*Computer Modelling and New Technologies, 2007, Vol.11, No.1, 73-83*
*Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia*

# FORECASTING OF RAIL FREIGHT CONVEYANCES IN EU COUNTRIES ON THE BASE OF THE SINGLE INDEX MODEL

***Diana Santalova***

*Riga Technical University, Faculty of Transport and Mechanical Engineering*
*Kalku Str. 1, Riga, LV-1658, Latvia*
*E-mail: Diana.Santalova@rtu.lv*

There are the regression models which describe rail freight conveyances of the member countries of the European Union considered in the investigation. The models contain such factors for each country as: total length of railways, gross domestic product per capita in Purchasing Power Standards and so on. All calculations are performed on the basis of the statistical data taken from EUROSTAT YEARBOOK 2005. Two estimation approaches are compared: the classical linear regression model and the single index model. Various tests for hypothesis of explanatory variables insignificance and model correctness have been carried out, and the cross-validation approach has been applied as well. The analysis has shown obvious advantage of the single index model.

***Keywords:*** *freight conveyances, forecasting, single index model*

## 1. Introduction

In this paper we consider the problem of forecasting of rail freight conveyances from the member countries of the European Union on the basis of EUROSTAT YEARBOOK 2005 data [5]. For that the *linear regression model* [4] and the *single index model* (SIM) [3] are used. The *object* of consideration is rail freight conveyance expressed in million tonne-kilometres. We call *observation* the data about an object for a concrete year from 1996 till 2000. The following countries were considered: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom. The main difficulty is to choose the set of convenient factors influencing the rail freight conveyances. 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 the ones, that give the best forecasts of conveyances. We use the following well known criteria for comparing the elaborating models: the coefficient of multiple determination $R^2$, Fisher's and Student's criteria and the residual sum of squares $R_0$ [1, 4]. The described below 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 paper a lot of attention is paid to this problem.

The paper is organized in a following way. First of all the used regression models are considered from theoretical point of view, then the used data are described. After that we consider the suggested group models for the forecasting of conveyances. The results of the carried out estimation and the comparative analysis of these models are presented as well.

## 2. Structure of the Used Models

In this research all investigated models are group models [1]. The main object of consideration is named an *object*. It is a freight conveyance from some EU country. The data about an object for a definite period of time is called *observation*. 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. In other words we are able to forecast rail freight conveyances for all considered countries using one and the same model.

With respect to used mathematical model we consider *linear regression models* and *semiparametric regression models*.

In general the regression model can be described as

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

where $Y_i$ is a dependent variable in the *i*-th observation, $m(\circ)$ is an unknown regression function, $x_i$ is a $d$ -dimensional vector of independent variables, $\varepsilon_i$ 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 $\sigma^2$ is an unknown constant and $\psi(x)$ is a known weighted function. Furthermore we have a sequence of independent observations $(Y_i, x_i)$, $x_i=(x_{i,1}, x_{i,2}, ..., x_{i,d})$, $i=1, 2, ..., n$. On that base we need to estimate the unknown function $m(x)$.

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

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

where $\beta^T=(\beta_0 \quad \beta_1 \quad \dots \quad \beta_d)$ is vector of unknown coefficients, $x_i=(1 \quad x_{i,1} \quad \dots \quad x_{i,d})^T$ is a vector of independent variables in *i*-th observation.

As it is known the forecasts obtained using the linear regression models are not very good. So, for rail conveyances forecasting we use the *single index regression model* [3] as well:

$$m(x_i)=g(\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+...+\beta_d x_{i,d})=g(\beta^T x_i), \qquad (3)$$

where $g(\circ)$ is an *unknown link function* of one dimensional variable and $\tau_i=\beta^T x_i$ is called an *index*.

## 3. Informative Base

For experiments we will use the below-described statistical data. All necessary data have been received from “The Statistical Office of the European Communities” electronic database (EUROSTAT) [5]. First of all, the variable of interest is the rail freight conveyance, expressed in million tonne-kilometres. Let us denote it by $t_0$.

The following factors have been selected as explanatory variables:

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

$t_2$ – Gross Domestic Product per capita in Purchasing Power Standards;

$t_3$ – comparative price level;

$t_4$ – total length of railways, in thousands of km;

$t_5$ – number of locomotives, in thousands;

$t_6$ – number of goods wagons, in thousands.

Let us comment on some of the described factors.

Gross Domestic Product is a measure for the economic activity. It is defined as the value of all goods and services produced less the value of any goods or services used in their creation. The volume index of GDP per capita in Purchasing Power Standards (PPS) for each country is expressed in relation to the European Union (EU-25) average set to equal 100.

Comparative price level is the ratio between Purchasing Power Parities (PPPs) and market exchange rate for each country.

## 4. Considered Models

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

The first model is a simple linear regression model (2). The dependent variable $Y^{(L1)}=t_0$ is conveyance of rail freight transport in millions tonne-kilometres. Note, that superscript by *Y* is introduced just for identification of models. Explanatory variables are $x_1=t_2$, $x_2=t_3$, $x_3=\frac{t_2}{t_3}$, $x_4=t_4$, $x_5=t_5$, $x_6=t_6$. The ratio $\frac{t_2}{t_3}$ enables us to see how these two factors in aggregate influence conveyances.

The second model is modification of the previous one. The dependent variable $Y^{(L2)}=\frac{t_0}{\sqrt{t_1}}$ is the ratio between the conveyance and the square root of the country area. Explanatory factors are $x_1=t_2, x_2=t_3, x_3=\frac{t_2}{t_3}, x_4=t_4, x_5=t_5, x_6=t_6$. In addition we introduce here the factor $t_7$, which is the index of the country area, by which we are able to consider gradation of the countries' areas. It is equal to 1 for relatively small countries (with areas less than 40 000 km$^2$ 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, because the areas of these countries are smaller than 40 000 km$^2$.

Finally we consider two variants of the Single Index Model (3). In the first variant the value of the dependent variable $Y^{(SIM1)} = \frac{t_0}{t_1}$ is the ratio between the conveyance and the country area for a concrete year. In the second variant the dependent variable $Y^{(SIM2)} = \frac{t_0}{\sqrt{t_1}}$ coincides with a dependent variable from the second linear Model L2.

The sets of explanatory variables for the models *SIM*1 and *SIM*2 coincide with the set for the first linear Model L1.

Thus, we have four regression models. Our task is to estimate the unknown coefficients $\beta$ for the models, to compare the suggested models and to choose the best ones taking in account their significance. All calculations are performed using Statistica 6.0 and MathCad 12 packages.

## 5. Estimation of the Linear Models

Firstly, we analyse all the suggested models in case of data smoothing. It means we estimate the unknown coefficients $\beta$ by all the observations. Thus, we are able to evaluate, how the considered models can only smooth the known conveyances and what variables have the greatest influence upon the conveyances.

Let us describe the obtained results.

The estimated Model L1 has the following form:

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

The estimates of the coefficients and calculated values of the Student's criterion for the Model L1 are presented in Table 1. Here $\tilde{\beta}_i$ is an estimate of $\beta_i$, *t*(68) is the calculated value of Student's criteria for 68 degrees of freedom, *p*-level is the error of the second kind (or level of insignificance of variable). The theoretical value of Student's criterion for 68 degrees of freedom and level of significance (or error of the first kind) α = 5% is equal to 1.67. Taking into account the fact that the hypothesis of *insignificance* of explanatory variable is tested, we can see that calculated value of Student's criterion exceeds its theoretical value for two variables only, i.e. these two variables cannot be recognized as insignificant. Thus, the most significant explanatory variables are $x_4$ and $x_6$, so, the greatest influence on conveyances is rendered by the total length of railways and the number of wagons. The positive sign for these variables corresponds to the physical sense of the regressors. The coefficient $R^2$ for this model is equal to 0.985 and the calculated value of Fisher's criterion is 383.69. The theoretical value of Fisher's 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's criterion we can conclude that the estimated Model L1 cannot be recognized as insignificant. So, Model L1 is adequate.

TABLE 1. Estimates of coefficients of Model L1 and their insignificance levels

| **Coefficients** | $\tilde{\beta}_i$ | ***t*(68)** | ***p*-level** |
|---|---|---|---|
| $\beta_0$ | -3 713 | 0.149195 | 0.881842 |
| $\beta_1$ | 118 | 0.480762 | 0.632229 |
| $\beta_2$ | 26 | 0.109604 | 0.913046 |
| $\beta_3$ | -11 769 | -0.462115 | 0.645474 |
| $\beta_4$ | 879 | 6.866741 | 0.000000 |
| $\beta_5$ | 549 | 0.799173 | 0.426973 |
| $\beta_6$ | 158 | 8.375650 | 0.000000 |

The estimated Model L2 is as follows:

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

The results of the analysis of Model L2 are presented in the Table 2. As we can see, almost all explanatory variables are recognized to be significant by Student's criterion. Only total length of railways does not influence the dependent variable. We obtain the positive signs for all significant variables with the exception of GDP; that means the positive correlation between these explanatory variables and the dependent variable. The coefficient $R^2$ for this model is equal to 0.985 and the calculated value of Fisher's criterion is 313.78. The theoretical value of Fisher's criterion for 7 and 67 degrees of freedom and level of significance α = 5% is 2.15, so, this regression model is significant as well.

TABLE 2. Estimates of coefficients of Model L2 and their insignificance levels

| **Coefficients** | $\tilde{\beta}_i$ | ***t*(68)** | ***p*-level** |
|---|---|---|---|
| $\beta_0$ | -120.4 | -3.00514 | 0.003732 |
| $\beta_1$ | -1.2 | -3.11117 | 0.002738 |
| $\beta_2$ | 1.4 | 3.55818 | 0.000692 |
| $\beta_3$ | 110.2 | 2.68390 | 0.009160 |
| $\beta_4$ | 0.2 | 1.03172 | 0.305913 |
| $\beta_5$ | 5.9 | 5.42836 | 0.000001 |
| $\beta_6$ | 0.3 | 9.33665 | 0.000000 |
| $\beta_7$ | 29.2 | 12.79621 | 0.000000 |

Figures 1 and 2 demonstrate how the investigated models smooth the observed true data. The observations are arranged in "country-year" order: every five points correspond to conveyances of 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 conveyances, expressed in thousands. It is obvious that both linear models show the similar smoothing.

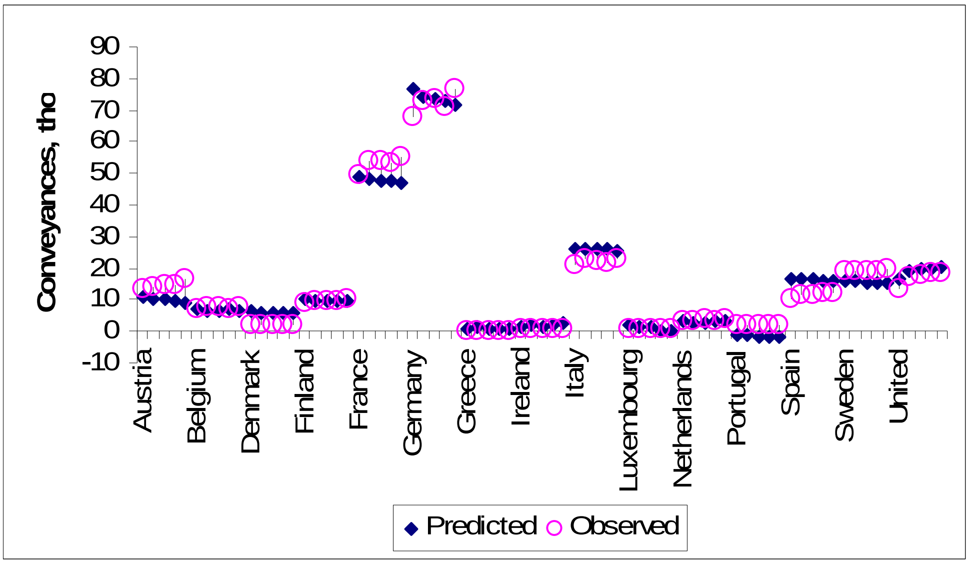

*Figure 1.* Smoothing by Model L1

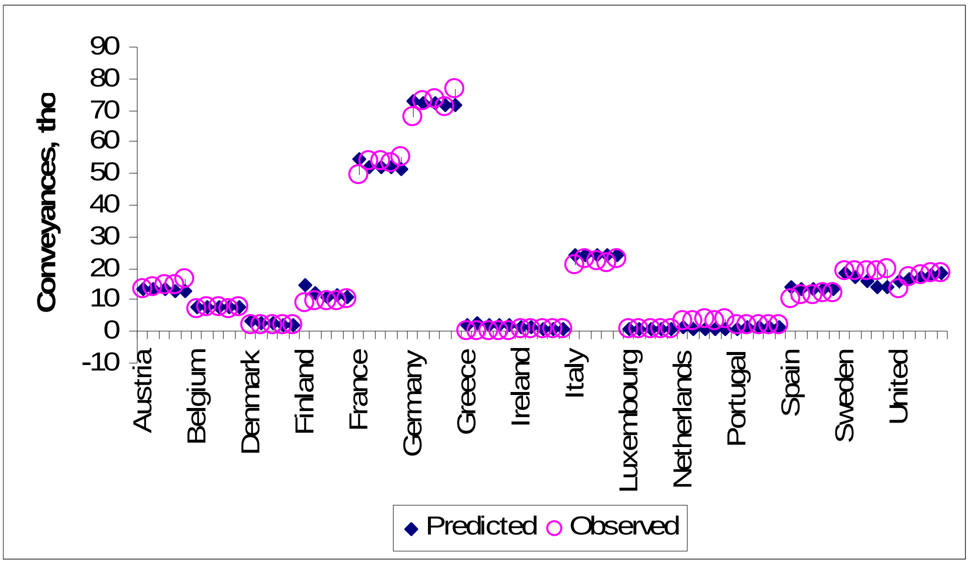

*Figure 2.* Smoothing by Model L2

## 6. Estimation of the Single Index Models

Now we will consider the suggested single index Models *SIM*1 and *SIM*2.

The estimation of these models consists of two steps: we have to estimate the unknown coefficients vector $\beta$ and the link function $g$. For the latter the Nadaraya-Watson kernel estimator can be applied [3]:

$$\tilde{g}(x) = \frac{1}{\sum_{i=1}^{n} K_h(\tau_i)} \sum_{i=1}^{n} K_h(\tau_i) Y_i \,, \tag{4}$$

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 $K_h(\circ)$ is the so-called *kernel function*.

We use the Gaussian function as $K_h(\circ)$:

$$K_h(\tau) = \frac{1}{h\sqrt{2\pi}} \exp\left( -\frac{1}{2}\left(\frac{\tau}{h}\right)^2 \right), \qquad -\infty < \tau < \infty, \tag{5}$$

where $h$ is a *bandwidth*.

The unknown parameter vector $\beta$ is estimated using the least squares criterion:

$$R(\beta) = \sum_{i=1}^{n} (Y_i - \tilde{g}(x_i))^2 \to \min_{\beta} \,. \tag{6}$$

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

$$\nabla R(\beta) = -2\sum_{i=1}^{n} \left( Y_i - \frac{\sum_{i=1}^{n} K_h(\tau_i) Y_i}{\sum_{i=1}^{n} K_h(\tau_i)} \right) \cdot \left( \sum_{i=1}^{n} K_h(\tau_i) \right)^{-2} \cdot \left( \frac{1}{h} \sum_{i=1}^{n} Y_i \frac{\partial}{\partial \tau_i} K_h(\tau_i) \cdot (Y_i - \tilde{Y}_i) \cdot x_i \right), \tag{7}$$

where

$$\widetilde{Y}_i = \sum_{j=1}^{n} K_h(\tau_j) Y_j \tag{8}$$

and

$$\frac{\partial}{\partial \tau_i} K_h(\tau_i) = -\frac{\tau_i}{h^2\sqrt{2\pi}} \exp\left( -\frac{1}{2}\left(\frac{\tau_i}{h}\right)^2 \right) \tag{9}$$

is the derivative of the Gaussian kernel.

We are able to compare single index models by the residual sum of squares $R_0$ only. We calculate the residual sum of squares as follows:

$$R_0 = \frac{1}{n-d}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \tag{10}$$

where $n$ is a number of observations, $d$ is a number of estimated coefficients, $Y_i$ is an observed value and $\hat{Y}_i = \widetilde{g}(x_i)$ is an estimated value.

The estimates of coefficients $\beta$ , i.e. the values of coefficients $\beta$ optimizing the object function (6), for both single index models have been obtained from the same starting point $\beta^{(0)}$ and with bandwidth $h = 7$ for *SIM*1 and $h = 6$ for *SIM*2. 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 MathCad12 package.

The estimated Model *SIM*1 has the following form:

$$\widehat{E}\left(Y^{(SIM1)}(x)\right) = \frac{\sum_{i=1}^{n} Y_i K_h\left(710(x - x_{1,i}) - 1\times10^3(x - x_{2,i}) + 18\times10^{-5}(x - x_{3,i}) + 758(x - x_{4,i}) + 155(x - x_{5,i}) - 2\times10^3(x - x_{6,i})\right)}{\sum_{i=1}^{n} K_h\left(710(x - x_{1,i}) - 1\times10^3(x - x_{2,i}) + 18\times10^{-5}(x - x_{3,i}) + 758(x - x_{4,i}) + 155(x - x_{5,i}) - 2\times10^3(x - x_{6,i})\right)}.$$

The estimated Model *SIM*2 can be written in the following way:

$$\widehat{E}\left(Y^{(SIM2)}(x)\right) = \frac{\sum_{i=1}^{n} Y_i K_h\left(716(x - x_{1,i}) - 1\times10^3(x - x_{2,i}) - 4(x - x_{3,i}) + 853(x - x_{4,i}) + 62(x - x_{5,i}) - 871(x - x_{6,i})\right)}{\sum_{i=1}^{n} K_h\left(716(x - x_{1,i}) - 1\times10^3(x - x_{2,i}) - 4(x - x_{3,i}) + 853(x - x_{4,i}) + 62(x - x_{5,i}) - 871(x - x_{6,i})\right)}.$$

Figures 3 and 4 represent smoothing by these models. Obviously, the estimates of conveyances almost in all observations coincide with the true conveyances.

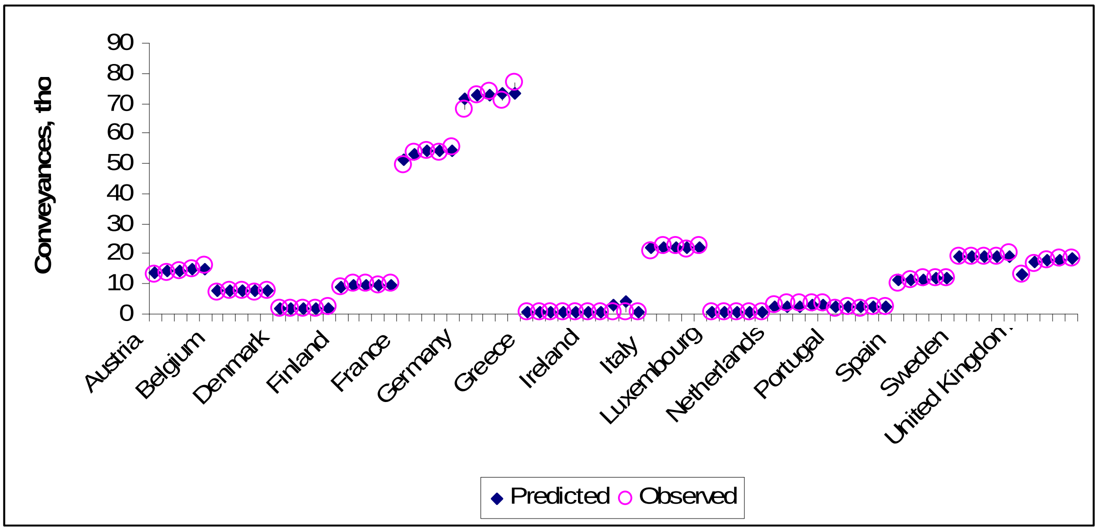

*Figure 3.* Smoothing by Model SIM1

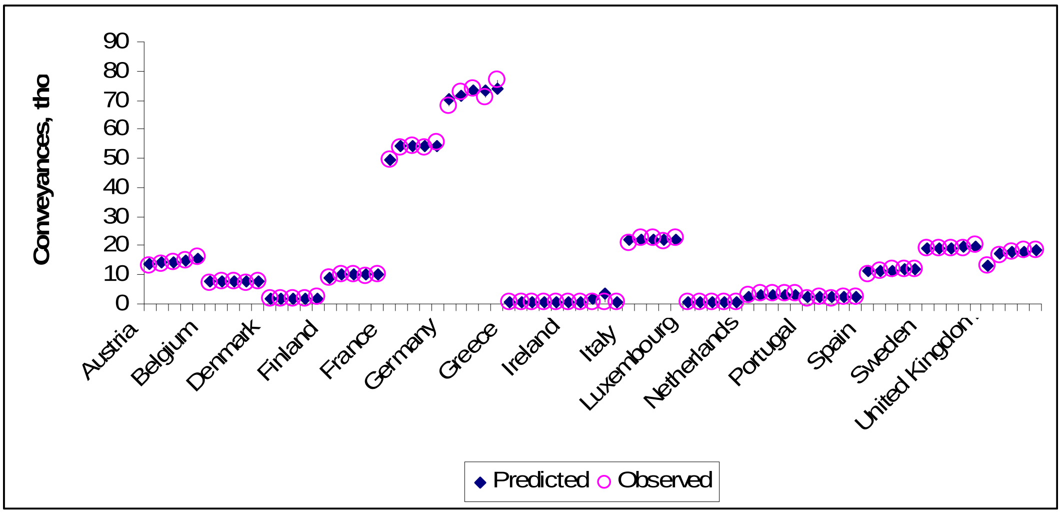

*Figure 4*. Smoothing by Model SIM2

As both linear models and both single index models give approximately similar results in data smoothing, we have to consider how precise the forecasts are which are given by analysing models. For this purpose we use the residual sum of squares $R_0$ (10). Table 3 involves the values of the residual sums of squares for all the models.

TABLE 3. Values of $R_0$ in case of smoothing

| Model | *L*1 | *L2* | *SIM*1 | *SIM*2 |
|---|---|---|---|---|
| $R_0$ | 11 543 065 | 4 830 576 | 894 265 | 565 407 |

So we can conclude that the linear Model L2 and the single index Model *SIM*2 have the minimum value of $R_0$ that means greater significance of these models in comparison with two others. As it was supposed, in general SIM gives the most precise estimates.

## 7. Cross-Validation Analysis

Now we will consider the suggested models from the other point of view. We use the *cross-validation approach*. That means we estimate the unknown coefficients $\beta$ for the models on the basis of a part of the data. Then using the obtained estimates of $\beta$ we forecast the conveyances for a remained part of the data and compare these forecasted conveyances with the real ones, i.e. we calculate $R_0$ for each model. Also the optimum value of bandwidth *h* is found for both single index models.

We estimate the coefficients $\beta$ on the basis of the period from 1996 till 1999 and perform the forecast for the year 2000. Table 4 contains the estimates of $\beta$ for the considered linear regression models. The signs of estimates correspond to physical sense of explanatory factors.

TABLE 4. Estimates of coefficients for the linear models

| Coefficients | *L*1 | *L*2 |
|---|---|---|
| $\beta_0$ | 4 154.4 | -119.7 |
| $\beta_1$ | 197.7 | 28.7 |
| $\beta_2$ | 45.4 | -1.2 |
| $\beta_3$ | -20 555.8 | 1.4 |
| $\beta_4$ | 898.5 | 110.0 |
| $\beta_5$ | 531.6 | 0.2 |
| $\beta_6$ | 148.1 | 6.0 |
| $\beta_7$ | – | 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 rail freight conveyances 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 conveyances. The true observed values of conveyances and the corresponding forecasts are displayed in Figures 5 and 6. We can see that Model L2 is more sensitive to the small conveyances which belong to the countries with small areas. Obviously this effect is achieved by using the above-mentioned additional gradation factor.

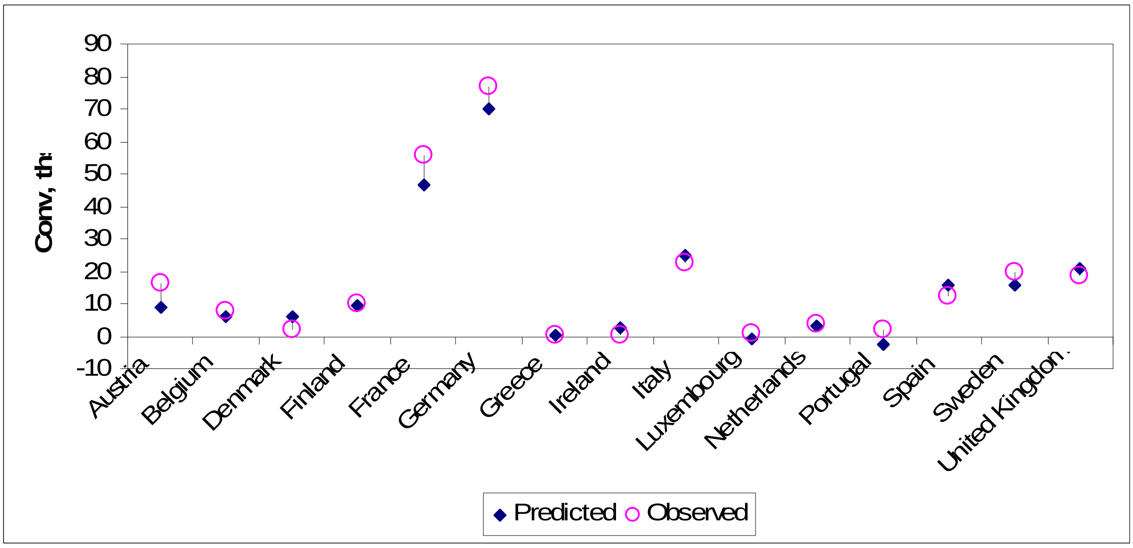

*Figure 5.* Forecasting by Model L1

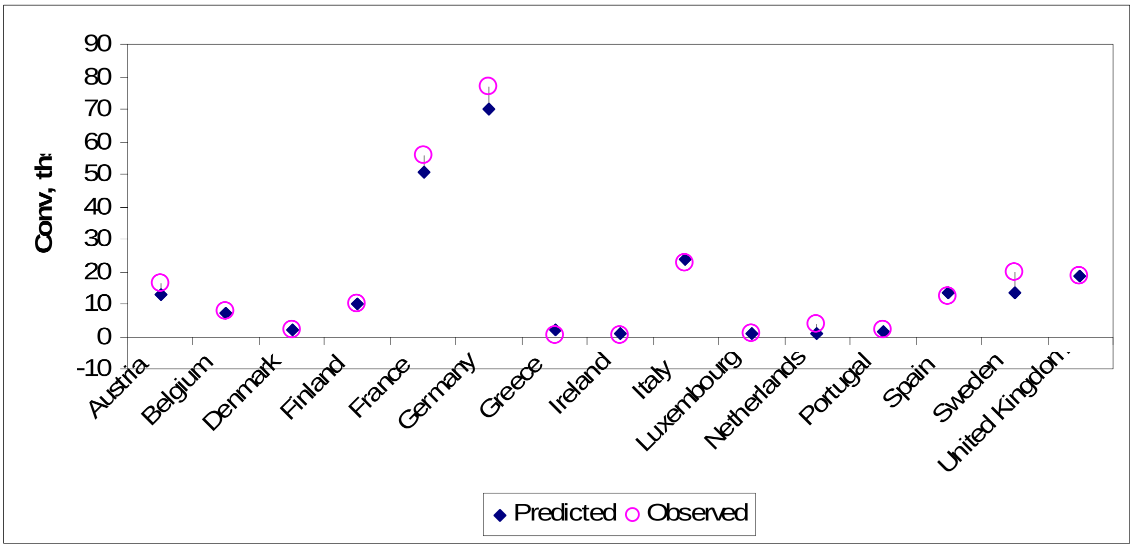

*Figure 6.* Forecasting by Model L2

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 $R_0$ (see [3]). The series of experiments was performed and the different estimates of $\beta$ and values of $R_0$ depending on various $h$ were obtained as well. The corresponding results for Models *SIM*1 and *SIM*2 are shown in Tables 5 and 6 respectively. We can see that all $\beta$ estimates differ from each other depending on $h$ in spite of the fact that they were obtained from the same initial value $\beta_0$. The values of $R_0$ (expressed in millions) corresponding to various $h$ for both SIM are represented in Table 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 7 and 8, respectively.

TABLE 5. The estimates of $\beta$ for *SIM*1

| Coefficients | Bandwidth ***h*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 6 | 7 | 8 | 9 | 10 | 15 | 20 |
| $\beta_1$ | 22.8 | 566.4 | 248.6 | 33 160.0 | 344.5 | 721.2 | 3 530.0 | 1 327.0 | -901.7 |
| $\beta_2$ | 26.6 | 299.9 | 216.7 | -22 420.0 | -95.8 | -24.1 | -512.5 | 358.2 | 1 304.0 |
| $\beta_3$ | -0.04 | 1.9 | -0.03 | 565.5 | 4.4 | 7.2 | 38.4 | 8.6 | -19.7 |
| $\beta_4$ | 0.13 | 257.9 | 88.9 | 19 870.0 | 120.5 | 207.1 | 1 310.0 | 550.0 | 1 011.0 |
| $\beta_5$ | 4x10$^{-5}$ | 62.9 | 17.9 | 3 996.0 | 25.3 | 37.7 | 174.4 | 119.2 | 198.3 |
| $\beta_6$ | 1x10$^{-5}$ | 885.7 | 252.2 | 56 310.0 | 356.9 | 572.7 | 3 832.0 | 3 058.0 | 724.6 |

TABLE 6. The estimates of $\beta$ for *SIM*2

| Coefficients | Bandwidth ***h*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 6 | 7 | 8 | 9 | 10 | 15 | 20 |
| $\beta_1$ | 29.3 | 859.9 | 962.7 | 1.7x10$^{3}$ | 1.2x10$^{3}$ | 697.5 | 618.7 | 757.1 | -4.5x10$^{3}$ |
| $\beta_2$ | 18.3 | 462.3 | 1.2x10$^{3}$ | 1.0x10$^{3}$ | 654.8 | 578.5 | 276.8 | 791.3 | 3.4x10$^{3}$ |
| $\beta_3$ | 0.1 | 3.4 | -2.4 | 7.5 | 6.0 | 1.9 | 3.3 | -0.2 | -73.6 |
| $\beta_4$ | -0.2 | 440.7 | 604.3 | 1.2x10$^{3}$ | 636.6 | 664.1 | 525.1 | 737.1 | 4.0E+3 |
| $\beta_5$ | 4x10$^{-5}$ | 103.7 | 49.0 | 97.1 | 61.0 | 24.1 | 12.1 | 17.3 | 916.8 |
| $\beta_6$ | 1x10$^{-5}$ | 1.5x10$^{3}$ | 690.0 | 1.4x10$^{3}$ | 859.7 | 851.7 | 672.9 | 1.2x10$^{3}$ | 1.9x10$^{4}$ |

TABLE 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 |

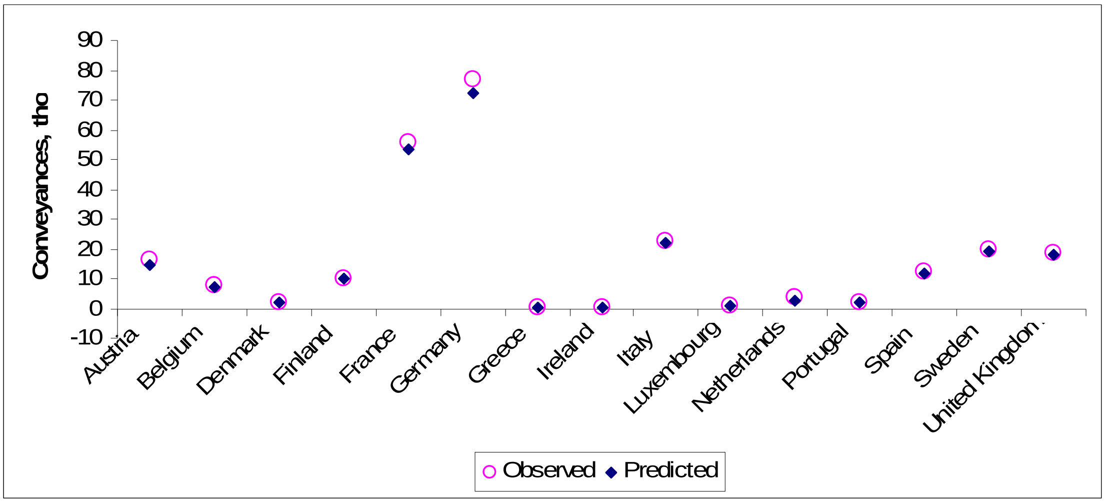

*Figure 7*. Forecasting by SIM1

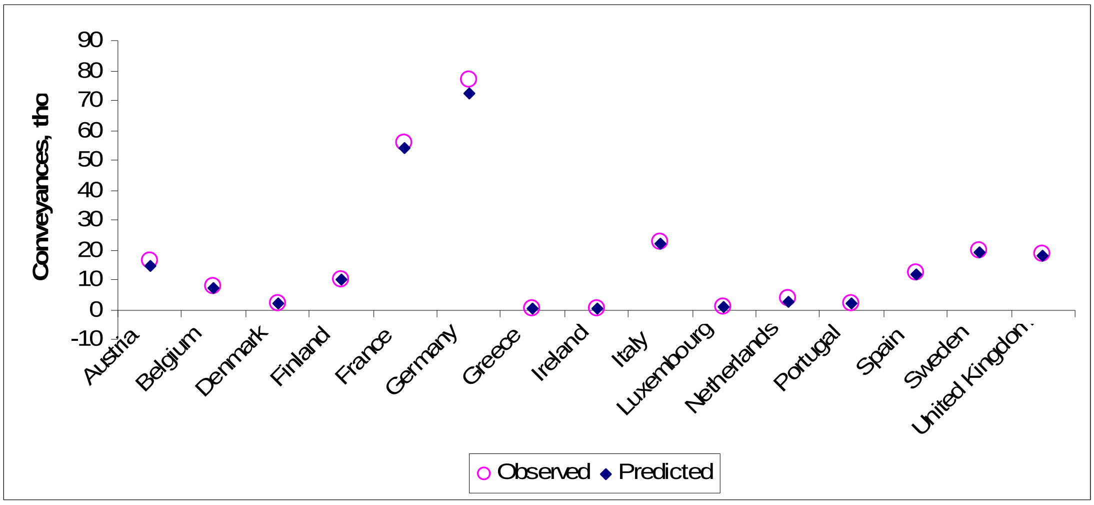

*Figure 8*. Forecasting by SIM2

Obviously, the forecasted values are very close to the observed values almost in all the observations.

Table 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 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 9 we can also visually evaluate behaviour of $R_0$ with respect to bandwidth $h$ for both single index models.

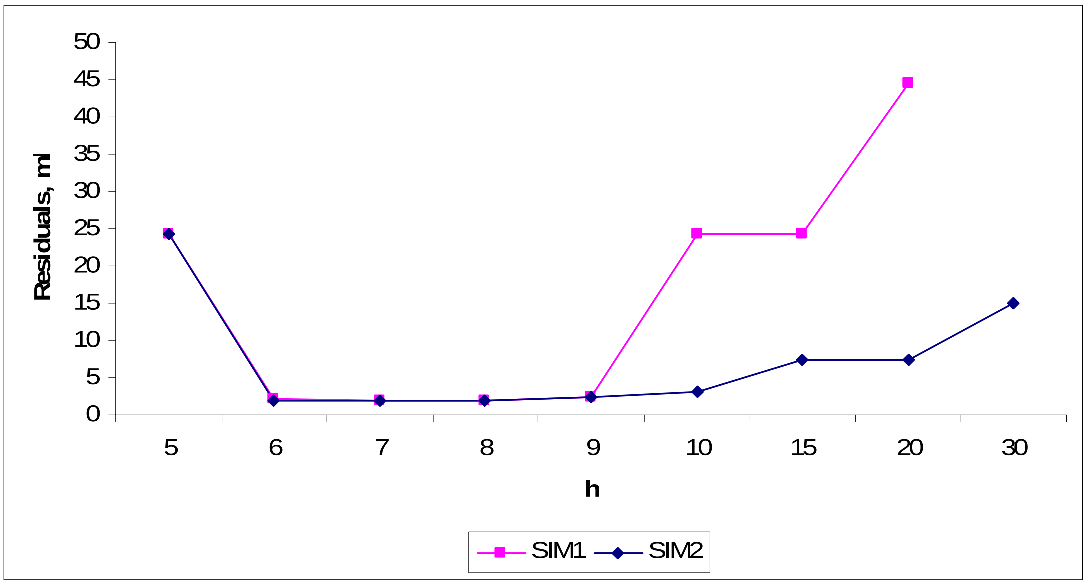

*Figure 9*. $R_0$ depending on bandwidth for SIMs

## Conclusions

In the presented paper two kinds of models for forecasting of inland rail freight conveyances are considered: 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. The efficiency of these models was investigated through the consideration of conveyances for the 15 member countries of the European Union. All the considered models include a great number of explanatory factors. The performed investigations show that the single index regression model gives more precise forecasts than classical methods of linear regression. For this purpose all the models have been estimated and compared by the criterion of the residual sum of squares in case of data smoothing and in case of forecasting as well, that required the cross-validation approach. Moreover, the optimal values of smoothing parameter *h* for the considered single index models have been obtained experimentally.

**References**

1. Andronov, A. M. et al. *Forecasting of the passenger conveyances on the air transport.* Moscow: Transport, 1983. (In Russian)
2. Andronov, A., Zhukovskaya, C., Santalova, D. On Mathematical Models for Analysis and Forecasting of the Europe Union Countries Conveyances. In: *Scientific Papers of the Riga Technical University, Information Technology and Management Science*. Riga: RTU 2006. (In printing)
3. Hardle, W., Muller, M., Sperlich, S., Werwatz, A. *Nonparametric and Semiparametric Models*. Berlin: Springer-Verlag, 2004.
4. Srivastava, M. S. *Methods of Multivariate Statistics.* New York: John Wiley & Sons Inc, 2002.
5. *EUROSTAT YEARBOOK 2005. The statistical guide to Europe. Data 1993–2004. EU, EuroSTAT, 2005* – http://epp.eurostat.ec.europa.eu