecasting the countries, regions, cities, the airports, stations, routes and directions can act. For a direction the pair of objects (cities, countries and so on) from which one is an origin point, and the second one is destination point (Origin-Destination pair, OD-pair, Correspondence Matrix, Trip Matrix) [77] is accepted.

Here the basic problem, in the author's opinion, is the knowledge of a technics of gathering and the analysis of corresponding statistics on transportations. In case of air transportation where the accounting of passengers is exact enough, two technics of gathering of statistics are used. The first technics concerns a case if indirect flights and transfers (or so-called indirect transit) are considered only. The second one takes into account direct transit flights. Let us state some crucial definition of air passenger transportations concerning to *Reference Manual on Air Transport Statistics* [37] and to *Glossary for Transport Statistics* [35].

Passengers on board are all passengers on board of the aircraft upon landing at the reporting airport or at taking off from the reporting airport. All revenue and non revenue passengers on board an aircraft during a flight stage. Includes direct transit pass counted at arrivals and departures.

Flight stage is the operation of an aircraft from take-off to its next landing.

On flight origin and destination. Traffic on a commercial air service identified by a unique flight number subdivided by airport pairs in accordance with point of embarkation and point of disembarkation on that flight. For passengers, freight or mail where the airport of embarkation is not known, the aircraft origin should be deemed to be the point of embarkation; similarly, if the airport of disembarkation is not known, the aircraft destination should be deemed to be the point of disembarkation.

The difference between on flight origin/destination and flight stage data can be illustrated by the following example: a flight is operated on a route New York-London-Paris 185 passengers travel from New York to London, 135 from New York to Paris and 75 from London to Paris. Thus in terms of on flight origin/destination data the figures recorded are 185 passengers New York-London, 135 passengers New York-Paris and 75 passengers London-Paris. New York would record the figures for New York-London and New York-Paris; London would record New York-London and London-Paris; Paris would record New York-Paris and London-Paris. In terms of flight stage data there are two flight stages and the figures reported by New York and London airports are: New York-London 320=(185+135)

passengers and by London and Paris airports are London-Paris 210=(135+75) passengers. The following diagrams give an example of reporting transport in datasets A1 (*Flight Stage*) and B1 (*Flight Origin and Destination*).

CASE 1

Journey from New York to London and then from London to Paris with 2 different Aeroplanes, i.e. 2 different flight numbers (Fig.2.1).

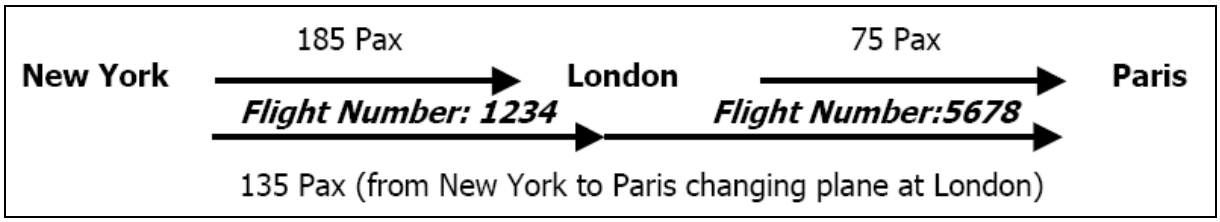

Fig.2.1. Scheme of journey with 2 flights

As we can see, in case of *transfer* or *indirect transit* passengers, the passengers figures reported in A1 are equal to the passenger figures reported in B1 (see Fig.2.2).

Reporting Airport	Next/Previous Airport	(A/D) Arrival/Depart.	A1 - Flight stage passengers	B1 - On Flight OD passengers	*True OD passengers (not to be reported)*
Reported by USA					
New York	London	D	320	320	*185*
Reported by UK					
London	New York	A	320	320	*185*
London	Paris	D	210	210	*75*
Reported by France					
Paris	London	A	210	210	*75*

Fig.2.2. Data to be reported for Case 1

CASE 2

Journey from New York to London and then from London to Paris with the same Aeroplane (same flight number), making a transit in London (Fig.2.3).

Fig.2.3. Scheme of journey with the same flight number

Reporting Airport	Next/Previous Airport	(A/D) Arrival/Depart.	A1 - Flight stage passengers	B1 - On Flight OD passengers	*True OD passengers (not to be reported)*
Reported by USA					
New York	London	D	320	**185**	*185*
New York	**Paris**	**D**	**-**	**135**	*135*
Reported by UK					
London	New York	A	320	**185**	*185*
London	Paris	D	210	**75**	*75*
Reported by France					
Paris	**New York**	**A**	**-**	**135**	*135*
Paris	London	A	210	**75**	*75*

Fig.2.4. Data to be reported for Case 2

Obviously, in case of *direct transit* passengers (Fig.2.4), the passenger figures reported in A1 are different from the passenger figures reported in B1.

Next let us give the main definition concerning rail transport according to *Regulation 1192/2003* and to *Glossary for Transport Statistics* [34], [35].

Transport of passengers by rail means the movement of passengers using railway vehicles between the place of embarkation and the place of disembarkation. The transport of passengers by metro, tram and/or light rail is excluded. In the case of data tables broken down by type of transport, the transit transport has not to be included in the international transport.

Transport of goods by rail means the movement of goods using railway vehicles between the place of loading and the place of unloading.

Number of passengers means the number of trips by rail passengers, where each trip is defined as the movement from the place of embarkation to the place of disembarkation, with or without transfers from one rail vehicle to another. If passengers use the services of more than one railway undertaking, when possible they should not be counted more than once. The calculation of passenger travel may not be exact. Passengers travelling from the point A to the point C with a changing of trains in the point B should be considered as one journey. Passengers who bought a ticket but did not travel should in principle not be counted. In practice there may be no clear and easy solution. Eurostat will accept different methodologies but they have to be explained by the Member States in a document.

Place of loading/embarkation (of the goods or passengers railway transport vehicle) is the place in which goods (passengers) are loaded (embarked) on the railway transport vehicle.

Passenger embarked is a passenger who boards a railway vehicle to be conveyed by it.

Place of unloading/disembarkation (of the goods or passengers railway transport vehicle) is the place in which goods (passengers) are unloaded (disembarked) from the railway transport vehicle. *Passenger disembarked* is a passenger alighting from a railway vehicle after having been conveyed by it.

Note, that a passenger transfer from one railway vehicle directly to another one, regardless of the railway undertaking, is not regarded as disembarkation/embarkation. Whenever during the transfer another mode of transport is used, this is to be regarded as disembarkation from a railway vehicle followed by a subsequent embarkation on a railway vehicle.

National transport means rail transport between two places (a place of loading/embarkation and a place of unloading/disembarkation) located in the reporting country. It may involve transit through a second country.

International transport total means rail transport between a place (of loading/embarkation or unloading/disembarkation) in the reporting country and a place (of loading/embarkation or unloading/disembarkation) in another country.

International transport-outgoing means rail transport between a place of loading/embarkation located in reporting country and a place unloading/disembarkation in another country.

International transport-incoming means rail transport between a place of unloading/disembarkation located in reporting country and a place loading/embarkation in another country.

Transit means rail transport through the reporting country between two places (a place of loading/embarkation and a place of unloading/disembarkation) outside the reporting country. Transport operations involving loading/embarkation or unloading/disembarkation of goods/passengers at the border of the reporting country from/onto another mode of transport are not considered as transit.

Example. A transit country is a country other than the country of loading or unloading. An example is a transport of goods, loaded in Germany and unloaded in Spain, that involves a transit through France. Consequently, France is the transit country. The type of transport is broken down as follows (Table 2.1).

As we can see, total freight rail transport includes transit, but total passenger transport does not include. It is noteworthy to underline, that reconciled classification allows having a quite clear notion of predicted indicators.

Table 2.1

Types of transport

No of column	Type of transport	Freight	Passengers
0	Total	0 = 1+2+5	0 = 1+2
1	National	1	1
2	International– total	2 = 3 + 4	2 = 3 + 4
3	International - outgoing	3	3
4	International - incoming	4	4
5	Transit	5	-

In the mathematical relation, however, it is indifferent that means under an analyzed indicator. As it has been noted in [121]: «In mathematical model the indicator acts as the dependent variable abstracted from the economic maintenance». The resulting indicators of forecasting are passenger turnover, freight turnover, passenger transportations (or departures) and freight transportations (or departures).

Depending on the time period of forecasting distinguish short-term forecasting (for a day, week up to a year), intermediate-term forecasting (from a year till five years) and long-range or long-term forecasting (over five years).

Now the classification of the factors influencing on passenger and freight transportations is resulted.

Depending on type of described information factors are divided on *quantitative* and *qualitative* (besides in various references is possible to meet such definitions, as *discrete*, *gradational, categorized* or *dummy* factors) [121].

The factors which values vary from observation to observation are related to the *quantitative* factors. For example, *gross domestic product per head*, *population density*, *monthly labour costs* and so on.

Qualitative factors have the same value for the whole group of observations, and take on some discrete values, according to in advance certain principle.

The elementary variant is inspected when the corresponding gradational variable can take on only two values, 0 and 1. Then an observation possessing given quality takes on U as value of the mentioned gradational variable, and an observation not possessing this quality takes on a zero. In other words, there are only two gradation of the given quality. For example, further the gradational variable which brings such information in model, whether the country has been entered the European Union till May, 1st, 2004, will be described. So-called old members of the EU receive U as value of the given variable, and new members, i.e.

entered the EU after May, 1st, 2004, receive zero value. Or on the contrary, since it is not critical. The number of gradations can be as much as wished big. For example, intensity of a specific transport type use in a country. Frequently gradations assignment to object is based on expert estimations.

According to the classification introduced in the promotional work of E.Zhukovska [8] and in [119], on character of the displayed information all influencing factors can be divided into following basic groups: *economical*, *social* and *structural*.

In particular, economical factors equally good describe both passenger and freight transportations. Social factors are more significant for passenger transportation analysis and forecasting. Structural factors, as usual, adequately assign freight transportations tendencies.

2.2. STATISTICAL DATA BASE FOR THE EU MEMBER STATES

In the course of search and gathering of statistics needed for the experiments 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 European Union from 1995 for 2008. The base contains the various factors across the Member States of the EU, influencing passenger and freight transportations, as well. Thereupon it would be desirable to describe briefly the main mission and functions of EuroSTAT.

First of all, The Statistical Office of the European Communities releases every year *The EuroSTAT Yearbook*, which contains statistics for the last year on all countries of Europe.

The EuroSTAT Yearbook and associated compendium publications span across either the full range or a large part of Eurostat statistical themes. These publications have been produced thanks to the cooperation of a number of sectoral specialists within EuroSTAT, who have gathered and analysed figures from the different European statistical offices in order to provide comparable and harmonised data for the European Union to use in the definition, implementation and analysis of Community policies.

As part of its dissemination policy, EuroSTAT has developed its website. All EuroSTAT publications are downloadable free of charge in PDF format from the website. Furthermore, EuroSTAT's databases are freely available, as are a set of pre-defined tables with the most frequently used and demanded data, including indicators that are contained within the EuroSTAT Yearbook.

Europe in figures – EuroSTAT Yearbook 2008, presents a comprehensive selection of statistical data on the European Union, its Member States and candidate countries. The Yearbook may be viewed as an introduction to European statistics and provides guidance to the vast range of data freely available from the EuroSTAT website.

Most data cover the period 1996-2006 and some indicators are provided for other countries such as candidate countries to the European Union, members of European Free Trade Association (further EFTA), Japan or the USA (subject to availability). With just over 500 statistical tables, graphs and maps, the yearbook treats the following areas: the economy, education, health, living conditions and welfare, the labour market, industry and services, agriculture, forestry and fisheries, international trade, transport, the environment, energy, science and technology and European regions. This edition's spotlight chapter covers Europe's ageing society and associated demographic challenges.

Each chapter (or sub-chapter) of the publication starts with a small introduction containing background information and policy relevance, as well as some details regarding the collection and interpretation of data; this is followed by a commentary on the data. The main focus of each chapter is a set of tables and graphs that have been selected to show the wide variety of data available for that particular topic; often these include information on how important benchmark indicators have developed during recent years within the EU, its Member States and the euro area.

The Yearbook is made on an annual basis and its content remains static until the next edition. In the meantime, online databases and pre-defined tables are being continually refreshed. In order to access the most recent data available, a EuroSTAT data code has been created for most indicators, in order to permit rapid access to online databases and supporting metadata [31], [32], [33].

Data comparability across countries is very high. This is ensured by the implementation of a common methodology. In addition, the so-called "mirror checks" allow comparing the data declared by partner reporting airports and find possible inconsistencies that are corrected as much as possible. Comparability over time is also very high. The present methodological approach has been applied for a number of years now and it is well understood and applied at airport and country level. So the analysis of the data over time produces very reliable results.

Only in the case of countries where there has been an increase in the number of reporting airports over time, the comparison of national aggregated data has to be taken with

care because the comparison is affected by the fact that more airports report data from one year to another. Anyhow, data availability over the time depends on each country.

In the frame of the data dissemination process, EuroSTAT has to calculate aggregates at intra-EU level (national, regional and intra-EU aggregates). It requires sometimes solving the problem of double counting. For each aggregate it is necessary to start at the airport level in order to identify the mirror declarations, i.e. the airport routes for which both airports report the volume, since these constitute the routes where the problem of double counting occurs. When calculating the total volume in such cases, only the departure declarations of the concerned airports have been taken into account. The problem of the double counting only appears for the calculation of the total passengers but not for the total arrivals (respectively total departures), which corresponds to the sum of the arrivals (respectively departures) at each domestic airport.

Concerning the total international extra-EU transport, the calculation is easier. It consists in the sum of all the declarations of the Member States to/from all the partner countries out of the European Union, as there is no double counting.

Data is collected on a monthly basis and then aggregated at quarterly and annual level. Only the airline information data is subject to confidentiality. The data providers deliver this information with a higher level of aggregation to avoid confidentiality constraints at national level. Completeness of data is high. There is an obligation of data provision for the Member States and, as a consequence, there are very few gaps in the data provision, at least since 2003 when the framework legal act came into force. The existing data collection on air transport statistics is well appreciated by the users. Accuracy, clarity and comparability are particularly indicated as good qualities of these data. European air transport statistics are a valuable resource to a wide range of users.

Users mainly request these data to properly monitor the development of air transport in the EU and other European countries, evaluate the impact of the air transport industry in the economy, quantify the importance of the transport flows of passengers and freight at intra-EU and extra-EU and assess the competition in the air transport market [34].

So, the first database contains the following factors selected as explanatory variables. In the brackets the short names of explanatory factors are noted, that afterwards will be used.

2.2.1. ECONOMICAL FACTORS

1. *Gross Domestic Product*, (*GDP*).

Gross Domestic Product is a measure for the economic activity. It reflects the total value of all goods and services produced less the value of goods and services used for intermediate consumption in their production. The volume index of GDP is expressed in relation to the European Union (EU-27) average set to equal 100. If the index of a country is higher than 100, this country's level of GDP per head is higher than the EU average and vice versa.

2. *Gross Domestic Product per capita in Purchasing Power Standards*, (*GDP per capita in PPS*).

   The volume index of GDP per capita in Purchasing Power Standards (PPS) is expressed in relation to the European Union (EU-27) average set to equal 100. If the index of a country is higher than 100, this country's level of GDP per head is higher than the EU average and vice versa. Basic figures are expressed in PPS, i.e. a common currency that eliminates the differences in price levels between countries allowing meaningful volume comparisons of GDP between countries.

3. *Growth rate of GDP volume – percentage change on previous year*, (*GDPrate*).

   The calculation of the annual growth rate of GDP volume is intended to allow comparisons of the dynamics of economic development both over time and between economies of different sizes. For measuring the growth rate of GDP in terms of volumes, the GDP at current prices are valued in the prices of the previous year and the thus computed volume changes are imposed on the level of a reference year; this is called a chain-linked series. Accordingly, price movements will not inflate the growth rate.

4. *Comparative Price Levels*, (*CPL*).

   Comparative Price Level is the ratio between Purchasing Power Parities (PPPs) and market exchange rate for each country. PPPs are currency conversion rates that convert economic indicators expressed in national currencies to a common currency, called Purchasing Power Standard (PPS), which equalizes the purchasing power of different national currencies and thus allows meaningful comparison. This ratio is shown in relation to the EU average (EU-27 = 100). If the index of the comparative price levels shown for a country is higher/lower than 100, the country concerned is relatively expensive/cheap as compared with the EU average.

5. *Inflation Rate*, (*IR*).

   As the given index the Consumer Price Index (CPI) which expresses relative

fluctuation of an average level of the prices of group of the goods and services (consumer's basket) for the certain period acts. A foodstuff, clothes, the electric power, the maintenance of a living accommodation and vehicles, health services, rest and formation enter into a basket of the goods and services.

6. *Labor Productivity per Hour Worked*, (*LPHW*).

   Labor productivity per hour worked is intended to give a picture of the productivity of national economies expressed in relation to the European Union (EU-27) average. If the index of a country is higher than 100, this country level of LPHW is higher than the EU average and vice versa. Expressing productivity per hour worked will eliminate differences in the full-time/part-time composition of the workforce.

7. *Employment Growth – Annual Percentage Change in Total Employed Population*, (*EGAP*).

   This indicator gives the change in percentage from one year to another of the total number of employed persons on the economic territory of the country or the geographical area.

8. *Trade Integration of Goods*, (*TI*).

   Average of imports and exports of the item goods of the balance of payments divided by GDP. If the index increases over time it means that the country/zone is becoming more integrated within the international economy.

9. *Annual Average EUR Exchange Rate versus National Currencies*, (*EURrates*).

10. *Prices on Petroleum Products, EUR per tonne*, (*PP*).

11. *Final Energy Consumption by Transport, 1000 toe* (*FECT*).

    Final energy consumption by transport covers the consumption of energy products in all types of transportation, i.e. rail, road, international and domestic air transport and inland navigation/coastal shipping, with the exception of maritime shipping.

12. *Consumption of electricity by industry, transport activities and households/services, GWh*, (*CEITH*).

    This consumption stands for final energy consumption. This means that the consumption in industry covers all industrial sectors with the exception of the energy sector, like power stations, oil refineries, coke ovens and all other installations transforming energy products into another form. Final energy consumption in transport covers mainly the consumption by railways and electrified urban transport systems. Final energy consumption in households/services covers quantities consumed

by private households, small-scale industry, crafts, commerce, administrative bodies, and services with the exception of transportation, agriculture and fishing.

2.2.2. SOCIAL FACTORS

1. *Monthly Labour Costs*, (*MLC*).

   The Monthly Labour Costs are specified as the relation of the general labour costs of all population employed of the country for a month to corresponding number of the population employed.
2. *Hourly Labour Costs*, (*HLC*).

   The Hourly Labour Costs are defined as total labour costs divided by the corresponding number of hours worked.
3. *Labour Productivity per Employment*, (*LPE*).
4. *Unemployment Rate*, (*UR*).

   The Unemployment Rate is determined as the relation of unemployed persons oldest than 15 years to total number of the active population of the country.

2.2.3. STRUCTURAL FACTORS

1. *Total Population*, (*TP*).

   The indicator of a population of the country on the beginning of considered year is used.
2. *Population Change*, (*PC*).

   It is the difference between the size of the population at the end and the beginning of a period. It is equal to the algebraic sum of natural increase and net migration (including corrections). There is negative change when both of these components are negative or when one is negative and has a higher absolute value than the other
3. *Country Area*, km^2 (*SQUARE*).
4. *Population Density*, (*PD*).

   Population density is determined as the relation of a mid-annual population of the country for certain year to the area of the considered country.
5. *Total Length of Railways*, *km* (*TOTLEN*).
6. *Number of Locomotives*, (*LOKOM*).
7. *Number of Good Wagons*, (*WAGONS*).

Statistical data has been collected on the Member States of the European Union, on Candidate Countries and on States of the EFTA. Let us list these countries in the order as in ISO-3166-alpha2 nomenclature, see also CIRCA and Ramon (Eurostat's Classification Server): Belgium (BE), Bulgaria (BG), Czech Republic (CZ), Denmark (DK), Germany (DE), Estonia (EE), Ireland (IE), Greece (GR), Spain (ES), France (FR), Italy (IT), Cyprus (CY), Latvia (LV), Lithuania (LT), Luxembourg (LU), Hungary (HU), Malta (MT), Netherlands (NL), Austria (AT), Poland (PL), Portugal (PT), Romania (RO), Slovenia (SI), Slovakia (SK), Finland (FI), Sweden (SE), United Kingdom (UK), Croatia (HR), Macedonia (MK), Turkey (TR), Iceland (IS), Liechtenstein (LI), Switzerland (CH), Norway (NO).

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

2.3. STATISTICAL DATA BASE FOR LATVIA

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 got from the *Central Statistical Bureau of Latvia* (LR CSP) and materials of the annual report of closed joint-stock company *Latvijas Dzelzceļš*. Thereupon first of all let us tell a few words about the *Central Statistical Bureau of Latvia*.

Central Statistical Bureau of Latvia is the main co-ordinator of statistical processes and the main performer of statistical information in Latvia. The mission of *Central Statistical Bureau of Latvia* is to provide internal and foreign users of statistical data with timely, exact, full, easily interpreted and international comparable statistical information on the economic, demographic, social and environmental phenomena and processes, using modern information technology decisions and the international experience.

Quality of the information prepared with its help is in conformity with the standarts of European Statistics Code of Practice. Owing *Central Statistical Bureau of Latvia* used principles, statistical data gathering, processing and dissemination are appropriately recorded, correspond to the international good practice and are available to the public access. In the course of preparation of data *Central Statistical Bureau of Latvia* uses such techniques of data gathering, alternative sources of the data, and methods of data processing and mathematical models, which allow reducing loading on respondents during preparation of surveys.

For dissemination of the statistical data the methods most convenient for users are used, i.e. Internet became a primary and priority kind of dissemination of the data. The

prepared statistical information is supplemented with comments on quality of the data. Level of the competence of *Central Statistical Bureau of Latvia* personnel allows expressing interests of Latvia in the international institutions. Used methods of personnel training provide fast acquisition of skills by new employees, and also provide improvement of professional skill to the present employees.

It would be desirable to notice that *Central Statistical Bureau of Latvia* has statistics on passenger and freight transportations only across all Latvia as a whole, while the data about transportations from regions of Latvia was necessary as well. Therefore we also had to fall back upon alternative sources of the statistical data. In this connection allow us to adumbrate them.

Joint-stock company *Latvijas Dzelzceļš* sells passenger and freight shipment services in Latvia and also in an international market, rolling stock repair services and manages railway infrastructure. The annual reports released by joint-stock company *Latvijas Dzelzceļš*, contain statistics about performed passenger and freight transportations both across Latvia, and behind its borders, and statistics about changes in the fixed assets and the rolling stock as well.

The Candidate is very grateful to Vasily Demidov, the head of *Latvijas Dzelzceļš* IT department. The statistics given by Vasily Demidov, has allowed making experiments on the analysis and forecasting of rail passenger departures from regions and cities of Latvia.

We have been planning to make experiments devoted to the analysis and forecasting of road passenger departures from regions and cities of Latvia. Here we have faced a problem since in Latvia the quantity of enterprises engaged in passenger road transportations performing is numerous. Accordingly, it is almost impossible to collect statistics and to systematise it. Besides not each enterprise-transport operator will agree to give such statistics, or it can appear the enterprise does not gather such statistics at all. There is *Road Transport Administration Ltd* which is engaged in statistics centralisation and ordering on passenger road transportations in Latvia.

The holder of *Road Transport Administration Ltd* state capital is the Ministry of Transport and Communication. Main functions of the Road Transport Administration Ltd are:

- Licensing of entrepreneurship in the field of road transport (commercial carriage by road);
- Co-ordination of access to the road transport market;

- Monitoring of compliance with requirements regulating road transport operations;
- Statistics gathering on the transported passengers and cargoes.

Unfortunately, *Road Transport Administration Ltd*, which should have the statistics on road transportations necessary for our experiments, has not placed this statistics at our disposal.

Thus, the second database contains sizes of rail passenger internal departures from districts and cities of Latvia from 2000 for 2003; and the various factors influencing sizes of those departures as well. In the brackets the short names of explanatory factors are noted, that afterwards will be used.

2.3.1. ECONOMICAL FACTORS

1. *Monthly Labour Costs*, (*MLC*), in LVL.
2. *Employed Population*, (*EP*).
3. *Number of Enterprises*, (*NE*).
4. *Number of New Building Projects*, (*NNBP*).
5. *Floor-space of New Building Projects, m^2*, (*FNBP*).
6. *Cost of New Building Projects*, (*CNBP*).

2.3.2. SOCIAL FACTORS

1. *Unemployed Population*, (*UP*).
2. *Number of Actual Criminal Acts*, (*NACA*).
3. *Number of Medical Officers*, (*NMO*).
4. *Number of Hospitals*, (*NHS*).
5. *Number of Beds in Hospitals*, (*NBH*).
6. *Number of General Education Institution* (at 1st of September), (*NGEI*).
7. *Number of Pre-School Education Institution* (at 1st of September), (*NPSEI*).
8. *Number of Students* (*Apprentices*) *in General Education Institution* (at 1st of September), (*NSGEI*).

2.3.3. STRUCTURAL FACTORS

1. *Total Population*, (*TP*).

2. *Area of Region, km*^{2} (*SQUARE*).
3. *Number of Cars*, (*NC*).
4. *Number of Buses*, (*NB*).
5. *Number of Railway Stations*, (*NRS*).
6. *Number of Hotels*, (*NHT*).

The database contains statistical data about following districts and cities of Latvia: Aizkraukles region, Aluksnes region, Balvu region, Bauskas region, Cesu region, Daugavpils, Daugavpils region, Dobeles region, Gulbenes region, Jekabpils region, Jelgava, Jelgavas region, Kraslavas region, Kuldigas region, Liepaja, Liepajas region, Limbazu region, Ludzas region, Madonas region, Ogres region, Preilu region, Rezekne, Rezeknes region, Riga, Rigas region, Jurmala, Saldus region, Talsu region, Tukuma region, Valkas region, Valmieras region, Ventspils, Ventspils region.

CONCLUSION

In the given Chapter we have described the sources of the statistical data used for two statistical databases generating. The content of these databases is subjected to the further statistical analysis and processing.

After all, the following structure of transportations being subjected to verification can be drawn:

1. By type of transportation
1.1. Passenger
1.2. Freight
2. By transportation mode
2.1. International
2.2. Internal (domestic)
3. By mode of vehicle
3.1. Rail transport
3.2. Air transport
4. By object of forecasting
4.1. Countries or regions of countries
4.2. OD-pair
5. By direction
5.1. Incoming
5.2. Outgoing

6. By type of forecasted indicator

6.1. Transportations or departures, in thousands of passengers or tonnes

6.2. Turnover, in tonne-km or passenger-km.

We accentuate such factors influencing transportations in the EU, as Gross Domestic Product per capita in Purchasing Power Standards, Comparative Price Levels, Inflation Rate, Unemployment Rate, Total Length of Railways, Number of Locomotives, Number of Good Wagons, Employment Growth, Final Energy Consumption by Transport and Consumption of Electricity by Industry and Transport.

Concerning rail departures in Latvia it is noteworthy to notice such factors, as Number of Enterprises, Number of General Education Institution, Total Population, Number of Buses and Number of Railway Stations.

3. MATHEMATICAL MODELS FOR FORECASTING AND METHODS OF THEIR ESTIMATION

In this Chapter the theoretical description of mathematical models used for analysis and forecasting of freight and passenger transportations is resulted, and known methods of their estimation are considered as well.

In the present promotional work such regression models are used, as multiple regression model, multivariate regression model, single index model, SURE-model and the modified gravity model (the corresponding classification is reviewed in the Introduction). Each of these five models can be both individual and group. Thereupon, first of all, we will give the characteristic of individual and group models.

3.1. INDIVIDUAL AND GROUP MODELS

Let in our disposal there are some objects in which relation it is necessary to predict one or several indicators. 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 the given work the objects of transportations forecasting are the Member States of the European Union and regions of Latvia. If one separately taken country, for example Latvia, is considered for a number of years, in this case the individual model is applicable. But if the same country is considered as one object from aggregate of the EU countries, then the group model is applicable.

Using individual model one take into account only that statistical data which are referred directly to the considered object. The analytical dependence deduced on the basis of individual model, is correct only for the considered object [121]. As a lack of individual model of forecasting is possible to notice that forecasting on the basis of individual model is not effective if in the future occurrence of factors which have not forced on the considered object earlier is expected.

As in group model of forecasting the statistical data for all the considered objects are simultaneously considered, such model is free from mentioned above restriction. In group model of forecasting the analytical dependence of the predicted indicator on values of predictors will be correct for all the considered objects. Presence of only one dependence for all the considered objects leads to necessity of essential increase in number of predictors with

the intent that each object's specific features influencing the predicted indicator, have to be reflected in the general developed model.

Dynamism, or adjustment feature, is the main advantage of group model as the data it contains hold fixed the development of various considered objects at different stages. Therefore for the majority of objects the data contain such characteristics which objects will get only in the future. It allows to obtain, in particular, the forecasts for such objects, i.e. for the countries or cities, on which there is no statistical data, or to take into account the influence of the factors which have not forsed on the objects earlier.

It is considered to be that individual models are more exact for short-term forecasting, but group models for long-term forecasting.

In the framework of the present Thesis only the group models have been used for passenger and freight transportations forecasting for the Member States of the EU and for regions of Latvia.

3.2. PARAMETRIC REGRESSION MODELS

3.2.1. MULTIPLE LINEAR REGRESSION MODEL

In the beginning we consider a *general view of regression model*. In general the regression model [120], [95], [127] can be described as

$$Y_i = m(x_i) + \varepsilon_i, \tag{3.1}$$

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), $i = 1, 2, ..., n$. Three conditions are supposed:

1) the random term has zero expectation, $E(\varepsilon) = 0$;
2) the variance $Var(\varepsilon) = \sigma^2 \psi(x)$, where σ^2 is an unknown constant and $\psi(x)$ is a known weighted function;
3) $Cov(Y_i, Y_j) = Cov(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j$.

On that basis we need to estimate the unknown function $m(\bullet)$.

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

$$m(x_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}\,, \tag{3.2}$$

where $\beta' = (\beta_0 \quad \beta_1 \quad \dots \quad \beta_d)$ is vector of unknown regression coefficients,

$j = 1, 2, ..., d$ is a number of predictors,

$x_i = (1 \quad x_{i,1} \quad \dots \quad x_{i,d})'$ is a vector of independent variables in *i*-th observation,

$i = 1, 2, ..., n$ is a number of observation.

To write the above model in the matrix notation [95], let

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

where $x_i = (1 \quad x_{i,1} \quad \dots \quad x_{i,d})'$. Then the multiple linear regression model can be written as

$$Y = X\beta + \varepsilon\,. \tag{3.3}$$

The above three assumptions can be written in the matrix notations as

1) $E(\varepsilon) = 0$,
2) $Cov(\varepsilon) = \sigma^2 I$,

known as Gauss-Markov conditions. The least squares estimate of β is obtained by minimizing the sum of squares

$$(Y - X\beta)'(Y - X\beta)$$

with respect to β. The last squares estimate of β is given by

$$\hat{\beta} = (X'X)^{-1} X'Y\,. \tag{3.4}$$

It can be shown that under the Gauss-Markov conditions

$$E(\hat{\beta}) = \beta \text{ and } Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1}.$$

That is, $\hat{\beta}$ is unbiased estimator of β. An unbiased estimator of σ^2 is given by

$$s^2 = \frac{1}{n-d}\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right). \tag{3.5}$$

3.2.2. MULTIVARIATE LINEAR REGRESSION MODEL

Multivariate regression model [95], [97] usually is used when we have to forecast two or more strongly correlated variables, and includes simultaneously many dependent variables. In sample surveys many variables are assigned values according to responces to a questionnaire. Variables such as age, socioeconomic status, religion and marital status are designated independent, whereas variables measuring questions such as authority patterns, desire for children, and homemaking skills are designated dependent. The relationship between one dependent variable and many independent variables can be measured by a linear regression analysis. In this case a linear regression on each dependent variable would be required to examine the effect of *d* independent variables on them.

For example, *d* independent variables could denote the concentration of *d* chemicals in a tank, and dependent variables $y_1, \ldots, y_p$ could denote the amounts of these chemicals in a fish in the tank after 1 week, 2 weeks, …, and *p* weeks. As we can see, in this example of Srivastava all the dependent variables have the same sense. It will be shown below, is possible to expand multivariate regression model application also on situation if *p* dependent variables describe factors of the different nature.

So, in the multivariate regression model the vectors y, β and ε become matrices, which we denote Y, B and E. So, multivariate regression model can be written in the following way:

$$\mathrm{Y} = \mathrm{XB} + \mathrm{E}, \tag{3.6}$$

besides the following notations are used:

$$\mathrm{Y}=\begin{pmatrix} y_{11} & y_{12} & \dots & y_{1p} \\ y_{21} & y_{22} & \dots & y_{2p} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{np} \end{pmatrix},\ \mathrm{X}=\begin{pmatrix} 1 & x_{11} & \dots & x_{1d} \\ 1 & x_{21} & \dots & y_{2d} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nd} \end{pmatrix},$$

$$\mathrm{B}=\begin{pmatrix} \beta_{01} & \beta_{02} & \dots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \dots & \beta_{1p} \\ \dots & \dots & \dots & \dots \\ \beta_{d1} & \beta_{d2} & \dots & \beta_{dp} \end{pmatrix},\ \mathrm{E}=\begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \dots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \dots & \varepsilon_{2p} \\ \dots & \dots & \dots & \dots \\ \varepsilon_{n1} & \varepsilon_{n2} & \dots & \varepsilon_{np} \end{pmatrix}.$$

Here Y is the matrix of observations on the dependent variables, X is the design matrix, B is the unknown matrix of regression coefficients, E is the matrix of random members. And *p* is the quantity of dependent variables, *d* is the quantity of predictors and *n* is the quantity of observations. Let us rewrite matrix E in the following way: $\mathrm{E}=(e_1,e_2,\dots,e_n)'$. It is supposed that errors are independent:

$$\mathrm{E}(e_i)=0\,;\mathrm{E}\left(e_i e_j'\right)=0\,,i\neq j\,. \tag{3.7}$$

The matrix Σ of covariance of errors $\{e_i\}$ can be expressed as

$$Cov\left(e_i e_j'\right)=\Sigma=(\sigma_{kl}),k,l=1,\dots p\,,\ i,j=1,\dots,n. \tag{3.8}$$

The least squares estimates for the B coefficients are obtained by solving the normal equations like for multiple regression models:

$$\widetilde{\beta}^{\langle k\rangle}=(X'X)^{-1}X'Y^{\langle k\rangle},\ k=1,\dots,p, \tag{3.9}$$

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

An unbiased estimate of Σ is given by

$$\widetilde{\Sigma}=\frac{1}{n-d}Y'\left[I-X(X'X)^{-1}X'\right]Y. \tag{3.10}$$

Now we can express the covariance between two estimates of unknown vectors of regression coefficients:

$$Cov\left(\widetilde{\beta}^{\langle k\rangle},\widetilde{\beta}^{\langle l\rangle}\right)=\sigma_{kl}\left(X'X\right)^{-1},\ k,l=1,\ldots p\,. \tag{3.11}$$

As the matrix Σ of covariance of errors is unknown, we can use its unbiased estimate $\widetilde{\Sigma}$. So,

$$Cov^{*}\left(\widetilde{\beta}^{\langle k\rangle},\widetilde{\beta}^{\langle l\rangle}\right)=\widetilde{\sigma}_{kl}\left(X'X\right)^{-1},\ k,l=1,\ldots p\,. \tag{3.12}$$

Further application of multivariate model for forecasting expectation of total transportations (i.e. total air freight and mail transportation intra- and extra-EU), and confidence limits construction for that expectation, is shown in Chapter 4 of the present promotional work.

3.2.3. SEEMINGLY UNRELATED REGRESSION EQUATION MODEL

The *seemingly unrelated regression equation model* (further *SURE-model*) can be considered as continuation of the multivariate model (3.6) for a case when the part of predictors coincides for all variables of interest (i.e. dependent variables), and the part does not coincide [115],[116]. Besides, the number of observations on each variable of interest can be variously, for instance, in situations, when we have missing data (for some reasons) for some objects for some time moments.

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

$$Y_{i,j}=\sum_{v=1}^{m_i}\beta_{i,v}x_{i,j}^{(v)}+Z_{i,j}, \tag{3.13}$$

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 (a disturbance) 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 $i = 1, 2, \ldots, G, j = 1, 2, \ldots, n_i$ $x_{i,j} = \left(x_{i,j}^{(1)}, x_{i,j}^{(2)}, \ldots, x_{i,j}^{m_i}\right)$ 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.

Obtaining an unbiased estimate for a covariance matrix for SURE-model and forecasting of the total passenger air transportations for the EU Member States using this model will be shown in Chapter 7.

In this Section only parametric models were described. In these models variables are not grouped by type, i.e. both quantitative and qualitative variables are included into model and are estimated on equal assumptions. Parametric models are fully determined up to a parameter vector. The fitted models can easily be interpreted and estimated accurately if the underlying assumptions are correct. If, however, they are violated then parametric estimates can be inconsistent and give a misleading picture of the regression relationship.

Next Section is devoted to the nonlinear (mostly semiparametric and nonparametric) models, which could be considered as more complicated from the one hand, but more flexible from the second hand.

3.3. NONLINEAR REGRESSION MODELS

3.3.1. GRAVITY MODELS

The first attempt of mathematical methods application for determining size of transport flow for two cities depending on distance between these cities and number of the population living in them, has been undertaken in the end of a XIX-th century (by

A. Wellington and *E. Lille*). The model allowing such determination has received the title *gravity model*. Generally that model can be represented in the following way [121]:

$$T_{ij} = K \frac{P_i P_j}{D_{ij}^2}, \tag{3.14}$$

besides T_{ij} is passenger transportation size between cities i and j; P_i and P_j are population living in cities i and j, correspondingly; D_{ij} is a distance between cities 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 (see Chapter 2, Section 2.1). So, in the late sixties of the last century the model (3.14) has been used for an estimation of volumes of the passenger transportation for a pair of cities, at the decision of a question of expediency to input between them a direct airline [28]. For example, in the given book [121] the model for forecasting of air passenger transportations in the directions containing Moscow is offered. The following kind of gravity model has been used as the basis of the considered model:

$$Y_i^* = \left(h_i^{(1)} h_i^{(2)}\right)^{\sum b_\nu^* t_i^{(\nu)}} l^{\sum b_\mu^*} l^{\sum b_\mu^* t_i^{(\mu)}}, \tag{3.15}$$

where i is number of year of forecast, besides $i = 1985$, 1990 and so on;

Y_i^* are forecasted transportations for corresponding year and direction, in tens of passengers;

$h_i^{(1)}$ and $h_i^{(2)}$ are populations of Moscow and the other city of the considered direction in the corresponding year, in thousands;

l is air tariff distance between cities of the considered direction, km;

$\left\{t_i^{(\bullet)}\right\}$ are predictors for the considered direction in the corresponding year;

$\left\{b_\nu^*\right\}$ are parameters estimation for predictors for producted population;

$\left\{b_\mu^*\right\}$ are parameters estimation for predictors for distances.

Together with the model (3.15), as competiting models the following four kinds of models were considered:

$$Y_i^* = \left(h_i^{(1)} + h_i^{(2)}\right)^{\sum_\nu b_\nu^* t^{(\nu)}} l^{\sum_\mu b_\mu^* t^{(\mu)}}; \tag{3.16}$$

$$Y_i^* = \left(h_i^{(1)} h_i^{(2)}\right)^{\alpha} l^{\gamma} \prod \left(t^{(\nu)}\right)^{b_\nu^*}; \tag{3.17}$$

$$Y_i^* = \left(h_i^{(1)} h_i^{(2)} / l^2\right)^{\sum_\nu b_\nu^* t^{(\nu)}}; \tag{3.18}$$

$$Y_i^* = \sum_\nu b_\nu^* t^{(\nu)} \left(h_i^{(1)} h_i^{(2)} / l^2\right). \tag{3.19}$$

Notations in the models (3.16) – (3.19) are similar to corresponding notations in the model (3.15). It is necessary to notice that, despite the seeming nonlinearity, these models can be easily led to a model, linear on parameters (see next Section).

In the present promotional work application of the modified gravity model for passenger railway departures estimation between capitals of Member States of the European Union (so called *Correspondence Matrix*) has been considered in Chapter 5.

3.3.2. Overview of Semiparametric and Nonparametric Models

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

Nonparametric models avoid restrictive assumptions of the functional form of the parametric regression function $m(\bullet)$. However, they may be difficult to interpret and yield inaccurate estimates if the number of predictors is large. *Semiparametric* models combine components of parametric and nonparametric models, keeping the easy interpretability of the former and retaining some of the flexibility of the latter. A lot of effort has been allocated to developing methods which reduce the complexity of high dimensional regression problems. This refers to the reduction of dimensionality as well as allowance for partly parametric modeling. The resulting models can be grouped together as so-called semi-parametric models.

The basis for many semiparametric regression models is the *generalized linear model* (*GLM*) [75], which is given by

$$m(x_i) = G\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_d x_{i,d}\right) = G\left(\beta^T x_i\right), \tag{3.20}$$

where $G(\bullet)$ denotes any *known* continuous function and is called the *link function* of one dimensional variable $\beta^T x_i$, which, in its turn, is called the *index function* or simply the *index*.

The estimation of the *GLM* is performed through an iterative algorithm. This algorithm, the *iteratively reweighed least squares* (*IRLS*) algorithm, applies weighted least squares to the adjusted dependent variable z in each iteration step:

$$\beta^{new} = \left(X^T W X\right)^{-1} X^T W Z \,. \tag{3.21}$$

For details see [53].

Prominent example of *GLM* are binary choice models (*logit* or *probit*) or count data model (Poisson regression). For instance, the logit model assumes that $G\left(\beta^T x_i\right)$ is the standard logistic cumulative distribution function for all x_i:

$$m(x_i) = \frac{1}{\exp\left(-\beta^T x_i\right)}. \tag{3.22}$$

If $G(\bullet)$ is identity, we are back in the classical linear model. Let us consider a quite common approach for investigating growth models. Here, the model is often assumed to be multiplicative instead of additive, i.e.

$$Y = \prod_{j=1}^{d} x_j^{\beta_j} \cdot \varepsilon, \;\; E\log(\varepsilon) = 0 \tag{3.23}$$

in contrast to

$$Y = \prod_{j=1}^{d} x_j^{\beta_j} + \xi, \;\; E\xi = 0 \,. \tag{3.24}$$

Depending on whether we have multiplicative or additive errors, we can transform model (3.23) to

$$E\{\log(Y) \mid x\} = \sum_{j=1}^{d} \beta_j \log\left(x_j\right) \tag{3.25}$$

and model (3.24) to

$$E\{Y \mid x\} = \exp\left\{\sum_{j=1}^{d} \beta_j \log\left(x_j\right)\right\}. \tag{3.26}$$

Equation (3.26) is equivalent to (3.22) with $G(\bullet) = \exp(\bullet)$. Equation (3.25), however, is a transformed model. Author has been used the regression model (3.23) with transformation (3.25) for analysis and forecasting sales volumes from wholesale warehouse [84]. As well transformation of (3.26) kind is used in the present work for gravity model modification.

In many applications a canonical partitioning of the explanatory variables exists. In particular, if there are categorical or discrete explanatory variables, we may want to keep them separate from the other design variables (or predictors). Note that only the continuous variables in the nonparametric part of the model cause the curse of dimensionality [128], [3].

There are the following kinds of non- and semi-parametric models.

Additive Model (*AM*). The standard additive model is a generalization of the multiple linear regression model (3.2) by introducing *unknown* one-dimensional nonparametric functions in the place of the linear components:

$$E(Y \mid x) = c + \sum_{j=1}^{d} g_j(x_j), \tag{3.27}$$

where c is intercept. Instead of estimating one function of several variables, as we do in completely nonparametric regression, we merely have to estimate d functions of one-dimensional variable x_j.

Partial Linear Model (*PLM*). Suppose we only want to model parts of the index linearly. This could be for analytical reasons or for reasons going back to economic theory. For instance, the impact of dummy variable $x_1 \in \{0,1\}$ might be sufficiently explaining by estimating the coefficient β_1. For the sake of clarity, let us now separate the d-dimensional vector of explanatory variables into $u = (u_1, \ldots, u_p)^T$ and $t = (t_1, \ldots, t_q)^T$. The regression is assumed to have a form:

$$E(Y \mid u, t) = u^T \beta + m(t), \tag{3.28}$$

where $m(\bullet)$ is an unknown multivariate function of the vector t. Thus, a *PLM* can be interpreted as a sum of a purely parametric part and a purely nonparametric part. Estimating β and $m(\bullet)$ involves the combination of both parametric and nonparametric techniques.

Generalized Additive Model (*GAM*). This models are based on the sum of *d* nonparametric functions of the *d* variables *x* (plus an intercept term). In addition, they allow for a known parametric link function $G(\bullet)$, that relates the sum of functions to the dependent variable:

$$E(Y \mid x) = G\left\{ c + \sum_{j=1}^{d} g_j(x_j) \right\}. \tag{3.29}$$

Generalized Partial Linear Model (*GPLM*). Introducing a link $G(\bullet)$ for a *PLM* yields the *generalized partial linear model*:

$$E(Y \mid u,t) = G\{u^T \beta + m(t)\}. \tag{3.30}$$

$G(\bullet)$ denotes the known link function as in the *GAM*. In contrast to the *GAM*, $m(\bullet)$ is possibly a multivariate nonparametric function of the variable *t*.

Generalized Additive Partial Linear Model (*GAPLM*). In high dimensions of *t* the estimate of the nonparametric function $m(\bullet)$ in the *GPLM* faces the same problems as the fully nonparametric multidimensional regression function estimates: the curse of dimensionality and the practical problem of interpretability. Hence, it is useful to think about a lower dimensional modeling of the nonparametric part. This leads to the *GAPLM* with an additive structure in the nonparametric component:

$$E(Y \mid u,t) = G\left\{ u^T \beta + \sum_{j=1}^{q} g_j(t_j) \right\}. \tag{3.31}$$

Here, the $g_j(\bullet)$ will be univariate nonparametric functions of the variables t_j. In the case of an identity function $G(\bullet)$ we speak of an additive partial linear model.

In the given promotional work mainly application of the *single index regression model* (further *SIM*) is investigated. In this connection allow us to describe this model. 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:

$$E(Y \mid x) = m(x) = g\{v_\beta(x)\}. \tag{3.32}$$

Here we have only one assumption that 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. In our investigation we use a linear combination:

$$m(x_i) = g(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}) = g(\tau_i), \tag{3.33}$$

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 our observations we estimate g by ordinary univariate nonparametric regression of Y on $v_{\hat{\beta}}(x)$. *W.Hardle* in [53] considers two methods for estimation of the *SIM*, i.e. *semiparametric least squares* (*SLS*) and *pseudo maximum likelihood estimation* (*PMLE*). Both these methods have the following idea in common: establish an appropriate object function to estimate β with parametric $\sqrt{n}$ rate. The main problem is to decide by which to replace the unknown link function g. Moreover, there can be variation in choice of optimization method.

As it is known, in the given work as an estimate of unknown link function of single index model the Nadaraya-Watson kernel estimator is used. Therefore in the following Section the concept of kernel estimator and its application for estimation of a regression function will be described in details. Such important questions, as problem of choosing of the optimal value of smoothing parameter will be taken up as well.

3.3.3. NADARAYA-WATSON KERNEL ESTIMATOR AS A MEAN OF REGRESSION FUNCTION ESTIMATION

In this Section the basic concepts of kernel estimators are stated before their use description in case of an unknown regression function estimation.

So, kernel density estimation can be considered as a generalization or improvement of the histogram, details see in [53]. The kernel density estimate at point x

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{3.34}$$

corresponds to the histogram bar height for the bin $[x - h/2, x + h/2)$ if the uniform kernel is used, besides h is a bandwidth, or a smoothing parameter. The uniform kernel and other kernels can be observed in Table 3.1.

Table 3.1

Kernel functions

Kernel	$K(u)$
Uniform	$\frac{1}{2} I(\lvert u\rvert \le 1)$
Triangle	$(1 - \lvert u\rvert) I(\lvert u\rvert \le 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2) I(\lvert u\rvert \le 1)$
Quartic (Biweight)	$\frac{15}{16}(1 - u^2)^2 I(\lvert u\rvert \le 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3 I(\lvert u\rvert \le 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right) I(\lvert u\rvert \le 1)$

It is necessary to underline that all kernel functions are approximately equivalent. First of all, kernel functions are usually probability density functions (further *pdf*), i.e. they integrate to one and $K(u) \ge 0$ for all u in the domain of K. An immediate consequence of $\int K(u)du = 1$ is $\int \hat{f}_h(x)dx = 1$, i.e. the kernel density estimator is a pdf as well.

Concerning to the question of optimal bandwidth h choosing, we are interested in the mean squared error (further *MSE*) since it combines squared bias and variance. In this connection the bias of the kernel density estimator is

$$Bias\{\hat{f}_h(x)\} = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2), \text{ as } h \to 0. \tag{3.35}$$

Here we denote $\mu_2(K) = \int s^2 K(s) ds$. Variance can be expressed as

$$Var\left\{\hat{f}_h(x)\right\} = \frac{1}{nh}\|K\|_2^2 f(x) + o\left(\frac{1}{nh}\right), \text{ as } nh \to \infty. \tag{3.36}$$

Here, $\|K\|_2^2$ is shorthand for $\int K^2(s) ds$, the squared L_2 norm of K.

Minimising the *MSE* allows getting a compromise between over- and undersmoothing. Figure 3.1 puts variance, bias and *MSE* onto one graph. At last, *MSE* can be expressed in the following way

$$MSE\left\{\hat{f}_h(x)\right\} = \frac{h^4}{4} f''(x)^2 \mu_2(K)^2 + \frac{1}{nh}\|K\|_2^2 f(x) + o(h^4) + o\left(\frac{1}{nh}\right). \tag{3.37}$$

The mean integrated squared error (further *MISE*) is given by

$$MISE\left\{\hat{f}_h\right\} = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2\|f''\|_2^2 + o\left(\frac{1}{nh}\right) + o(h^4), \tag{3.38}$$

where $h \to 0$ and $nh \to \infty$. Ignoring higher order terms in (3.38) allow obtaining of asymptotical mean integrated squared error (further *AMISE*):

$$AMISE\left(\hat{f}_h\right) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2\|f''\|_2^2. \tag{3.39}$$

It is noteworthy to note, that *MISE* and *AMISE* are used for canonical kernel and canonical bandwidth determination, see [53], [72], [102].

In [53] two most frequently used approaches for optimal smoothing parameter (i.e. bandwidth *h*) selection are proposed, the *plug-in method* and the *method of cross-validation*. Authors assure they have still not found a way to select the bandwidth that is both applicable in practice as well as theoretically desirable. For more complete treatments of plug-in and cross-validation methods, see e.g. [52] and [79].

So, the basic concepts about kernel functions and features of their application for estimation of unknown pdf are known, we can pass to their use in the regression analysis. It

would be desirable to note, why so steadfast attention was given to properties of kernel functions with reference to pdf estimation. The point is that kernel regression estimators inherit features of described above kernel density estimators. So, further we intend to tell about the Nadaraya-Watson kernel regression estimator, which has been widely used in the framework of the present promotional work.

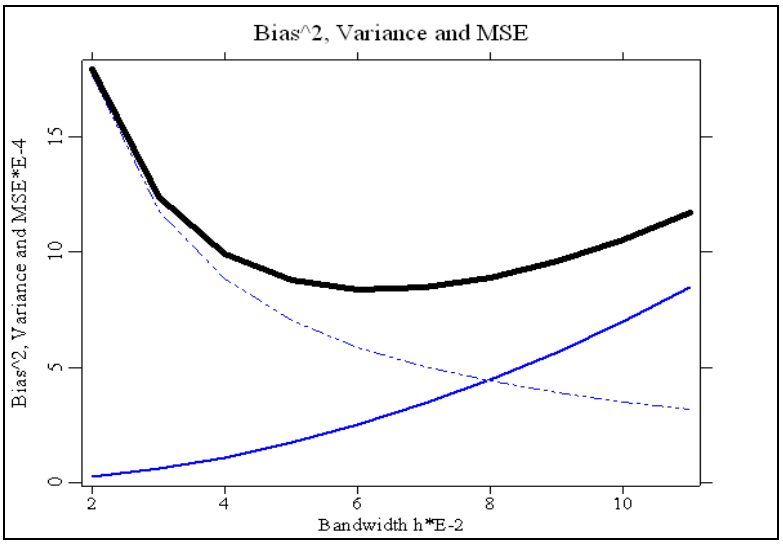

Fig.3.1. Squared bias part (thin solid), variance part (thin dashed) and *MSE* (thick solid) for kernel density estimate

Next only case of univariate kernel regression will be stated, because described above and widely used in this promotional work single index model requires this kind of kernel regression.

First of all, recall that in general a regression function can be represented in the form

$$Y = m(x),$$

where $m(\bullet)$ is a function in the mathematical sense.

Secondly, recall two concepts – *conditional expectation* and *conditional expectation function*. Let X and Y be two random variables with joint probability density function $f(x, y)$. The conditional expectation of Y given that $X = x$ is defined as

$$E(Y \mid X = x) = \int y \cdot f(y \mid x) dy = \int y \frac{f(x, y)}{f_X(x)} dy = m(x), \qquad (3.40)$$

where $f(y|x)$ is the conditional pdf of Y given $X = x$, and $f_X(x)$ is the marginal pdf of X. The mean function might be quite nonlinear even for simple-looking densities.

Note that $E(Y \mid X = x)$ is a function of x alone. Consequently, we may abbreviate this term as $m(x)$. If we vary x we get a set of conditional expectations. This mapping from x to $m(x)$ is called the conditional expectation function and is often denoted as $E(Y \mid X)$. This tells us how Y and X are related "on average". Next our step is $m(\bullet)$ estimation.

We assume that both X and Y are random variables with joint pdf $f(x, y)$. The natural sampling scheme in this setup is to draw a random sample from the bivariate distribution that is characterized by $f(x, y)$. That is, we randomly draw observations of the form $\{X_i, Y_i\}$, $i = 1, \ldots, n$. Before the sample is drawn, we can view the n pairs $\{X_i, Y_i\}$ as i.i.d. pairs of random variables. This sampling scheme will be referred to as the *random design* [53].

The derivation of the estimator in the *random design* case starts with the definition of conditional expectation:

$$m(x) = E(Y \mid X = x) = \int y \frac{f(x, y)}{f_X(x)} dy = \frac{\int y \cdot f(x, y) dy}{f_X(x)}. \tag{3.41}$$

Given that we have observations of the form $\{X_i, Y_i\}$, $i = 1, \ldots, n$, the only unknow quantities on the right hand side of (3.41) are $f(x, y)$ and $f_X(x)$. From all discussed in the previous section we know how to estimate pdf. Consequently, we plug in kernel estimates for $f(x, y)$ and $f_X(x)$ in (3.41). Estimating $f_X(x)$ is straightforward. To estimate $f(x, y)$ we employ the multiplicative kernel estimator with product kernel [53]:

$$\hat{f}_{h,g}(x, y) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(\frac{x - X_i}{h}\right) K_g\left(\frac{y - Y_i}{g}\right). \tag{3.42}$$

Hence, for the numerator of (3.41) we get

$$\begin{aligned} \int y \hat{f}_{h,g}(x, y) dy &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \int \frac{y}{g} K_g\left(\frac{y - Y_i}{g}\right) dy = \\ &= \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \int (sg + Y_i) K(s) ds = \\ &= \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) Y_i, \end{aligned} \tag{3.43}$$

where we used the facts that kernel functions integrate to 1 and are symmetric around zero. Plugging in leads to the Nadaraya-Watson estimator introduced by Nadaraya and Watson in 1964 [76], [104]:

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{n^{-1}\sum_{j=1}^{n} K_h(x - X_j)}, \tag{3.44}$$

which is the natural extension of kernel estimation to the problem of estimating an unknown conditional expectation function.

The following points are noteworthy. Rewriting (3.44) as

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{K_h(x - X_i)}{n^{-1}\sum_{j=1}^{n} K_h(x - X_j)}\right)Y_i = \frac{1}{n}\sum_{i=1}^{n} W_{hi}(x)Y_i \tag{3.45}$$

reveals that the Nadaraya-Watson estimator can be seen as a weighted local average of the response variables Y_i (note $\frac{1}{n}\sum_{i=1}^{n} W_{hi}(x) = 1$). In fact, the Nadaraya-Watson estimator shares this weighted local average property with several other smoothing techniques, e.g. *k*-nearest-neighbor and spline smoothing.

Note that just as in kernel density estimation the bandwidth *h* determines the degree of smoothness of $\hat{m}_h$, see Figure 3.2. To motivate this, let *h* go to either extreme. If $h \to 0$ then $W_{hi}(x) \to n$ if $x = X_i$ and is not defined elsewhere. Hence, at an observation X_i, $\hat{m}_h(X_i)$ converges to Y_i, i.e. we get an interpolation of the data. On the other hand if $h \to \infty$ then $W_{hi}(x) \to 1$ for all values of *x* and $\hat{m}_h(X_i) \to \bar{Y}$, i.e. the estimator is a constant function that assigns the sample mean of *Y* to each *x*. Choosing *h* so that a good compromise between over- and undersmoothing is achieved, is once again a crucial problem.

What happens if the denominator of $W_{hi}(x)$ is equal to zero. The numerator is also equal to zero, and the estimate is not defined. This can happen in regions of sparse data.

CONCLUSION

In the Chapter 3 theoretical foundation of regression models used in the given promotional work has been 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 Sub-section is dedicated to gravity models.

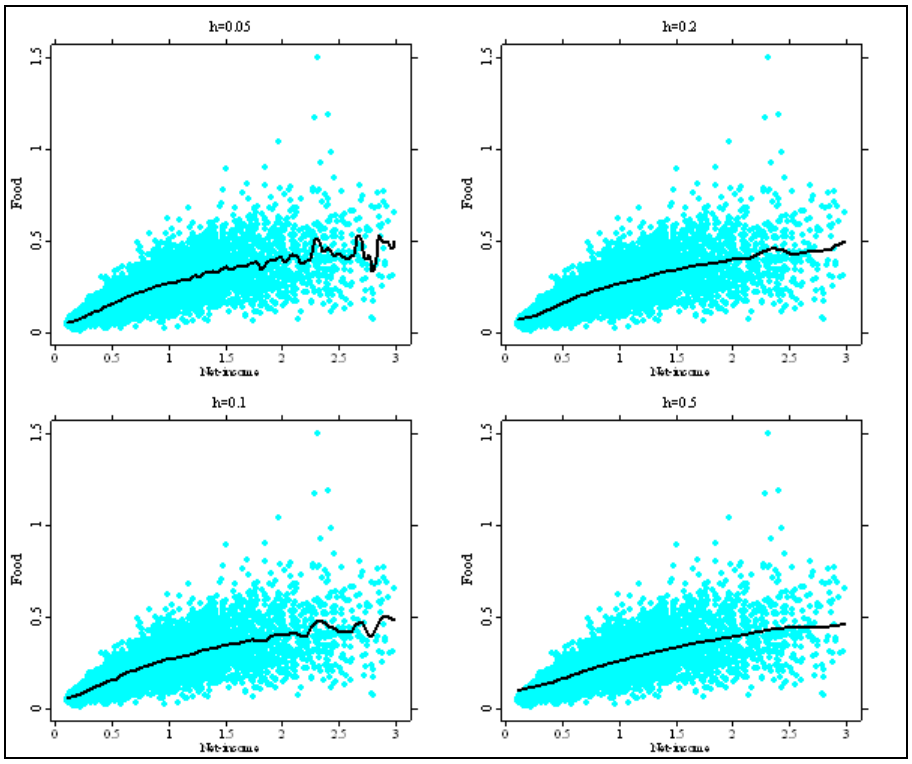

Fig.3.2. Four kernel regression estimates for the 1973 U.K. Family Expenditure data with bandwidths $h = 0.05$, $h = 0.1$, $h = 0.2$ and $h = 0.5$.

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 efficiency of single index model, the kernel estimators and their properties is described. In particular, the Nadaraya-Watson kernel estimator, applied for estimation of SIM, is considered in details.

The next four chapters are devoted to the methodology of elaboration and development of models and methods for analysis and forecasting passenger and freight transportations, and to the investigation of models adequacy and methods efficiency. Chapters are allocated correspondingly to the classification of the used models stated in Introduction.

4. ANALYSIS AND FORECASTING INTERNATIONAL AIR FREIGHT TRANSPORTATIONS FOR THE EU MEMBER STATES

In this Chapter analysis and forecasting of freight and mail air international transportations by the Member States of the European Union on the basis of parametric regression models are considered. The used statistical data and the procedure of obtainting of upper confidence limit for the total forecast of two kinds of transportations are considered. The best versions of elaborated models for forecasting air freight and mail transportations are shown as well. The detailed analysis of the obtained results is given, i.e. the calculated variance of total forecast and the obtained corresponding confidence limits.

4.1. PROBLEM SETTING

In the framework of the present research the freight and mail air international transportations are divided into two parts in relation to the borders of the European Union. Thus there are two forecasted variables: internal and external freight and mail air international transportations. After those 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.

The corresponding regression models contain main social-economical factors affected internal and external transportation for each country. Two different approaches are considered: the classical linear multiple regression model and the multivariate regression model. Various tests for hypothesis of explanatory variables insignificance and model correctness are carried out. The total forecast of internal and external transportations is obtained for 2007 year. The upper confidence limits for that forecast (at various confidence levels) are calculated as well. The purpose of research is the experimental proof of significance of total forecast.

The task of the current research is elaboration of regression models for the freight and mail air international transportations analysis and forecasting. In this connection the following problems are considered:

1. Analysis and *separately* forecasting of external and internal freight and mail international transportations by the Member States of the European Union.
2. Analysis and *total* forecasting of external and internal freight and mail international transportations by the Member States of the European Union.

For the first problem the multiply regression model is used. In other words, we intend to construct two different multiply models, containing the same social-economical factors. It is necessary to find out, how the same factors influence different kind of transportations. Our assumption concerning that question is such: the considered factors should influence internal and external transportations differently, i.e. probably we will obtain opposite signs by the same estimated coefficients. For the second problem the multivariate regression model is used. The unknown coefficient estimation and covariance matrix of errors estimation procedures for both kinds of models have been represented by package Statistica 6.0 means and own software developed in the MathCad 13 environment.

Let us consider models for *transportation forecasting*. The main object of consideration, named *object,* is the 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 [121], [14]. The main part of factors selected for the present experiments has been taken from the generated statistical data base for the EU (see Chapter 2, Section 2.3). There are used the time factor and new gradation factors in the present research. Let us specify the considered factors:

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 ton;

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

t_7 - gradation of countries under value of forecasted parameter (*GradY*);

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*);

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*).

Two last factors are forecasted values. Let us comment them, using the *Reference Manual on Air Transport Statistics*, issued at 7 November 2008 [37].

The *Reference Manual on Air Transport Statistics* contains detailed methodological information as well as background information on the implementation of the legal acts and on how data are processed and disseminated by EuroSTAT.

First of all, *freight and mail* is identified as scheduled or non-scheduled air service performed by aircraft carrying revenue loads other than revenue passengers. It excludes flights carrying one or more revenue passengers and flights listed in published timetables as open to passengers.

By *FrM_Intra* and *FrM_Extra* we denote freight and mail taken *on board* by a country for a certain year and transported within or behind the borders of the European Union, correspondingly. *On board* means all freight and mail on board of the aircraft upon landing at the reporting airport or at taking off from the reporting airport. All freight and mail on board an aircraft during a flight stage. Includes direct transit freight and mail (counted at arrivals and departures). Includes Express services and diplomatic bags. Excludes passenger baggage.

Additional definitions of the terms used in the frame of the statistics on air transport are available in the *Glossary on Air Transport Statistics* [35]. The main principles of EuroSTAT methodology for statistical data gathering and processing are stated in Chapter 2.

The next Section of this Chapter is devoted to the procedure of obtainting of upper confidence limit for the total forecast of two kinds of transportations.

4.2. General Structure of Considered Models and Evaluation Procedure

In the current investigation two kind of mathematical models are used. First of them is well known *multiple regression model*, second one is *multivariate regression model*. The corresponding models are described in Chapter 3, Section 3.2. Let us remind that multivariate model in such treatment of Srivastava resolves use of only same set of explanatory variables for all dependent variables. Besides, Srivastava interprets multivariate model thus when all the dependent variables have the same sense. For example, it might be values of some characteristic taken in equidistant time intervals. How it will be shown in the present Chapter, is possible to expand multivariate regression model application also on situation if p dependent variables describe factors of the different nature.

In our experiment the analysis and forecasting freight and mail air international transportations both Intra-EU and Extra-EU, depending on the same explanatory factors will allow us to show, how they differently influence these two kinds of transportations.

Now let us deduce the formula for obtaining the total forecast of two kinds of transportations. We rewrite the formula for covariance between two estimates of unknown vectors of regression coefficients:

$$Cov\left(\widetilde{\beta}^{\langle k\rangle},\widetilde{\beta}^{\langle l\rangle}\right)=\sigma_{kl}\left(X'X\right)^{-1},\ k,l=1,\dots p. \tag{4.1}$$

Here $\widetilde{\beta}^{\langle k\rangle}$ and $\widetilde{\beta}^{\langle l\rangle}$ are estimated separately using formula (3.9), and σ_{kl} is a corresponding element from the matrix of covariance of errors Σ. As the matrix Σ of covariance of errors is unknown, we can use its unbiased estimate $\widetilde{\Sigma}$:

$$\widetilde{\Sigma}=\frac{1}{n-d}Y'\left[I-X\left(X'X\right)^{-1}X'\right]Y. \tag{4.2}$$

So,

$$Cov^{*}\left(\widetilde{\beta}^{\langle k\rangle},\widetilde{\beta}^{\langle l\rangle}\right)=\widetilde{\sigma}_{kl}\left(X'X\right)^{-1},\ k,l=1,\dots p. \tag{4.3}$$

In our case we have two dependent variables, i.e. internal and external freight and mail international air transportations. First we construct two several multiple models for separately forecasting these values of interest. Further we construct the common multivariate model for forecasting of total interest. For forecasting of total interest we have to obtain the covariance matrixes of coefficients. These matrixes allow us calculating the variance of sum of forecasts of analysed dependent variables. After that we are able to estimate the confidence limit for forecasted expectations of both transportations sum. Note, that if we consider multivariate model, in our case is necessary to calculate the joint covariance matrix for two estimated vectors of regression coefficients $\widetilde{\beta}^{\langle 1\rangle}$ and $\widetilde{\beta}^{\langle 2\rangle}$.

Let we obtain two several forecasts $\widetilde{\mathrm{Y}}^{\langle 1\rangle}=\mathrm{X}\widetilde{\beta}^{\langle 1\rangle}$ and $\widetilde{\mathrm{Y}}^{\langle 2\rangle}=\mathrm{X}\widetilde{\beta}^{\langle 2\rangle}$. Thus 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:

$$E\left(\widetilde{S}\right)=E\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle}\right)+E\left(\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right)=\mathrm{X}\left(\widetilde{\beta}^{\langle 1\rangle}+\widetilde{\beta}^{\langle 2\rangle}\right), \tag{4.4}$$

$$D\left(\widetilde{S}\right)=D\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle}\right)+D\left(\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right)+2Cov\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle},\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right). \tag{4.5}$$

The first and second terms of (4.5) are

$$D\left(\widetilde{\mathrm{Y}}^{\langle 1\rangle}\right)=D\left(\mathrm{X}\widetilde{\beta}^{\langle 1\rangle}\right)=\mathrm{X}\cdot Cov\left(\widetilde{\beta}^{\langle 1\rangle}\right)\cdot\mathrm{X}' \tag{4.6}$$

and

$$D\left(\widetilde{\mathrm{Y}}^{\langle 2\rangle}\right)=D\left(\mathrm{X}\widetilde{\beta}^{\langle 2\rangle}\right)=\mathrm{X}\cdot Cov\left(\widetilde{\beta}^{\langle 2\rangle}\right)\cdot\mathrm{X}', \tag{4.7}$$

where $Cov\left(\widetilde{\beta}^{\langle 1\rangle}\right)$ and $Cov\left(\widetilde{\beta}^{\langle 2\rangle}\right)$ are covariance matrixes of vectors $\widetilde{\beta}^{\langle 1\rangle}$ and $\widetilde{\beta}^{\langle 2\rangle}$ respectively, calculated according formula (4.1). The third term of (4.5) can be defined as follows:

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

$$=\mathrm{X}\cdot E\left(\left(\widetilde{\beta}^{\langle 1\rangle}-\beta^{\langle 1\rangle}\right)'\cdot\left(\widetilde{\beta}^{\langle 2\rangle}-\beta^{\langle 2\rangle}\right)\right)\cdot\mathrm{X}'. \tag{4.8}$$

The member $E\left(\left(\widetilde{\beta}^{\langle 1\rangle}-\beta^{\langle 1\rangle}\right)'\cdot\left(\widetilde{\beta}^{\langle 2\rangle}-\beta^{\langle 2\rangle}\right)\right)$ is the joint covariance matrix of two vectors of coefficients, calculated by formula (4.1). On the base of all listed above the upper confidence limit for total forecast $\overline{E}\left(\widetilde{\widetilde{S}}\right)=E(S)$ corresponding to confidence probability γ is $\left(0,\overline{S}_{\gamma}\right)$, and

$$\overline{S}_{\gamma}=E\left(\widetilde{S}\right)+\sqrt{D\left(\widetilde{S}\right)}\cdot\Phi^{-1}(\gamma), \tag{4.9}$$

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

Now we consider analysed regression models for internal and external transportations forecasting.

4.3. INVESTIGATED MODELS FOR TRANSPORTATIONS FORECASTING

In this section we consider the models for internal and external air freight and mail international transportations forecasting. Note that the suggested models are the *group* models [14], i.e. we forecast the transportations for all the considered countries using the same sets of

the explanatory variables and the same estimates of coefficients. We would ask to pay attention to what objects, i.e. Member States of the EU, are chosen for experiment. First, it is old members of the European Union. These countries have well developed economy, moreover, they have settled and approximately identical tendencies in all kinds of transportations, both passenger and freight and mail. Secondly, it is new members of the European Union. Basically these countries have poorly developed and unstable, somewhere «overheated» economy. The following countries were selected to the present experiment: 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, it is necessary to underline the fact, that not all objects of research have observations on all considered years.

We hope the models offered in the given Section of the present promotional work will allow us to forecast internal and external air freight and mail transportations with equal success for all objects: with small and big transportations, with strongly differing population density, with the various areas and different term of membership in the European Union. We are able to reach it by introduction of some gradation variables which are discussed below.

The first and second models are multiple linear regression models (3.2). 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. Note, that superscript by Y is introduced just for identification of models. Explanatory variables are $x_1 = t_1$, $x_2 = t_2$, $x_3 = t_3$, $x_4 = t_4$, $x_5 = t_5$, $x_6 = t_6$, $x_7 = t_7$, $x_8 = t_8$, $x_9 = t_9$, $x_{10} = t_{10}$.

Notwithstanding that forecasted parameter is transportations in tonnes per thousand of inhabitants, it does not take into account such important characteristics, as area of countries, duration of membership in the EU and so on. Let us explain the sense of the used gradations.

The first gradation t_7 is gradation of countries under value of forecasted parameter. It has been 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. Thus, we can hold corresponding observations in the model. We suppose that indroducing into the model such gradation factor, in its turn, enables us to predict large and small transportations using one same model. We are relieved of necessity of selection of two

different models for large and small transportations [62], [63]. The factor t_7 is equal to 1 for countries with large specific transportations (equal or larger than 20 tonnes per thousand of inhabitants), and it is equal to 0 for countries with transportations smaller than 20 tonnes per thousand of inhabitants.

The second gradation t_8 is gradation of countries under population density. It is equal to 1 for countries with population density equal or larger than 200 inhabitants per km^2 (Belgium, Germany, Netherlands and the United Kingdom), and it is equal to 0 for countries with population density smaller than 200 inhabitants per km^2 (i.e. for other considered countries).

The third gradation t_9 is gradation of countries under area. It is equal to 1 for countries with areas equal or less than 40 000 km^2, and it is equal to 0 for countries with areas larger than 40 000 km^2. This index is equal to 1 for Belgium.

The fourth gradation t_{10} is gradation of countries under the duration of membership in the EU. This index is equal to 1 for new members of the EU (Czech Republic, Estonia, Hungary, Lithuania, Poland and Slovakia) and is equal to 0 for former members of the EU (i.e. for other considered countries).

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 (3.6) 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.

Further we perform two steps: considering of internal and external transportations.

4.4. EVALUATION OF INTERNAL TRANSPORTATIONS

Let us present the results of the first model estimation. For the experiments data from 2001 till 2006 have been used. During the experiment 94 observations have been processed. The estimates of the coefficients and Student criterion values are resulted in Table 4.1. Here $\widetilde{\beta}_i^{\langle 1 \rangle}$ is an estimate of $\beta_i^{\langle 1 \rangle}$, $t(83)$ is the calculated value of Student criterion for 83 degrees of freedom, p-level is the error of second kind (or level of insignificance of variable). The theoretical value of Student criterion for 83 degrees of freedom and level of significance (or error of first kind) $\alpha = 45\%$ is equal to 0.76.

Table 4.1

Estimation results for the first model

Variable	Factor	$\widetilde{\beta}_i^{\langle 1 \rangle}$	$t(83)$	p-level
x_9	GradArea	16.47	6.40056	0.000000
x_1	Year	0.60	2.74044	0.007511
	Intercept	-1192.52	-2.72571	0.007825
x_8	GradMember	-7.10	-2.51648	0.013779
x_5	PP	-0.01	-2.17534	0.032449
x_2	TI	0.09	1.72584	0.088098
x_7	GradY	1.99	1.03248	0.304845
x_4	IR	-0.17	-0.62978	0.530567
x_3	EurRate	-0.00	-0.60482	0.546952
x_{10}	GradPD	-0.52	-0.30429	0.761669
x_6	GDP_PPS	0.00	0.09337	0.925833

Taking into account the fact that the hypothesis of *insignificance* of explanatory variable is tested, we can see that most significant factors affected internal air transportations are all considered ones except GDP in PPS and gradation under population density. The signs by estimated coefficients correspond to the physical sense of considered factors. For example, the negative sign by estimated coefficient of variable GradMember makes it clear that the countries new members of the EU do not make large freight and mail transportations, that corresponds to true. Unexpectedly GDP in PPS has been appeared as the most insignificant variable, and that fact causes bewilderment. Here, of course, negative role has played the fact that considered model is group one, intended for forecasting of transportations for all the EU countries, both high developed and low. New members of the EU, taking into account their overheated economy, have exaggerated Gross National Product. Such countries have GDP higher, than developed countries do, and at the same time insignificantly small transportations in comparison with them. Most likely, this reason has affected the model in such a manner. But the fact that the model has reflected such a feature of economy of developing countries could be recognized as one of advantages of considered model. The signs by other estimated coefficients do not cause any objections.

The coefficient R^2 for this model is equal to 0.82 and the Fisher criterion is 38.43. The theoretical value of Fisher criterion for 10 and 83 degrees of freedom and level of significance $\alpha = 5\%$ is equal to 1.95. Comparing the theoretical and calculated values of Fisher criterion we can conclude that first model cannot be recognized as insignificant. The residual sum of squares R_0 [95] is equal to 2 202.

The estimated model is the following:

$$\widetilde{E}\left(\mathrm{Y}^{\langle 1\rangle}(x)\right) = -1193 + 0.6x_1 + 0.09x_2 - 0.001x_3 - 0.17x_4 - 0.01x_5 +$$
$$+1.99x_7 - 7.1x_8 + 16.47x_9 . \qquad (4.10)$$

Note, that equation (4.10) contains most significant explanatory variables. At the Figure 4.1 we can visually evaluate, how estimated model smoothes experimental data. The observations are arranged in "country-year" order: every point corresponds to some country transportation during the analysed period from 2001 till 2006. Moreover, countries are sorted in alphabetical order. Horizontal axis reflects the observations, arranged in the above-mentioned order. Vertical axis reflects the corresponding transportations, expressed in thousands of tonnes. Notation: not all countries declared in the previous section have data about corresponding transportations for all six years. In spite of this fact, obviously the first model shows quite good smoothing.

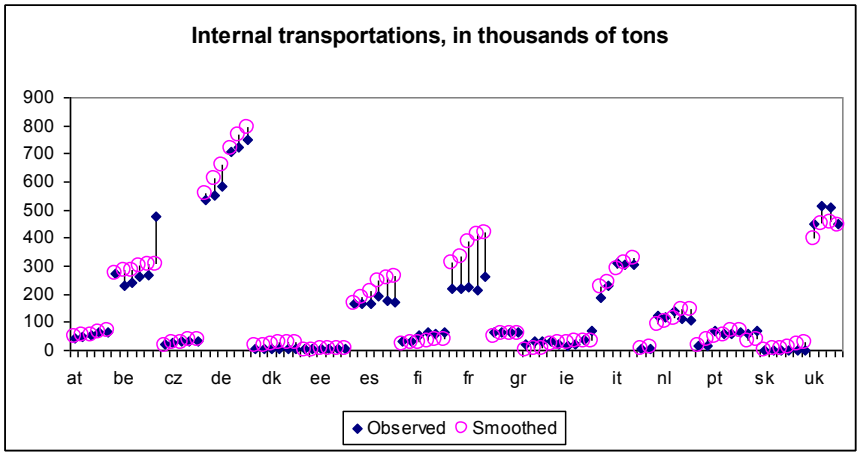

Fig.4.1. Observed and smoothed internal transportations

Table 4.2

Forecasted internal transportations, in thousands of tonnes

Country	*2006**	*2006*	*2007*	*Country*	*2006**	*2006*	*2007*
Austria	64.72	69.67	75.06	Ireland	70.41	32.57	39.44
Belgium	478.75	307.63	318.76	Italy	306.89	326.13	373.12
Czech	34.56	37.16	51.24	Lithuania	6.42	8.99	14.12
Denmark	5.69	26.46	31.03	Netherlands	105.44	146.01	156.53
Estonia	5.52	6.11	7.83	Poland	17.69	35.49	64.13
Finland	62.91	35.64	39.64	Slovakia	0.78	24.24	26.72
France	261.96	419.57	472.59	Spain	173.70	261.76	301.49
Germany	752.61	791.24	850.40	Sweden	68.67	38.05	46.30
Greece	61.67	60.48	72.59	UK	452.67	445.58	491.57
Hungary	22.97	24.32	37.16				

We have estimated internal air freight and mail international transportations for several countries-members of EU for 2007. Table 4.2 contains corresponding forecasted values of dependent variable $Y^{(1)}$ for 2007 (column 2007). Table contains *observed* data for 2006 (column 2006*) and forecasts for 2006 as well (column 2006) for seeing that there are no greater distinctions between 2007 and 2006, i.e. the tendency is kept. Note the fact that data in interesting affected parameters for 2007 were accessible only for these countries.

4.5. EVALUATION OF EXTERNAL TRANSPORTATIONS

Let us present the results of the second model estimation. For the experiments data from 2001 till 2006 have been used. 94 observations have been processed. The estimates of the coefficients and Student criterion values are resulted in Table 4.3. Here $\widetilde{\beta}_i^{\langle 2\rangle}$ is an estimate of $\beta_i^{\langle 2\rangle}$, $t(83)$ is the calculated value of Student criteria for 83 degrees of freedom, p-level is the error of second kind (or level of insignificance of variable). The theoretical value of Student criterion for 83 degrees of freedom and level of significance (or error of first kind) $\alpha = 30\%$ is equal to 1.04.

Taking into account the fact that the hypothesis of insignificance of explanatory variable is tested, we can see that most significant factors affected external air transportations are all considered ones except trade integration of goods and annual average inflation rate. The signs by estimated coefficients correspond to the physical sense of considered factors. Here we observe that GDP in PPS is significant variable in this model, and the sign by its estimated coefficient is positive, as well as should be. The coefficient R^2 for this model is equal to 0.95 and the Fisher criterion is 150.88. The theoretical value of Fisher criterion for 10 and 83 degrees of freedom and level of significance $\alpha = 5\%$ is equal to 1.95. Comparing the theoretical and calculated values of Fisher criterion we can conclude that first model cannot be recognized as insignificant. The residual sum of squares R_0 is equal to 21 989.

The estimated equation is such:

$$\widetilde{E}\left(\mathrm{Y}^{\langle 2\rangle}(x)\right) = -3394 + 1.7x_1 - 0.001x_3 - 0.02x_5 + 0.07x_6 - 52.5x_7 - \\ - 6.6x_8 + 9.7x_9 + 69.3x_{10}. \qquad (4.11)$$

Table 4.3

Estimation results for second model

Variable	Factor	$\widetilde{\beta}_i^{\langle 2\rangle}$	$t(83)$	p-level
x_7	GradY	-52.45	-17.3288	0.000000
x_{10}	GradPD	69.34	25.8921	0.000000
	Intercept	-3394.40	-4.9450	0.000004
x_1	Year	1.70	4.9551	0.000004
x_5	PP	-0.02	-3.5823	0.000573
x_9	GradArea	9.65	2.3907	0.019081
x_8	GradMember	-6.55	-1.4802	0.142593
x_6	GDP_PPS	0.07	1.3577	0.178251
x_3	Eur	-0.00	-1.0497	0.296887
x_2	TI	0.05	0.6399	0.523968
x_4	IR	-0.13	-0.3002	0.764794

Note, that equation (4.11) contains most significant explanatory variables. At the Figure 4.2 we can visually evaluate, how estimated model smoothes experimental data. Every point corresponds to transportations of some country during the analysed period from 2001 till 2006.

We have estimated external air freight and mail international transportations for several countries-members of EU for 2007. Table 4.4 contains corresponding forecasts for 2006 and 2007 (columns 2006 and 2007 respectively) and observed external transportations for 2006 as well (column 2006*).

Note the fact that the model describing external transportations gives a larger absolute error (i.e. $R_0 = 21\ 989$) in comparison with the model describing internal transportations ($R_0 = 2\ 202$). Therefore the forecasts of external transportations, received for 2007, for certain are less exact. Here we can draw the following conclusion that all chosen explanatory factors in a greater measure approach for the description and forecasting of internal transportations, rather than external.

The next steps of our investigation are considering of forecasted expectations of both transportations sums for each country for 2007 and calculating of upper confidence limits for corresponding forecasts.

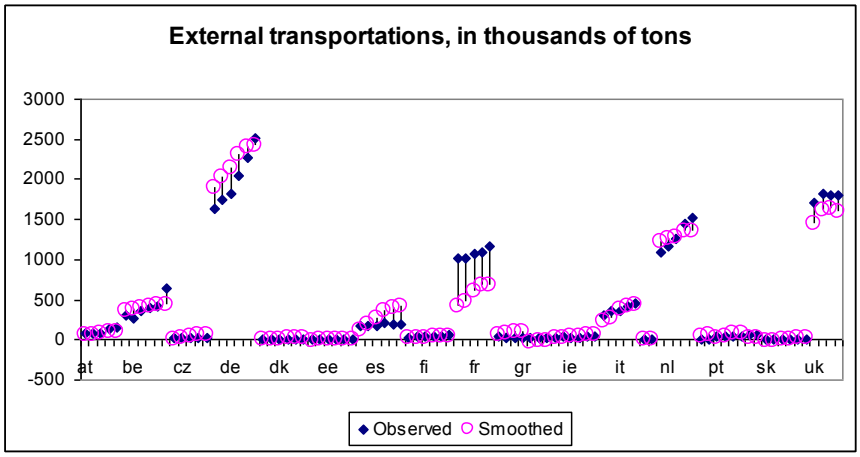

Fig.4.2. Observed and smoothed external transportations

Table 4.4

Forecasted external transportations, in thousands of tonnes

Country	*2006**	*2006*	*2007*	*Country*	*2006**	*2006*	*2007*
Austria	163.21	104.55	121.53	Ireland	41.23	53.70	69.99
Belgium	646.20	435.33	463.03	Italy	457.16	436.58	576.77
Czech	22.49	53.47	81.97	Lithuania	6.25	12.95	22.01
Denmark	1.00	29.25	41.61	Netherlands	1 515.91	1 353.06	1 388.74
Estonia	4.53	9.97	13.13	Poland	14.33	69.14	129.74
Finland	55.51	48.53	61.52	Slovakia	4.58	26.20	34.65
France	1 159.45	686.25	834.43	Spain	195.27	411.72	509.72
Germany	2 515.30	2 422.69	2 593.99	Sweden	70.81	28.36	48.37
Greece	28.93	94.46	125.31	UK	1 795.74	1 599.09	1 756.88
Hungary	41.92	16.90	44.73				

4.6. EVALUATION OF TOTAL TRANSPORTATIONS

For further analysis we will use the results obtained above (see Tables 4.1 – 4.4). Additionally, we must estimate the covariance matrix of errors (3.10). It has been calculated on the base of data from the considered period from 2001 till 2006:

$$\widetilde{\Sigma} = \begin{bmatrix} 9.12 & 5.24 \\ 5.24 & 22.45 \end{bmatrix}. \qquad (4.12)$$

Then on the base of (4.12) we are able to obtain covariance matrices of each estimated vectors of regression coefficients $\widetilde{\beta}^{\langle 1 \rangle}$ and $\widetilde{\beta}^{\langle 2 \rangle}$, and the joint covariance matrix of both vectors of coefficients, calculated by (4.1).

Table 4.5 contains obtained on the base of covariance matrices (4.1) variances of forecasted internal and external transportations, calculated by (4.6) and (4.7) (columns Var_1

and Var_2 respectively), the joint covariance (4.8) of forecasted sum of both transportations (columns *Cov*) and the variance of total forecast of both transportations (4.5) for 2007 (column Var_{2007}). Note that values contained in Table 4.5 are given for specific transportations.

The given Chapter has been devoted to elaboration of regression models for Member States of the Europe Union freight and mail air international transportations analysis and forecasting. The air international transportations are divided into two parts, thus there are two forecasted variables: internal and external transportations in relation to the EU. Two types of models have been investigated: multiple regression model and multivariate regression model. The last model is applied for forecasting of sums of mentioned transportations by means of upper confidence limits.

Table 4.5

Variance of forecasted total specific transportations for 2007

Country	Var_1	Var_2	*Cov*	Var_{2007}	*Country*	Var_1	Var_2	*Cov*	Var_{2007}
Austria	1.05	2.58	0.60	4.84	Ireland	1.75	4.32	1.01	8.08
Belgium	2.07	5.08	1.19	9.53	Italy	2.08	5.13	1.20	9.61
Czech	1.67	4.11	0.96	7.69	Lithuania	1.39	3.41	0.80	6.39
Denmark	1.27	3.12	0.73	5.84	Netherlands	2.32	5.72	1.34	10.72
Estonia	1.28	3.16	0.74	5.91	Poland	2.44	6.02	1.41	11.27
Finland	0.64	1.58	0.37	2.95	Slovakia	1.37	3.36	0.79	6.30
France	0.98	2.41	0.56	4.51	Spain	1.10	2.70	0.63	5.06
Germany	1.47	3.63	0.85	6.79	Sweden	2.10	5.18	1.21	9.70
Greece	1.05	2.58	0.60	4.84	UK	1.62	3.99	0.93	7.47
Hungary	1.02	2.50	0.58	4.68					

Table 4.6 contains the mentioned above total forecasts for 2007 and for 2006 (4.4) (columns 2007 and 2006 respectively) and sums of observed internal and external transportations for 2006 (column 2006*). We have calculated the upper confidence limits (4.9) with different confidence probabilities γ for obtained total forecasts for 2007 for each country. The corresponding data are presented in Table 4.6 as well. At Figure 4.3 we can see represented in Table 4.6 observed and forecasted transportations.

CONCLUSION

Gotten results show remarkable evidence of proposed approach. However, individual models would be preferable for forecasts for several 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 of transportations sums.

In this connection it is supposed to improve the present research by means of SURE-model. Multivariate and SURE-model have been applied for forecasting of the total passenger air intra-EU and extra-EU transportations (in the framework of the Master thesis performed under supervision of the Candidate, [1]). The results of corresponding investigation are described in Chapter 7.

Table 4.6

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

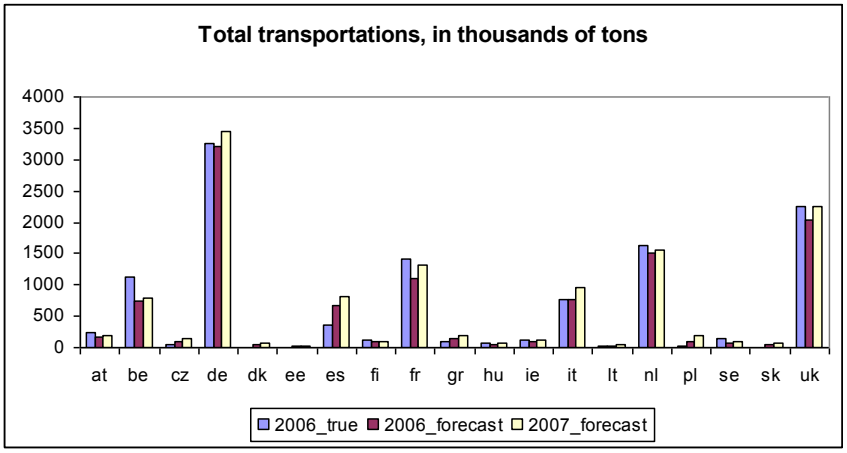

Fig.4.3. Total observed and forecasted transportations

5. EVALUATION OF RAILWAY PASSENGER CORRESPONDENCES BETWEEN MEMBER STATES OF THE EUROPEAN UNION

In the current Chapter the nonlinear regression model for forecasting passenger departures (or correspondences) between various geographical points is described. The suggested nonlinear model has been developed on the basis of the gravity model (Chapter 3). Unknown parameters of the model are estimated using aggregated data when only the information on the departed from each geographical point passenger's quantity is available. For this purpose an original algorithm has been elaborated. As the estimation efficiency criterion the weighted sum of residual squares has been used. The elaborated estimation procedure has been tested on the real data.

5.1. PROBLEM SETTING

So, let us describe the problem. We have n corresponding geographical points (which may be towns, countries, stations and so on) with numbers $i = 1, 2, \ldots n$. For the point i, one is known inhabitants (citizen) number h_i and m numerical characteristics (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. Besides let the point i be considered as *origin point*, and another, l, be considered as *destination point*. In addition, we know the quantity of the departed passengers Y_i from the point i during considered time interval, that is a random variable.

Our aim is to estimate correspondence size $Y_{i,l}$ for all pairs of points (i,l), precisely the quantity of the departed passengers from the point i to the point l. The matrix of $Y_{i,l}$ is to be called *the correspondence matrix*, or *Origin-Destination matrix* (*OD-matrix*, see [77]). Let us denote an estimate of $Y_{i,l}$ by $Y^*_{i,l}$. Moreover, it is required that all estimated correspondences $Y^*_{i,l}$ would be positive $(Y^*_{i,l} > 0)$ for $i \neq l$, a correspondence from any point in itself would be equal to zero, $Y^*_{i,i} = 0$; and also correspondence from the origin point in the destination point would be to equal correspondence from the destination point in the origin point, $Y^*_{i,l} = Y^*_{l,i}$.

As the mathematical model for the particular correspondence (i,l) for $i \neq l$ we use the following expression:

$$Y_{i,l} = \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}\right), \tag{5.1}$$

where a, $\alpha = (\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_m)^T$ and $\beta = (\beta_1 \quad \beta_2 \quad \dots \quad \beta_m)^T$ are unknown regression parameters, τ and θ are unknown form parameters, $c_{(i)} = (c_{i,1} \quad \dots \quad c_{i,m})$ and $g_{(i,l)} = (c_{i,1}c_{l,1} \quad \dots \quad c_{i,m}c_{l,m})$ are m–vector-rows, $\{V_{i,l}\}$ are independent identically distributed random variables with zero mean and unknown variance σ^2.

Note that the case $\theta = 1$ and $\tau = 2$ corresponds to the so-called gravity model (described in Chapter 3).

As a corollary of this model we get the following presentation for the quantity of the departed passengers from the point *i*:

$$Y_i = \sum_{\substack{l=1 \\ i \neq l}}^{n} Y_{i,l} = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta + V_{i,l}\right). \tag{5.2}$$

Now we must estimate unknown parameters on the basis of fixed values $\{Y_i\}$. Such a problem was considered earlier in the literature. Besides, for that purpose usually the entropy approach is used. But there are many obtained estimates of $Y_{i,l}^*$ equal to zero that is inaccessible. We use regression theory [95]. For that we need to investigate the distribution and the expectation of $Y_{i,l}$.

5.2. Analysis of Random Variables Distribution

We suppose that $V_{i,l}$ has normal distribution. Then $Z_{i,l} = \exp(V_{i,l})$ has the log-normal distribution [91] with characteristics

$$E(Z_{i,l}) = E(\exp(V_{i,l})) = \exp\left(\frac{1}{2}\sigma^2\right),$$
$$D(Z_{i,l}) = D(\exp(V_{i,l})) = \exp(\sigma^2)(\exp(\sigma^2) - 1).$$

Therefore, for $i \neq l$

$$E(Y_{i,l}) = \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right)\exp(\frac{1}{2}\sigma^2), \tag{5.3}$$

$$\begin{aligned} D(Y_{i,l}) &= \frac{(h_i h_l)^{2\theta}}{(d_{i,l})^{2\tau}} \exp\left(2\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right)\right)\exp(\sigma^2)(\exp(\sigma^2) - 1) = \\ &= (\exp(\sigma^2) - 1)(E(Y_{i,l}))^2. \end{aligned} \tag{5.4}$$

Analogous formulae have places for $\{Y_i\}$:

$$E(Y_i) = \sum_{l=1}^{n} E(Y_{i,l}) = \exp\left(\frac{1}{2}\sigma^2\right)\sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(a + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right), \tag{5.5}$$

$$D(Y_i) = (\exp(\sigma^2) - 1)\sum_{l \neq i} (E(Y_{i,l}))^2. \tag{5.6}$$

5.3. THE LEAST SQUARES ESTIMATES OF UNKNOWN PARAMETERS

We wish to use the expectation (5.5) for the estimation of the unknown parameters θ, τ, a, α, β and σ^2. But one cannot identify both parameters a and σ^2 simultaneously. So, let us introduce the united parameter $\widetilde{a} = a + \frac{1}{2}\sigma^2$ and rewrite (5.5) as

$$E(Y_i) = \sum_{\substack{l=1 \\ l \neq i}}^{n} \frac{(h_i h_l)^{\theta}}{(d_{i,l})^{\tau}} \exp\left(\widetilde{a} + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta\right). \tag{5.7}$$

As a criterion of estimates efficiency we use weighted least squares sum:

$$R(\gamma, w) = \sum_{i=1}^{n} w_i \left(Y_i - E(Y_i)\right)^2, \tag{5.8}$$

where $\gamma = \left(\theta \quad \tau \quad \widetilde{a} \quad \alpha^T \quad \beta^T\right)^T$ and $w = (w_1 \quad w_2 \quad \ldots \quad w_n)^T$ is a vector of weights.

For a minimization of (5.7) we use the gradient method. Let

$$\nabla R(\gamma, w) = \left(\frac{\partial}{\partial \theta} R \quad \frac{\partial}{\partial \tau} R \quad \frac{\partial}{\partial a} R \quad \frac{\partial}{\partial \alpha} R \quad \frac{\partial}{\partial \beta} R \right)^T . \tag{5.9}$$

If the weights w do not depend on the parameters, then

$$\nabla R\left(\begin{pmatrix} \theta \\ \tau \\ \tilde{a} \\ \alpha \\ \beta \end{pmatrix}, w \right) = -2 \begin{pmatrix} \sum_{i=1}^{n} w_i (Y_i - E(Y_i)) \sum_{l=1}^{n} \ln(h_i h_l) \frac{(h_i h_l)^{\theta}}{d_{i,l}^{\tau}} \exp(f_{(i,l)}) \\ -\sum_{i=1}^{n} w_i (Y_i - E(Y_i)) \sum_{l=1}^{n} \ln(d_{i,l}) \frac{(h_i h_l)^{\theta}}{d_{i,l}^{\tau}} \exp(f_{(i,l)}) \\ \sum_{i=1}^{n} w_i (Y_i - E(Y_i)) \sum_{l=1}^{n} \frac{(h_i h_l)^{\theta}}{d_{i,l}^{\tau}} \exp(f_{(i,l)}) \\ \sum_{i=1}^{n} w_i (Y_i - E(Y_i)) \sum_{l=1}^{n} \frac{(h_i h_l)^{\theta}}{d_{i,l}^{\tau}} \exp(f_{(i,l)}) (c_{(i)} + c_{(l)})^T \\ \sum_{i=1}^{n} w_i (Y_i - E(Y_i)) \sum_{l=1}^{n} \frac{(h_i h_l)^{\theta}}{d_{i,l}^{\tau}} \exp(f_{(i,l)}) g_{(i,l)}{}^T \end{pmatrix}, \tag{5.10}$$

where $f_{(i,l)} = (\tilde{a} + (c_{(i)} + c_{(l)})\alpha + g_{(i,l)}\beta)$.

In this case (w_i = const), the gradient method quickly gives estimates θ^*, τ^*, $\tilde{a}^*$, α^*, β^*. Otherwise, the weights contain the unknown parameters. Therefore we must use an iterative procedure and successively recalculate estimates of the weights and the parameters. To avoid such complicacy, another estimation approach, based on the maximum likelihood estimator (MLE), is used [9], [10].

5.4. ESTIMATION OF PARAMETERS a AND σ^2

According to (5.3) and (5.7) we have the estimates for $i \neq l$:

$$E(Y_{i,l})^* = \frac{(h_i h_l)^{\theta^*}}{(d_{i,l})^{\tau^*}} \exp(\tilde{a}^* + (c_{(i)} + c_{(l)})\alpha^* + g_{(i,l)}\beta^*), \tag{5.11}$$

$$E(Y_i)^* = \sum_{\substack{l=1 \\ i \neq l}}^{n} E(Y_{i,l})^* = \sum_{\substack{l=1 \\ i \neq l}}^{n} \frac{(h_i h_l)^{\theta^*}}{(d_{i,l})^{\tau^*}} \exp(\tilde{a}^* + (c_{(i)} + c_{(l)})\alpha^* + g_{(i,l)}\beta^*). \tag{5.12}$$

Analogously from (5.6) we get

$$D(Y_i)^* = \left(\exp\left(\sigma^{2*}\right)-1\right)\sum_{l=1}^{n}\left(E\left(Y_{i,l}\right)^*\right)^2. \tag{5.13}$$

At other hand we can estimate the variance of Y_i with respect to the variance definition

$$D(Y_i) = E(Y_i - E(Y_i))^2.$$

Using $E(Y_i)^*$ as the estimate $E(Y_i)$, we have an alternative estimate of $D(Y_i)$:

$$D(Y_i)^{**} = \left(Y_i - E(Y_i)^*\right)^2. \tag{5.14}$$

Here we suppose a weak dependence between Y_i and $E(Y_i)^*$ because the last is calculated on base of many $\{Y_l\}$.

Now the variance parameter σ^2 can be estimated using the equalization of both values (5.13) and (5.14). By summing ones for $i = 1, \ldots, n$, we get

$$\sum_{i=1}^{n} D(Y_i)^* = \sum_{i=1}^{n} D(Y_i)^{**}$$

or

$$\left(\exp\left(\sigma^{2*}\right)-1\right)\sum_{i=1}^{n}\sum_{\substack{l=1\\ l\neq i}}^{n}\left(E\left(Y_{i,l}\right)^*\right)^2 = \sum_{i=1}^{n}\left(Y_i - E(Y_i)^*\right)^2 .$$

Therefore

$$\sigma^{2*} = \ln\left\{1+\left(2\sum_{i=1}^{n-1}\sum_{l=i+2}^{n}\left(E\left(Y_{i,l}\right)^*\right)^2\right)^{-1}\sum_{i=1}^{n}\left(Y_i - E(Y_i)^*\right)^2\right\}. \tag{5.15}$$

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

5.5. BALANCING PROCEDURE

Often one requires that the statistical data $\{Y_i\}$ and the estimates $\{Y^*_{i,l}\}$ have been balanced:

$$\sum_{l=1}^{n} Y^*_{i,l} = Y_i \;\; i = 1,\ldots,n. \tag{5.16}$$

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

$$\widetilde{Y}_{i,l} = \delta_i Y^*_{i,l} \delta_l, \;\; i,l = 1,\ldots,n. \tag{5.17}$$

To calculate the coefficients $\{\delta_i\}$ we have the following nonlinear system:

$$\delta_i \sum_{l=1}^{n} Y^*_{i,l} \delta_l = Y_i, \;\; i = 1,\ldots,n,$$

$$\delta_i = \left(\sum_{l=1}^{n} Y^*_{i,l} \delta_l \right)^{-1} Y_i, \;\; i = 1,\ldots,n.$$

The experience shows that a solution is easily determined by the successive approaches method. For the *k*-th iteration we have:

$$\delta_i^{(k)} = \left(\sum_{l=1}^{i-1} Y^*_{i,l} \delta_l^{(k)} + \sum_{l=i+1}^{n} Y^*_{i,l} \delta_l^{(k-1)} \right)^{-1} Y_i, \;\; i = 1,\ldots,n, \tag{5.18}$$

where $\{\delta_i^{(0)}\}$ are initial values and we take in mind that $Y^*_{i,i} = 0$.

The iterations are ended, when a difference between two last values of $\delta^{(k)} = \left(\delta_1^{(k)} \;\; \delta_2^{(k)} \;\; \ldots \;\; \delta_n^{(k)}\right)$ is less then prescribed precision $\varepsilon > 0$.

5.6. NUMERICAL EXAMPLE

As we have already known, many models, allowing to estimate correspondences between origin point (or point of embarkation) and destination point (or point of disembarkation), frequently make specific demands to the statistical data [12]. First of all, the considered system should be closed, i.e. the sum of all passengers departed (passengers embarked) on all considered origin points should be equal to the sum of the passengers arrived (passengers disembarked) on all considered destination points of a route. Besides, transportation (correspondence) from point i to point j should be equal to the so called *inverse* transportation, i.e. from the point j to the point i. Such condition can be satisfied, only if destination points are terminal points, instead of transshipment points. The next requirement is equality to zero of transportation from point i to itself. Here it is necessary to know, at what level of generalization the statistics about transportations is collected, i.e. what are an origin and a destination – any city, station, airport or the whole country.

So, we apply the suggested approach for the passenger railway transportations estimation between 23 Member States of the European Union, further called *countries*. So here these countries play the part of the geographical points. The following countries have been selected for the present investigation: Belgium, Bulgaria, Czech Republic, Denmark, Germany, Ireland, Greece, Spain, France, Italy, Latvia, Lithuania, Luxembourg, Hungary, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Sweden, and United Kingdom. Such coutries as Malta and Cyprus have not been considered as they do not have railways at all. Estonia and Finland have not been included in the experiment as well because EuroSTAT database does not contain any rail passenger statistics on them.

Taking into account our previous investigation [14] the following characteristics of the countries 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 is outside the fold. It can take on only two values: 0 for the old, former Member States of the EU, and 1 for the new

Member States of the EU. 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 each considered country about air, sea and rail international passenger transportations for the 2007 year have been analysed [31].

As distances between points, the distances between capitals of countries have been taken. As a possible measure of distance there also could be considered distances between the weighted average coordinates of countries [11].

In the present experiment 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 5.1). The statistical data for the year 2007 have been obtained from the EuroSTAT database [31]. Recall that in accordance with the common EuroSTAT methodology passenger rail international outgoing transportations do not include transit passengers.

Let us show a part of the data available in our disposal. The rows of our completed correspondence matrix represent the coutries of embarkation; the columns represent the countries of disembarkation. So, our correspondence matrix should be symmetric (see Fig.5.1).

But as it was mentioned in Chapter 2 of the present promotional work, despite seeming uniformity and reliability of the data taken from such reliable origin, as EuroSTAT database is, even there can be different interpretations.

First, the requirement that considered system should be closed (as stated above) is not satisfied for all the considered countries. Secondly, the requirement that transportation (correspondence) from point i to point j should be equal to transportation from the point j to the point i does not satisfied as well for couple of countries. For example, for the pair „Belgium-France" correspondence is 1852 thousands of passengers, i.e. embarked in Belgium and disembarked in France. According to our model assumptions and EuroSTAT Regulation No 1192/2003 [34] maintained that so called „mirror check" takes place during data gathering and verification, for the pair „France-Belgium" we should have the same correspondence. Analysing the present data, we have 1734 thousands of passengers, i.e. embarked in France and disembarked in Belgium. Most likely, such discrepancies in data are observed because of different metodologies of passenger counting used by the Member States of the EU.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
		geo	*eu27* E	*be* Bel	*bg* Bul	*cz* Cze	*dk* Der	*de* Ger	*ie* Irela	*gr* Gre	*es* Spa	*fr* Fran	*it* Italy	*lv* Latv	*lt* Lithu	*lu*
	eu27 Europea		0	3188	69	1016	4974	5072	347	18	329	5184	1599	2	7	2
	be Belgium		0	0	0	0	0	0	0	0	0	1852	0	0	0	6
	bg Bulgaria		0	0	0	1	0	0	0	18	0	0	0	0	0	0
	cz Czech Repu		0	0	0	0	14	0	0	0	0	0	12	0	0	0
	dk Denmark		0	0	0	0	0	0	0	0	0	0	0	0	0	0
	de Germany (		0	235	1	136	263	5072	0	0	0	545	372	0	0	2
	ie Ireland		0	0	0	0	0	0	0	0	0	0	0	0	0	0
	gr Greece		0	0	52	0	0	0	0	0	0	0	0	0	0	0
0	*es* Spain		0	2	0	0	0	0	0	0	0	315	24	0	0	0
1	*fr* France		0	1734	0	0	1	0	0	0	231	3	839	0	0	1
2	*it* Italy		0	1	0	0	0	0	0	0	20	777	0	0	0	0
3	*lv* Latvia		0	0	0	0	0	0	0	0	0	0	0	0	2	0
4	*lt* Lithuania		0	0	0	0	0	0	0	0	0	0	0	2	0	0
5	*lu* Luxembour		0	410	0	0	0	0	0	0	0	238	0	0	0	0
6	*hu* Hungary		0	0	2	29	0	0	0	0	0	0	19	0	0	0
7	*nl* Netherlanc		0	805	0	1	6	0	0	0	0	265	1	0	0	1
8	*at* Austria		1575	0	1	205	11	0	0	0	0	8	298	0	0	1
9	*pl* Poland		0	0	0	58	1	0	0	0	0	0	3	0	5	0
0	*pt* Portugal		0	0	0	0	0	0	0	0	78	0	0	0	0	0

Fig.5.1. Fragment of the true correspondence matrix

Therefore it is necessary to estimate 15 parameters. The above described estimation procedure (for $w_i = 1$), developed in the MathCad 13 environment [126], gives the following values of estimated parameters:

$$\theta^* = 0.788,\ \ \tau^* = 2.786,\ \ \widetilde{a}^* = -10.01,$$

$$\alpha^* = \begin{pmatrix} 0.074 & 0.061 & 0.077 & 0.063 & 0.078 & 0.33 \end{pmatrix}^T,$$

$$\beta^* = \begin{pmatrix} 0.197 & 0.194 & 0.152 & 0.054 & 0.097 & 0.104 \end{pmatrix}^T.$$

The corresponding value of criteria (5.8) R = 1.8×10^5. The multiple correlation coefficient [95] is equal to 0.97. The estimate σ^{2*} is calculated by formula (5.15): $\sigma^{2*} = 0.065$. Now using the estimate $\widetilde{a}^* = -10.01$ we find

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

Observed (Y_i) and estimated (Y_i^*) departures from each country in thousands of passengers are represented in the Table 5.1. The correction coefficients δ_i have been obtained as well.

Table 5.1

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 5.2 contain a part of estimated and true correspondences for some of analysed countries. As we can see, the considerable part of the true data is inaccessible. Corresponding data are denoted by double point.

Let us perform light analysis of obtained results and give some comments on it. For example, for the pair “France-Germany” the estimated correspondence is equal to 153 thousands of passengers, i.e. embarked in France, disembarked in Germany. The inverse correspondence is 155 thousands of passengers, i.e. embarked in Germany, disembarked in France. Unfortunately, we are prevented from seeing true correspondence for this pair and, in its turn, from comparing with the estimated one. In spite of that fact, obviously the suggested model is able to keep theoretical assumptions stated above and tendencies in transportations between the Member States of the EU. The next proof of such keeping is estimated correspondence from Germany to Germany, which is equal to zero. According the stated model (5.1), such correspondences $Y_{i,i}^*$ should be equal to zero. True correspondence here is equal to 5072 thousands of passengers, that contradiscts to that assumption. Analysing Table 5.1, we can see, that estimated total departure for Germany is very close to the true one. This fact proofs correctness of balancing condition (5.17). In other words, we have obtained that have assumed, making such model.

Table 5.2

Several estimated correspondences

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

Analyzing the results, after all 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.

CONCLUSION

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.

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

The given Chapter of the present Thesis is devoted to the analysis and forecasting of rail transportation volumes for the Member States of the European Union and Latvia on the basis of the single index model. In this connection first Section of this Chapter is dedicated to the question of SIM estimation. Other two Sections contain descriptions of corresponding numerical experiments.

6.1. ESTIMATION OF THE SINGLE INDEX MODEL

In the present Section the elaborated method of the single index model estimation is stated. Developed procedures of choosing of the most efficient SIM and linear regression model are explained as well.

Let us start with estimation of the unknown coefficients of the SIM. In details SIM is described in the Chapter 3. Recall that general formula of the single index model is the following:

$$E(Y \mid x) = m(x) = g\{v_\beta(x)\}. \tag{6.1}$$

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

$$m(x_i) = g(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_d x_{i,d}) = g(\tau_i), \tag{6.2}$$

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:

$$\tilde{g}(x) = \frac{1}{\sum_{i=1}^{n} K_h(\tau_i)} \sum_{i=1}^{n} K_h(\tau_i) Y_i \,, \tag{6.3}$$

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*. Nadaraya-Watson estimator and kernel functions are described in the Chapter 3, Section 3.3.

In our investigation we use only the Gaussian function as $K_h(\bullet)$:

$$K_h(\tau) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\tau}{h} \right)^2 \right), \; -\infty < \tau < \infty, \tag{6.4}$$

where h is a *bandwidth*. Importance of appropriate bandwidth choosing is described in the Chapter 3 as well.

The unknown coefficients vector β is estimated using the OLS criterion:

$$R(\beta) = \sum_{i=1}^{n} (Y_i - \tilde{g}(x_i))^2 \to \min_{\beta} \,. \tag{6.5}$$

For such minimization we will use the *gradient method* [4], [46]. So, first of all, we need to derive the formula of gradient of criterion (6.5) with respect to the unknown coefficients vector β. Let us deduce it:

$$\nabla R(\beta) = -2 \sum_{i=1}^{n} \left(Y_i - \frac{\sum_{i=1}^{n} K_h(\tau_i) Y_i}{\sum_{i=1}^{n} K_h(\tau_i)} \right) \times$$

$$\times \frac{\left(\sum_{i=1}^{n} \frac{\partial}{\partial \tau_i} K_h(\tau_i) Y_i \right) \cdot \frac{x_i}{h} \cdot \sum_{i=1}^{n} K_h(\tau_i) - \left(\sum_{i=1}^{n} \frac{\partial}{\partial \tau_i} K_h(\tau_i) \right) \cdot \frac{x_i}{h} \cdot \sum_{i=1}^{n} K_h(\tau_i) Y_i}{\left(\sum_{i=1}^{n} K_h(\tau_i) \right)^2}. \tag{6.6}$$

After uncomplicated expressions we have the formula of the corresponding gradient:

$$\nabla R(\beta) = -2\sum_{i=1}^{n}\left(Y_i - \frac{\sum_{i=1}^{n} K_h(\tau_i) Y_i}{\sum_{i=1}^{n} K_h(\tau_i)} \right) \cdot \left(\sum_{i=1}^{n} K_h(\tau_i) \right)^{-2} \cdot \left(\frac{1}{h} \sum_{i=1}^{n} \frac{\partial}{\partial \tau_i} K_h(\tau_i) \cdot \left(Y_i \sum_{i=1}^{n} K_h(\tau_i) - \widetilde{Y} \right) \cdot x_i \right), \quad (6.7)$$

where

$$\widetilde{Y} = \sum_{i=1}^{n} K_h(\tau_i) Y_i \quad (6.8)$$

and

$$\frac{\partial}{\partial \tau_i} K_h(\tau_i) = -\frac{\tau_i}{h^2 \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\tau_i}{h} \right)^2 \right) \quad (6.9)$$

is the first derivative of the Gaussian kernel (6.4).

As we can see, the formula of the used gradient has quite complicated analytical view. Thereupon it is possible to imagine that expressions for the second derivatives of the used criterion (i.e. Hessian) will have more complicated analytical appearance [46]. It is a reason why optimization methods of the second order have not been used.

Before to state the procedure of choosing of the most significant SIM, it is important to make notice about in how way are we able to compare single index models among them and with parametric ones. Hardle in [53] describes the special hypothesis for testing whether the estimated model is GLM or a true SIM. According to our reckoning, this procedure is not described distinctly enough. On the other hand, *Y.Xia*, *W.K.Li*, *H.Tong*, and *D.Zhang* proposed a goodness-of-fit test for single-index models with unknown link function [111].

For simplification of procedure of the choice of the most significant model, and also in view of the fact that we are interested to obtain the most exact forecasts, we suggest comparing parametric and semiparametric models by the *residual sum of squares* R_0. We calculate the residual sum of squares in such a manner:

$$R_0 = \frac{1}{n-d} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2, \quad (6.10)$$

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 $\hat{\beta}^T x_i$.

Now, we are able to state the procedure of the choice of the most significant single index model on the basis of the *gradient method*.

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 [53], [102].
3. Choose the adequate accuracy level ε for the criterion (6.5) optimal value calculation and the adequate accuracy level η for the gradient (6.7) 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 $\nabla R(\beta^{(n,i)})$:

$$\theta^{(i)} = \frac{\nabla R(\beta^{n,(i)})}{\sqrt{\nabla R(\beta^{(n,i)})^T \cdot \nabla R(\beta^{(n,i)})}}. \tag{6.11}$$

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. [46]).
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 [46]:

$$\beta^{(n,i)} = \beta^{(n,i-1)} - \nu^{(i)} \cdot \theta^{(i)}. \tag{6.12}$$

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 step 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 single index model efficiency supposes comparing SIM with linear models (having the similar set of factors), let us describe the procedure of the choice of the most significant linear model.

Begin of procedure II.

1. The analysis of the factors influencing depending variable. Selection of the most suitable factors and their inclusion in model. Set sequence number of the considered model $n = 0$.
2. $n = n + 1$. Obtaining of the unknown coefficient estimate $\beta^{(n)}$ by means of LS method.
3. Test for conformity of signs by coefficients estimates $\beta^{(n)}$ to the physical sense of corresponding factors. If signs by coefficient estimates do not correspond to to the physical sense of factors, then go to step 12.
4. Calculation of multiple determination coefficient $R^2_{(n)}$ [95], [120], [127]. If $R^2_{(n)}$ does not exceed some limit (for example, 0.8) then go to step 12.
5. Test of hypothesis of regression insignificance using Fisher criterion [120], [127]. If hypothesis of regression insignificance is not rejected, then go to step 12.
6. Test of hypothesis of *k*-th factor insignificance using Student criterion [120], [127]. If there are factors for which hypothesis of insignificance is not rejected, then exclude these factors from model and go to step 2.

7. Construct equation of the best chosen model, obtaining the estimates of the forecasted values.
8. Calculate the residual sum of squares $R_0^{(n)}$.
9. If $n = 1$ then go to step 10. Else go to step 11.
10. Current $R_0^{(n)}$ is recognized as optimal value of the residual sum of squares (so called *record* $RE^{(n)}$) and go to step 12.
11. If current $R_0^{(n)}$ is less than last record $RE^{(n)}$ then recognize it as new record, set $RE^{(n)} \leftarrow R_0^{(n)}$.
12. If among analysed factors still are not included in the model, then go to step 1.
13. Declare current model with number n as the best chosen model.

End of procedure II.

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. In this connection let us mention and define so called *approximation criteria*, used for analysis of regression models efficiency.

Approximation criteria are the criteria based on direct comparison of the obtained estimates with the observed data. Correctness of the made forecast can be checked up only after a while – for this purpose it is necessary to wait the new data and to compare them with settlement values. However, it is very important to know the forecast quality directly at its drawing up. Therefore comparison of the obtained forecasts with the observed data one spends on the retrospective period, i.e. under the available statistical data [101].

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, i.e. on all the observations from the existing statistical sample. 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.

Cross-validation is a statistical practice of division of the data sample on subsets in such a manner that the initial analysis is carried out on one subset while other subset is kept for the subsequent use and serves for validation of the made analysis [101].

The cross-validation approach is applied when the subsequent data samples are dangerous, expensive, are spaced apart on time or are simply impossible to be got.

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.

So, taking into account such stated above model estimation approaches as smoothing and cross-validation, we are going to test the efficiency of our suggested models in two cases: existing data smoothing and forecasting. There is a further point to be made here, that earlier described procedures of choice of the most siginificant models are intended for case of smoothing only. However, these procedures are able to be easily arranged for the case of cross-validation too. For that purpose steps from 2 till 11 of the procedure I and steps from 2 till 6 of the procedure II have to be designated to regression model coefficients estimation on the basis of a part of data. Then steps 12-13 of the procedure I and steps 7-8 of the procedure II are aimed for forecasting of values of the dependent variable corresponding to observations of second part of the data sample. After that we can perform comparing of the models in sense of the residual sum of squares R_0 (6.10).

In the next Section we show the results of numerical experiment inducted to the analysis and forecasting of turnover of international rail freight transport for the Member States of the European Union using single index model and linear model.

6.2. Analysis and Forecasting of Turnover for International Rail Freight Transport for the EU Member States

In the corresponding collective research [14] the linear regression models, generalized regression models and semiparametric regression models have been considered. *C.Zhukovskaya* has analyzed the generalized regression models in her promotional work [8] as well for air passenger transportations forecasting for the Member States of the EU.

In the present Section we perform analysis and forecasting of international rail freight transport turnover applying single index regression model and multiple linear regression

model. The estimation of SIM is carried out using the method developed in the framework of the present Thesis. Efficiency of SIM is investigated according to stated procedures (described in the previous Section of this Chapter). The results of the performed research show that turnover forecasts obtained on the basis of SIM are more precise in comparison with forecasts obtained by use of linear regression models.

6.2.1. PROBLEM SETTING

The main object of consideration named *object* is a Member State of the European Union performing certain turnover of international rail freight transport, further named *country*. We call as *observation* a data about object for an actual time moment of analyzed period. All below considered models are group models (Chapter 3, Section 3.1), i.e. we forecast the volumes of turnover for all the considered countries using the same sets of the explanatory variables and corresponding statistical data.

The task of research is to construct various regression models, i.e. models with different combinations of explanatory factors, and then to choose from them ones given the best forecasts of turnover of international rail freight transport. The main difficulty is to choose the set of convenient factors influencing the turnover.

So, we consider volumes of turnover (see definition in Chapter 2) of international rail freight transport, expressed in million tonne-km (see definition in Chapter 2), for the set of Member States of the European Union. The following 15 countries had been selected for the experiment: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom. As we can see, only so-called old Member States of the EU are included in the experiment.

These countries have a high standard of living, well developed industry and infrastructure, and a high level of export as well. It gives us hope that suggested group models will reflect tendencies in fluctuations of turnover of international rail freight transport equally adequately for all these countries.

The analyzed time period is from 1996 to 2000. So, for each country for each considered year we have the volumes of turnover of international rail freight transport and the volumes of factors influencing turnover as well. So, for 15 countries and 5 years we have 75 observations.

First of all, the variable of interest, i.e. mentioned above turnover, is denoted by t_0. Let us rewrite some factors from the general data base of factors (see Chapter 2, Section 2.2), which are used in this experiment as explanatory variables:

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

t_2 – Gross Domestic Product per capita in Purchasing Power Standards (*GDP_PPS*);

t_3 – Comparative Price Level (*CPL*);

t_4 – total length of railways (*TOTLEN*), in thousands of km;

t_5 – number of locomotives (*LOKOM*), in thousands;

t_6 – number of goods wagons (*WAGONS*), in thousands;

t_7 – index of country area (*GradAREA*).

Last factor GradAREA is introduced especially for models considered in the current experiment.

6.2.2. SUGGESTED MODELS

Now we describe four investigated regression models. Two of them are linear regression models and other two are SIM.

The first model is a multiple linear regression model (3.2). The dependent variable $Y^{(L1)} = t_0$ is raw (not-specified) turnover. Note, that superscript by Y is introduced just for identification of models. 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 ratio t_2/t_3 enables us to see how these two factors in aggregate influence the volumes of turnover.

The second model is the modification of the previous one. The dependent variable $Y^{(L2)} = t_0/\sqrt{t_1}$ is the ratio between the turnover and the squared root of the country area for an actual year. Explanatory factors are $x_1 = t_2$, $x_2 = t_3$, $x_3 = t_2/t_3$, $x_4 = t_4$, $x_5 = t_5$, $x_6 = t_6$. In addition we introduce into this model new gradational factor t_7, which is the index of the country area. Due to this factor we are able to divide countries by its areas. It is equal to 1 for relatively small countries (with areas less than or equal to 40 000 km^2), and it is equal to 0 for countries with areas larger than 40 000 km^2. For example, this index is equal to 1 for Belgium, Luxembourg and Austria, as the areas of these countries are less than 40 000 km^2.

Finally we consider two variants of the single index model (6.3). In the first variant the value of the dependent variable $Y^{(SIM1)} = t_0/t_1$ is the ratio between the turnover and the

country area for an actual year. In the second variant the dependent variable $Y^{(SIM2)} = t_0 / \sqrt{t_1}$ coincides with a dependent variable from the second linear model $L2$. The sets of explanatory factors for the models $SIM1$ and $SIM2$ coincide with the set for the first linear model $L1$.

Thus, we have four regression models. Our tasks are:

1) estimation of the unknown coefficients β for all the models;
2) calculating of the main statistical criteria for the models, i.e. multiple determination coefficient R^2, Fisher and Student criteria [120], [127];
3) comparing the suggested models and choosing the best ones taking in account their significance.

For estimation and comparing of stated four models we use procedures I and II, described in the previous Section of present Chapter. As we intend to investigate the efficiency of the suggested models not only in case of smoothing, but also in case of forecasting, the described above cross-validation approach is used as well. Especially for the single index model the series of experiments is carried out with the aim to determine the optimal value of bandwidth h. In the present work a lot of attention is paid to this problem.

Firstly, we analyse all the suggested models in case of data smoothing. It means we estimate the unknown coefficients β by all the observations. Thus, we are able to evaluate, how considered models can only smooth the known turnover and what factors have the greatest influence upon the turnover. Next Subsection is devoted especially to this problem.

6.2.3. ESTIMATION OF THE LINEAR MODELS

In this Subsection we present the results of estimation of the offered linear models in case of data smoothing.

The estimated model $L1$ can be written in the following form:

$$\hat{E}\left(Y^{(L1)}(x)\right) = -3\,713 + 118x_1 + 26x_2 - 11\,769x_3 + 879x_4 + 549x_5 + 158x_6 . \qquad (6.13)$$

The estimates of the coefficients and calculated values of the Student criterion for the model $L1$ are presented in Table 6.1. Here $\hat{\beta}_i$ is an estimate of β_i, $t(68)$ is the calculated value of Student criterion for 68 degrees of freedom, p-level is the error of second kind (or level of insignificance of variable). The theoretical value of Student criterion for 68 degrees of freedom and level of significance (or error of first kind) $\alpha = 5\%$ is equal to 1.99. Taking into

account the fact that the hypothesis of *insignificance* of explanatory variable is tested [120], [127], we can see that calculated value of Student criterion exceeds its theoretical value for two variables only, i.e. these two variables cannot be recognized as insignificant. The most significant explanatory factors are x_4 and x_6, so, the greatest influence on the turnover render the total length of railways and the number of good wagons. The signs of the coefficients for these variables correspond to their physical sense – as it was supposed, the signs by these estimated coefficients are positive.

Against expectation, GDP_PPS and CPL are insignificant factors, besides CPL is more insignificant. Notwithstanding estimated coefficients are positive. It means, the higher are the values of GDP_PPS and of CPL, the higher is the export of goods made by a country, and, as a consequence, the higher is the turnover of rail freight transport. Relation of GDP_PPS to CPL is insignificant as well, and estimated coefficient is negative. It means, if GDP_PPS exceeds CPL, the less is turnover, and vise versa.

The coefficient R^2 for this model is equal to 0.985 and the calculated value of Fisher criterion is 383.69. The theoretical value of Fisher criterion for 6 and 68 degrees of freedom and level of significance α = 5% is equal to 2.23. Comparing the theoretical and calculated values of Fisher criterion we can conclude that the estimated model $L1$ as a wole cannot be recognized as insignificant. So, model $L1$ is adequate.

The estimated model $L2$ is the following:

$$\hat{E}\left(Y^{(L2)}(x)\right) = -120.4 - 1.2x_1 + 1.4x_2 + 110.2x_3 + 0.2x_4 + 5.9x_5 + 0.3x_6 + 29.2x_7 . \quad (6.14)$$

Table 6.1

Results of model $L1$ estimation

Factors	$\hat{\beta}_i$	$t(68)$	p-level
Intercept	-3 713	-0.149195	0.881842
GDP_PPS	118	0.480762	0.632229
CPL	26	0.109604	0.913046
GDP_PPS/CPL	-11 769	-0.462115	0.645474
TOTLEN	879	6.866741	0.000000
LOKOM	549	0.799173	0.426973
WAGONS	158	8.375650	0.000000

The results of model *L*2 analysis are presented in the Table 6.2. As we can see, almost all explanatory variables are recognized to be significant by Student criterion. Only total length of railways doesn't influence the ratio between the turnover and the squared root of country area.

We obtain the positive signs for all significant variables except GDP; that means the positive correlation between these explanatory variables and the dependent variable. The coefficient by GDP is negative. It could be explained by many reasons. First of all, the positive sign by factor t_7 shows that the smaller country area the bigger transportations. These countries, Belgium, Luxembourg and Austria, have big value of GDP, but relatively small absolute transportations in comparison with other countries. We assume that such warp has been occurred exactly because of allocation of these countries in separate group. Notice that estimated coefficient by relation of GDP_PPS to CPL has the sign opposite to the sign by such relation in the previous model.

The coefficient R^2 for this model is equal to 0.985 and the calculated value of Fisher criterion is 313.78. The theoretical value of Fisher criterion for 7 and 67 degrees of freedom and level of significance $\alpha = 5\%$ is 2.15, so, this regression model is significant as well.

Figures 6.1 and 6.2 demonstrate how the investigated linear models smooth the observed true data. Here and further observations are arranged in "country-year" order: every five points correspond to turnover for some country during the analysed period from 1996 till 2000, i.e. for five years. Moreover, countries are sorted in alphabetical order. Horizontal axis reflects the number of observations, arranged in the above-mentioned order. Vertical axis reflects the corresponding turnover, expressed in thousands of tonne-km.

Table 6.2

Results of model *L*2 estimation

Factors	$\hat{\beta}_i$	*t*(67)	*p*-level
Intercept	-120.4	-3.00514	0.003732
GDP_PPS	-1.2	-3.11117	0.002738
CPL	1.4	3.55818	0.000692
GDP_PPS/CPL	110.2	2.68390	0.009160
TOTLEN	0.2	1.03172	0.305913
LOKOM	5.9	5.42836	0.000001
WAGONS	0.3	9.33665	0.000000
GradAREA	29.2	12.79621	0.000000

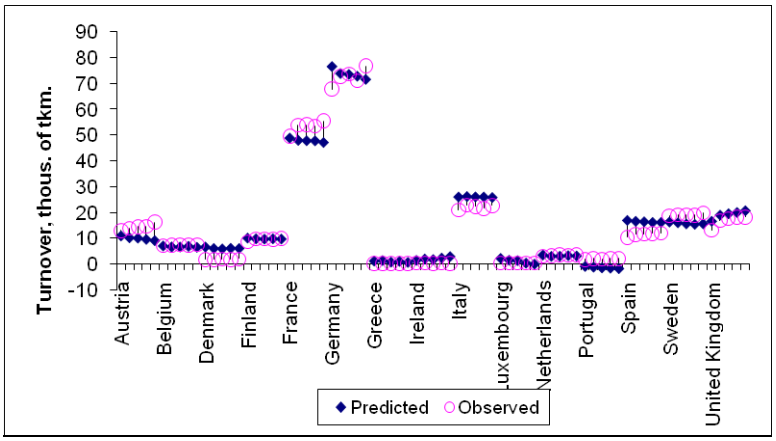

Fig.6.1. Smoothing by the model $L1$

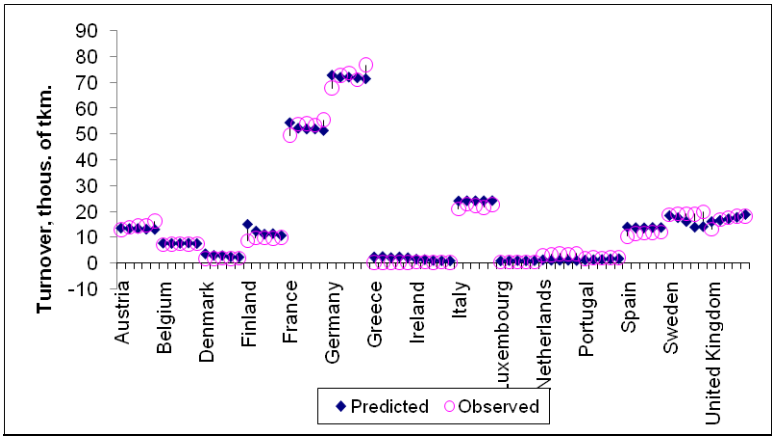

Fig.6.2. Smoothing by the model $L2$

It is obvious that both linear models show the similar smoothing. Next Subsection contains results of estimation of considered single index models in case of smoothing.

6.2.4. Estimation of the Single Index Models

Now we will consider suggested single index models $SIM1$ and $SIM2$.

The estimation of these models consists of two steps: we have to estimate the unknown coefficients vector β and the link function g. The corresponding estimation method and procedure of choosing of the most significant single index model are described in the previous Section 6.1.

The estimates of coefficients β , i.e. the values of coefficients β optimizing the object function (6.5), for both single index models have been obtained from the same starting point $\beta^{(0)}$ and with bandwidth $h = 7$ for $SIM1$ and $h = 6$ for $SIM2$. Note that these values of

bandwidth are optimal and have been obtained as a result of the series of experiments using our own program written in MathCad 13 environment.

The estimated model *SIM*1 and *SIM*2 can be rewritten in the following form:

$$\widehat{E}\left(Y^{(*)}(x)\right)=\frac{\sum_{i=1}^{n} Y_i K_h\left((x-x_i)^T\hat{\beta}\right)}{\sum_{i=1}^{n} K_h\left((x-x_i)^T\hat{\beta}\right)}, \qquad (6.15)$$

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

Figures 6.3 and 6.4 represent smoothing by these models. Obviously, the estimates of turnover almost in all observations coincide with the true turnover.

Is necessary to underline, the third and fourth models are nonlinear, in this connection the estimated coefficients can not be easy interpreted, like in the linear models.

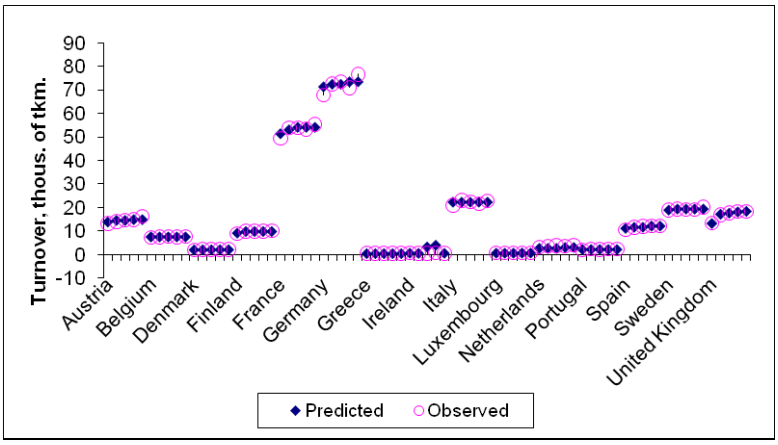

Fig.6.3. Smoothing by model *SIM*1

As both linear models and both single index models give approximately similar results in data smoothing, we have to consider how precise are the forecasts given by analysed models. For this purpose we use the residual sum of squares R_0 (6.10). Table 6.3 involves the values of the residual sums of squares for all the models.

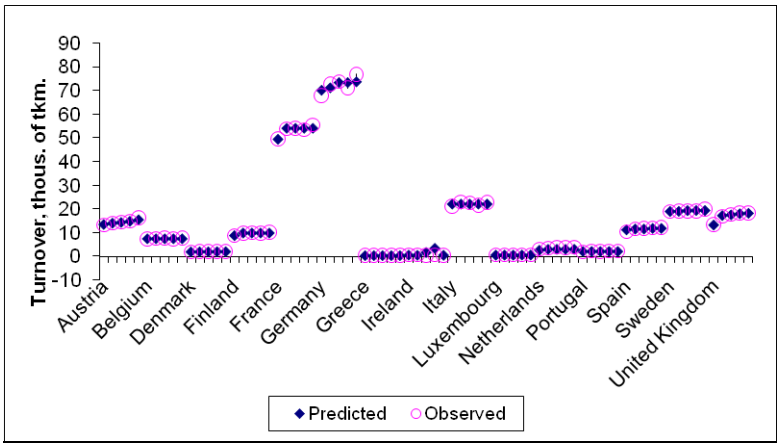

Fig.6.4. Smoothing by model *SIM*2

Table 6.3

Values of R_0 in case of smoothing

Model	*L*1	*L*2	*SIM*1	*SIM*2
R_0	11 543 065	4 830 576	894 265	565 407

So we can conclude that the linear model *L*2 and the single index model *SIM*2 have the minimum value of R_0 that means the greater significance of these models in comparison with two other. As it was supposed, in general SIM gives the most precise estimates of the volumes of the turnover.

6.2.5. CROSS-VALIDATION ANALYSIS

Now we will consider the suggested models from the other point of view. We use the *cross-validation approach* (see Section 6.1 of the current Chapter). That means we estimate the unknown coefficients β for the models on the basis of a part of the data. Then using the obtained estimates of β we forecast the volumes of the turnover for a remained part of the data. After that we compare these forecasted volumes of the turnover with the real ones, i.e. we calculate R_0 for each model. Also the optimal value of bandwidth h (see Chapter 3, Section 3.3.2) is found for both single index models.

We estimate the coefficients β on the basis of the period form 1996 till 1999 and perform the forecast for the 2000. Table 6.4 contains the estimates of β for considered linear regression models. The signs of estimates correspond to physical sense of explanatory factors.

Table 6.4

Estimates of coefficients for the linear models

Factors	Estimates of coefficients	
	$L1$	$L2$
Intercept	4 154.4	-119.7
GDP_PPS	197.7	28.7
CPL	45.4	-1.2
GDP_PPS/CPL	-20 555.8	1.4
TOTLEN	898.5	110.0
LOKOM	531.6	0.2
WAGONS	148.1	6.0
GradAREA	-	0.3

The residual sum of squares R_0 for model $L1$ is 18 509 464 and for model $L2$ is 8 941 875. Obviously, forecasts of the turnover of international rail freight transport obtained by the second linear model have to be much better than those obtained by the first one. Moreover, the first linear model gives negative forecasts of some small volumes of the turnover, in particular, for Portugal and Luxembourg. The true observed values of turnover and the corresponding forecasts are displayed on Figures 6.5 and 6.6. We can see that model $L2$ is more sensitive to the small volumes of the turnover, which belong to the countries with small areas. Obviously this effect is achieved by using the above-mentioned additional gradation factor GradAREA.

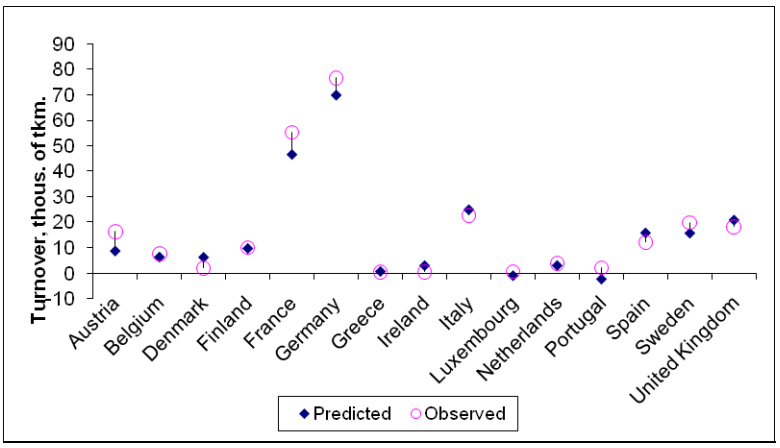

Fig.6.5. Forecasting by model $L1$

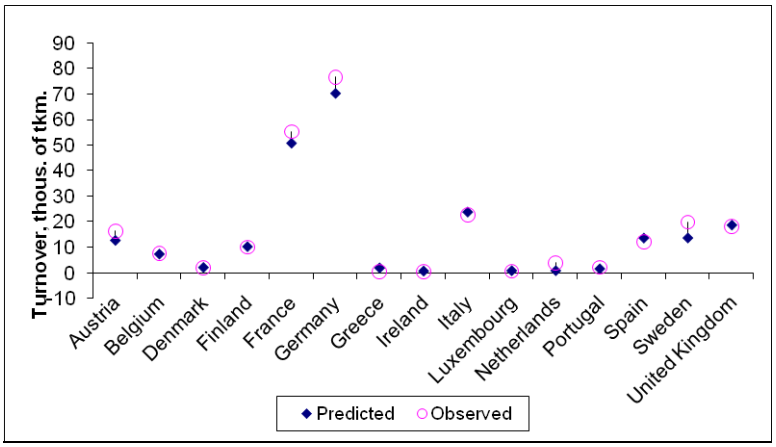

Fig.6.6. Forecasting by model *L*2

Now we will analyze SIM in detail. We begin with a choice of the bandwidth size. Our task is to find the optimal value of bandwidth h_0 that gives a minimal value of R_0. The series of experiments was performed and the different estimates of β and values of R_0 depending of various h were obtained as well. The corresponding results for models *SIM*1 and *SIM*2 are shown in Tables 6.5 and 6.6 respectively. We can see that all β estimates differ from each other depending on h in spite of the fact that they were obtained from the same initial value β_0.

The values of R_0 (expressed in millions) corresponding to various h for both SIM are represented in Table 6.7. Thus, the best result for R_0 is achieved for $h_0 = 7$ and $h_0 = 8$ for *SIM*1 and for $h_0 = 6$ for *SIM*2. As it was supposed the sum of squared residuals increases if h is bigger and smaller than the optimal value. The forecasted conveyances by *SIM*1 with $h_0 = 7$ and by SIM_2 with $h_0 = 6$ and observed conveyances are shown on the Figures 6.7 and 6.8 respectively.

Table 6.5

The estimates of β for *SIM*1

Factors	Bandwidth *h*								
	1	5	6	7	8	9	10	15	20
Intercept	22.8	566.4	248.6	33 160.0	344.5	721.2	3 530.0	1 327.0	-901.7
GDP_PPS	26.6	299.9	216.7	-22 420	-95.8	-24.1	-512.5	358.2	1 304.0
CPL	-0.04	1.9	-0.03	565.5	4.4	7.2	38.4	8.6	-19.7
GDP_PPS/CPL	0.13	257.9	88.9	19 870.0	120.5	207.1	1 310.0	550.0	1 011.0
TOTLEN	$4x10^{-5}$	62.9	17.9	3 996.0	25.3	37.7	174.4	119.2	198.3
LOKOM	$1x10^{-5}$	885.7	252.2	56 310.0	356.9	572.7	3 832.0	3 058.0	724.6

Table 6.6

The estimates of β for *SIM*2

Factors	Bandwidth *h*								
	1	5	6	7	8	9	10	15	20
Intercept	29.3	859.9	962.7	$1.7x10^3$	$1.2x10^3$	697.5	618.7	757.1	$-4.5x10^3$
GDP_PPS	18.3	462.3	$1.2x10^3$	$1.0x10^3$	654.8	578.5	276.8	791.3	$3.4x10^3$
CPL	0.1	3.4	-2.4	7.5	6.0	1.9	3.3	-0.2	-73.6
GDP_PPS/CPL	-0.2	440.7	604.3	$1.2x10^3$	636.6	664.1	525.1	737.1	4.0E+3
TOTLEN	$4x10^{-5}$	103.7	49.0	97.1	61.0	24.1	12.1	17.3	916.8
LOKOM	$1x10^{-5}$	$1.5x10^3$	690.0	$1.4x10^3$	859.7	851.7	672.9	$1.2x10^3$	$1.9x10^4$

Table 6.7

The values of R_0 for *SIMs*

	Bandwidth *h*								
	1	5	6	7	8	9	10	15	20
*SIM*1	676.9	24.3	2.0	1.9	1.9	2.5	24.3	24.3	44.5
*SIM*2	676.9	24.2	1.9	1.9	1.9	2.4	3.0	7.5	7.3

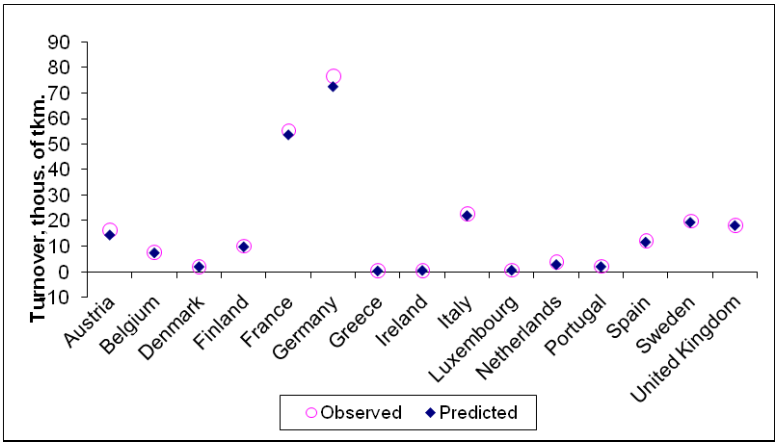

Fig.6.7. Forecasting by *SIM*1

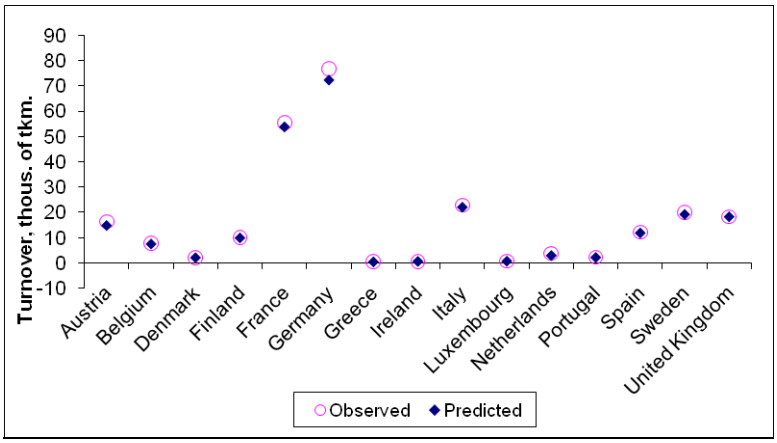

Fig.6.8. Forecasting by *SIM*2

Obviously, the forecasted values are very close to the observed values almost in all the observations. Table 6.8 contains the values of R_0 for four investigated models. As we can see, values of R_0 for single index models are in a number of orders less than for linear models. This fact gives evidence of greater accuracy of SIM models.

Table 6.8

The values of R_0 in case of forecasting

Model	*L*1	*L*2	*SIM*1	*SIM*2
R_0	18 509 464	8 941 875	1 894 237	1 896 287

From Figure 6.9 we can also visually evaluate behavior of R_0 with respect to bandwidth h for both single index models.

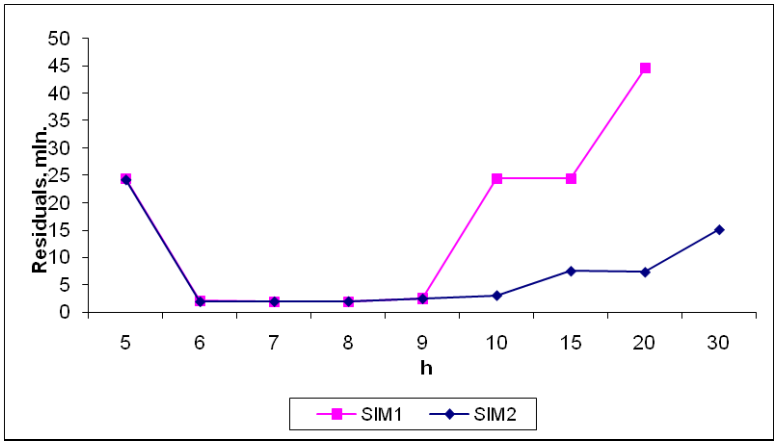

Fig.6.9. R_0 depending on bandwidth for *SIM*s

In fact, the meaning of used residual sum of squares R_0 (6.10) coincides with the meaning of MSE – mean squared error, see (3.37). MSE consists of squared bias and variance and depends on bandwidth. Besides, the bias in direct ratio depends on a bandwidth, and the variance depends from h in inverse proportion. Figure 3.1 clearly demonstrates that fact. So, the error R_0 also at first should decrease depending on a bandwidth, reaches some minimum value, and then should increase again. In Figure 6.9 we see that the criterion R_0 behaves in accuracy under this scenario. It proves that our experiments have been spent correctly and the received results are authentic.

CONCLUSION

There were considered two kinds of models for analysis and forecasting of turnover of international rail freight transport: linear regression model and single index regression model. Four different regression models were constructed and tested, two of them are linear regression models and two others are single index models. For the estimation of unknown coefficients in case of SIM the Nadaraya-Watson estimator and Gaussian kernel function were used (see details in Section 6.1). The efficiency of these models was investigated through the consideration of turnover for the 15 Member States of the European Union.

The advantage of considered models comparing with the models presented in such papers as [15], [19], [20], [23], [24], [57], [92] consists in including the greater number of the used factors. The performed investigations show that the single index regression model gives more precise forecasts than classical methods of linear regression. All the models have been estimated and compared by the criterion of the residual sum of squares R_0, which required the cross-validation approach. Moreover, the optimal values of smoothing parameter h for considered single index models have been obtained experimentally.

6.3. ANALYSIS AND FORECASTING OF THE INLAND RAIL PASSENGER TRANSPORTATIONS FROM THE REGIONS OF LATVIA

Comprehensive planning of transport company activity demands presence of set of the mathematical models that adequately describe functioning of the railway and application of various mathematical methods for forecasting of rail passenger transportations. The models considered in the present Section allow analyzing passenger flows and forecasting the demand for passenger transportation along some particular sectors of the railway on the whole territory of Latvia.

Principal aim of the research is to construct and improve the linear models and the single index models, and then to choose from them one, which gives the better forecasts of rail passenger transportations from the regions of Latvia.

The main problem consists of inconsistency of volumes of transportations. Transportations from the biggest cities exceed transportations from other cities and districts in times of orders. The reason can be that inhabitants of various cities and districts use the railway in the different purposes. For example, inhabitants of such big regions as Riga, Ogres rajons, Daugavpils rajons, and Jurmala, use railway transport (to be exact, electric trains), as usual municipal transport. At the same time, inhabitants of such cities as Ventspils or Gulbene, use railway transport for long-distance trips, for example, in the capital. In this connection the main problem of the presented research is to construct the group parametric and semi-parametric models adequately and equally well describing both large and small transportations.

For that purpose many researches have been spent [62], [63]. Not all of them were equally successful. First of all, it is necessary to choose the set of regressors, significantly describing both big and small transportations. In our numerous researches observations have been divided onto two groups, i.e. depending on transportation size. In this connection model describing only small transportations has been developed [63]. Many attempts to exclude outliers from the statistical sample with respect to different reasons have been carried out as well. Finally the best results have been achieved in case of outliers excluded according to Mahalanobis distance [87].

Various tests for hypothesis of explanatory variables insignificance and model correctness have been led, and the cross-validation approach has been carried out as well in all the considered experiments (see Chapter 6, Section 6.1). We have used the residual sum of squares R_0 for comparing the elaborated parametric and semiparametric models [95], (Section 6.1). Especially for the single index model the series of experiments is carried out with the aim to determine the optimal value of bandwidth h [14], [53], [85]. The analysis has shown obvious preference of the single index model.

6.3.1. PROBLEM SETTING

In the course of the suggested research the linear regression models and the single index models have been developed and investigated. Theoretical description of the used models is given in Chapter 3, Sections 3.2 and 3.3. 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.

In the present research all investigated models are group models (see Chapter 3, Section 3.1), [121], i.e. we forecast transportations for all the considered regions using the same sets of the explanatory variables and corresponding statistical data.

The task of research is to draft the best regression models, i.e. models adequately and equally well describing both large and small transportations.

So, we consider volumes of internal rail passenger transportations (see definition in Chapter 2), expressed in thousands of passenger, for the regions of Latvia. The following 33 regions had been selected for the experiment: Aizkraukles rajons, Aluksnes rajons, Balvu rajons, Bauskas rajons, Cesu rajons, Daugavpils, Daugavpils rajons, Dobeles rajons, Gulbenes rajons, Jekabpils rajons, Jelgava, Jelgavas rajons, Kraslavas rajons, Kuldigas rajons, Liepaja, Liepajas rajons, Limbazu rajons, Ludzas rajons, Madonas rajons, Ogres rajons, Preilu rajons, Rezekne, Rezeknes rajons, Riga, Rigas rajons, Jurmala, Saldus rajons, Talsu rajons, Tukuma rajons, Valkas rajons, Valmieras rajons, Ventspils, Ventspils rajons.

The analyzed time period is from 2000 to 2003. However, not every region has observed volume of transportation and volumes of influencing factors for each considered year. So, for 33 regions and 4 years we have 113 observations.

The variable of interest is the inland rail passenger transportations, expressed in thousands of passengers. Let us denote it by t_0. The considered explanatory factors, or predictors, taken from generated database, described in Section 2.3, are:

t_1 – the population density, in thousands of residents per a unit of territory, *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 of uneployed residents per a unit of territory, *UP*/*SQUARE*;

t_5 – the number of general education institution per a unit of territory, *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*.

In the present Section we intend to describe almost all the carried out experiments, containing outliers removing and statistical data splitting.

Now we describe the investigated regression models. The first model is the simple linear regression model (3.2) and the second model is the single index model (6.2). The dependent variables $Y^{(1)}$ in the linear model and $Y^{(2)}$ in single index model are inland rail passenger transportations, in thousands of passengers. Explanatory variables in both models are all eight mentioned above.

So, we have two regression models and our task is to estimate the unknown coefficients β for the models, to compare the suggested models and to demonstrate the preference of semiparametric model.

Many different experiments have been performed; results of four from them are shown in the next sub-Sections. Each experiment was carrying out on the base of common scheme, which is clearly discernible in Section 6.2. In other words, each experiment contains estimation of both models in case of smoothing, probable improvements of the models, after that both models are estimated in case of forecasting using described in Section 6.1 cross-validation approach. We intend to show only results without surplus comments.

6.3.2. RESULTS OF EXPERIMENTS FOR FULL DATA

In the framework of the first experiment all 113 observations according to all 33 regions have been analyzed. Perhaps, there are outliers among them. Besides some regions have observations on all the considered period, others don't have.

Smoothing

The estimates of the coefficients and values of the Student criterion for the linear model are resulted in Table 6.9. R_0 for this model is 861 202, coefficient R^2 is equal to 0.95 and the Fisher criterion is 224, so, this model is adequate.

Thus, variables UP/SQUARE, NGEI/SQUARE and NB/SQUARE are less significant than others by Student criterion, and then we use the *Backward Stepwise* mode of Statistica 6.0 package [123], [125], which allows us sequentially including only most significant predictors in the considering linear model. At last, analyzing model contains only five predictors.

Table 6.9

Results of the linear model estimation

Factors	$\hat{\beta}_i$	$t(104)$	p-level
Intercept	-271.50	-0.48300	0.630113
TP/1000*SQUARE	-7.84	-3.48196	0.000729
NE/SQUARE	332.72	13.91136	0.000000
NE/1000*TP	170.76	4.00629	0.000116
UP/SQUARE	0.15	0.05058	0.959759
NGEI/SQUARE	16 823.21	1.58356	0.116330
NB/SQUARE	-126.54	-0.88860	0.376268
NB/1000*TP	-464.56	-3.30525	0.001303
NRS	143.59	9.54924	0.000000

In Table 6.10 the results of this modification of the considered linear model are shown. R_0 for this model is 884 141, coefficient R^2 is equal to 0.94 and the Fisher criterion is 358 and all the predictors are significant by the Student criterion, so, this modification of the considered model is adequate and even more significant in sense of the Fisher criterion than the full variant with 8 predictors (see Table 6.9). Equation for the best chosen linear model can be written as follow:

$$\hat{E}\left(Y^{(1)}(x)\right) = -75 - 4x_1 + 323x_2 + 168x_3 - 476x_7 + 146x_8. \qquad (6.16)$$

We can see the signs by the predictors in both variants of the investigated linear regression model are the same. It testifies to obvious and steady enough influence of the chosen factors on inland rail transportations on regions of Latvia.

Table 6.10

Results of the linear model estimation after modification

Factors	$\hat{\beta}_i$	$t(107)$	p-level
Intercept	-75.48	-0.17186	0.863871
TP/1000*SQUARE	-4.17	-9.40339	0.000000
NE/SQUARE	322.65	14.79656	0.000000
NE/1000*TP	168.42	4.54257	0.000015
NB/1000*TP	-475.56	-5.37702	0.000000
NRS	145.71	9.82005	0.000000

Estimation of coefficients β , i.e. the values of coefficients β optimizing the object function (6.5), for the single index model has carried out with different bandwidths. Note that in the single index model only the most significant predictors are included, i.e. TP/1000*SQUARE, NE/SQUARE, NE/1000*TP, NB/1000*TP and NRS.

Table 6.11 contains the estimates of unknown coefficients, calculated with different bandwidths. As we can see, the signs by one and the same predictors are different. Table 6.12 contains the R_0 for this model depending on h. We can conclude, that the best results in sense of minimizing R_0 can be obtained with h = 1. Taking into account the different signs by predictors and the similar values of R_0, we are able to choose the more appropriate variant of regression.

Table 6.11

Results of the single index model estimation

Factor	Bandwidth h			
	1	30	50	100
TP/1000*SQUARE	9 183	41 650	71 800	444 300
NE/SQUARE	-395	10 930	581	-7 259
NE/1000*TP	7 375	34 100	11 280	17 790
NB/1000*TP	-3 523	-3 763	-425	-2 048
NRS	0.1	100 800	53 080	117 700

Table 6.12

The values of R_0 for the SIM

Bandwidth h			
1	30	50	100
133 556	747 012	1 644 348	1 926 936

So, the best chosen single index model corresponds to h = 1 and can be presented in the form (6.15), where vector of estimated coefficients is $\hat{\beta}^T = (9183 \quad -395 \quad 7375 \quad -3523 \quad 0.1)$. We can conclude the chosen single index model in case of smoothing gives more precise results than the linear model in sense of R_0.

Cross-validation approach

We estimate the unknown coefficients β for the models on the base of period from 2000 till 2002. Then using the obtained estimates of β we forecast the transportations for the

2003 and compare these forecasted transportations with true ones, i.e. we calculate R_0 for both models. The optimum value of bandwidth h is found for the single index model as well.

The Table 6.13 contains the estimates of β for considered linear regression model. Note that for cross-validation approach only the most significant variables are included in the model. The signs of estimates coincide with signs for the case of smoothing (see Table 6.10) and correspond to physical sense of explanatory factors.

Table 6.13

Results of the linear model estimation

Factors	$\hat{\beta}_i$	$t(84)$	p-level
Intercept	-166.78	-0.33377	0.739387
TP/1000*SQUARE	-3.90	-7.63842	0.000000
NE/SQUARE	311.36	12.28271	0.000000
NE/1000*TP	143.53	3.47051	0.000823
NB/1000*TP	-406.11	-3.99505	0.000138
NRS	139.99	8.41602	0.000000

Coefficient R^2 is equal to 0.94 and the Fisher criterion is 245, so, the investigated linear model is adequate in the case of cross-validation as well. The residual sum of squares is 1 181 086. Unfortunately, this model gives negative forecasts for seven cases, i.e. about in 30% of observations. The true observed values of transportations and the corresponding forecasts are displayed at Figure 6.10.

If we construct the similar plots separately for regions with large transportations (i.e. for Riga, Jurmala, Rigas region and Ogres region, see Figure 6.11) and for regions with small transportations, i.e. for all except Riga, Jurmala, Rigas region and Ogres region (see Figure 6.12), we shall visually evaluate in the more convenient scale, how the investigated linear model predicts the transportations. Further it will be shown, that worst forecasts, even negative ones, correspond to the observations with biggest values of Mahalanobis distance.

Now we analyze the single index model in details. We begin with a choice of the bandwidth size. Our task is to find the optimal value of bandwidth h_0 that gives a minimal value of R_0. The series of experiments was performed and the different estimates of β and values of R_0 depending of various h were obtained as well.

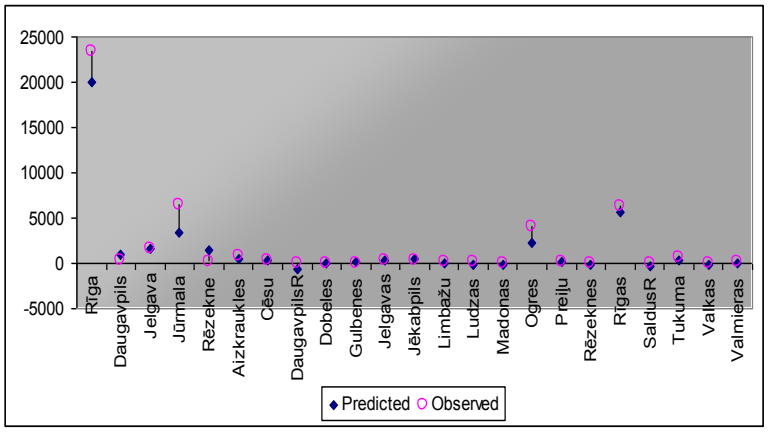

Fig.6.10. Forecasting by the linear model

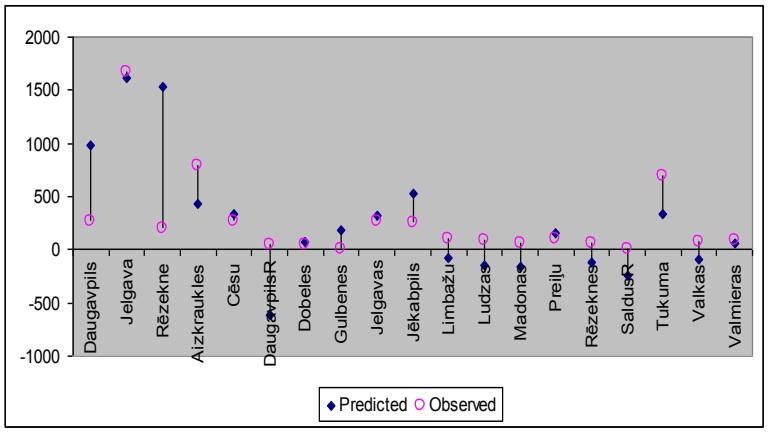

Fig.6.11. Forecasting of small transportations by the linear model

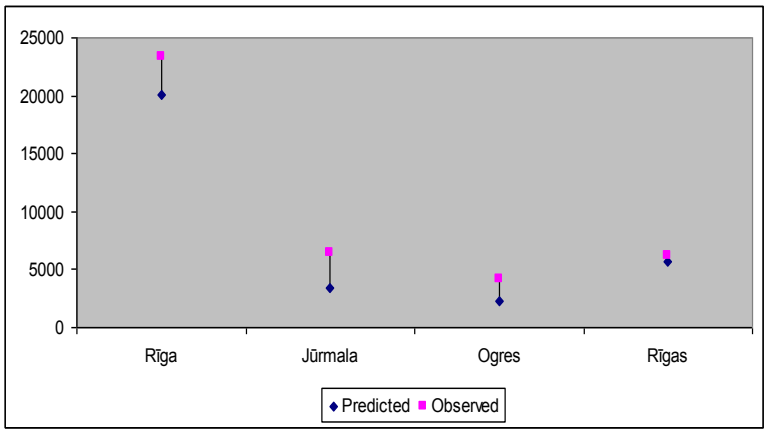

Fig.6.12. Forecasting of large transportations by the linear model

The corresponding results for the analysed single index model are shown in Table 6.14. We can see all the estimates of β differs from each other depending on h in spite of

they were obtained for the same start value of β. The values of R_0 corresponding to various h are resulted in Table 6.15. Thus, the best result for R_0 is achieved for h = 30. As it was supposed the residual sum of squares increases if h is bigger and smaller than optimal value of h. The forecasted transportations by the single index model with h_0 = 30 and observed transportations as well are shown on the Figure 6.13.

Table 6.14

The estimates of β for the SIM

Factors	Bandwidth h			
	1	30	50	100
TP/1000*SQUARE	971	134 500	276 900	86 910
NE/SQUARE	-99	3 975	3 992	-1 657
NE/1000*TP	118	8 273	41 770	3 356
NB/1000*TP	-396	-1 287	355	-981
NRS	0.1	24 270	201 400	24 920

Table 6.15

The values of R_0 for the SIM

Bandwidth h			
1	30	50	100
25 610 975	466 016	466 062	3 043 535

Obviously, the forecasted values are very close to the observed values almost in all the observations. Comparing represented on the Figures 6.10 and 6.13 forecasts, we can conclude, the single index model produces no negative forecasts. Moreover, the SIM shows more precise forecasts for regions with large transportations that the linear one.

We are able again to construct plots for regions with large transportations (see Figure 6.14) and for regions with small transportations (see Figure 6.15). Thus, we can see in the more convenient scale, the single index model predicts the transportations more efficient.

We have collected the values of R_0 for both investigated models in cases of smoothing and cross-validation in Table 6.16. As we can see, the R_0 values are less for SIM in all the considered cases.

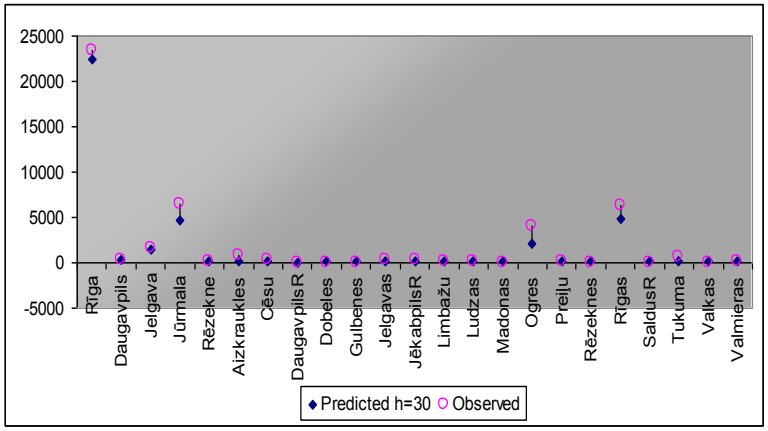

Fig.6.13. Forecasting by the single index model

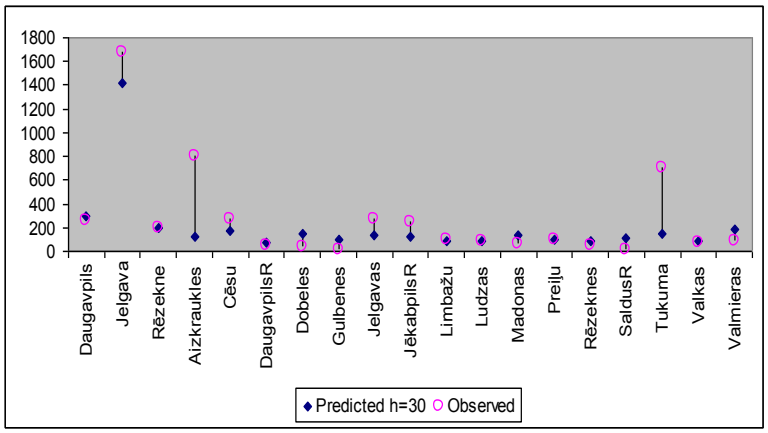

Fig.6.14. Forecasting of small transportations by the single index model

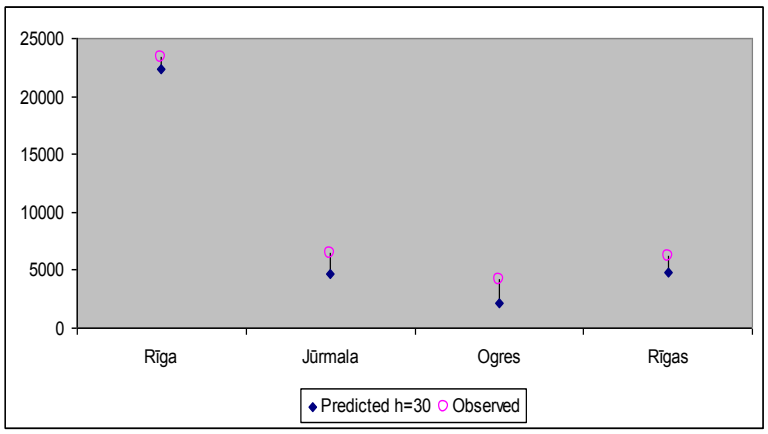

Fig.6.15. Forecasting of large transportations by the single index model

Table 6.17 contains the observed and forecasted transportations obtained in case of cross-validation by both investigated models for the analyzing period (i.e. 2003) and the relative error in % for each observation δ_i, calculated as following:

$$\delta_i = \frac{Y_i - \hat{Y}_i}{Y_i} \cdot 100\%. \tag{6.17}$$

Table 6.16

The values of R_0

	Smoothing	Cross-Validation
Linear Model	884 141	1 181 086
SIM	133 556	466 016

Table 6.17

Comparative results in case of cross-validation

	Observed values	Forecasts		Residuals		Relative error in %	
		Linear Model	SIM	Linear Model	SIM	Linear Model	SIM
Riga	23323.05	20067.43	22370.00	3255.62	953.05	13.96	4.09
Daugavpils	259.85	979.57	301.83	-719.73	-41.98	-276.98	-16.16
Jelgava	1676.11	1611.31	1412.00	64.80	264.11	3.87	15.76
Jurmala	6436.22	3360.92	4682.00	3075.30	1754.22	47.78	27.26
Rezekne	192.50	1532.21	193.77	-1339.71	-1.27	-695.95	-0.66
Aizkraukles r.	795.41	427.85	120.98	367.56	674.43	46.21	84.79
Cesu r.	272.15	332.88	174.91	-60.72	97.25	-22.31	35.73
Daugavpils r.	47.71	-618.34	72.59	666.05	-24.88	1395.97	-52.14
Dobeles r.	43.10	78.08	142.96	-34.98	-99.86	-81.17	-231.72
Gulbenes r.	10.25	183.36	94.47	-173.11	-84.22	-1689.53	-822.01
Jelgavas r.	267.40	315.14	132.57	-47.74	134.83	-17.85	50.42
Jekabpils r.	251.74	522.49	126.90	-270.75	124.84	-107.55	49.59
Limbazu r.	96.13	-71.44	85.70	167.57	10.43	174.32	10.85
Ludzas r.	90.78	-152.40	84.19	243.18	6.59	267.88	7.26
Madonas r.	66.30	-165.41	130.82	231.70	-64.52	349.50	-97.33
Ogres r.	4108.32	2269.00	2149.00	1839.32	1959.32	44.77	47.69
Preilu r.	104.37	155.95	100.04	-51.59	4.33	-49.43	4.14
Rezeknes r.	53.34	-119.31	81.32	172.65	-27.98	323.67	-52.46
Rigas r.	6219.64	5660.02	4813.00	559.62	1406.64	9.00	22.62
Saldus r.	9.36	-249.69	115.56	259.05	-106.20	2768.52	-1134.99
Tukuma r.	698.92	332.84	149.75	366.08	549.17	52.38	78.57
Valkas r.	74.60	-97.24	89.15	171.84	-14.55	230.34	-19.50
Valmieras r.	86.74	62.98	185.40	23.76	-98.66	27.39	-113.74

Thus, the single index model has given more precise forecasts for the biggest transportations, i.e. for Riga and Jurmala, and less precise forecasts for Rigas and Ogres regions, than the linear one. R_0 for SIM in 2.5 times is less, than for linear one. There is a further point to be made here, that linear model in 30 % of observations gives negative forecasts. These are regions with small transportations, further we will see that to these regions corresponds the biggest values of Mahalanobis distance. After all, semiparametric approach has given better results than classical parametric approach in cases of smoothing and forecasting.

6.3.3. RESULTS OF EXPERIMENTS FOR RESTRICTED DATA

In the framework of the second experiment 88 observations according to 22 regions have been analyzed. We would like to underline, that only regions on which observations over all 4 considered years present are taken.

Smoothing

The estimates of the coefficients and calculated values of the Student criterion for the linear model are presented in Table 6.18. The theoretical value of Student criterion for 79 degrees of freedom and level of significance (or error of first kind) $\alpha = 5\%$ is equal to 1.99. R_0 for this model is 714 291, coefficient R^2 is equal to 0.96 and the calculated value of Fisher criterion is 259. The theoretical value of Fisher criterion for 8 and 79 degrees of freedom and level of significance $\alpha = 5\%$ is 2.06. Comparing the theoretical and calculated values of Fisher criterion we can conclude that the estimated model cannot be recognized as insignificant.

Thus, variables UP/SQUARE and NB/1000*TP are less significant than others by Student criterion. At last, analyzing model contains only six predictors.

In Table 6.19 the results of this modification of the considered linear model are shown. R_0 for this model is 823 311, coefficient R^2 is equal to 0.95 and the Fisher criterion is 306, and all the predictors are significant by the Student criterion. So, this modification of the considered model is adequate and even more significant in sense of the Fisher criterion than the full variant with 8 predictors (see Table 6.18). Equation for the linear model can be written as follow:

$$\hat{E}\left(Y^{(1)}(x)\right) = -2094 - 19x_1 + 389x_2 + 83x_3 + 75995x_5 - 886x_6 + 155x_8. \qquad (6.18)$$

Table 6.18

Results for the linear model estimation

Factors	$\hat{\beta}_i$	$t(79)$	p-level
Intercept	-1002.5	-1.74985	0.084027
TP/1000*SQUARE	-23.7	-4.76075	0.000009
NE/SQUARE	525.5	5.85782	0.000000
NE/1000*TP	164.2	3.63926	0.000486
UP/SQUARE	81.0	1.85968	0.066653
NGEI/SQUARE	67788.5	3.98316	0.000150
NB/SQUARE	-937.5	-3.26015	0.001644
NB/1000*TP	-391.4	-2.85571	0.005484
NRS	149.8	9.96456	0.000000

Table 6.19

Results of the linear model estimation after modification

Factors	$\hat{\beta}_i$	$t(81)$	p-level
Intercept	-2093.9	-5.33258	0.000001
TP/1000*SQUARE	-19.3	-5.43468	0.000001
NE/SQUARE	389.3	15.91297	0.000000
NE/1000*TP	82.5	2.11533	0.037475
NGEI/SQUARE	75994.9	4.59043	0.000016
NB/SQUARE	-886.2	-5.66157	0.000000
NRS	155.0	10.03094	0.000000

Table 6.20 contains results of SIM estimation depending on different bandwidths. Table 6.21 contains the R_0 for this model depending on h.

Table 6.20

Results of the SIM estimation

Factors	Bandwidth h				
	0.5	1	2.5	5	10
TP/1000*SQUARE	646	848	382	12 470	222 400
NE/SQUARE	-82	-73	52	495	4 379
NE/1000*TP	-159	226	1 294	3 195	19 140
NGEI/SQUARE	0.1	0.1	0.25	-0.04	-0.8
NB/SQUARE	-165	-188	-21	1 533	23 260
NRS	0.1	0.1	19.5	841	15 540

Table 6.21

The values of R_0 for the SIM

Bandwidth h				
0.5	1	2.5	5	10
166 373	171 435	202 523	277 330	328 141

So, the best chosen single index model corresponds to $h = 1$ and can be presented in the form (6.15), where vector of estimated coefficients is $\hat{\beta}^T = (848 \quad -73 \quad 226 \quad -0.1 \quad -188 \quad 0.1)$.

Cross-validation approach

Time period used for coefficient estimation is from 2000 till 2002, forecasting has been led for 2003 year.

Table 6.22 contains the estimates of β for considered linear regression model. Coefficient R^2 is equal to 0.96 and the Fisher criterion is 225, so, the investigated linear model is adequate in the case of cross-validation as well. The residual sum of squares R_0 is 1335775.

The corresponding results for the analysed single index model are shown in Table 6.23. The values of R_0 corresponding to various h are resulted in Table 6.24.

Table 6.22

Results of the linear model estimation

Factors	$\hat{\beta}_i$	$t(59)$	p-level
Intercept	-2135.1	-4.84245	0.000010
TP/1000*SQUARE	-21.4	-4.91007	0.000008
NE/SQUARE	380.5	13.58575	0.000000
NE/1000*TP	74.7	1.73310	0.088303
NGEI/SQUARE	86962.6	4.24508	0.000078
NB/SQUARE	-928.0	-5.00188	0.000005
NRS	145.8	8.58770	0.000000

Table 6.23

The estimates of β for the SIM

Factors	Bandwidth h		
	20	30	35
TP/1000*SQUARE	4 132	21 620	163 100
NE/SQUARE	63	235	-1 667
NE/1000*TP	518	2 216	16 530
NGEI/SQUARE	0.1	3	6
NB/SQUARE	35	132	815
NRS	3 029	7 522	61 320

Table 6.24

The values of R_0 for the SIM

Bandwidth h		
20	30	35
25 590 350	909 120	1 017 500

As we can see, the best SIM corresponds to $h = 30$.

We have collected the values of R_0 for both investigated models in cases of smoothing and cross-validation in Table 6.25. As we can see, the R_0 values are less for SIM in all the considered cases.

Table 6.25

The values of R_0

	Smoothing	Cross-Validation
Linear Model	823 311	1 335 774
SIM	171 435	909 120

We calculate R_0 separately for regions with large transportations (i.e. for Riga, Jurmala, Rigas region and Ogres region) and for regions with small transportations, i.e. for all regions except Riga, Jurmala, Rigas region and Ogres regions (Table 6.26).

Table 6.26

Separately calculated R_0

Model	Part of R_0	
	Large transp.	Small transp.
SIM	4 699 152	66 900
Linear Model	3 334 732	891 562

Table 6.27 contains the observed and forecasted transportations obtained in case of cross-validation by both investigated models for the analyzing period (i.e. 2003) and the relative error.

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, as it does not give negative forecasts at all (see Table 6.27). Thus, the single index model has given more precise forecasts for the largest transportations, i.e. for Riga and Ogres region, and less precise forecasts for Rigas region and Jurmala, than the linear one.

Table 6.27

Comparative results in case of cross-validation

	Observed values	Forecasts		Residuals		Relative error	
		Linear Model	SIM	Linear Model	SIM	Linear Model	SIM
Aizkraukles r.	795.41	296.64	338.19	498.77	457.22	0.63	0.57
Cesu r.	272.15	771.72	338.60	-499.56	-66.45	-1.84	-0.24
Daugavpils	259.85	999.52	277.47	-739.67	-17.63	-2.85	-0.07
Daugavpils r.	47.71	-795.07	337.52	842.78	-289.80	17.66	-6.07
Dobeles r.	43.10	314.98	338.47	-271.88	-295.38	-6.31	-6.85
Gulbenes r.	10.25	-81.50	337.85	91.74	-327.61	8.95	-31.97
Jekabpils r.	251.74	309.10	338.29	-57.36	-86.54	-0.23	-0.34
Jelgavas r.	267.40	325.13	338.40	-57.73	-71.00	-0.22	-0.27
Jurmala	6436.22	4865.52	4568.00	1570.70	1868.22	0.24	0.29
Limbazu r.	96.13	-594.46	337.73	690.59	-241.60	7.18	-2.51
Ludzas r.	90.78	-163.73	337.68	254.51	-246.90	2.80	-2.72
Madonas r.	66.30	499.63	338.17	-433.33	-271.87	-6.54	-4.10
Ogres r.	4108.32	1898.51	587.40	2209.81	3520.92	0.54	0.86
Preilu r.	104.37	-174.73	338.00	279.10	-233.63	2.67	-2.24
Rezekne	192.50	3769.23	193.77	-3576.73	-1.27	-18.58	-0.01
Rezeknes r.	53.34	-158.76	337.67	212.11	-284.33	3.98	-5.33
Riga	23323.05	21161.15	22370.00	2161.90	953.05	0.09	0.04
Rigas r.	6219.64	5072.99	4805.00	1146.65	1414.64	0.18	0.23
Saldus r.	9.36	-153.85	338.08	163.21	-328.72	17.44	-35.13
Tukuma r.	698.92	102.14	338.52	596.78	360.40	0.85	0.52
Valkas r.	74.60	-343.48	337.74	418.08	-263.14	5.60	-3.53
Valmieras r.	86.74	108.26	338.70	-21.52	-251.96	-0.25	-2.90

The residual sum of squares R_0 for SIM in 1.47 times is less, than for linear one. The received results show necessity of construction of separate models for forecasting of the large and small transportations.

6.3.4. RESULTS OF EXPERIMENTS FOR SLIGHT FLOWS

In this experiment 95 observations on 28 regions are analyzed. Observations with small transportations are considered only.

Smoothing

In the process of carrying out calculations the number of prediction variables was decreased from 8 to 5. At last, only five most significant predictors are included in this model, such as NE/SQUARE, NE/1000*TP, UP/SQUARE, NGEI/SQUARE and NRS. R_0 for this

model is 42 430, coefficient R^2 is equal to 0.48 and the calculated value of Fisher criterion is 16.58. The theoretical value of Fisher criterion for 5 and 89 degrees of freedom and level of significance $\alpha = 5\%$ is 2.32. So, this model can be recognized as adequate.

Equation for the estimated linear model is such:

$$\hat{E}\left(Y^{(1)}(x)\right) = -255 - 82x_2 + 16x_3 - 34x_4 + 13033\,x_5 + 18x_8. \qquad (6.19)$$

Now we discuss results of estimation of investigated single index model. The best chosen single index model with $h = 0.5$ and $R_0 = 8\,877$ can be written as (6.15), where vector of estimated coefficients $\hat{\beta}^T = (12 \quad 265 \quad 78 \quad 0.3 \quad 0.1)$.

Cross-validation

Time period used for coefficient estimation is from 2000 till 2002, forecasting has been led for 2003 year.

Table 6.28 contains the estimates of β for considered linear regression model. Coefficient R^2 is equal to 0.42 and the Fisher criterion is 10.29, the residual sum of squares is 79 029. So, the investigated linear model is adequate in the case of cross-validation as well. The observed transportations and the corresponding forecasts for the tested period are displayed at Figure 6.16.

The corresponding results for the analysed single index model are shown in Table 6.29. The values of R_0 corresponding to various h are resulted in Table 6.30. The forecasted transportations by the single index model with $h_0 = 2$ and observed transportations as well are shown on the Figure 6.17.

Table 6.28

Results of the linear model estimation

Factors	$\hat{\beta}_i$	$t(70)$	p-level
Intercept	-223.30	-1.83	0.071314
NE/SQUARE	-74.35	-2.52	0.014008
NE/1000*TP	14.24	1.51	0.135366
UP/SQUARE	-27.89	-6.13	0.000000
NGEI/SQUARE	11258.10	5.05	0.000003
NRS	15.57	1.64	0.104623

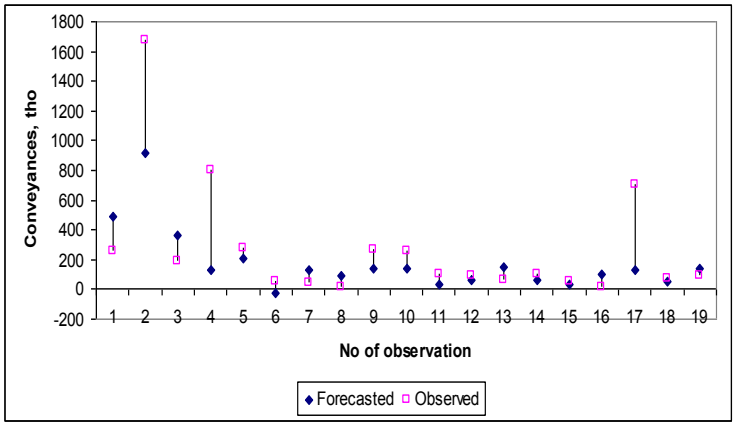

Fig.6.16. Forecasting by the linear model

Table 6.29

The estimates of β for the SIM

Factors	Bandwidth h				
	1	2	3	5	10
NE/SQUARE	-6	-38	121	2 462	6 253
NE/1000*TP	330	2 555	3 177	7 730	-15 700
UP/SQUARE	12	180	148	35 180	62 860
NGEI/SQUARE	0.2	0.05	0.7	3	619
NRS	0.1	1 561	3 110	13 380	25 750

Table 6.30

The values of R_0 for the SIM

Bandwidth h				
1	2	3	5	10
166 415	40 992	50 454	46 194	79 453

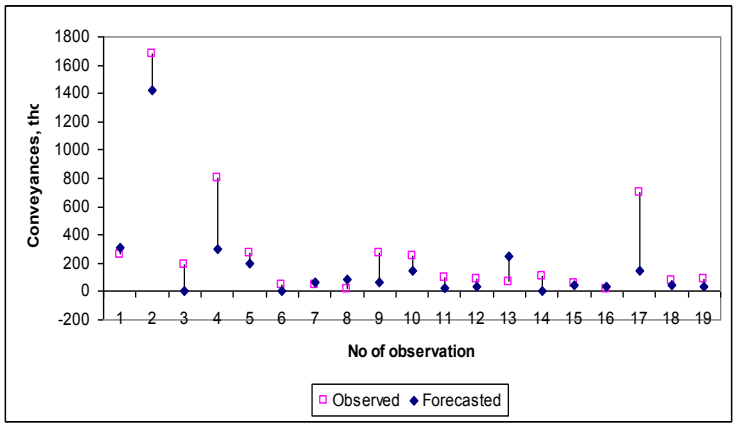

Fig.6.17. Forecasting by the single index model

Obviously, the forecasted values are very close to the observed values almost in all the observations. Moreover, the SIM shows more precise forecasts for regions with large transportations that the linear one.

The values of R_0 for both investigated models in cases of smoothing and cross-validation analysis are collected in Table 6.31. As we can see, the R_0 values are less for SIM in all the considered cases.

Table 6.31

The values of R_0

	Smoothing	Cross-Validation
Linear Model	42 430	79 029
SIM	8 877	40 992

In the present experiment regression models for the analysis and forecasting of small transportations are shown. The following Sub-section is devoted to improvement of considered during previous experiments models for the purpose of forecasts efficiency increasing. Outliers will be removed on the basis of Mahalanobis distance.

6.3.5. REMOVAL OF OUTLIERS CORRESPONDING TO MAHALANOBIS DISTANCE

In this experiment 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 from the statistical sample according to Mahalanobis distance.

Results after improvements are illustrated and comparative analysis of models is shown.

Smoothing

The estimates of the coefficients and calculated values of the Student criterion for the linear model are presented in Table 6.32. The theoretical value of Student criterion for 82 degrees of freedom and level of significance (or error of first kind) $\alpha = 5\%$ is equal to 1.99. Calculated value of Student criterion exceeds its theoretical value for all variables exclude UP/SQUARE, i.e. all these variables cannot be recognized as insignificant. Data in the table are arranged in order of decreasing of variables significance. R_0 for this model is 801 111, coefficient R^2 is equal to 0.96 and the calculated value of Fisher criterion is 258. The theoretical value of Fisher criterion for 8 and 82 degrees of freedom and level of significance $\alpha = 5\%$ is 2.05. Comparing the theoretical and calculated values of Fisher criterion we can

conclude that the estimated model cannot be recognized as insignificant. So, this model is adequate. Equation for the linear model can be written in the following way:

$$\hat{E}\left(Y^{(1)}(x)\right) = -20x_1 + 436x_2 + 163x_3 + 64418x_5 - 734x_6 - 398x_7 + 147x_8. \qquad (6.20)$$

Table 6.32

Results of the linear model estimation

Factors	$\hat{\beta}_i$	$t(82)$	p-level
NRS	146.8	9.8454	0.0000
NE/SQUARE	435.9	7.2432	0.0000
TP/1000*SQUARE	-20.1	-4.6139	0.0000
NGEI/SQUARE	64417.9	3.8097	0.0003
NE/1000*TP	162.6	3.5982	0.0005
NB/1000*TP	-397.6	-2.9033	0.0047
NB/SQUARE	-733.6	-2.8749	0.0051
Intercept	-923.9	-1.6149	0.1102
UP/SQUARE	36.9	1.2843	0.2026

Now the gradation variable GRAD is introduced. It takes value 1 for regions which have obvious large transportations, i.e. many times mentioned above Riga, Rigas rajons, Jurmala and Ogres rajons, and takes value 0 for others. The following results were obtained: R_0 = 546 895; R^2 = 0.97; F = 339.15; $F(9, 81)$ = 1.99. Table 6.33 contains results of model estimation.

Table 6.33

Results of the linear model estimation after modification

Factors	$\hat{\beta}_i$	t(81)	p-level
NE/SQUARE	618.01	10.37811	0.000000
GRAD	2742.63	6.16807	0.000000
TP/1000*SQUARE	-20.39	-5.66098	0.000000
NB/SQUARE	-921.14	-4.32218	0.000044
UP/SQUARE	121.74	4.30780	0.000046
NRS	62.79	3.43493	0.000937
NGEI/SQUARE	38524.91	2.61912	0.010521
NE/1000*TP	80.88	2.04862	0.043738
NB/1000*TP	-165.62	-1.39920	0.165571
Intercept	-520.11	-1.08597	0.280715

Equation for the linear model after modification has the following form:

$$\hat{E}\left(Y^{(1)}(x)\right) = -20x_1 + 618x_2 + 81x_3 + 122x_4 + 38525x_5 - 921x_6 + 63x_8 + 2743x_9. \quad (6.21)$$

Now we discuss results of estimation of investigated single index model. Table 6.34 contains the R_0 for this model depending on h. Table 6.35 contains the estimates of unknown coefficients, calculated with different bandwidths.

Table 6.34

Values of R_0 for SIM

Bandwidth h		
1	5	10
165 843	268 740	317 251

Table 6.35

Results of SIM estimation

Factors	Bandwidth h		
	1	5	10
TP/1000*SQUARE	8 959	35 570	24 450
NE/SQUARE	-386	1 412	477
NE/1000*TP	7 195	9 192	1 935
NGEI/SQUARE	0.1	0.327	0.0036
NB/1000*TP	-3 437	9 267	4 559
NRS	0.1	2 500	1 603

The best chosen single index model with $h^* = 1$ can be written as (6.15), where vector of estimated coefficients is $\hat{\beta}^T = (8959 \quad -386 \quad 7195 \quad 0.1 \quad -3437 \quad 0.1)$.

As it has been already known, in the framework of the previous experiment the models well describing small transportations were constructed. In the given research an attempt to construct the models adequately describing both large and small transportations has been taken. For this purpose we are going to remove outliers from the data set on the basis of Mahalanobis distance.

Mahalonobis square distance [Srivastava 2002] shows the distance between each observation and the mean of the observations:

$$D_i = (x_i - \bar{x}) \cdot S^{-1} \cdot (x_i - \bar{x})', \quad (6.22)$$

where x_1, …, x_n is the sample of d-dimensional vectors (observations), $\bar{x}$ is the vector of means of each column of matrix X, S is the sample covariance matrix. Mahalanobis distance is superior to Euclidean distance, because it takes the distribution of the point's correlation into account.

Table 6.36 demonstrates observations with largest values of Mahalanobis distance (i.e. outliers). Last column of Table 6.36 contains corresponding transportations. Table 6.37 contains results of linear model estimation after removal of outliers.

Table 6.36

Outliers

Region	Year	D	Transp.
Daugavpils	2000	45.98	246.47
Rezekne	2000	35.39	37.40
Jelgava	2002	33.68	1424.74
Rezekne	2003	33.36	192.50
Riga	2003	29.15	23323.05
Riga	2000	27.78	16252.34
Riga	2002	25.87	22368.90
Riga	2001	23.97	19029.73
Rezekne	2002	23.74	193.77
Rezekne	2001	22.53	143.82
Daugavpils	2001	20.78	246.17
Jurmala	2000	10.21	839.86
Jurmala	2001	10.45	6218.92

Table 6.37

Estimation of the linear model after removal of outliers

Factors	$\hat{\beta}_i$	$t(69)$	p-level
NRS	128.1	13.5537	0.0000
NB/1000*TP	-361.0	-4.3788	0.0000
NE/SQUARE	2307.4	4.3114	0.0000
UP/SQUARE	-392.7	-4.1716	0.0001
NB/SQUARE	-1206.7	-3.4162	0.0011
NE/1000*TP	59.4	1.9859	0.0510
Intercept	919.8	1.5576	0.1239
NGEI/SQUARE	-26512.5	-0.8851	0.3792
TP/1000*SQUARE	-1.3	-0.1499	0.8813

For this model the calculated Fisher criterion is 91, R^2 = 0.91, R_0 = 259 096. As we can see, R_0 is smaller than in previous experiment. In other words the mistake of smoothing has been decreased after removal of outliers from data. Please pay attention to the fact, that removal of outliers has changed the significance of variables.

Regression equation containing only most significant variables for modified linear model is such:

$$\hat{E}\left(Y_M{}^{(1)}(x)\right) = 2307\,x_2 + 59x_3 - 393x_4 - 1207\,x_6 - 361x_7 + 128x_8. \qquad (6.23)$$

In the similar way we can modify suggested SIM. Estimation has been lead after removal of the same outliers from the data set (see Table 6.36). As well as in the previous experiment, the most significant variables by results of estimation of linear model have been included in SIM, i.e. NE/SQUARE, NE/1000*TP, UP/SQUARE, NB/SQUARE, NB/1000*TP and NRS. Table 6.38 contains the estimated coefficients, calculated with different bandwidths. Table 6.39 demonstrates the R_0 for this model depending on h.

Table 6.38

Estimation of modified SIM

Factors	Bandwidth h		
	1	5	10
NE/SQUARE	56	8 480	292
NE/1000*TP	820	63 980	2 499
UP/SQUARE	24	95 890	305
NB/SQUARE	4	7 966	87
NB/1000*TP	114	1 368	-927
NRS	0.1	44 880	7 849

Table 6.39

Values of R_0 for SIM

Bandwidth h		
1	5	10
10 053	74 032	474 028

The best chosen single index model with h^* = 1 can be written as (6.15), where vector of estimated coefficients is $\hat{\beta}_M{}^T = (56 \quad 820 \quad 24 \quad 4 \quad 114 \quad 0.1)$.

So, removal of outliers according to Mahalanobis distances has improved models in sense of reduction of a mistake of smoothing (i.e. R_0) in time of orders.

Cross-validation

Time period used for coefficient estimation is from 2000 till 2002, forecasting has been led for 2003 year. Table 6.40 contains the estimates of β for considered modified linear regression model. The signs of estimates coincide with signs for the case of smoothing and correspond to physical sense of explanatory factors. Coefficient R^2 is equal to 0.91 and the Fisher criterion is 94, so, the investigated linear model is adequate in the case of cross-validation as well. The R_0 is 739 151.

Table 6.40

Results of the linear model estimation for C-V

Factors	$\hat{\beta}_i$	$t(52)$	p-level
NRS	110	10.8784	0.0000
UP/SQUARE	-525	-10.0051	0.0000
NE/SQUARE	3863	8.1543	0.0000
NB/1000*TP	-416	-5.6836	0.0000
Intercept	2 692	5.5176	0.0000
NGEI/SQUARE	-117 031	-5.1087	0.0000

Regression equation for of the linear model for C-V can be written as:

$$\hat{E}\left(Y_{C-V}{}^{(1)}(x)\right) = 2692 + 3863\,x_2 - 525\,x_4 - 117031\,x_5 - 416\,x_7 + 110\,x_8. \qquad (6.24)$$

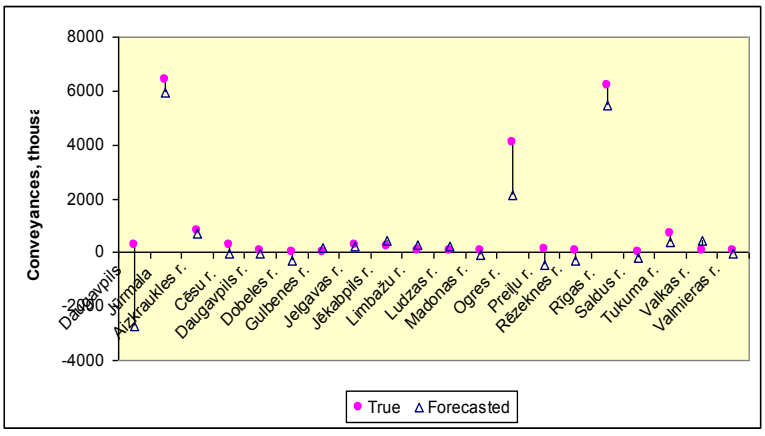

Fig.6.18. Forecasting by the Linear Model

Figure 6.18 demonstrates values of transportations for some regions of Latvia, forecasted by considered linear model. Unfortunately, this model gives negative forecasts for 9 objects, i.e. in 45% of observations.

The corresponding results for the analysed single index model are shown in Table 6.41. The values of R_0 corresponding to various h are resulted in Table 6.42. Thus, the best result for R_0 is achieved for h = 1. The forecasted transportations by the single index model with h^* = 1 and observed transportations as well are shown on the Figure 6.19.

The best chosen single index model with h^* = 1 can be written as (6.15), where vector of estimated coefficients is $\hat{\beta}_{C-V}{}^T = (283 \quad -380 \quad 4 \quad 2709 \quad 0.1)$. Obviously SIM gives more exact forecasts of transportations in comparison with linear model.

Table 6.41

Results of SIM estimation for C-V

Factors	Bandwidth h		
	1	5	10
NE/SQUARE	282.6	1 060 000	1 876
UP/SQUARE	-380.0	2 215 000	3 171
NGEI/SQUARE	4.1	11 820	16
NB/1000*TP	2 709	675 200	-4 972
NRS	0.1	5 653 000	54 650

Table 6.42

Values of R_0 for SIM

Bandwidth h		
1	5	10
147 810	256 774	720 403

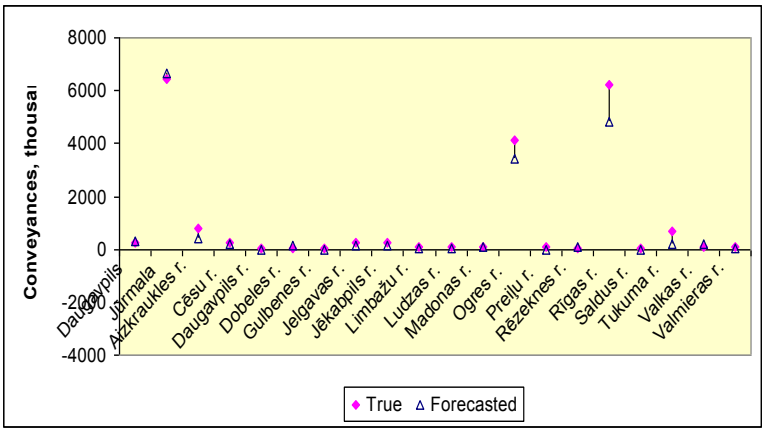

Fig.6.19. Forecasting by the single index model

We have collected the values of R_0 for both investigated models in cases of smoothing and cross-validation in Table 6.43. As we can see, the values of R_0 for SIM are less in all the considered cases.

Table 6.43

The values of R_0

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

CONCLUSION

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 obtained. As the result two group models were compared: the multiple linear regression model and the single index model. Various tests for hypothesis of explanatory variables insignificance and models correctness have been lead, and the cross-validation approach has been carried out as well. 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. In other words the semiparametric approach has given better results than classical parametric approach.

[2] including variable GRAD

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

Seemingly Unrelated Regression Equation model (SURE-model) proposed by Zellner in 1962 is appropriate and useful for a wide range of applications in econometrics, logistics and other areas. 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 but about a part of them only. An unbiased estimate for a covariance matrix of the model is obtained [13]. 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 stateded on the basis of practical results of Master Thesis of S.Bogdanova, performed under supervision of the Candidate [1].

As supplementary to the present promotional work, the weighed sum of all-share indexes NASDAQ OMX for three Baltic countries in case of incompleteness of statistical data is evaluated using the stated approach. The results of corresponding research are described in Appendix 3.

7.1. PROBLEM SETTING

Let us state the problem. 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 fix 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 linear-regression equation

$$Y_{i,j} = \sum_{v=1}^{m_i} \beta_{i,v} x_{i,j}^{(v)} + Z_{i,j}, \qquad (7.1)$$

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 (a disturbance) 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 $i = 1, 2, \ldots, G$ and $j = 1, 2, \ldots, n_i$, $x_{i,j} = (x_{i,j}^{(1)}, x_{i,j}^{(2)}, \ldots, x_{i,j}^{m_i})$ is a known constant vector, $Y_{i,j}$ is the fixed value. On this basis the unknown parameters of the regression model $\{\beta_{i,v}\}$, $\{c_{i,i'}\}$, where $\sigma_i^2 = c_{i,i}$, should be estimated.

Our final aim is to obtain the prognosis of the expectation of the sum

$$W(t) = \sum_{i=1}^{G} Y_{i,j(t)} = \sum_{i=1}^{G} \left(\sum_{v=1}^{m_i} \beta_{i,v} x_{i,j(t)}^{(v)} + Z_{i,j(t)} \right), \tag{7.2}$$

for future time moment *t*.

In fact the described model generalizes the seemingly unrelated regression model that was introduced by Zellner in [115], [116] and has many applications in the econometrics and in the other areas. For the last model $n_i = n$, $t_{i,j} = t_{i',j}$ for all *i*, *i*'. If in addition to the last model $x_{i,j} = (x_{i,j}^{(1)}, x_{i,j}^{(2)}, \ldots, x_{i,j}^{m_i}) = x_{i',j} = (x_{i',j}^{(1)}, x_{i',j}^{(2)}, \ldots, x_{i',j}^{m_i})$ for all *i*, *i*', then we have the multivariate linear regression model [95].

This Chapter is organized as following. The second Section contains a procedure of coefficient $\{\beta_{i,v}\}$ estimation and sum (7.2) prediction. In the third Section the covariance matrix of the coefficient's estimates is calculated. It allows us to estimate the variance of predicted sum. Two last sections contain numerical and econometric examples.

At first we need to estimate the unknown coefficients $\{\beta_{i,v}\}$. For that we use the ordinary least square method. Let X_i be a $(n_i \times m_i)$-matrix of the independent variables for the *i*-th object. We suppose that the rank of X_i equals m_i. Unknown coefficients $\{\beta_{i,v}\}$ are estimated using the well known formula

$$\beta_i^* = \left(X_i^T X_i\right)^{-1} X_i^T Y_i, \tag{7.3}$$

where $\beta_i^* = (\beta_{i,1}^*, \beta_{i,2}^*, ..., \beta_{i,m_i}^*)^T$ and $Y_i = (Y_{i,1}, Y_{i,2}, ..., Y_{i,n_i})^T$ are the vectors of the estimates and the dependent variables. Let us underline, that for estimation of the unknown coefficients for the multivariate model (3.6) the same formula is used.

The last formula gives the unbiased estimate of $\beta_i = (\beta_{i,1}, \beta_{i,2}, ..., \beta_{i,n_{ii}})^T$. It allows us to get unbiased estimate for the expectation of the sum $W(t)$ (7.2):

$$W^*(t) = \sum_{i=1}^{G} Y_{i,j(t)}^* = \sum_{i=1}^{G} x_{i,j(t)} \beta_i^* = \sum_{i=1}^{G} \left(\sum_{v=1}^{m_i} \beta_{i,v}^* x_{i,j(t)}^{(v)} \right). \tag{7.4}$$

We would like to comment a reason of choosing ordinary least square method. It allows us to get the simple explicit formula (7.3) for the coefficients. In other case, when the generalized least square method is used, it is necessary to apply complex iterated two-step estimation procedure [100], [103], [115].

Now our aim is the calculation of the variance of this estimate.

7.2. COVARIANCES OF THE ESTIMATES

According to formula (7.3), the covariance matrix of two coefficient vectors β_i^* and β_l^* is calculated by formula

$$Cov(\beta_i^*, \beta_l^*) = (X_i^T X_i)^{-1} X_i^T Cov(Y_i, Y_l) X_l (X_l^T X_l)^{-1}. \tag{7.5}$$

Also we need to calculate the covariance $Cov(Y_i, Y_l)$. Let $D^{(i,l)}$ is $(n_i \times n_l)$-matrix for which $D_{j,f}^{(i,l)} = 1$ if $t_{i,j} = t_{l,f}$ and $D_{j.f}^{(i,l)} = 0$ otherwise.

Let us recall that 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

$$Cov(Y_i, Y_l) = c_{i,l} D^{(i,l)}.$$

Now we are able to rewrite the formula (7.5) in the following form:

$$Cov(\beta_i^*, \beta_l^*) = c_{i,l}\left(X_i^T X_i\right)^{-1} X_i^T D^{(i,l)} X_l \left(X_l^T X_l\right)^{-1}. \quad (7.6)$$

Our next task is the estimation of unknown covariance $\{c_{i,l}\}$. For that we try to use usual estimator of the least squares:

$$c_{i,l}^* = \frac{1}{v_{i,l}} \left(Y_i - X_i \beta_i^*\right)^T D^{(i,l)} \left(Y_l - X_l \beta_l^*\right), \quad (7.7)$$

where $v_{i,l}$ is a constant that is determined by a condition of the unbiased estimator. To define the constant $v_{i,l}$ it is necessary to calculate expectation of the estimate $c_{i,l}^*$. We have:

$$\begin{aligned} E\left(c_{i,l}^*\right) &= E\left(\frac{1}{v_{i,l}} \left(Y_i - X_i \beta_i^*\right)^T D^{(i,l)} \left(Y_l - X_l \beta_l^*\right) \right) = \\ &= \frac{1}{v_{i,l}} E\left(\left(Y_i - X_i \left(X_i^T X_i\right)^{-1} X_i^T Y_i\right)^T D^{(i,l)} \left(Y_l - X_l \left(X_l^T X_l\right)^{-1} X_l^T Y_l\right) \right) = \\ &= \frac{1}{v_{i,l}} E\left(Y_i^T \left(I_i - X_i \left(X_i^T X_i\right)^{-1} X_i^T\right) D^{(i,l)} \left(I_l - X_l \left(X_l^T X_l\right)^{-1} X_l^T\right) Y_l \right), \end{aligned}$$

where I_i and I_l are identity matrixes of the rank n_i and n_l correspondingly.
Since

$$\left(I_i - X_i (X_i^T X_i)^{-1} X_i^T\right) Y_i = \left(I_i - X_i (X_i^T X_i)^{-1} X_i^T\right)\left(X_i \beta_i + Z_i\right) = \left(I_i - X_i (X_i^T X_i)^{-1} X_i^T\right) Z_i,$$

then

$$E\left(c_{i,l}^*\right) = \frac{1}{v_{i,l}} E\left(Z_i^T \left(I_i - X_i \left(X_i^T X_i\right)^{-1} X_i^T\right) D^{(i,l)} \left(I_l - X_l \left(X_l^T X_l\right)^{-1} X_l^T\right) Z_l \right). \quad (7.8)$$

Let us introduce the following notation: $R_j(i)^T$ is the *j*-th row ($R_j(i)$ is the *j*-th column) 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:

$$f(l,i,j)=\sum_{v=1}^{n_l} vD_{j,v}^{(i,l)}\ . \tag{7.9}$$

Then

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

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

The last formula shows that

$$v_{i,l}=\sum_j R_j(i)^T D^{(i,l)}R_{f(l,i,j)}(l). \tag{7.10}$$

With this value of the constant $v_{i,l}$ formula (7.10) gives the unbiased estimate of the covariance $c_{i,l}$.

Note for $\sigma_i^2=c_{i,i}$ we have $l=i,\ f(l,i,j)=j,\ D^{(i,i)}=I_{n_i}$, and the sum in the formula (7.10) equals the trace of the matrix A^2, where $A=I_i-X_i\left(X_i^TX_i\right)^{-1}X_i^T$. Besides A is an idempotent matrix, therefore its trace is equal to the rank $\rho(A)$. It is known [95] that $\rho(A)+\rho(I-A)=n_i$, and $\rho(I-A)=\rho\left(X_i\left(X_i^TX_i\right)^{-1}X_i^T\right)=m_i$. Therefore, $v_{i,i}=n_i-m_i$ and for the final estimate we get the well known expression

$$\sigma_i^2*=c_{i,i}*=\frac{1}{n_i-m_i}\left(Y_i-X_i\beta*_i\right)^T\left(Y_i-X_i\beta*_i\right). \tag{7.11}$$

A trivial case is interesting too. Let we intent to estimate a covariance for two random continue variables with numbers 1 and 2. At that we have two marginal samples of sizes n_1 and n_2 correspondingly, among them for *k* pairs observation time moments coincide, $k\le\min\{n_1,\ n_2\}$. For our notations we have: $m_1=m_2=1$, X_1 and X_2 are n_1–dimensional and n_2–dimensional unit vectors, $\left(X_i^TX_i\right)^{-1}=1/n_i$, $X_iX_i^T$ is square identity matrix of dimension n_i.

Further

$$I - X_i\left(X_i^T X_i\right)^{-1} X_i^T = \frac{1}{n_i}\begin{bmatrix} n_i - 1 & -1 & \ldots & -1 \\ -1 & n_i - 1 & \ldots & -1 \\ \ldots & \ldots & \ldots & \ldots \\ -1 & -1 & \ldots & n_i - 1 \end{bmatrix}_{n_i \times n_i},$$

$$R_j(i)^T = \frac{1}{n_i}\left(-1 \quad \ldots \quad -1 \quad n_i - 1 \quad -1 \quad \ldots \quad -1\right),$$

where the values $n_i - 1$ in the last vectors is in the *j*–th place.

Without loss of generality, we assume the observations are enumerated in such way, that the first initial observations coincide for both variables: $f(i, l, j) = j$. Then the matrix $D^{(1,2)}$ of dimension $n_1 \times n_2$ has the following form:

$$D^{(1,2)} = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix},$$

where I_k is *k*-dimensional identity matrix, another matrixes consist from zero.

Therefore for $i, l \le k$

$$R_j(1)^T D^{(1,2)} R_{f(i,l,j)}(2)^T = R_j(1)^T D^{(1,2)} R_j(2)^T =$$
$$= \frac{1}{n_1}\frac{1}{n_2}\left(\overbrace{1+1+\ldots+1}^{k-1 \ times} = 1 + (n_1 - 1)(n_2 - 1)\right) = \frac{1}{n_1 n_2}\left(k + n_1 n_2 - n_1 - n_2\right).$$

Finally

$$v_{i,l} = \sum_j R_j(i)^T D^{(i,l)} R_{f(l,i,j)}(l) = \sum_{j=1}^{k} R_j(i)^T D^{(i,l)} R_j(l) = \frac{k}{n_1 n_2}\left(k + n_1 n_2 - n_1 - n_2\right).$$

If $n_1 = n_2 = k = n$ then we have well known formula

$$v_{i,l} = \frac{n}{nn}\left(n + nn - n - n\right) = n - 1.$$

Now the variance of the sum (7.4) is calculated by usual way:

$$Var\left(W^*(t)\right) = Var\left(\sum_{i=1}^{G} x_{i,j(t)}\beta_i^*\right) = \sum_{i=1}^{G} x_{i,j(t)} Cov(\beta_i^*) x_{i,j(t)}^T + \\ + 2\sum_{i=1}^{G-1}\sum_{l=i+1}^{G} x_{i,j(t)} Cov(\beta_i^*, \beta_l^*) x_{l,j(t)}^T. \tag{7.12}$$

Here $Cov(\beta_i^*, \beta_l^*)$ is calculated by formula (7.6) and the covariance matrix of the vector β_i^* is calculated by the well known formula

$$Cov(\beta_i^*) = c_{i,i}\left(X_i^T X_i\right)^{-1} = \sigma_i^2\left(X_i^T X_i\right)^{-1}. \tag{7.13}$$

In the next Section the numerical example [13] based on the generated data is considered.

7.3. NUMERICAL EXAMPLE

Let we have two objects ($G = 2$) with numbers $i = 1, 2$. The first object is examined $n_1 = 5$ times, at the time moments $t_{1,1} = 1,\ t_{1,2} = 2,\ t_{1,3_i} = 4,\ t_{1,4} = 6,\ t_{1,5_i} = 9$. The second object is examined $n_2 = 7$ times, at the time moments $t_{2,1} = 1,\ t_{2,2} = 3,\ t_{2,3_i} = 4,\ t_{2,4} = 5,$ $t_{2,5_i} = 6,\ t_{2,6} = 7,\ t_{2,7_i} = 9$. It is supposed that the dependent variables $\{Y_{1,j}, Y_{2,j}\}$ are formed by the following linear-regression equations:

$$Y_{1,j} = \beta_{1,1} + \beta_{1,2} j + \beta_{1,3} j^2 + Z_{1,j},\ \ j = 1,...,5,$$
$$Y_{2,j} = \beta_{2,1} + \beta_{2,2} j + \beta_{2,3} j^2 + \beta_{2,4}\frac{1}{j} + Z_{2,j},\ \ j = 1,...,7.$$

Let us calculate unique normalized constant $v_{1,2}$ using formula (7.10). The matrixes X_1, X_2 and $D^{(1,2)}$ have the following forms:

$$X_1 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 9 & 81 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 9 & 1/3 \\ 1 & 4 & 16 & 1/4 \\ 1 & 5 & 25 & 1/5 \\ 1 & 6 & 36 & 1/6 \\ 1 & 7 & 49 & 1/7 \\ 1 & 9 & 81 & 1/9 \end{pmatrix}, \quad D^{(1,2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Matrixes $R_1 = \left(I_1 - X_1 (X_1^T X_1)^{-1} X_1^T\right)$ and $R_2 = \left(I_2 - X_2 (X_2^T X_2)^{-1} X_2^T\right)$ are

$$R_1 = \begin{pmatrix} 0.283 & -0.409 & 0.014 & 0.177 & -0.065 \\ -0.409 & 0.660 & -0.212 & -0.096 & 0.057 \\ 0.014 & -0.212 & 0.539 & -0.440 & 0.099 \\ 0.177 & -0.096 & -0.440 & 0.484 & -0.126 \\ -0.065 & 0.057 & 0.099 & -0.126 & 0.034 \end{pmatrix},$$

$$R_2 = \begin{pmatrix} 0.059 & -0.203 & 0.104 & 0.055 & 0.012 & -0.015 & -0.012 \\ -0.203 & 0.698 & -0.364 & -0.178 & -0.04 & 0.044 & 0.043 \\ 0.104 & -0.364 & 0.363 & -0.23 & 0.043 & 0.164 & -0.08 \\ 0.055 & -0.178 & -0.23 & 0.734 & -0.254 & -0.195 & 0.069 \\ 0.012 & -0.04 & 0.043 & -0.254 & 0.59 & -0.412 & 0.061 \\ -0.015 & 0.044 & 0.164 & -0.195 & -0.412 & 0.526 & -0.112 \\ -0.012 & 0.043 & -0.08 & 0.069 & 0.061 & -0.112 & 0.03 \end{pmatrix}.$$

The function $f(2, 1, j)$ values are given in Table 7.1.

Table 7.1

The function $f(2, 1, j)$ values

j	1	2	3	4	5
$f(2,1,j)$	1	0	3	5	7

The calculations according to formula (7.10) show that $v_{1,2} = 0.491$. We wish to remark that the result is not the integer number as it takes place in the usual regression analysis.

Now we wish to compare variances estimates for two cases: 1) the random variables $Y_{1,j}$ and $Y_{2,j}$ are independent; 2) they are dependent.

For the first case with respect to (7.12) and (7.13), we have

$$Var(W^*(t)) = Var\left(\sum_{i=1}^{2} x_{i,j(t)}\beta_i^*\right) = \sum_{i=1}^{2} x_{i,j(t)} Cov(\beta_i^*) x_{i,j(t)}^T =$$

$$= \sum_{i=1}^{2} \sigma_i^2 x_{i,j(t)} \left(X_i^T X_i\right)^{-1} x_{i,j(t)}^T. \qquad (7.14)$$

For the second case

$$Var(W^*(t)) = Var\left(\sum_{i=1}^{2} x_{i,j(t)}\beta_i^*\right) = \sum_{i=1}^{2} \sigma_i^2 x_{i,j(t)} \left(X_i^T X_i\right)^{-1} x_{i,j(t)}^T + \qquad (7.15)$$

$$+ 2x_{1,j(t)} Cov(\beta_1^*, \beta_2^*) x_{2,j(t)}^T.$$

Let $t = 10$, then

$$x_{1,j(10)} = (1 \quad 10 \quad 100), \quad x_{2,j(10)} = (1 \quad 10 \quad 100 \quad 0.1).$$

Let $\sigma_1^2 = c_{1,1} = 2$, $\sigma_2^2 = c_{2,2} = 5$, $c_{1,2} = \rho\sigma_1\sigma_2$, where ρ is the correlation coefficient. Then for the first case a calculation by formula (7.14) gives $Var(W^*(10)) = 16.703$. For the second case the values of this variance are presented in Table 7.2 as a function of the correlation coefficient ρ.

Table 7.2

The variance $Var(W^*(10))$ values as a function of ρ

(four joint observations)

ρ	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
Var	4.838	7.474	10.11	12.75	15.39	18.02	20.66	23.30	25.93	28.57

Table 7.3

The variance $Var(W^*(10))$ values as a function of ρ

(three joint observations)

ρ	-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
Var	7.369	9.359	11.35	13.33	15.32	17.31	19.30	21.29	22.28	25.26

We see that the dependence changes the variance values very sufficiently. It is due to a big number of joint observations (four from five for the variable Y_1). Obviously the less is the

number of the joint observations the less is this dependence. Table 7.3 contains corresponding results for three joint observations when the variable Y_1 is fixed at the time moments $t_{1,1} = 1$, $t_{1,2} = 2$, $t_{1,3} = 4$, $t_{1,4} = 8$, $t_{1,5} = 9$.

We can conclude that it is necessary to attach great importance to considered phenomena of dependence in the given statistical data. In the next Section applying of the suggested approach on the real data is considered.

7.4. FORECASTING TOTAL AIR PASSENGER TRANSPORTATIONS FOR THE EU

We demonstrate the suggested approach on an example of forecasting total air passenger transportations for the EU Member States. In the framework of the present research the passenger air international transportations are divided into two parts in relation to the borders of the European Union. Thus there are two forecasted variables: internal and external passenger air international transportations. Let us denote them by Intra-EU and Extra-EU. 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. Both dependent variables are measured in thousand of passengers. The model has 7 independent variables. Full their desciption is stated in Section 2.2. 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, percentage 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*) in GWh.

It should be stressed that selected independent variables are the most significant ones with respect to primary research, carried out in the framework of [1].

We have statistical data about analyzed objects for period from 2000 till 2007. Our aim is to investigate, how great number of missing observations influence the variance of sum of interest. 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. Our suggested approach does not deform statistical data. 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.

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 (7.15). Besides covariances $Cov(\beta_i^*, \beta_l^*)$ were calculated by formula (7.6).

The rule of creation of matrixes *R* and *D* and of function $f(l, i, j)$ has been shown in the previous Section, therefore in the given Section we will not present them because of their big dimension. In the present example we have two matrixes *R* (R_1 and R_2), one matrix $D^{(1,2)}$ and one normalizing constant $v_{1,2}$.

Let us present the results of statistical data estimation. The estimation procedure has been developed in MathCad 13 environment [125], [126]. We would like to notice, that calculations of normalizing constant have been taking a lot of time because of many operations with matrixes of quite high dimension.

Approach 1. Estimated regression coefficients are:

$$\beta_1^* = \begin{pmatrix} 0.02 \\ 49.59 \\ -21.32 \\ 4.15 \\ -0.66 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} 8.79 \\ -28.35 \\ 28.82 \\ 0.22 \\ 164.29 \end{pmatrix}.$$

The estimated covariance matrix of errors is

$$\widetilde{\Sigma} = \begin{pmatrix} 1.684 \cdot 10^8 & 1.687 \cdot 10^7 \\ 1.687 \cdot 10^7 & 8.234 \cdot 10^6 \end{pmatrix}.$$

The matrix of estimated parameters covariances is:

$$CovBeta(1,2) = \begin{pmatrix} 4.311 & -3.597 & -2.113 & -7.385 \times 10^{-3} & -142.335 \\ -6.376 & 28.659 & 1.486 & -3.774 \times 10^{-3} & 85.121 \\ -1.979 & 0.076 & 1.703 & 1.889 \times 10^{-3} & 87.678 \\ 0.013 & -4.827 \times 10^{-3} & -0.013 & 1.728 \times 10^{-4} & -1.28 \\ -9.52 \times 10^{-3} & -3.511 \times 10^{-3} & 7.442 \times 10^{-3} & -2.347 \times 10^{-5} & 0.623 \end{pmatrix}.$$

Approach 2. Estimated regression coefficients are:

$$\beta_1^* = \begin{pmatrix} -0.19 \\ 49.12 \\ -21.09 \\ 4.14 \\ -0.66 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} 11.93 \\ -30.78 \\ 29.68 \\ 0.21 \\ 102.48 \end{pmatrix}.$$

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

The estimated covariance matrix of errors is

$$\widetilde{\Sigma} = \begin{pmatrix} 1.668 \cdot 10^8 & 8.538 \cdot 10^6 \\ 8.538 \cdot 10^6 & 8.994 \cdot 10^6 \end{pmatrix}.$$

The matrix of estimated parameters covariances is:

$$CovBeta(1,2) = \begin{pmatrix} 16.483 & -5.691 & -5.156 & -0.052 & -564.61 \\ -15.243 & 77.989 & 3.583 & -0.028 & 140.635 \\ -4.447 & -0.346 & 9.067 & -0.024 & 241.092 \\ -0.012 & 0.172 & -0.035 & 6.566 \times 10^{-4} & -12.028 \\ -0.025 & -0.089 & 6.613 \times 10^{-3} & 6.424 \times 10^{-5} & 5.264 \end{pmatrix}.$$

The results of the total air passenger transportations forecasting are presented in Table 7.4 and Table 7.5. The first column of Table 7.4 contains the real 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 (for details see Chapter 4), by both approaches as well. In Table 7.5 are represented values of residual sum of squares R_0, received in case of smoothing and forecasting (for details see Section 6.1).

Table 7.4

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 7.5

R_0 values

	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

In this Chapter some generalization of known seemingly unrelated regression equation model has been described. Estimation procedures for regression coefficients and covariance matrix of dependent variables are elaborated and tested. The stated approach has been verified for the generated and the real data. As a result the forecasts of total air passenger transportations for 2008 for 26 Member States of the EU are obtained, and upper bounds of 95% confidence limits are calculated as well. The obtained results show the advantage of the offered approach in comparison with one used now. Nevertheless, it is necessary to notice, that variance calculation using stated approach is complicated enough in case of use MathCad 13 environment.

CONCLUSIONS

1. Promotional work is devoted to development and application of modern statistical methods for the transportations volumes analysis and forecasting for the European Union 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 and practical works

on the subject “Mathematical Methods of Traffic Flow Analysis and Forecasting” for the second year foreign students of bachelors studies programme of the Riga Technical University 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.

REFERENCES

1. Bogdanova S. Pasažieru aviopārvadājumu prognozēšana Eiropas Savienībai uz SURE-modeļa bāzes. Maģistra darbs. – Rīga: Rīgas Tehniskā universitāte, 2009. – 85 lpp.
2. Demidovs V. Dzelzceļa transporta lēmumatbalsta sistēmu informatīvās nodrošināšanas modeļu izstrāde. Promocijas darbs. – Rīga: Transporta un sakaru institūts, 2006. – 147 lpp.
3. Jackiva I. un Santalova D. Faktoru analīzes izmantošana transporta nozares attīstības tendenču noteikšanai ES valstīs// RTU zinātniskie raksti: Mašīnzinātne un Transports. – Rīga: RTU, 2005. – pp. 139-147.
4. Kļaviņš D. Optimizācijas metodes ekonomikā I, II. Mācību līdzeklis. Otrais izdevums. – R.: Datorzinības centrs, 2003. – 272 lpp.
5. Latvijas Statistikas Gadagrāmata 2006. – Rīga: LR CSP, 2007. – 302 lpp.
6. Santalova D. Vairumtirdzniecības noliktavas pārdošanu lieluma regresijas modelis// RTU zinātniskie raksti: Mašīnzinātne un Transports. – Rīga: RTU, 2005. – pp. 67-73.
7. Valsts A/S "Latvijas dzelzceļš" gada pārskats 2003. – Rīga: VAS "Latvijas dzelzceļš", 2004. – 37 lpp.
8. Žukovska J. Pasažieru aviopārvadājumu plūsmu prognozēšana. Promocijas darbs. – Rīga: Rīgas Tehniskā universitāte, 2008. – 177 lpp.
9. Andronov A. On Some Approach to an Estimation of Correspondence Matrix of Transport Network// Proceedings of the International Conference "Mathematical Methods for Analysis and Optimisation of Information Telecommunication Networks". – Minsk: Belarusian State University, 2009. – pp. 261-267.
10. Andronov A. Maximal Likelihood Estimates for Modified Gravitation Model by Aggregated Data// Proceedings of the 6th St. Petersburg Workshop on Simulation. – St. Petersburg, St. Petersburg State University, 2009. – pp. 1016-1021.
11. Andronov A. and Santalova D. On Nonlinear Regression Model for Correspondence Matrix of Transport Network// ASMDA-2009 Selected papers. L.Sakalauskas, C.Skiadas and E.K.Zavadskas (Eds.). – Vilnius, 2009. – pp. 90-94.
12. Andronov A. and Svirchenkov A. On some Modification of Seemingly Unrelated Regression Equations Model// Proceedings of the International Conference Statistical Methods for Biomedical and Technical Systems. Edited by Filia Vonta. – Limassol, Cyprus, 2006. – pp. 195 – 200.

13. Andronov A., Zhukovskaya C. and Santalova D. On Mathematical Models for Analysis and Forecasting of the Europe Countries Conveyances// RTU Zinātniskie raksti, Datorzinātne, 5. Sērija, Informācijas tehnoloģijas un vadības zinātne, 28. Sējums. – Rīga: RTU, 2007. – pp. 96 – 106.
14. Baublys A. The Econometric Models of Forecasting of the Transport Flows// Computer Modelling and New Technologies. – 2006. – Vol. 10, No. 4. – pp. 53-60.
15. Ben-Akiva M. and Lerman S. Discrete Choice Analysis: Theory and Applications to Travel Demand, sixth ed. Cambridge, MA: The MIT Press, 1994. – 416 p.
16. Bonneu M., Delecroix M. and Malin E. Semiparametric versus nonparametric estimation in single index regression model: A computational approach// Computational Statistics. – 1993. – No. 8. – pp. 207-222.
17. Bowman A. and Azzalini A. Applied Smoothing Techniques for Data Analysis. – Oxford: Oxford University Press, 1997. – 208 p.
18. Butkevičius J., Mazūra M., Ivankovas V. and Mazūra S. Analysis and forecast of the dynamics of passenger transportation by public land transport// Transport. – Vilnius: Tehnika, 2004. – Vol XIX, No. 1. – pp. 3 – 8. / Internet. – http://www.lib.berkeley.edu/ITSL/services.html
19. Butkevičius J. and Vyskupaitis A. Development of Passenger Transportation by Lithuanian Sea Transport// Transport and telecommunication. (In Procedings of International Conference RelStat'04). – Riga, 2005. – Vol. 6, No. 2. – pp. 274 – 279. / Internet. – http://www.tsi.lv/Transport&Telecommunication/v61_en/part2/art10.pdf.
20. Chatfield C. The Analysis of Time Series: An Introduction. 5th edition. – Chapman and Hall, 1996. – 304 p.
21. Cosslett S. Distribution-free maximum likelihood estimation of the binary choice model// Econometrica. – 1983. – No. 51. – pp. 765-782.
22. Cokasova A. Air Rail Intermodality from Passenger Perspective. // In Proceedings of the 19th Dresden Conference on Traffic and Transportation Science. – Dresden: 2003a, online. / Internet. – http://vwitme011.vkw.tu-dresden.de/TrafficForum /vwt_2003/beitraege/VWT19proceedings_contribution_103.1-103.17.pdf
23. Cokasova A. Modelling of Air-Rail Intermodality from Passenger Perspective at Major European Airports. Ingeneer Thesis. Paris: University of Zilina & EUROCONTROL, 2003b. – 109 p. / Internet. – www.eurocontrol.fr/Newsletter/2003/March/Intermodality/FINAL_THESIS.pdf

24. Deaton A. and Muellbauer J. Economics and Consumer Behavior. – Cambridge: Cambridge University Press, 1980. – 464 p.
25. Denby L. Smooth regression functions// Statistical report. – AT&T Bell Laboratories, 1986. – No 26.
26. Dobson A. J. An Introduction to Generalized Linear Models. Second edn. – London: Chapman and Hall, 2001. – 240 p.
27. Doganis J. Traffic forecasting and gravity model// Flight International. – 1966. – No. 20. – pp. 547 – 554.
28. Draper N. and Smith H. Applied Regression Analysis. Third ed. – New York: John Wiley and Sons, 1998. – 706 p.
29. European Commission: Eurostat Regional Yearbook 2008. Luxembourg, Office for Official Publications of the European Communities, 2008 — 187 pp. Internet. – http://epp.eurostat.ec.europa.eu/portal/page/portal/publications/regional_yearbook.
30. European Commission: Europe in figures – Eurostat Yearbook 2008. Data 1996 – 2006. EU, EuroSTAT, 2008. / Internet. – http://epp.eurostat.ec.europa.eu/portal/page /portal/publications/ eurostat_yearbook
31. European Commission: Europe in figures – Eurostat Yearbook 2006 – 2007. Data 1995 – 2005. EU, EuroSTAT, 2007. / Internet. – http://epp.eurostat.ec.europa.eu/portal /page/portal/product_details/publication?p_product_code=KS-CD-06-001
32. European Commission: Europe in figures – Eurostat Yearbook 2005. Data 1993–2004. EU, EuroSTAT, 2005. / Internet. – http://epp.eurostat.ec. europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-CD-05-001
33. European Commission: Guidelines for the implementation of the Regulation 91/2003 Version 5.1. – 147 p. –. Internet. – http://circa.europa.eu/Public/irc/ dsis/transport/library?l=/02_rail/data_monitoring/guidelines_implementatio/_EN_1.0_ &a=d. Last update 08/02/2006.
34. European Commission: Illustrated Glossary for Transport Statistics. 4st edition. Final version 14/07/2009. – 185 p. – Internet. – http://www.internationaltransportforum.org /Pub/pdf/09GloStat.pdf
35. European Commission: Panorama of Transport 2009. Luxembourg, Office for Official Publications of the European Communities, 2009. — 185 pp. Internet. –

http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-DA-09-001/EN/KS-DA-09-001-EN.PDF

36. European Commission: Reference Manual on Air Transport Statistics, Version 5. – 277 p. – Internet. – http://epp.eurostat.ec.europa.eu/portal/page/portal/transport/documents/Aviation%20Reference%20Manual%20version%205.pdf. Last update 06.05.2009.
37. Eurostat News Release: Euro area GDP down by 2.5% and EU27 GDP down by 2.4%. No 100/2009 – 8 July 2009. / Internet. – http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/2-08072009-AP/EN/2-08072009-AP-EN.PDF
38. Eurostat News Release: Euro area inflation estimated at -0.6%. No 111/2009 – 31 July 2007. / Internet. – http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/2-31072009-AP/EN/2-31072009-AP-EN.PDF
39. Eurostat News Release: Euro area unemployment up to 9.4%. No 112/2009 – 31 July 2009. / Internet. – http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-31072009-BP/EN/3-31072009-BP-EN.PDF
40. Eurostat News Release: Euro area unemployment up to 8.2%. No 25/2009 – 27 February 2009. / Internet. – http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-27022009-AP/EN/3-27022009-AP-EN.PDF
41. Farnum N.R. and Staton L.W. Quantitative Forecasting Methods. – Wadsworth Publishing Company, 1989. – 656 p.
42. Fengler M.R. Semiparametric Modeling of Implied Volatility. – Springer, 2005. – 244 p.
43. Fuss M., McFadden D. and Mundlak Y. A survey of functional forms in the economic analysis of production// Production Economics: A Dual Approach to Theory and Applications, M. Fuss & D. McFadden (Eds.). – Amsterdam: North-Holland, 1978. – pp. 219-268.
44. Gallant A. and Nychka D. Semi-nonparametric maximum likelihood estimation// Econometrica. – 1987. – Vol. 55, No. 2. – pp. 363-390.
45. Gill P., Murray W. and Wright M. Practical Optimization. – London: Academic Press, 1981. – 402 p.

46. Green P. J. and Silverman B. W. Nonparametric Regression and Generalized Linear Models// Monographs on Statistics and Applied Probability. – London: Chapman and Hall, 1994. – Vol. 58. – 184 p.
47. Green P. J. and Yandell B. S. Semi-parametric generalized linear models// Proceedings 2nd International GLIM Conference, Lecture Notes in Statistics 32. – New York: Springer, 1985. – Vol. 32. – pp. 44-55.
48. Han A. Non-parametric analysis of a generalized regression model// Journal of Econometrics. – 1987. – No. 35. – pp. 303-316.
49. Hardin J. and Hilbe J. Generalized Linear Models and Extensions. – Stata Press, 2001. – 245 p.
50. Härdle W. Applied Nonparametric Regression// Econometric Society Monographs. – Cambridge University Press, 1990. – No. 19. – 352 p.
51. Härdle W. Smoothing Techniques, With Implementations in S. – New York: Springer, 1991. – 261 p.
52. Härdle W., Muller M., Sperlich S, Werwatz A. Nonparametric and Semiparametric Models. Springer: Berlin – Heidelberg – New York, 2004. – 328 p. / Internet. – http://www.quantlet.com/mdstat/scripts/spm/html/spmhtml.html.
53. Härdle W. and Stoker T. M. Investigating smooth multiple regression by the method of average derivatives// Journal of the American Statistical Association. – 1989. – No. 84. – pp. 986-995.
54. Hastie T. J. and Tibshirani R. J. Generalized Additive Models// Monographs on Statistics and Applied Probability. – London: Chapman and Hall, 1990. – No. 43. – 352 p.
55. Hristache A., Juditski M. and Spokiony V. Direct estimation of the index coefficients in a single-index model// Annals of Statistics. – 2001. – No. 29. – pp. 596 – 623.
56. Hunt Ü. Forecasting of the Railway Freight Volume: Approach to Estonian Railway to Arise Efficiency// Transport. – Vilnius: Tehnika, 2003. – Vol XVIII, No. 6. – pp. 255 – 258. / Internet. – http://www.transport.vgtu.lt/en/?page=3&pub=2109
57. IATA economic briefing: Passenger and freight forecasts 2006 to 2010. September 2006. / Internet. – http://www.iata.org/NR/rdonlyres/D4F7A43D-DE5B-4FFF-BE72-16731F286402/0/traffic_forecast_2006_2010.pdf
58. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models// Journal of Econometrics. – 1993. – No. 58. – pp. 71-120.

59. Andronov A. and Kashurin A. On a Problem of Spatial Arrangement of Service Stations// Computer Modelling and New Technologies. – 2007. – Vol. 11, No. 1. – pp. 31 – 37.
60. Keele L.J. Semiparametric Regression for the Social Sciences. – Wiley, 2008. – 230 p.
61. Kong E. and Xia Y. Variable Selection for the Single-index Model// Biometrika. – 2007. - Vol. 1, No. 94. – pp. 217-229.
62. Kopytov E. and Demidovs V. Virtual Data Models in Anticipatory System of Railway Transportation// International Journal of Computing Anticipatory Systems, Daniel M. Dubois (Eds.). - CHAOS, University of Liege, Institute of Mathematics, 2006. – Vol. 19. – pp. 135-145.
63. Kopytov E. and Santalova D. Application of the Single Index Model for Forecasting of the Inland Conveyances// Recent Advances in Stochastic Modelling and Data Analysis, Christos H. Skiadas (Eds.). – Singapore: World Scientific Publishing Co Pte Ltd., 2007. – pp. 268-276.
64. Lehman E.L. Theory of Point Estimation. – New York: John Wiley and Sons, 1983. – 506 p.
65. Leontief W. Introduction to a theory of the internal structure of functional relationships// Econometrica. – 1947. – No. 15. – pp. 361-373.
66. Leontief W. A note on the interrelation of subsets of independent variables of a continuous function with continuous first derivatives// Bulletin of the American Mathematical Society. – 1947. – No. 53. – pp. 343-350.
67. Linton O. Efficient estimation of generalized additive nonparametric regression models// Econometric Theory. – 2000. – Vol. 16, No. 4. – pp. 502-523.
68. Linton O. and Nielsen J. P.A kernel method of estimating structured nonparametric regression based on marginal integration// Biometrika. – 1995. – No. 82. – pp. 93-101.
69. Lyumkis V. and Tarasov A. Estimation of Risk Insurance in Practice of Work of Actuarials// Computer Modelling and New Technologies, 2008, Vol.12, No.3, 16–21.
70. Makridakis S., Wheelwright S. and Hyndman R. Forecasting: Methods and Applications. – New York: John Wiley and Sons, 1998. – 656 p.
71. Manski C. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator// Journal of Econometrics. – 1985. – No. 3. – pp. 205-228.
72. Marron J. S. and Nolan D. Canonical kernels for density estimation// Statistics & Probability Letters. – 1988. – Vol. 7, No. 3. – pp. 195-199.

73. Masry E. and Tjøstheim D. Non-parametric estimation and identification of nonlinear arch time series: strong convergence properties and asymptotic normality// Econometric Theory. – 1995. – No. 11. – pp. 258-289.
74. Masry E. and Tjøstheim D. Additive nonlinear ARX time series and projection estimates// Econometric Theory. – 1997. – No. 13. – pp. 214-252.
75. McCullagh P. and Nelder J. Generalized Linear Models. – London: Chapman and Hall, 1989. – 511 p.
76. Nadaraya E. A. On estimating regression// Theory of Probability and its Applications. – 1964. – No. 10. – pp. 186-190.
77. Ortuzar J. de D. and Willumsen L. G. Modelling Transport. – New York: John Wiley and sons, 2001. – 514 p.
78. Pagan A. and Ullah A. Nonparametric Econometrics. – Cambridge University Press, 1999. – 444 p.
79. Park B. U. and Turlach B. A.Practical performance of several data driven bandwidth selectors// Computational Statistics. – 1992. – No. 7. – pp. 251-270.
80. Preparation of Guidelines on the Application of the Core Set of Environmental Indicators in Eastern Europe, the Caucasus and Central Asia. Workshop on the Application of Environmental Indicators. Chisinau, United Nations Economic Commission for Europe, 2004. – 15 p. / Internet. – http://www.unece.org/env/europe/monitoring/Chisinau.Jul04/Indicators%20Transport.e.pdf
81. Road Transport Forecasts 2008. Results from the Department for Transport's National Transport Model. / Internet. – http://www.dft.gov.uk/pgr/economics/ntm/roadtransportforcasts08/ executivesummary
82. Robinson P. M. Semiparametric econometrics: A survey// Journal of Applied Econometrics. – 1988. – No. 3. – pp. 35-51.
83. Ruppert D., Wand M. P. and Carroll R. J. Semiparametric Regression. – Cambridge University Press, 1990.
84. Santalova D. Regression Model of Sales Volume from Wholesale Warehouse// Proceedings of the 13th International Conference on Analytical and Stochastic Modelling Techniques and Applications. – Bonn, 2006. – pp. 133 – 137.
85. Santalova D. Forecasting of Rail Freight Conveyances in EU Countries on the Base of the Single Index Model// Computer Modelling and New Technologies. – Riga: TSI, 2007. – Vol. 11, No. 1. – pp. 73-83.

86. Santalova D. The Use of Multivariate Regression Model to Forecast the Freight Air Transportations in the Members Countries of the Europe Union// Proceedings of The International Conference „Modelling of Business, Industrial and Transport Systems – 2008". – Riga: TSI, 2008. – pp. 258 – 266.
87. Santalova D. Forecasting of Passenger Conveyances in Latvian Regions Applying Semiparametric Regression Models// RTU zinātniskie raksti: Mašīnzinātne un Transports. – Riga: RTU, 2008. – In appearance.
88. Silverman B. W. Density Estimation for Statistics and Data Analysis// Monographs on Statistics and Applied Probability. – London: Chapman and Hall, 1986. – No. 26. – 170 p.
89. Simonoff J. Smoothing Methods in Statistics. – New York: Springer, 1996.
90. Simonoff J. and Tsai, C.-L. Score Tests for the Single Index Model// Technometrics. – American Statistical Association and the American Society for Quality, 2002. – Vol. 44, No. 2. – pp. 142 – 151. / Internet. – http://archive.nyu.edu/bitstream /2451/14790/1/SOR-2000-5.pdf
91. Sleeper A. Six Sigma Distribution Modeling. – New York: McGraw-Hill, 2007. – 448 p.
92. Šliupas T. Annual Average Daily Traffic Forecasting Using Different Techniques // Transport. – Vilnius: Tehnika, 2006. – Vol XXI, No 1. – pp. 38 – 43. / Internet. – http://www.transport.vgtu.lt/en/?page=3&pub=2707
93. Speckman P. E. Regression analysis for partially linear models// Journal of the Royal Statistical Society. – 1988. – Series B, No. 50. – pp. 413-436.
94. Spissu E., Pinjari A.R., Pendyala R.M. et al. A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel// Transportation. – Springer Science+Business Media, LLC, 2009. – No. 36. – pp. 403–422.
95. Srivastava M. Methods of Multivariate statistics. – New York: John Wiley and Sons, 2002. – 702 p.
96. Statistical Yearbook of Latvia 2003. – Riga: Central Statistical Bureau of Latvia, 2003. – 272 p.
97. Stevens J. Applied Multivariate Statistics for the Social Sciences. – New York: Lawrence Erlbaum Associates, 2002. – 699 p.
98. Taneja N. Airline Traffic Forecasting: A Regression Analysis Approach. – Lexington Books, 1978. – 230 p.

99. Tjøstheim D. and Auestad B. Nonparametric identification of nonlinear time series: Projections// Journal of the American Statistical Association. – 1994. – No. 89. – pp. 1398-1409.
100. Turkington D.A. Matrix Calculus and Zero-One Matrices: Statistical and Econometric Application. – Cambridge: Cambridge University Press, 2002. – 206 p.
101. Upton G. and Cook I. A Dictionary of Statistics. Oxford University Press, 2006. – xxx p.
102. Ushakov N. Some Inequalities for the Mean Integrated Squared Error of Multivariate Kernel Density Estimators// Proceedings of the International Conference "Mathematical Methods in Reliability". – Moscow, 2009. – pp. 376-379.
103. Velu R. and Richards J. Seemingly unrelated reduced-rank regression model// J. Statistical Planning and Inference. – 2008. – No. 138. – pp. 2837 – 2846.
104. Watson G. S. Smooth regression analysis// Sankhyā. – 1964. – Series A, No. 26. – pp. 359-372.
105. Wecker W. and Ansley C. The signal extraction approach to nonlinear regression and spline smoothing// Journal of the American Statistical Association. – 1983. – No. 78. – pp. 351-365.
106. Weisberg S. Applied Linear Regression. – New York: John Wiley and Sons, 1985. – 336 p.
107. Weisberg S. and Welsh A. H. Adapting for the missing link// Annals of Statistics. – 1994. – No. 22. – pp. 1674-1700.
108. Wheatcroft S. and Lipman G. European Liberalization and World Air Transport: Towards a Transnational Industry. – London: Economist Intelligence Unit, 1990. – 213 p.
109. Wong H, Ip W.C. and Zhang R. Varying-coefficient single-index model// Computational Statistics & Data Analysis. – January 2008. – No. 52, Issue 3. – pp. 1458-1476.
110. Xia Y., Tong H., Li W.K. et al. An Adaptive Estimation of Dimension Reduction Space (with Discussions)// Roy. Statist. Soc. B. – 2002. – No. 64. – pp. 363-410.
111. Xia Y., Li W. K., Tong H. et al. A Goodness-of-fit Test for Single-Index Models// Statistica Sinica. – 2004. – No. 14. – pp. 1-39.

112. Xia Y. Asymptotic Distributions for two Estimators of the single-index model// Econometric Theory. – Cambridge University Press, 2006. – Vol. 6, No. 22. – pp. 1112-1137.
113. Xue L., Zhu L. Empirical Likelihood Semiparametric Regression Analysis for Longitudinal Data// Biometrika. – 2007. – Vol. 4, No. 94. – pp. 921-937. / doi:10.1093/biomet/asm066.
114. Yatchew A. Semiparametric Regression for the Applied Econometrician. – Cambridge University Press, 2003. – 236 p.
115. Zellner A. An efficient method of estimating seemingly unrelated regression some exact finite sample// J. American Statistical Association. – 1962. – No. 57. – pp. 348-368.
116. Zellner A. Estimators for estimating seemingly unrelated regressions equations and tests for aggregation bias// J. American Statistical Association. – 1963. – No. 58. – pp. 977-992.
117. The NASDAQ OMX Group, Inc. official web-site http://www.baltic.omxnordicexchange.com
118. Bonneu M. et Delecroix M. Estimation semiparamétrique dans les modèles explicatifs conditionnels à indice simple// Cahier de gremaq. – GREMAQ, Université Toulouse I., 1992. – 92.09.256.
119. Benítez R. Factores determinantes de la demanda de transporte aéreo y modelos de previsión// Boletín económico de ICE, Información Comercial Española. – 2000. – Nº 2652, pags. 41 – 48. / Internet. – http://www.revistasice.com/estudios/Documen/bice/2652/BICE26520203.PDF
120. Андронов А.М., Копытов Е.А., Гринглаз Л.Я. Теория вероятностей и математическая статистика: Учебник для ВУЗов. – СПб, Питер, 2004. – 461 с.
121. Андронов А.М., Хижняк А.И., Швацкий И. Е. и др. Прогнозирование перевозок пассажиров на воздушном транспорте. – М.: Транспорт, 1983. – 183 с.
122. Афанасьева Е. Оценивание моделей процессов транспорта и логистики на базе интенсивных компьютерных методов статистики. Расширенный автореферат промоционной работы. – Рига: Институт транспорта и связи, 2006. – 50 с.
123. Боровиков В. П.. STATISTICA: Искусство анализа данных на компьютере (2-ое издание). – СПб.: Питер, 2003. – 700 с.
124. Леман Э. Проверка статистических гипотез. – М.: Наука, 1979. – 408 с.

125. Люмкис В.Д., Яцкив И.В. Пакеты прикладных программ для научных исследований. Ч. 2. Методические указания по выполнению лабораторных работ на базе пакета STATISTICA . - Рига: Институт транспорта и связи, 2003. – 63 с.
126. Ракитин В.И. Руководство по методам вычислений и приложения MATHCAD. – М.: Физмалит, 2005. – 246 с.
127. Яцкив И.В. Методы компьютерной обработки статистических данных. – Рига: Институт транспорта и связи, 2002. – 204 с. В эл. виде.
128. Яцкив И.В. Многомерный статистический анализ: классификация и снижение размерности. Учебное пособие по курсу "Компьютерная статистика". 2-е изд., исправленное и доп. - Рига: Институт транспорта и связи, 2005. – 207 с.

APPENDIXES

APPENDIX 1

LIST OF CONFERENCES

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 Conference 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", authors: 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. 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). Vilnius, Lithuania, June 30 – July 3, 2009. Report: „On Nonlinear Regression Model for Correspondence Matrix of Transport Network", authors: Andronov A., Santalova D.

APPENDIX 2

ECONOMETRICAL STUDY OF SURE-MODEL

The corresponding report "On Some Generalization of Seemingly Unrelated Regression Equation Models" was presented on the 2nd International Workshop on Computational and Financial Econometrics (CFE-2008), which took place at University of Neuchatel, Switzerland.

We demonstrate the approach, suggested in Chapter 7, by an example of forecasting weighted sum of NASDAQ OMX Group all-share indexes for three Baltic countries, i.e. OMX Tallinn (further OMXT) index for Estonia, OMX Riga (OMXR) index for Latvia and OMX Vilnius (OMXV) index for Lithuania [117]. The indexes are weighted with ratios of Gross Domestic Product (GDP) a_1, a_2, a_3 of each considered country for 2007.

The NASDAQ OMX Group, including Baltic Market stock exchanges in Tallinn, Riga and Vilnius, is the world's largest exchange company. It delivers trading, exchange technology and public company services across six continents and more than 50 countries. The Baltic Market is positioned at the heart of Baltic Sea region, Europe's fastest growing market. An excellent business environment provides solid growth opportunities for new and existing companies, generating exceptional opportunities for investors.

All-share indexes are intended to display the general level of economy and its changes for the countries. The basic attention is turned to the equities representing index, without taking into attention the requirements to liquidity and stability. Considered three indexes are calculated in currency of a corresponding stock exchange. OMXT index is calculated in euro, OMXV index in lits and OMXR index in lats. In carried out experiments all three indexes are recalculated in euro.

So, we have three objects ($G = 3$) with numbers $i = 1, 2, 3$. Each object has one own dependent variable: for the first object dependent variable Y_1 is the value of the index OMXT for Estonia, for the second object Y_2 is the value of the index OMXR for Latvia and for the third object Y_3 is the value of the index OMXV for Lithuania. The model has 4 independent variables. The first three of them are common for all objects: X_1 is time factor; X_2 is EUR/USD daily exchange rate; X_3 is EUR/GBP daily exchange rate. The last variable X_4 are EUR/EEK, EUR/LVL and EUR/LTL daily exchange rates for the first, second and third objects, where EEK, LVL and LTL are the national currencies of Estonia, Latvia and Lithuania correspondingly.

We have statistical data about analyzed objects from 3 March, 2008 till 30 June, 2008, total 81 observations for each object. Our aim is to investigate, how much missing data influence the variance of weighted sum of interest.

We suppose that 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 way.

Both these approaches deform statistical data and forecasting results. Our suggested approach does not deform statistical data. To analyze the efficiency of the stated approach we fix some observations on which data will be absent. Table A3.1 contains the corresponding dates for all considered objects.

We use three approaches of data processing:

1. all data are presented and processed (i.e. above mentioned 81 observations);
2. some observations are absent (see Table A3.1) and statistical data for all corresponding time moments are extracted for all the objects;
3. some observations are absent (see Table A3.1) and suggested approach is applied.

Table A3.1

Dates on which observations are absent

Estonia OMXT	**Latvia OMXR**	**Lithuania OMXV**
14.03.2008	10.03.2008	10.03.2008
31.03.2008	25.03.2008	11.03.2008
09.04.2008	31.03.2008	25.03.2008
16.04.2008	04.04.2008	14.04.2008
21.04.2008	21.04.2008	21.04.2008
25.04.2008	02.05.2008	02.05.2008
09.05.2008	05.05.2008	05.05.2008
15.05.2008	20.05.2008	26.05.2008
05.06.2008	05.06.2008	27.05.2008
06.06.2008	17.06.2008	06.06.2008
17.06.2008		

As efficiency criterion *the estimate of variance of weighted sum of interest* is used. For that aim we process the data for analyzed period and make forecast for 01 July, 2008. We wish to compare estimates of variances when the random variables $Y_{1,j}$, $Y_{2,j}$ and $Y_{3,j}$ are independent, and when they are dependent.

For the first case we have according to (7.14) and (7.15):

$$Var(W^*(t)) = Var\left(\sum_{i=1}^{3} a_i x_{i,j(t)} \beta_i^*\right) = \sum_{i=1}^{3} a_i^2 \sigma_i^2 x_{i,j(t)} \left(X_i^T X_i\right)^{-1} x_{i,j(t)}^T . \quad \text{(A3.1)}$$

For the second case

$$\begin{aligned} Var(W^*(t)) = Var\left(\sum_{i=1}^{3} a_i x_{i,j(t)} \beta_i^*\right) = \sum_{i=1}^{3} a_i^2 \sigma_i^2 x_{i,j(t)} \left(X_i^T X_i\right)^{-1} x_{i,j(t)}^T + \\ + 2a_1 a_2 x_{1,j(t)} Cov(\beta_1^*, \beta_2^*) x_{2,j(t)}^T + 2a_2 a_3 x_{2,j(t)} Cov(\beta_2^*, \beta_3^*) x_{3,j(t)}^T + \\ + 2a_1 a_3 x_{1,j(t)} Cov(\beta_1^*, \beta_3^*) x_{3,j(t)}^T , \end{aligned} \quad \text{(A3.2)}$$

where covariances $Cov(\beta_i^*, \beta_l^*)$ were calculated by formula (7.6).

The weights a_i (i = 1, 2, 3) in formulas (A3.1) and (A3.2) are ratios of GDP for considered countries for 2007, i.e. for Estonia a_1 = 0.37, for Latvia a_2 = 0.31 and for Lithuania a_3 = 0.32.

The rule of creation of matrixes R and D and of function $f(l, i, j)$ has been shown in the Section 7.3, therefore now we will not present them because of their big dimension. In the present example we have three matrixes R (R_1, R_2 and R_3), three matrixes D ($D^{(1,2)}$, $D^{(2,3)}$, $D^{(1,3)}$) and three normalizing constants $v_{1,2}$, $v_{2,3}$ and $v_{1,3}$.

Let us present the results of statistical data estimation. The estimation procedure has been developed in MathCad 13 environment. We would like to notice, that calculations of normalizing constant have been taking a lot of time because of many operations with matrixes.

Approach 1. Estimated regression coefficients are:

$$\beta_1^* = \begin{pmatrix} -0.77 \\ 84.58 \\ -474.58 \\ 56.74 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} -0.50 \\ -115.56 \\ 287.26 \\ 722.65 \end{pmatrix}, \ \beta_3^* = \begin{pmatrix} -0.64 \\ 196.86 \\ -1.06 \times 10^3 \\ 290.39 \end{pmatrix}.$$

The estimated covariance matrix of errors is

$$\widetilde{\Sigma} = \begin{pmatrix} 58.39 & -3.21 & 43.24 \\ -3.21 & 54.56 & -13.20 \\ 43.24 & -13.20 & 67.92 \end{pmatrix}.$$

Approach 2. Estimated regression coefficients are:

$$\beta_1^* = \begin{pmatrix} -0.75 \\ 90.81 \\ -598.35 \\ 62.33 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} -0.48 \\ -100.33 \\ 270.64 \\ 706.19 \end{pmatrix}, \ \beta_3^* = \begin{pmatrix} -0.62 \\ 196.54 \\ -1.17 \times 10^3 \\ 315.97 \end{pmatrix}.$$

The estimated covariance matrix of errors is

$$\widetilde{\Sigma} = \begin{pmatrix} 55.12 & 1.78 & 40.40 \\ 1.78 & 54.98 & -11.70 \\ 40.40 & -11.70 & 68.60 \end{pmatrix}.$$

Approach 3. Estimated regression coefficients are:

$$\beta_1^* = \begin{pmatrix} -0.76 \\ 89.83 \\ -525.26 \\ 58.72 \end{pmatrix}, \ \beta_2^* = \begin{pmatrix} -0.50 \\ -108.94 \\ 296.79 \\ 698.12 \end{pmatrix}, \beta_3^* = \begin{pmatrix} -0.63 \\ 186.60 \\ -1.10 \times 10^3 \\ 304.84 \end{pmatrix}.$$

The normalizing constants are $v_{1,2} = 60.01$, $v_{2,3} = 61.99$ and $v_{1,3} = 58.05$.

The estimated covariance matrix of errors is

$$\widetilde{\Sigma} = \begin{pmatrix} 54.78 & 1.48 & 39.76 \\ 1.48 & 54.78 & -10.55 \\ 39.76 & -10.55 & 68.18 \end{pmatrix}.$$

The results of the estimation are presented in Table A3.2. We ask to pay attention, that true weighted sum of interest at 1 July, 2008 is 511.70.

Table A3.2

Variance estimation results

Approach No	Forecasted weighted sum of interest at 1 July 2008	Estimated variance (independent)	Estimated variance (dependent)
1.	510.71	2.58	3.49
2.	509.95	2.98	4.41
3.	510.06	2.78	3.74

Firstly, we must note that dependence between the indexes of interest for various objects essentially changes the estimate of the variance (see two last columns of Table A3.2). Therefore this fact must certainly be taken into account. Secondly, the variance estimates for the suggested approach are closer to the most successful case (it is the first approach) than for the third approach. It demonstrates the obviously advantage of the stated approach for obtaining the total interest in case of incompleteness of statistical data.