ON MATHEMATICAL MODELS FOR ANALYSIS AND FORECASTING OF THE EUROPE UNION COUNTRIES CONVEYANCES

МАТЕМАТИЧЕСКИЕ МОДЕЛИ ДЛЯ АНАЛИЗА И ПРОГНОЗИРОВАНИЯ ПЕРЕВОЗОК В СТРАНАХ ЕВРОПЕЙСКОГО СОЮЗА

MATEMĀTISKIE MODEĻI PĀRVADĀJUMU ANALĪZEI UN PROGNOZĒŠANAI EUROPAS SAVIENĪBAS VALSTĪS

Alexander Andronov, Catherine Zhukovskaya and Diana Santalova
Aleksandrs Andronovs, Jekaterīna Žukovska un Diāna Santalova
Александр Андронов, Екатерина Жуковская и Диана Санталова

Keywords: freight conveyances, passenger conveyances, forecasting, single index model, generalized linear model

1. Introduction

The aim of our research is to elaborate mathematical models of the passenger and freight conveyances.

In this investigation the following problems are considered:

1. Forecasting of passenger conveyances from European Union (EU) countries.
2. Forecasting of passenger flows between EU countries.
3. Forecasting of freight conveyances from EU countries.
4. Forecasting of freight flows between EU countries.

The following mathematical models are considered: linear regression model, generalized linear regression model and semiparametric regression model. In this article the used information base and the short theoretical description of the used mathematical models are considered, also the concrete models for forecasting air passenger and freight conveyances are suggested as well as the detailed analysis of the given results.

2. Informative Base

The main information is available in the information base Intra-EU. The basic data have been received from the electronic database "The Statistical Office of the European Communities" (EUROSTAT) [7], but as some data were incomplete it was necessary to search data on the separate statistical sites of the EU countries.

The following factors have been selected as explanatory variables:

t_1 - area of the country (SQUARE), thousands of km^2;
t_2 - total population (TP), millions of inhabitants;
t_3 - hourly labour cost (HLC), euro;
t_4 - monthly labour cost (MLC), euro;
t_5 - gross domestic product (GDP), millions of euro;
t_6 - GDP "per capita" in Purchasing Power Standards (PPS);
t_7 - labour productivity per hour worked (LPHW);
t_8 - labour productivity per employment (LPE);
t_9 - unemployment rate (UR);
t_{10} - comparative price level (CPL);
t_{11} - total length of railways (TOTLEN), km;
t_{12} - number of locomotives (LOKOM);
t_{13} - number of wagons (WAGONS).

Let us comment some of the described factors:

a) the volume index of GDP is expressed in relation to the European Union (EU-25) 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;
b) 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 (EU25 = 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.

3. The Models for Forecasting of Conveyances

Let us consider models *for conveyances forecasting.* The main object of consideration is named *object.* Sometimes it is a passenger or freight conveyances from EU countries, sometimes it is a passenger or freight flows between EU countries etc. We call as *observation* a data about object for concrete time.

We talk about the *individual model* if one object corresponds to another object for various observations, and about the *group model* if one corresponds to various objects.

With respect to used mathematical model the two latter models are regression ones. We will consider *linear regression models*, *generalized regression models* and *semiparametric regression models*.

We use the nonparametric regression

$$Y^{(k)}(x) = m(x) + \varepsilon, \tag{1}$$

where $Y^{(k)}$ is a dependent variable in the k-th considered model (transportation indicator of interest), $m(\circ)$ is an unknown regression function, x is a d-dimensional vector of independent variables (regressors), ε is a random term.

It is supposed that the random term has zero expectation ($E(\varepsilon) = 0$) and the variance $Var(\varepsilon) = \sigma^2 \psi(x)$ where σ^2 is an unknown constant and $\psi(x)$ is a known weighted function. Furthermore we have a sequence of independent observations $(Y_i^{(k)}, x_i)$, $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$, $i = 1, 2, ..., n$. On that base we need to estimate the unknown function $m(x)$.

In the simplest case *the linear regression model* is used that is described as

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

where $\beta_0, \beta_1, \beta_2, ..., \beta_d$ are unknown coefficients.

But such model gives good results very seldom. To improve the latter so called *generalized linear regression model* is applied:

$$m(x_i) = G(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}), \tag{3}$$

where $G(\circ)$ is the *known link function* of one dimensional variable.

Finally the *single index regression model* gives the best results:

$$m(x_i) = g\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_d x_{i,d}\right), \tag{4}$$

where $g(\circ)$ is an *unknown link function* of one dimensional variable. As $g(\circ)$ function the *kernel function* is considered usually [4].

4. Forecasting of air passenger conveyances

In the given research it was originally planned to consider all 25 EU countries, but at data gathering we have collided with shortage of data on many countries, especially it concerned new members of EU. Therefore, we have analyzed 11 EU countries, such as Austria, Belgium, Finland, France, Germany, Greece, Italy, Netherlands, Portugal, Spain and the United Kingdom.

We forecast the volumes of air passenger conveyances expressed in number of passengers carried during the years 1996 - 2005. For each country of each year we have the volumes of all factors mentioned above and the volumes of conveyances of air passenger carried, therefore 120 observations are valid. All the considered models are the group models.

The big number of various models differed among themselves both as structure of factors and their combinations has been constructed and investigated. In the given work three presented models are regression ones.

The first considered model is the simple linear regression model (2). The regressand $Y^{(1)} = m(x)$ is the number of air passenger carried, and the regressors are $x_1 = t_1$, $x_2 = \frac{t_2}{t_1}$, $x_3 = t_4$, $x_4 = t_5$, $x_5 = t_7$, $x_6 = t_9$.

In the second model as the regressand we considered the ratio between the total number of air passenger carried and the number of inhabitants of the country $Y^{(2)} = \frac{Y^{(1)}}{t_2}$, and as the regressors we used the following variables: $x_1 = t_1$, $x_2 = \frac{t_2}{t_1}$, $x_3 = t_4$, $x_4 = t_5$, $x_5 = t_6$, $x_6 = t_7$, $x_7 = t_9$.

For the third model the generalized linear regression model (3) has been used. Here the value of the regressand is the same with the regressand from the second model, i.e. $Y_i^{(3)} = \frac{Y^{(1)}}{t_2}$, and we used the same regressors as for the second model. The GLM is expressed in the following way:

$$m(x_i) = h_i \frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{1 + \exp\left(\sum_j \beta_j x_{i,j}\right)}, \; i = 1, 2, ..., n, \tag{5}$$

where h_i is the total population in the i-th country.

The unknown parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_k)'$, is estimated by the use of the least squares criterion:

$$R(\beta) = \sum_{i=1}^{n} \left(Y_i - \tilde{Y}_i\right)^2 \to \min_{\beta}, \tag{6}$$

where Y_i and $\tilde{Y}_i$ are observed and calculated values of Y.

Minimum of this criterion we find by means of the well-known gradient method. For this purpose we use the gradient with the respect to the unknown parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_k)'$:

$$\nabla R(\beta) = -2\sum_{i=1}^{n-1} \left(Y_i - h_i \frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{1 + \exp\left(\sum_j \beta_j x_{i,j}\right)} \right) \cdot h_i \cdot \frac{\exp\left(\sum_j \beta_j x_{i,j}\right)}{\left(1 + \exp\left(\sum_j \beta_j x_{i,j}\right)\right)^2} \cdot x_i \tag{7}$$

where x_i is vector-columns of the independent variables.

The corresponding computer program has been completed in MathCad 12.

Now allow us to consider the results of estimating and testing. The first model gives the following estimate for $E(Y)$:

$$\hat{E}\left(Y^{(1)}(x)\right) = 793.5 + 2.5x_1 + 3790.6\,x_2 + 345.5\,x_3 + 0.047\,x_4 - 25.6\,x_5 - 15.3\,x_6.$$

The estimated unknown parameters and Student criterion values for explanatory variables for the first model are resulted in the Table 1. The most significant explanatory variables are x_1, x_2, x_3 and x_6, so, the greatest influence on the air passenger conveyances renders: the area of the country, the density of the country population, the monthly labour coast, the gross domestic product and labour productivity per hour worked. The positive and the negative signs for all regressors in this model correspond to physical sense of regressors. The coefficient R^2 for this model is equal to 0.81 and the Fisher criterion is 61.82, so, this model is adequate (the significant level here and in further is 5%).

Table 1.

Results for the first model

Variable	Factor	B	t(92)	p-level
	Intercept	793.5	7.02719	0.000000
x_1	SQUARE	2.5	8.17163	0.000000
x_2	TP/SQUARE	3790.6	9.02128	0.000000
x_3	MLC	345.5	6.20054	0.000000
x_4	GDP	0.047	0.74657	0.457225
x_5	LPHW	-25.6	-8.78115	0.000000
x_6	UR	-15.3	-1.81660	0.072534

The second model has the following form:

$$\hat{E}\left(Y^{(2)}(x)\right) = 40.7 + 0.058\,x_1 + 96.4\,x_2 + 12.1\,x_3 + 0.013\,x_4 + 0.26\,x_5 - 0.5\,x_6 - 0.6\,x_7 .$$

The estimated parameters and Student criterion values for explanatory variables are shown in the Table 2. All explanatory variables in this model are significant and the signs for all regressors correspond to physical sense of regressors. The coefficient R^2 for this model is equal to 0.83 and the Fisher criterion is 73.46, so, this model is adequate.

Table 2.

Results for the second model

Variable	Factor	B	t(91)	p-level
	Intercept	40.7	10.5025	0.000000
x_1	SQUARE	0.058	9.2779	0.000000
x_2	TP/SQUARE	96.4	11.2493	0.000000
x_3	MLC	12.1	9.5535	0.000000
x_4	GDP	0.013	10.4362	0.000000
x_5	GDP_PPS	0.26	4.1199	0.000083
x_6	LPPHW	-0.5	-7.6185	0.000000
x_7	UR	-0.6	-3.2292	0.001727

The third model has the following form:

$$\hat{E}\left(Y^{(3)}(x)\right) = h\frac{\exp\left(944 - 10^{-3}x_1 + 2\times10^{-4}x_2 + 7\times10^{-3}x_3 + 0.035x_4 - 0.053x_5 + 0.041x_6 + 0.042x_7\right)}{1+\exp\left(944 - 10^{-3}x_1 + 2\times10^{-4}x_2 + 7\times10^{-3}x_3 + 0.035x_4 - 0.053x_5 + 0.041x_6 + 0.042x_7\right)}.$$

The observed and predicted values of air passenger conveniences are shown in the figures 1-3. We can see that some predicted values for the first model lies in the negative area. The second and the third models don't contain this inconvenience. Note that for latter ones the recalculated predicted values and residuals were analyzed, i.e. the obtained values were multiplied by the value of the country population.

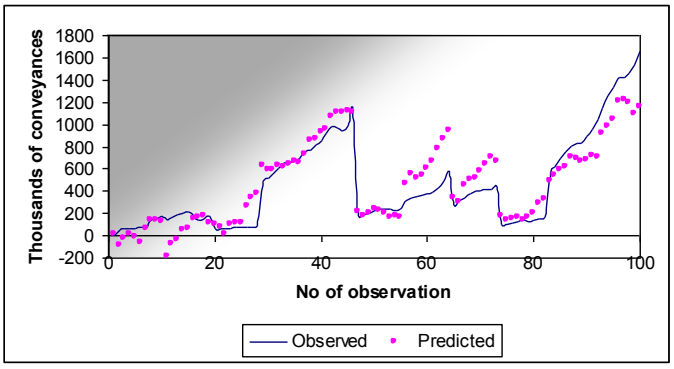

Figure 1. Plot of observed and predicted values for the first model in order to Country-Year

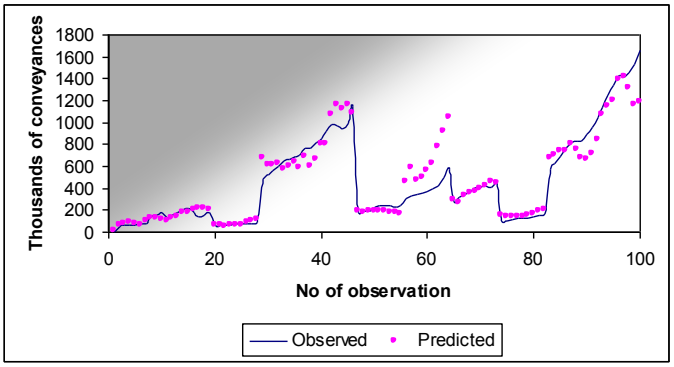

Figure 2. Plot of recalculated observed and predicted values
for the second model in order to Country-Year

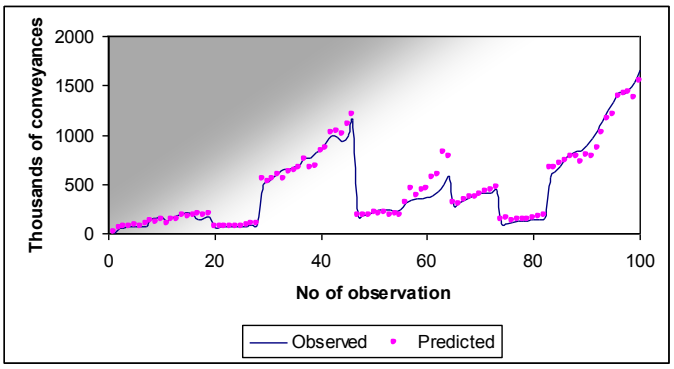

Figure 3. Plot of recalculated observed and predicted values
for the third model in order to Country-Year

The analysis of figures 2 and 3 allows concluding that the using of generalized linear regression model the improve results. Also we analyzed the plots of residuals for all the models. In the Table 3 the squared residuals sums for all considered models are shown.

Table 3.

Sums of the squared residuals for all models

	Residuals		
	1st model	2nd model	3rd model
R_0/n	2 781 745	1 675 399	539 206

We can conclude that the generalized linear regression model allows obtaining the best result among the considered models.

5. Forecasting of freight conveyances

We consider conveyances of rail freight transport, expressed in million tkm, for the set of countries-members of the European Union. Note that the suggested models are the group models, i.e. we forecast the volumes of conveyances for all the considered countries using the same sets of the explanatory variables and corresponding data. The following countries were selected: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom. The considered period is from 1996 to 2000.

So, for each country for each year we have the volumes of all factors mentioned above and the volumes of conveyances of rail freight transport too. So, for 15 countries and 5 years we have 75 observations.

Now let us describe three investigated regression models.

The first model is the simple linear regression model (2). The dependent variable $Y^{(1)} = m(x)$ is the raw conveyances of rail freight transport in millions tkm. Explanatory variables are $x_1 = t_6,\ x_2 = t_{10},\ x_3 = \frac{t_6}{t_{10}},\ x_4 = t_{11},\ x_5 = t_{12},\ x_6 = t_{13}$. The ratio $\frac{t_6}{t_{10}}$ enables us to see how these two factors in aggregate influence conveyances.

The second model is the modification of the previous one. The dependent variable $Y^{(2)} = \frac{Y^{(1)}}{\sqrt{t_1}}$ is the ratio between the raw conveyances and the square root of the country area. Explanatory factors are $x_1 = t_6,\ x_2 = t_{10},\ x_3 = \frac{t_6}{t_{10}},\ x_4 = t_{11},\ x_5 = t_{12},\ x_6 = t_{13},\ x_7 = t_{14}$. We introduce here the additional factor t_{14}, which is the index of the country area. This index was introduced in the model to consider gradation of the countries' areas. It is equal to 1 for relatively small countries with areas less or equal than 40 000 km^2, and it is equal to 0 for countries with areas bigger than 40000 km^2. As we can see, these models are parameter linear ones.

Finally we consider the Single Index Regression Model (4). Here the value of the dependent variable in the i-th observation $Y_i^{(3)} = \frac{Y^{(1)}}{t_1}$ is the ratio between the volumes of conveyances and the country area for concrete year. The set of explanatory variables coincides with the set from the first model.

So, we have three regression models and our aim is to estimate unknown parameters and to test the correctness of models. It allows us to choose the best model.

The semiparametric model (4) contains the *unknown link function* $g(\tau)$, where τ is so called *index*: $\tau = x^T\beta$. The *Nadaraya-Watson estimator* [4] for estimating $g(\tau)$ is used:

$$\tilde{g}(\tau_i) = \frac{1}{\sum_{i=1}^{n} K_h(\tau_i)} \sum_{i=1}^{n} K_h(\tau_i) Y_i\,, \tag{8}$$

where $\tau_i = (x - x_i)^T\beta$ is the value of index for the i-th observation, β is vector of unknown parameters, $K_h(\circ)$ is so called *kernel function*.

As $K_h(\circ)$ we use the Gaussian function:

$$K_h(\tau_i) = \begin{cases} \frac{1}{c\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\tau_i}{h}\right)^2\right), & |x - x_i| < h; \\ 0, & |x - x_i| \geq h, \end{cases} \tag{9}$$

where ***h*** is a *bandwidth* and c is normalising constant.

The unknown parameter vector $\beta = (\beta_1, \beta_2, ..., \beta_6)'$ is estimated by use of the least squares criterion:

$$R(\beta) = \sum_{i=1}^{n} (Y_i - \tilde{g}(\tau_i))^2 \to \min_{\beta} . \quad (10)$$

For that we use the gradient method. The corresponding gradient is the following:

$$\nabla R(\beta) = -2\sum_{i=1}^{n} \left(Y_i - \frac{\sum_{i=1}^{n} K_h(\tau_i) Y_i}{\sum_{i=1}^{n} K_h(\tau_i)} \right) \cdot \frac{\left(\sum_{i=1}^{n} Y_i \frac{\partial}{\partial \tau_i} K_h(\tau_i) \cdot \frac{x_i}{h} \right) \cdot K_h(\tau_i) - \sum_{i=1}^{n} K_h(\tau_i) Y_i \cdot \left(\sum_{i=1}^{n} \frac{\partial}{\partial \tau_i} K_h(\tau_i) \cdot \frac{x_i}{h} \right)}{\left(\sum_{i=1}^{n} K_h(\tau_i) \right)^2} , \quad (11)$$

where

$$\frac{\partial}{\partial \tau_i} K_h(\tau_i) = -\frac{\tau_i}{c\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\tau_i}{h} \right)^2 \right) \quad (12)$$

is the derivative of the Gaussian function.

The first model can be written in the following form:

$$\hat{E}\left(Y^{(1)}(x)\right) = -3\,713 + 118x_1 + 26x_2 - 11\,769x_3 + 0.9x_4 + 0.5x_5 + 0.2x_6 .$$

The estimates of the parameters and Student criterion for the first model are resulted in Table 4. The most significant explanatory variables are x_4 and x_6, so, the greatest influence on conveyances renders the total length of railways and the number of wagons. The positive sign for these variables corresponds to physical sense of the regressors. The coefficient R^2 for this model is equal to 0.97 and the Fisher criterion is 383.69, so, this model is adequate.

Table 4.

Results for the first model

Variable	Factor	B	t(68)	p-level
	Intercept	-3 713	-0.149195	0.881842
x_1	GDP	118	0.480762	0.632229
x_2	CPL	26	0.109604	0.913046
x_3	GDP/CPL	-11 769	-0.462115	0.645474
x_4	TOTLEN	0.9	6.866741	0.000000
x_5	LOKOM	0.5	0.799173	0.426973
x_6	WAGONS	0.2	8.375650	0.000000

The second model is the following:

$$\hat{E}\left(Y^{(2)}(x)\right) = -120 - 1.2x_1 + 1.4x_2 + 110x_3 + 8\times10^{-5}x_4 + 6\times10^{-3}x_5 + 6\times10^{-4}x_6 + 29x_7 .$$

The results for the second model are presented in the Table 5. 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 raw conveyances and the square 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 R^2 for this model is equal to 0.97 and the Fisher criterion is 313.00, so, this regression is significant.

Table 5.

Results for the second model

Variable	Factor	B	t(67)	p-level
	Intercept	-120	-3.00514	0.003732
x_1	GDP	-1.2	-3.11117	0.002738
x_2	CPL	1.4	3.55818	0.000692
x_3	GDP/CPL	110	2.68390	0.009160
x_4	TOTLEN	8×10^{-5}	1.03172	0.305913
x_5	LOKOM	6×10^{-3}	5.42836	0.000001
x_6	WAGONS	6×10^{-4}	9.33665	0.000000
x_7	GRAD	29	12.79621	0.000000

The third model has the following form:

$$\hat{E}\left(Y^{(3)}(x)\right)=\frac{\sum_{i=1}^{n} Y_i K_h\left(10^{-5}+0.02\left(x-x_{1,i}\right)-2.9\times10^{-5}\left(x-x_{2,i}\right)+2.2\times10^{-5}\left(x-x_{3,i}\right)+0.7\left(x-x_{4,i}\right)+0.08\left(x-x_{5,i}\right)-0.7\left(x-x_{6,i}\right)\right)}{\sum_{i=1}^{n} K_h\left(10^{-5}+0.02\left(x-x_{1,i}\right)-2.9\times10^{-5}\left(x-x_{2,i}\right)+2.2\times10^{-5}\left(x-x_{3,i}\right)+0.7\left(x-x_{4,i}\right)+0.08\left(x-x_{5,i}\right)-0.7\left(x-x_{6,i}\right)\right)}.$$

Figures 4-6 demonstrate the way the investigated models smooth the observations. It is obvious that the third model shows the best smoothing.

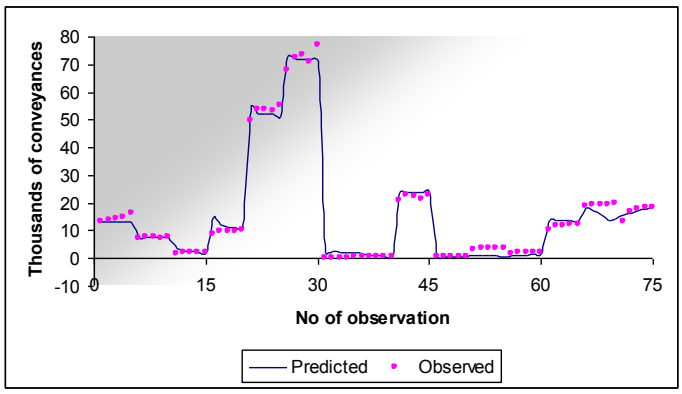

Figure 4. Smoothing for the first model

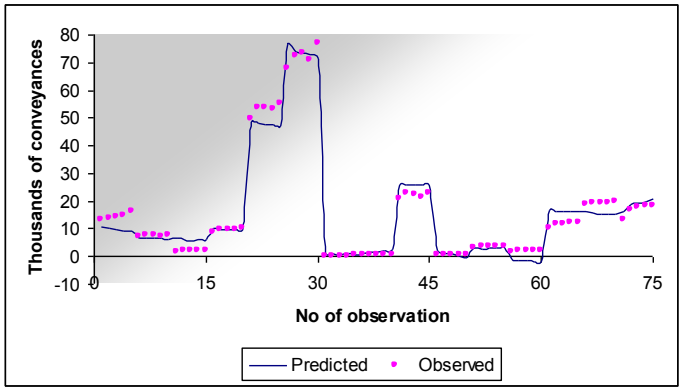

Figure 5. Smoothing for the second model

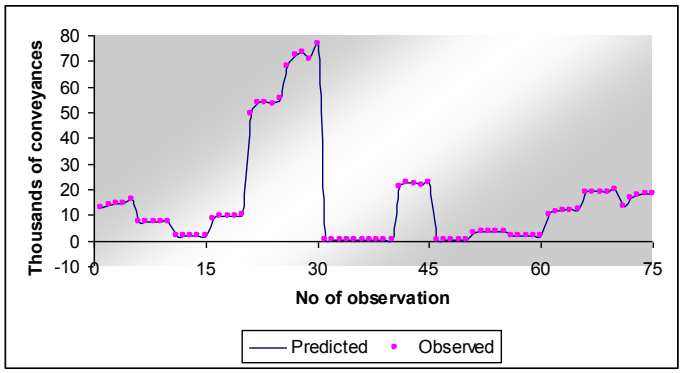

Figure 6. Smoothing for the third model

The sums of squared residuals for the all models in the Table 6 were calculated as well.

Table 6.

Sums of the squared residuals for all models

	Residuals		
	1st model	2nd model	3rd model
R_0/n	11 543 065	4 830 576	894 265 565 407

After analysing the obtained results we can conclude that the single index model is the most significant.

6. Conclusions

In the presented paper various models for the forecasting of passenger and freight conveyances are considered: linear regression model, generalized linear regression model and single index regression model. Efficiency of these models is investigated by the consideration of conveyances for the Europe Union countries. The advantage of considered models comparing with the models presented in such papers as [5] (autoregression integrated moving average models) and [2, 3, 6] (multiple regression models) consists in including the greater number of the used factors. Moreover the performed investigations show that the generalized

linear regression model and the single index regression model give more exact forecasts than classical methods of linear regression.

7. References

1. Andronov A.M. etc. Forecasting of the air passenger conveyances on the air transport. // Transport, Moscow, 1983. (In Russian).
2. Butkevicius J., Mazura M., Ivankovas V., Mazura S. Analysis and forecast of the dynamics of passenger transportation by public land transport. In TRANSPORT – 2004, Vol XIX, No 1, 3-8.
3. Butkevicius J., Vyskupaitis A. Development of passenger transportation by Lithuanian sea transport. // In Proceedings of International Conference RelStat'04, Transport and Telecommunication, Vol.6. N 2, 2005.
4. Hardle W., Muller M., Sperlich S., Werwatz A. Nonparametric and Semiparametric Models. // Springer, Berlin, 2004.
5. Hunt U. Forecasting of railway freight volume: approach of Estonian railway to arise efficiency. // In TRANSPORT – 2003, Vol XXVIII, No 6, 255-258.
6. Sliupas T. Annual average daily traffic forecasting using different techniques. // In TRANSPORT – 2006, Vol XXI, No 1, 38-43.
7. EUROSTAT YEARBOOK 2005. The statistical guide to Europe. Data 1993–2004. EU, EuroSTAT, 2005. URL: http://epp.eurostat.ec.europa.eu

Alexander Andronov, Dr.sc.ing. Professor of the Department of Transport System Management Mathematical Support at Riga Technical University. His professional interests include Stochastic Processes, Mathematical Statistic, Optimisation Methods, Discrete Mathematics and their applications. He is the author of about 200 scientific publications and 10 books. Alexander Andronov is a member of American Statistical Association, e-mail: Aleksandrs.Andronovs@rtu.lv

Catherine Zhukovskaya, M.sc.ing., the post graduate student of the Institute of Transport Vehicles Technology at Riga Technical University, e-mail: kat_zuk@hotmail.com

Diana Santalova, M.sc.ing., the post graduate student of the Institute of Transport Vehicles Technology at Riga Technical University, e-mail: Diana.Santalova@rtu.lv

Abstract. *New methods of the forecasting of passenger and freight conveyances are considered. These methods are based on generalized linear regression model and single index regression model. Elaborated methods are used for the forecasting of the conveyances for the Europe Union countries. The performed investigations show that the suggested methods give more exact forecasts than the classical methods of linear regression.*

Abstrakts. *Jaunas metodes pasažieru un kravu pārvadājumu prognozēšanai ir apskatītas. Metodes ir balstītas uz vispārināta lineāras regresijas modeļa un uz viena indeksa regresijas modeļa. Izstrādātas metodes ir pielietotas Eiropas Savienības valstu pārvadājumu prognozēšanai. Veiktie pētījumi parādā, ka metodes dod precīzākos rezultātus, nekā klasiskās lineāras regresijas metodes.*

Абстракт. *Рассматриваются новые методы прогнозирования перевозок пассажиров и грузов. Методы базируются на обобщённой модели линейной регрессии и на одно-индексной модели регрессии. Разработанные методы применяются для прогнозирования перевозок для стран Европейского союза. Проведенные исследования показывают, что предлагаемые методы дают более точные прогнозы, чем классические методы линейной регрессии.*