Scientific Journal of Riga Technical University
Transport and Engineering. Intelligent Transport Systems

2010
Volume 34

Forecasting of Passenger Conveyances in Latvian Regions Applying Semiparametric Regression Models

Diana Santalova, *Riga Technical University*

Abstract. **In this paper the regression models used for description and forecasting the rail passenger conveyances of the regions of Latvia are considered. Two estimation approaches were compared: parametric and semi-parametric (single index model). Various tests for hypothesis of explanatory variables insignificance and model correctness have been lead, and the cross-validation has been carried out as well. The obtained results have shown obvious preference of the single index model.**

Keywords: **single index model, regression model, passenger conveyances, forecasting, Mahalanobis distance**

I. Introduction

The considered problem is forecasting of rail passenger conveyances from the regions of Latvia on the basis of data taken from statistical books [3], [8]. For that the multiple linear regression model [7] and the single index model (SIM) [4] are used. The *object* of consideration is inland rail passenger conveyances expressed in hundreds. We call as *observation* a data about object for concrete year from 2000 till 2003. For the experiments 91 observations were chosen. Conveyances for some regions, such as Riga, Jurmala and Ogre, exceed conveyances for other regions in times of orders. The task of research is to construct the models adequately describing both large and small conveyances. We use the following criteria for comparing the elaborated models: the coefficient of multiple determinations R^2, Fisher and Student criteria and the residual sum of squares *RSS* [1], [7]. Improvement of models by removal of outliers from data set according to Mahalanobis distance has been done. Described below cross-validation approach is used for evaluating models in case of forecasting. Especially for the single index model the series of experiment is carried out with aim to determine the optimal value of bandwidth *h*.

The paper is organized as follow. First of all the used regression models are considered from theoretical point of view, then the used experimental data are described. After that we consider the suggested group models for conveyances forecasting. Results of carried out estimations are illustrated and comparative analysis of models is shown as well.

II. Structure of Used Models

In this research all investigated models are group models [1], [2]. The main object of consideration is named an *object.* It is a passenger conveyance from some region of Latvia. 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 passenger conveyances for all considered regions of Latvia using one and the same model.

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

General regression model can be described as

$$Y_i = m(x_i) + \varepsilon_i \ , \qquad (1)$$

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 and ε_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 σ^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:

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

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 linear regression models are not very precise (as usual). So, for rail conveyances forecasting we use the *single index regression model* [4] as well:

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

where $g(\bullet)$ is an *unknown link function* of one dimensional variable and $\tau_i = \beta^T x_i$ is called an *index*. Here we assume only that unknown function $m(x)$ is a smooth function.

As $g(\bullet)$ function the *kernel function* usually is considered [4]. Therefore we need to estimate the unknown coefficients vector β and the link function g. For the latter, the *Nadaraya-Watson* kernel estimator can be applied:

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

where $\tau_i=(x-x_i)^T\beta$ is the value of index for the i -th observation, Y_i is the value of dependent variable for the i-th observation and $K_h(\bullet)$ is so called *kernel function*.

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

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

where **h** is a *bandwidth*.

The unknown parameter vector β is estimated by use of the least squares criterion:

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

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

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

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

where

$$\tilde{Y}_i=\sum_{j=1}^{n}K_h(\tau_j)Y_j\ , \qquad (8)$$

and

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

is the derivative of the Gaussian kernel.

We will compare single index models by the residual sum of squares *RSS* only. We calculate the residual sum of squares following:

$$RSS=\frac{1}{n-d}\sum_{i=1}^{n}(Y_i-\tilde{g}(x_i))^2, \qquad (10)$$

where n is the number of observations, d is the number of estimated coefficients, Y_i is observed value and $\tilde{g}(x_i)$ is estimated value.

III. Models suggested for conveyances forecasting

All the needed data have been obtained from the Statistical Yearbook of Latvia 2003 and Annual Report of State Joint-Stock Company “Latvijas dzelzceļš” for 2003. First of all, the forecasted parameter is the inland rail passenger conveyances, expressed in hundreds of passengers. Let us denote it by t_0.

The considered explanatory factors, or predictors, are:

t_1 – population density (PD);
t_2 – number of enterprises per a unit of territory (ED1);
t_3 – number of enterprises per 1000 residents (ED2);
t_4 – density of the unemployed population (UD);
t_5 – number of schools per a unit of territory (SD);
t_6 – number of buses per a unit of territory (BD1);
t_7 – number of buses per 1000 residents (BD2);
t_8 – number of railway stations (SN).

In the brackets the short names of explanatory factors are noted, that afterwards will be used for description of models estimation results.

Now we describe the investigated regression models.

First model is the simple linear regression model (2). The dependent variable $Y^{(1)}=t_0$ is inland rail passenger conveyances. Independent variables are all eight predictors mentioned above. So, we have $x_1=t_1$, $x_2=t_2$, $x_3=t_3$, $x_4=t_4$, $x_5=t_5$, $x_6=t_6$, $x_7=t_7$ and $x_8=t_8$.

In the single index model (3) dependent variable $Y^{(2)}=t_0$ is inland rail passenger conveyances as well. The sets of independent variables coincide with the set from the linear model.

So, we have two regression models and our task is to estimate the unknown coefficients β for both models, to compare the suggested models and to prove the preference of semiparametric model. All the calculations are spent using Statistica 6.0 and MathCAD 12 packages.

IV. Results of model estimation

Let us describe the obtained results. We have used the Forward Stepwise mode of Statistica 6.0 package, which allows sequentially including step by step most insignificant predictors into the considered linear model.

The estimates of the coefficients and calculated values of the Student criterion for the linear model are presented in Table 1. Here and further $\hat{\beta}_i$ is an estimate of β_i , $t(82)$ is the calculated value of Student criteria for 82 degrees of freedom, p-level is the error of second kind (or level of insignificance of variable). The theoretical value of Student criterion for 82 degrees of freedom and level of significance (or error of first kind) α = 5% is equal to 1.99. Taking into account the fact that the hypothesis of *insignificance* of explanatory variable is tested, we can see that calculated value of Student criterion exceeds its theoretical value for all variables exclude UD, i.e. all these variables cannot be recognized as insignificant. Data in the table are arranged in order of decreasing of variables significance. The signs for significant variables correspond to physical sense of the predictors. It testifies to obvious and steady enough influence of the chosen factors on inland rail conveyances on regions of Latvia.

RSS 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 α = 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.

Figure 1 demonstrates how the investigated linear model smooths conveyances. Here and below the observations are arranged in order to region-year: each point corresponds to conveyances of some region during analysing period from 2000 till 2003. Moreover, regions are sorted in alphabetical order. Horizontal axis reflects to the number of observation, arranged in mentioned above order. Vertical axis reflects to the corresponding conveyances, expressed in thousands. It is obvious that considered linear model shows enough good smoothing, but somewhere produces negatives estimates (about in 30% of observations).

At last, equation for the linear model can be written in the following way:

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

TABLE 1

RESULTS OF LINEAR MODEL ESTIMATION

Variable	$\hat{\beta}_i$	t(82)	p-level
SN	146.8	9.8454	0.0000
ED1	435.9	7.2432	0.0000
PD	-20.1	-4.6139	0.0000
SD	64417.9	3.8097	0.0003
ED2	162.6	3.5982	0.0005
BD2	-397.6	-2.9033	0.0047
BD1	-733.6	-2.8749	0.0051
Intercept	-923.9	-1.6149	0.1102
UD	36.9	1.2843	0.2026

Now we discuss results of estimation of investigated single index model. Estimation of coefficients β, i.e. the values of coefficients β optimizing the object function (6), 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. PD, ED1, ED2, SD, BD2 and SN.

For SIM estimation we have used our own programs written in MathCad12 package.

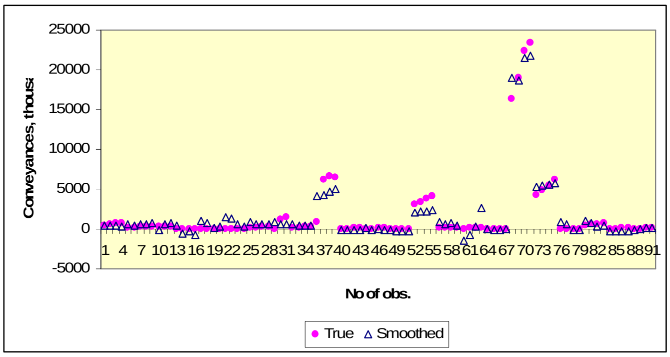

Fig. 1. Smoothing by Linear Model

Table 2 contains the *RSS* for this model depending on *h*. Table 3 contains the estimates of unknown coefficients, calculated with different bandwidths. As we can see, the signs by one and the same predictors are different. Obviously the residual sum of squares tends to zero with *h* → 0, it means, our regression function tends to the interpolation of data. We obtain more smoothed regression curve with *h* → ∞. We can conclude, that the best results in sense of minimizing *RSS* can be obtained with $h^* = 1$.

TABLE 2

VALUES OF *RSS* FOR SIM

Bandwidth *h*		
10	5	1
317 251	268 740	165 843

TABLE 3

RESULTS OF SIM ESTIMATION

$\hat{\beta}_i$	Bandwidth *h*		
	10	5	1
PD	24 450	35 570	8 959
ED1	477	1 412	-386
ED2	1 935	9 192	7 195
SD	0.0036	0.327	0.1
BD2	4 559	9 267	-3 437
SN	1 603	2 500	0.1

The best chosen single index model with $h^* = 1$ can be written as:

$$\hat{E}\left(Y^{(2)}(x)\right) = \frac{\sum_{i=1}^{n} Y_i K_h\left((x - x_i)^T \hat{\beta}\right)}{\sum_{i=1}^{n} K_h\left((x - x_i)^T \hat{\beta}\right)}, \tag{11}$$

where vector of estimated coefficients $\hat{\beta}^T = (8959 \quad -386 \quad 7195 \quad 0.1 \quad -3437 \quad 0.1)$.

Figure 2 represents smoothing by the best chosen single index model. Obviously, the estimates of conveyances almost in all observations are very close to the true conveyances.

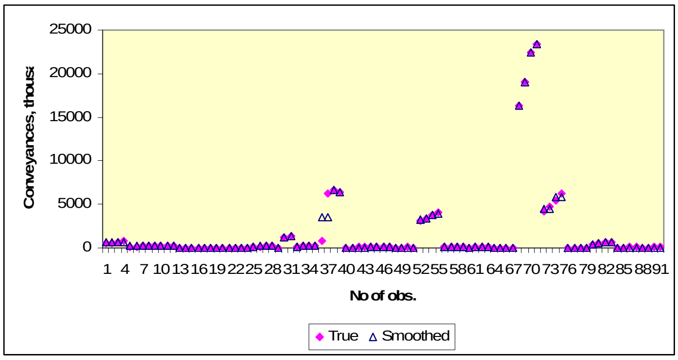

Fig. 2. Smoothing by SIM

So, we can conclude the chosen single index model with h^* = 1 more precise smoothes analysed data than the linear model in sense of *RSS*.

V. Removal of outliers corresponding to Mahalanobis distance

In [5] the models well describing only small conveyances were constructed. The given research is attempted to construct the models adequately describing both large and small conveyances. For this purpose we are going to remove outliers from the data set on the basis of Mahalanobis distance. Mahalonobis square distance [7] shows the distance between each observation and the mean of the observations:

$$D_i = (x_i - \bar{x}) \cdot S^{-1} \cdot (x_i - \bar{x})', \qquad (12)$$

where $x_1, \ \ldots, \ x_n$ 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 4 demonstrates observations with largest values of Mahalanobis distance (i.e. outliers). Last column of Table 4 contains corresponding conveyances.

So, next our experiment consists of our models estimation without outliers. Table 5 contains results of linear model estimation without outliers. For this model the calculated Fisher criterion is 91, R^2 = 0.91, *RSS* = 259096. As we can see, *RSS* 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, removal of outliers has changed the significance of variables.

TABLE 4
Outliers

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

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

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

TABLE 5
Estimation of modified linear model

Variable	$\hat{\beta}_i$	t(69)	p-level
SN	128.1	13.5537	0.0000
BD2	-361.0	-4.3788	0.0000
ED1	2307.4	4.3114	0.0000
UD	-392.7	-4.1716	0.0001
BD1	-1206.7	-3.4162	0.0011
ED2	59.4	1.9859	0.0510
Intercept	919.8	1.5576	0.1239
SD	-26512.5	-0.8851	0.3792
PD	-1.3	-0.1499	0.8813

Figure 3 demonstrates smoothing by the modified linear model.

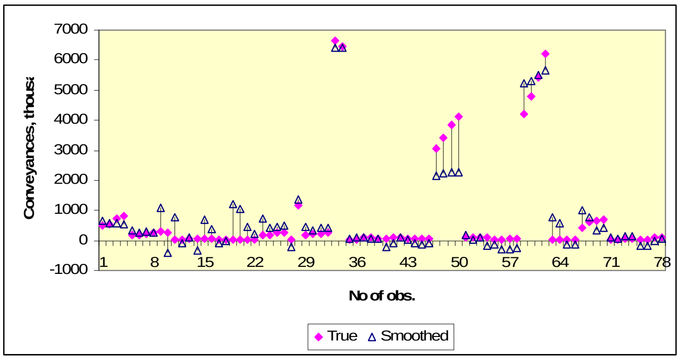

Fig. 3. Smoothing by Modified Linear Model

TABLE 6

VALUES OF *RSS* FOR SIM

Bandwidth *h*		
10	5	1
474 028	74 032	10 053

In the similar way we can modify suggested SIM. Estimation has been lead after removal of the same outliers from the data set (Table 4). 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. ED1, ED2, UD, BD1, BD2 and SN. Table 6 demonstrates the *RSS* for this model depending on *h*. Table 7 contains the estimated coefficients, calculated with different bandwidths.

The best chosen single index model with $h^* = 1$ can be written as (11), where vector of estimated coefficients $\hat{\beta}_M{}^T = (56 \quad 820 \quad 24 \quad 4 \quad 114 \quad 0.1)$.

TABLE 7

ESTIMATION OF MODIFIED SIM

$\hat{\beta}_i$	Bandwidth *h*		
	10	5	1
ED1	292	8 480	56
ED2	2 499	63 980	820
UD	305	95 890	24
BD1	87	7 966	4
BD2	-927	1 368	114
SN	7 849	44 880	0.1

Visually analyzing how modified SIM smoothes the conveyances (Figure 4) we can conclude that estimates almost coincide with true conveyances, in comparison with smoothing of linear model.

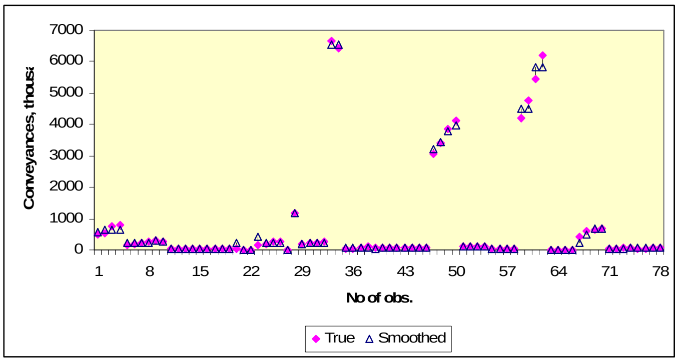

Fig. 4. Smoothing by Modified SIM

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

VI. CROSS-VALIDATION ANALYSIS

Now we consider modified in the previous section models from the other point of view. We use the *cross-validation (C-V) approach*. That means we estimate the unknown coefficients β for the models on the basis of period from 2000 till 2002. Then using the obtained estimates of β we forecast the conveyances for the 2003 year and compare these forecasted conveyances with true ones, i.e. we calculate *RSS* for both models. The optimum value of bandwidth *h* is found for the single index model as well.

Table 8 contains the estimates of β for considered modified linear regression model. The signs of estimates coincide with signs for the case of smoothing (see Table 5) and correspond to physical sense of explanatory factors.

TABLE 8

LINEAR MODEL ESTIMATION FOR C-V

Variable	$\hat{\beta}_i$	t(52)	p-level
SN	110	10.8784	0.0000
UD	-525	-10.0051	0.0000
ED1	3863	8.1543	0.0000
BD2	-416	-5.6836	0.0000
Intercept	2 692	5.5176	0.0000
SD	-117 031	-5.1087	0.0000

Regression equation for C-V can be written as:

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

Figure 5 demonstrates values of conveyances 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.

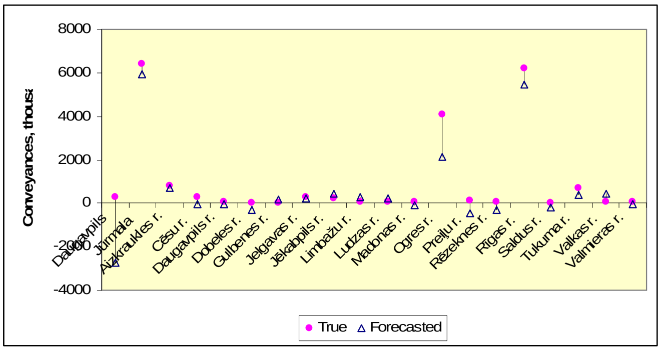

Fig. 5. Forecasting by the Linear Model

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 residual sum of squares is 739151. The true observed values of conveyances and the corresponding forecasts are displayed at Figure 3.

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 h^* of bandwidth that gives a minimal value of *RSS* [4]. The series of experiments was performed and the different estimates of β and values of *RSS* depending of various *h* were obtained as well. The corresponding results for the analysed single index model are shown in Table 9. We can see that estimates of β differs from each other depending on *h* in spite of they were obtained from the same starting point β_0. The values of *RSS* corresponding to various *h* are resulted in Table 10. Thus, the best result for *RSS* is achieved for *h* = 1. The forecasted conveyances by the single index model with h^* = 1 and observed conveyances are shown on the Figure 6.

The best chosen single index model with h^* = 1 can be written as (11), where vector of estimated coefficients $\hat{\beta}_{C-V}{}^T = (283 \quad -380 \quad 4 \quad 2709 \quad 0.1)$.

Obviously SIM gives more exact forecasts of conveyances in comparison with linear model. We have collected the values of *RSS* for both investigated models in cases of smoothing and cross-validation in Table 11. As we can see, the values of *RSS* for SIM are less in all the considered cases.

TABLE 9

SIM ESTIMATION FOR C-V

$\hat{\beta}_i$	**Bandwidth *h***		
	10	5	1
ED1	1 876	1 060 000	282.6
UD	3 171	2 215 000	-380.0
SD	16	11 820	4.1
BD2	-4 972	675 200	2 709
SN	54 650	5 653 000	0.1

TABLE 10

VALUES OF RSS FOR SIM

Bandwidth *h*		
10	5	1
720 403	256 774	147 810

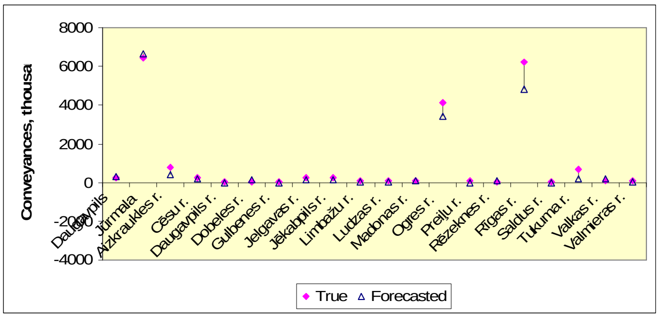

Fig. 6. Forecasting by the Single Index Mode

TABLE 11

THE VALUES OF *RSS*

Model	**Smoothing**		**Cross Validation**
	before	**after**	
	removal of outliers		
LM	801 111	259 096	739 151
SIM	165 843	10 053	147 810

As we can see, the linear model smoothes and forecasts small conveyances more precise. Single index model gives no negative estimates and forecasts.

VII. CONCLUSIONS

In the course of the suggested research several models of multiple regression, which allow evaluating the influence of the main social-economic factors on the volumes of passenger conveyances 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. 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.

REFERENCES

1. **Andronov A.M.** et al. 1983. *Forecasting of the passenger conveyances on the air transport*. Transport, Moscow. (In Russian).
2. **Andronov A.; Zhukovskaya C. and Santalova D**. 2006. “On Mathematical Models for Analysis and Forecasting of the Europe Union Countries Conveyances”. In *RTU zinātniskie raksti, serija Datorzinātne, sējums 28 – Informācijas Tehnologija un Vadības Zinātne.* ISSN 1407-7493. RTU, Riga, 96 - 106.

3. Central Statistical Bureau of Latvia. 2003. *Statistical Yearbook of Latvia 2003*. Riga.
4. **Hardle W.; Muller M.; Sperlich S. and Werwatz A.** 2004. *Nonparametric and Semiparametric Models*. Springer-Verlag, Berlin.
5. **Kopytov E. and Santalova D.** 2007. "Application of the Single Index Model for Forecasting of the Inland Conveyances". In: *Recent Advances in Stochastic Modelling and Data Analysis*. Christos H. Skiadas (Eds.). World Scientific Publishing Co Pte Ltd., Singapore.
6. **Santalova D**. 2007. "Forecasting of Rail Freight Conveyances in EU Countries on the Base of the Single Index Model". In *Computer Modelling and New Technologies*. Vol. 11, No 1. TSI, Riga, 73-83.
7. **Srivastava M. S.** 2002. *Methods of Multivariate Statistics*. John Wiley & Sons Inc., New York.
8. State Joint-Stock Company "Latvijas dzelzceļš". 2004. *Annual Report of State Joint-Stock Company "Latvijas dzelzceļš" for 2003*. SJSC "Latvijas dzelzceļš", Riga. (In Latvian).

Diana Santalova was born in Riga, Latvia. She graduated from Riga Aviation University in 1998, having received a degree of master of Computer Technologies. She works at Riga Technical University since 1998. Subject of her promotional work is "Semi-parametric regression models for forecasting of passenger flows in the Baltic countries". Her e-mail address is Diana.Santalova@rtu.lv

Diāna Santalova. Pasažieru dzelzceļa pārvadājumu prognozēšana Latvijas reģionos ar pusparametriskiem regresijas modeļiem
Šajā rakstā tiek apskatīti Latvijas rajonu un pilsētu regresijas modeļi dzelzceļa pasažieru pārvadājumu aprakstīšanai un prognozēšanai. Statistiskie dati tika savākti uz tādu avotu bāzes kā LR Centrālā statistikas pārvalde (LR CSP) un VAS „Latvijas dzelzceļš" Gadagrāmata. Pārvadājumi no Rīgas, Jūrmalas un Ogres vairākkārt pārsniedz pārvadājumus no citām pilsētām un rajoniem. Šajā ziņā pētījuma galvenā problēma ir parametrisko un pusparametrisko modeļu izstrāde, kas adekvāti apraksta gan lielos, gan mazos pārvadājumus. Tādam nolūkam tika veikti daži pētījumi. Ne visi no tiem bija vienlīdz veiksmīgi. Galvenais iemesls varētu būt tāds, ka vairāku pilsētu un rajonu iedzīvotāji izmanto dzelzceļa transportu dažādiem nolūkiem. Piemēram, tādu lielu pilsētu kā Rīga, Ogre un Jūrmala iedzīvotāji izmanto dzelzceļa transportu kā pilsētas municipālo transportu. Bet tādu pilsētu kā Ventspils iedzīvotāji izmanto dzelzceļa transportu starppilsētu braucieniem, piemēram, uz galvaspilsētu. Pirmkārt ir nepieciešams izvēlēties regresoru kopu, kas adekvāti apraksta gan lielus, gan mazus pārvadājumus. Iepriekšējos pētījumos bija izdarīti mēģinājumi izslēgt izņēmumus no statistiskās izlases saskaņā ar dažādām pazīmēm. Statistikas dati arī bija sadalīti divās daļās, t.i., atbilstoši pārvadājumu lielumam. Sakarā ar to tika mēģināts sameklēt atsevišķus modeļus, t.i., atsevišķi lielos un mazos pārvadājumus. Šajā pētījumā izņēmumi bija izslēgti uz Mahalanobija attāluma bāzes. Šajā pētījumā visi apskatītie modeļi ir grupāli. Tika salīdzināti divi novērtēšanas paņēmieni: parametriskais un pusparametriskais (viena indeksa modelis). Analizējamie modeļi tika pārbaudīti izlīdzināšanas gadījumā un prognozēšanas gadījumā. Otrajam gadījumam bija pielietots īpašs,t.i., kross-validācijas paņēmiens. Iegūtie rezultāti parādīja viena indeksa modeļa neapšaubāmu priekšrocību.

Диана Санталова. Прогнозирование железнодорожных пассажирских перевозок в районах Латвии с помощью полупараметрических регрессионных моделей
В данной статье рассматриваются регрессионные модели, используемые для описания и прогнозирования пассажирских железнодорожных перевозок из городов и районов Латвии. Статистические данные о перевозках из главных семи городов и 26 районов Латвии были собраны на основе информации, полученной из Центрального Статистического Бюро Латвии (LR CSP) и Годового отчета Latvijas dzelzceļš. Перевозки из больших городов (Рига, Юрмала и Огре) превышают перевозки из других городов и районов во много раз. В этой связи главная проблема исследования состоит в том, чтобы построить групповые параметрические и полупараметрические модели, одинаково адекватно описывающие как большие, так и малые перевозки. В этой связи были проведены некоторые исследования. Не все они были одинаково успешны. Причина может состоять в том, что жители городов и районов Латвии используют железную дорогу в различных целях. Например, жители Риги, Огре и Юрмалы используют железнодорожный транспорт как обычный муниципальный транспорт. Напротив, жители таких городов, как Вентспилс, используют железнодорожный транспорт для полноценных дальних поездок, например, в столицу. Прежде всего, необходимо выбрать набор факторов, адекватно описывающих большие и малые перевозки. В предыдущих исследованиях были осуществлены попытки исключить выбросы из статистической выборки согласно различным признакам. Также наблюдения были поделены на две группы, то есть соответственно величине перевозок. В связи с этим были подобраны две отдельные модели. В данном исследовании выбросы были исключены согласно расстоянию Махаланобиса. В этом исследовании все рассматриваемые модели являются групповыми. Сравнивались два подхода оценивания: параметрический и полупараметрический (одноиндексная модель). Рассматриваемые модели были протестированы в случае сглаживания и в случае прогнозирования. Для последнего был использован подход кросс-валидации. Полученные результаты показали очевидное преимущество полупараметрической одноиндексной модели.