Markov-Modulated Linear Regression: a case study of coaches' delay time

NADEZDA SPIRIDOVSKA
Mathematical Methods and Modelling Department
Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019
LATVIA
Spiridovska.N@tsi.lv http://www.tsi.lv

Abstract: - This research presents alternating Markov-modulated linear regression application for analysis of delays of regional buses (coaches). Markov-modulated linear regression suggests that the parameters of regression model vary randomly in accordance with external environment. The latter is described as a continuous-time homogeneous irreducible Markov chain with known parameters. For each state of the environment the regression model parameters are estimated. Data on weather conditions in the Ventspils city provided by the Latvian Environment, Geology and Meteorology Centre database is used for the environment description: two states are assumed: "no precipitation" and "precipitation". The model of the external environment is tested for the markovian properties. Actual data on coaches' trip times is provided by the Riga International Coach Terminal. Data is analysed by means of descriptive statistics. Different experiments are carried out and the application of Markov-modulated linear regression model on given sample showed adequate results indicating the validity of the model.

Key-Words: - External environment, Markov-modulated linear regression, trip time, delay time analysis

1 Introduction

Markov-modulated linear regression suggests that the parameters of regression model vary randomly in accordance with the external environment. The latter is described as a continuous-time homogeneous irreducible Markov chain with known parameters.

Previous investigations [1, 2] were executed on artificial data (simulation analysis) and showed that in case of small sample estimated parameters, they considerably deviated from true ones (what was explained by insufficient sample size and big randomness of the external environment) and in case of big sample the estimated parameters were very close too true ones, but still convergence to true values was very slow.

Due to cooperation with the Riga International Coach Terminal (RICT), it became possible to put the proposed Markov-modulated linear regression model into practice using real data. The RICT is a leader in the area of passenger bus transportation services in Latvia (151803 routes per 2015 year). RICT serves approximately more than 1.860 million of passengers per year [4]. Punctuality and accurate adherence to a timetable are ones of the most significant factors affecting the services quality level. RICT management annually conducts punctuality analysis as part of quality management system. In this research coaches delay time on the route Ventspils-Riga is analysed.

Section 2 provides brief description of Markov-modulated linear regression. Section 3 presents data chosen for the external environment description. Section 4 contains data description of coaches' delay time. And section 5 presents modelling results: implementation of Markov-modulated linear regression.

2 Markov-Modulated Linear Regression: the Main Idea

The idea of combining in particular way linear regression models and Markov-chain based models was put forward by professor Alexander Andronov and was first described in [1,2].

Generally, application of probabilistic-statistical models presupposes invariability of parameters throughout the process of model consideration. In this case it refers to the regression model parameters, i.e., the regression coefficients. However, in practice these parameters usually vary randomly attesting to "random environment" in which investigated object is constantly changing. Allowing for this fact it is necessary to consider developing models to ensure model adequacy to more realistic conditions [3].

Markov-Modulated linear regression model assumes that the external environment is described by an irreducible Markov chain with continuous time, parameters of which are known. Let us consider the main idea of the model.

The full description of the Markov-modulated linear regression model can be found in [1,2]. Let us see the model in matrix notation.

$$Y(t) = (Y_1(t_1), \dots, Y_n(t_n))^T = \begin{pmatrix} \vec{t}_1 \otimes x_1 \\ \vec{t}_2 \otimes x_2 \\ \dots \\ \vec{t}_n \otimes x_n \end{pmatrix} vec\,\beta +$$
$$+ diag\left(\sqrt{t_1}, \sqrt{t_2}, \dots, \sqrt{t_n}\right) Z \qquad (1)$$

where $Y_i(t)$ are scale responses which are time-additive ($Y_i(0) = 0$), n is the number of observations, the $1 \times m$ vector $\vec{t}_i = \left(t_{i,1}, \dots, t_{i,m}\right)$, which component $t_{i,j}$ means a sojourn time for response Y_i in the state j (it is supposed that model operates in the *external environment J(.)*, which has final state space $S = \{s_j, j = 1, \dots, m\}$, for the fixed state $s_j \in S$, j = 1,…, m, (note that $t_i = t_{i,1} + \dots + t_{i,m}$), the $n \times m$ matrix $T = (\vec{t}_i^T, \dots, \vec{t}_i^T)^T$, $\otimes$ is Kronecker product, $(x_{i,1}, x_{i,2}, \dots, x_{i,k})$ is $1 \times k$ vector, the $k \times m$ matrix $\beta = (\beta_1, \dots, \beta_m) = (\beta_{v,j})$ of unknown parameters, *vec* operator *vec(A)* of matrix *A*, the *n*-dimensional diagonal matrix $diag(v)$ with the vector v on the main diagonal, $Z = (Z_i)$ is the $n \times 1$ vector, where $Z_i(t)$ is Brown motion scale disturbance (Z_i are independently, identically normally distributed with mean zero and constant variance σ^2).

This is the case of the generalized linear regression model. The whole trajectory of the environment J(.) is unknown and the estimated conditional average sojourn time is used instead of unknown sojourn times $T_{i,j}$ in the state s_j. Then unknown parameters β are estimated as follows:

$$vec\tilde{\beta} = \left(\sum_{i=1}^{n} \frac{1}{t_i} (\vec{t}_i^T \vec{t}_i) \otimes (x_i^T x_i)\right)^{-1} \cdot$$
$$\cdot \begin{pmatrix} t_1^{-1} \vec{t}_1 \otimes x_1 \\ t_2^{-1} \vec{t}_2 \otimes x_2 \\ \dots \\ t_n^{-1} \vec{t}_n \otimes x_n \end{pmatrix}^T Y \qquad (2)$$

All necessary formulas for a calculation of the conditional average sojourn time that allows to get the needed estimates are provided in previous researches [1, 2].

3 External Environment: Data Description

Weather conditions in the city of Ventspils were chosen as an external environment. It was assumed that there are two states of the external environment: “No precipitation” or “dry” weather conditions and “Precipitation” or “wet” (alternating states). Division was made subjectively, but based on research objectives, namely, all weather conditions that worsen visibility and may influence different transport indicators (punctuality in this case) belong to the second group (“wet”), and, accordingly, the rest belongs to the first group (“dry”). The included data about weather conditions was obtained from the Latvian Environment, Geology and Meteorology Centre (LEGMC) database. General information regarding this resource is paraphrased from the LEGMC homepage www.meteo.lv, while details related to the included weather data are presented below.

Meteorological observations are carried out by LEGMC at 33 observation stations (1 station per 1500km^2), which are stationary and located over the territory of Latvia. Stations location is optimal to provide a sufficiently detailed description of Latvia weather conditions and climate. For this research data from the station named “Ventspils” which is located on the west coast of Latvia in the Ventspils city was used. All available data can be downloaded from LEGMC website in an excel format.

Monthly data about “Past weather conditions 1” was processed from 1986 to 2017 inclusively (except years 2010, 2011, 2012 – data was not available). Data was selected by parameter “Past weather conditions 1”, which contains codes from 0 to 9, with measurement points every 3 hours. Due to LEGMC explanations the results of visual observations are recorded in coded form. The codes are represented in Table 1.

Table 1. Codes of selected parameter

	Parameter: past weather conditions 1 and 2
0	The amount of clouds is less than 5 points between observation boundaries, clear
1	The amount of clouds has changed from < 5 to > = 5 points between observation boundaries
2	The amount of clouds covers > 5 points between observation boundaries
3	All kinds of drift storms (snow drifting close to the ground with or without snow falling) between the observation boundaries. Duststorms or sandstorms between observation boundaries.

4	Mist or ice fog between observation boundaries. Smog between observation boundaries. Visibility <1 km
5	Drizzle between observation boundaries
6	Rain between observation boundaries
7	Snow, snow pellets, needle ice or ice pellets, rain with snow between observation boundaries
8	Heavy precipitation (heavy snowfall, downpour, snow or ice grains, hail) between observation boundaries
9	Thunderstorm with or without precipitation between observation boundaries

All weather condition codes were divided into two groups mentioned above: “No precipitation” or “dry” weather conditions: codes from 0 to 2, and “Precipitation” or “wet”: codes from 3 to 9.

The Visual Basic code was written to divide the available codes (from 0 – 9) into two groups and to calculate the duration of a sojourn time in each state (multiplying the number of consecutive values in each group by 3 (in hours)). Thus, obtaining data to estimate the distribution of the sojourn time in each state.

It is natural to assume that different seasons will have distinctive characteristics of the transition intensities from state to state. At the first stage it was decided to look into the autumn months: September, October and November. Descriptive and inferential analysis was carried out by means of statistical software package Statistica 12.0. Table 2 represents descriptive statistics of the sojourn time in each state for each autumn month.

Table 2. Descriptive statistics of the sojourn time in each state for each month

Variable	Valid N	Mean	Sum	Max	Std.Dev.
Sept/NoPr	507	29.81065	15114	372	43.38
Sept/Prec	492	11.76220	5787	99	10.30010
Oct/NoPr	557	23.39677	13032	237	33.66277
Oct/Prec	548	14.32117	7848	93	11.60828
Nov/NoPr	628	16.98726	10668	141	21.06123
Nov/Prec	617	16.55105	10212	105	14.76343

It is evident from the Table 2 and the Fig.1 that the average sojourn time in the "dry" state decreases from the first autumn month to the last, which seems to be natural and therefore can serve as a validation of the conceptual model.

Other issues to be addressed: is it possible to combine all three months to describe the behaviour of the external environment? Are transition intensities the same for all autumn months? Answering these questions requires that homogeneity analysis should be carried out. The null hypothesis (in words) in general can be stated as follows: the sojourn time in particular state (“dry” or “wet”) is identical for Month1 and Month2. More formally, H_0: $F_1(x_1,\dots,x_n) = F_2(y_1,\dots,y_n)$. Two nonparametrical tests were used for homogeneity analysis at the significance level of 0.01: Mann-Whitney U test and Kolmogorov-Smirnov two-sample test. The results for different combinations of parameters are shown in the Table 3.

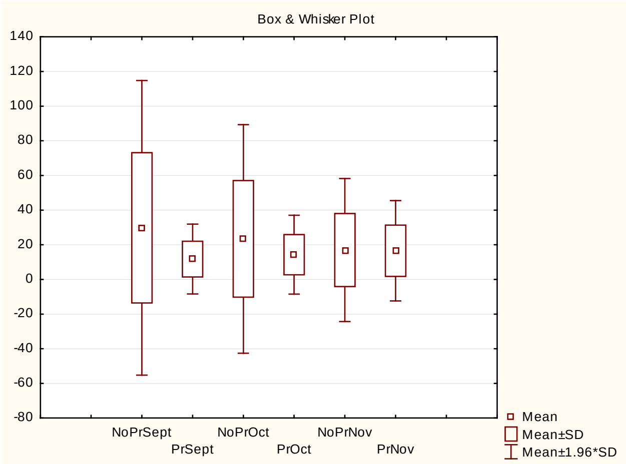

Fig.1. Box&Whisker plot of the sojourn time in each state for each month

Table 3. Homogeneity testing results

H_0: State, Month1 vs Month2	Mann-Whitney Test Z/ **p-value**	Kolmogorov-Smirnov test/ **p-value**	Interpretation of the results
“No prec”, Sept vs Oct	2.981851/ **0.002865**	0.096/ **p < .025**	Reject H_0
“Precipitation”, Sept vs Oct	-3.80089/ **0.000144**	-0.116/ **p < .005**	Reject H_0
“No prec”, Oct vs Nov	2.196350 **/0.028068**	0.071567 **/p < .10**	**Accept H_0**
“Precipitation”, Oct vs Nov	-2.21623 **/0.026676**	-0.058231/ **p > .10**	**Accept H_0**
“No prec”, Sept vs Nov	5.305069 **/0.000000**	0.159151 **/p < .001**	Reject H_0
“Precipitation”, Sept vs Nov	-6.06640 **/0.0000**	-0.174144 **/p < .001**	Reject H_0

Based on the obtained results, it was decided to carry out an experiment on the combined sample with the weather data for October and November.

Two criteria were selected to check markovian properties of the described external environment: distribution of the sojourn time in each state (which is supposed to be exponential) and independence of the observations’ pairs (memoryless property).

Visualization of the sojourn time in each state for each month by means of histogram (Fig.2) showed satisfactory results and suggests that the distribution of the sojourn time is indeed exponential. It is necessary to test hypothesis about the sojourn time distribution in each state. The null hypothesis (in words) in general can be stated as follows: the sojourn time in particular state (“No precipitation” or “Precipitation”) has an exponential distribution. More formally, H_0: $F_{emp}(x_1,\ldots,x_n) = F_{exp}(x_1,\ldots,x_n)$. Two nonparametrical tests were used in distribution fitting procedure: Chi-square test and Kolmogorov-Smirnov test. The results for different combinations of parameters at the significance level 0.01 are shown in the Table 4.

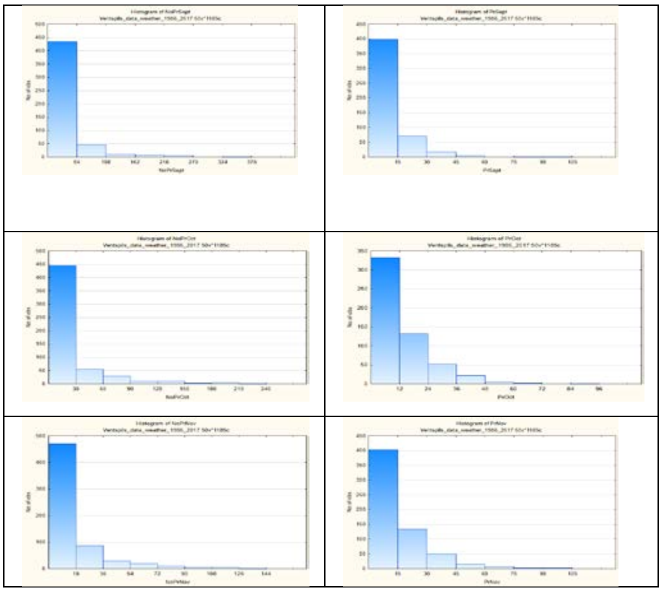

Fig.2. Histograms of the sojourn time in each state for each month

Table 4. Distribution fitting results

H_0: State, Month	Chi-Square test / **p-value**	Kolmogorov-Smirnov test/ **p-value**	Interpretation of the results
“No prec”, September	14.65/ 0.00013	0.16369/ **p < 0.01**	Reject H_0
“Precipitation”, September	6.90211/ **0.14115**	0.22513/ **p < 0.01**	**Accept H_0**
“No prec”, October	50.93942/ 0.00000	0.17619 p < 0.01	Reject H_0
“Precipitation”, October	10.58989/ **0.10191**	0.18899/ **p < 0.01**	**Accept H_0**
“No prec”, November	36.37629/ **0.00000**	0.16189/ p < 0.01	Reject H_0
“Precipitation”, November	5.72381/ **0.33403**	0.16793/ **p < 0.01**	**Accept H_0**
“No prec”, Oct+Nov	104.582/ 0.00000	0.15746/ p < 0.01	Reject H_0
“Precipitation”, Oct+Nov	10.01666/ **0.07476**	0.17595/ p < 0.01	**Accept H_0**

Distribution fitting procedure showed that for each considered autumn month the null hypothesis about exponentiality of the sojourn time in the state “Precipitation” cannot be rejected for the given sample and at the chosen significance level. However, the same null hypothesis for the state “No precipitation” should be rejected. Even though the results obtained were partially negative, it was decided that the data is relevant for describing the external environment in the context of this experiment. Moreover, it does not seem to be a problem, since author’s current studies prove using an approximation of arbitrary nonnegative density by a convolution of exponential densities.

Correlation analysis indicated that there is no linear dependence between the observation pairs "No precipitation – Precipitation" for each month. The results are presented in Table 5.

Table 5. Correlation analysis results

Variable N=492	Correlations (Casewise deletion of missing data)			
	Means	Std.Dev.	NoPrSept	PrSept
NoPrSept	29.85366	43.45491	1.000000	**0.057174**
PrSept	11.76220	10.30010	**0.057174**	1.000000
N=548	Means	Std.Dev.	NoPrOct	PrOct
NoPrOct	23.28285	33.16281	1.000000	**0.098189**
PrOct	14.32117	11.60828	**0.098189**	1.000000
N=617	Means	Std.Dev.	NoPrNov	PrNov
NoPrNov	17.12966	21.20255	1.000000	**-0.033114**
PrNov	16.55105	14.76343	**-0.033114**	1.000000

Data on means and standard deviations are partly different from the data presented in Table 2 (3 out of 6 values for each indicator differ insignificantly). It stems from deliberate exclusion of the data for which no corresponding pair was found in the analysis of the relations between the two variables.

In general, the test results of the external environment model for Markovian property are subject to different interpretations but has been considered as admissible ones for the experimental purposes.

4 Coaches’ Delay Time: Data Description

Ventspils-Riga route was selected for the analysis of coaches’ delay time. Depending on the day of the week, from Monday to Sunday, there are 16, 15, 14, 14, 14, 18 and 13 scheduled runs, correspondingly. The scheduled duration of the run is also a variable and ranges from 180 to 255 minutes. The RICT management provided data on scheduled and actual departure and arrival time of the coaches as well as the record date and capacity of a coach. This data covers the period from 2012 to 2017. For example, the data for 2012 contains 5414 records. The above-mentioned statistical data was made suitable for the analysis. Since in the Markov-modulated linear regression model the dependent variable is time-additive, the delays were summed for each day of the week, a total of 365 observations were obtained for 2013, 2014, 2015 and 2017 years, and 366 observations for 2012 and 2016 leap years, respectively. Since at the first stage of the analysis only autumn months are considered, the sample size has naturally decreased. The given sample can also be analysed as time series with various patterns characteristic of this type of observation, but this task goes beyond the scope of the current study.

One of the main principles of RICT management is the provision of quality services. Punctuality is an essential component of the quality system. The bulk of the delays is within acceptable limits, however in any system unforeseen circumstances may arise and cause schedule shifts (and, as a result, delays) of coaches. Thus, according to the results of a survey of coaches' drivers about the factors having negative impact on adherence to a timetable (RICT, internal procedure D07, 2017), 70% indicated traffic jams in Riga, 16% - weather conditions, 3% - technical condition of the coach, 8% - coach route timetable, and 3% indicated other reasons.

Table 6 presents the average total delays for each day of the week (according to the 6 years sample). According to the results for all autumn months, the least successful day from the point of view of punctuality is Friday, and the most successful ones are Saturday and Sunday, which can also be regarded as confirmation of the data validity. Firstly, since many studies show that the traffic intensity is higher on Fridays, for example, [5, 6], and, secondly, due to the general human experience that confirms this fact.

Table 6. Average delay time of coaches distributed by days

	Average delay time, minutes						
	Mon	**Tue**	**Wed**	**Thu**	**Fri**	**Sat**	**Sun**
Sept N=180	44.3 N=26	36.8 N=26	35.4 N=25	34.5 N=25	78.6 N=26	14.6 N=26	18.0 N=26
Oct N=186	37.1 N=27	21.8 N=27	21.4 N=27	22.2 N=27	52.2 N=26	12.6 N=26	10.8 N=26
Nov N=180	19.4 N=25	24.1 N=25	22.4 N=26	29.6 N=26	47.8 N=26	8.8 N=26	10.0 N=26

5 Modelling results

4 main experiments were carried out. The structure of the input data is the same for all experiments, only the values differ. To apply the model, it is necessary to have the following initial data: the matrix of the state transition intensities (λ), the matrix with the regressors' values (X), the vector of observation durations (τ), the vector with the external environment initial states (I), and the vector of dependent variable values (Y).

A random environment has two states (m = 2). The transition intensities from state i to state j are calculated as reciprocals to the sojourn time. The days of the week serve as regressors. The number of regressors is seven: six dummy variables, and a constant term. Days of the week are represented as dummy variables, "Monday" serves as a key variable.

5.1 Experiment 1

Data on coaches' delay times for the month of September from 2013 to 2017 was used for experiment 1. (Weather data for 2012 is not available. Consequently, there is no possibility to plot a vector with external environment initial states; therefore, data on delays for 2012 is excluded from consideration).

Transition rates from state i to state j are given by the transition matrix:

$$\lambda := \begin{pmatrix} 0 & 0.033545058 \\ 0.085018144 & 0 \end{pmatrix}.$$

Stationary probabilities of states are as follows:

$$\pi = (0.717 \quad 0.283)^T$$

Dimensions of given matrices and vectors: $\boldsymbol{X}_{150x7}$, $\boldsymbol{Y}_{150x1}$, $\boldsymbol{\tau}_{150x1}$, $\mathbf{I}_{150x1}$.

We begin with the estimates for the simple linear regression (ordinary weighted least squares) with 7 regressors: $vec(\tilde{\beta})$ = (44.318 -7.136 -6.556 -6.747 30.182 -28.604 -26.985), with noticeably large residual sum of squares RSS = 242700 and determination coefficient R-squared = 0.17.

Further we use supposed approach and get estimations with respect the external environment. Since we have two states and seven independent variables, the number of unknown parameters β equals to 14.

$vec(\tilde{\beta})$ = (4.217 -1.254 -0.663 -0.854 2.124 -3.498 -2.587 -0.055 1.92 -1.093 0.042 1.642 2.242 -0.024) with smaller RSS = 230700 and higher determination coefficient R-squared = 0.211. RSME= 39.346.

Compering to Monday (which is key variable) for all other days (except Friday) according to coefficients which have negative sign, delays are smaller (for the first state "No precipitation"), which seems adequate according to Table 6. For the second state "Precipitation", the coefficients appear with more random signs (what is less explainable).

5.2 Experiment 2

Since explicit validation set is not available 6-fold cross-validation technique was used for assessing accuracy of model prediction. Each validation set consists of $n = 25$ observations.

Transition rates from state i to state j and stationary probabilities of states are the same as in experiment 1.

Dimensions of given matrices and vectors: $\boldsymbol{X}_{125x7}$, $\boldsymbol{Y}_{125x1}$, $\boldsymbol{\tau}_{125x1}$, $\mathbf{I}_{125x1}$.

Table 7 contains the results of 3 iterations of model estimation.

Table 7. Estimations of unknown parameters, 3 iterations

Parameter	Iteration 1	Iteration 2	Iteration 3
$\tilde{\beta}_{00}$	4.613	5.419	4.144
$\tilde{\beta}_{01}$	-1.033	-1.917	-1.069
$\tilde{\beta}_{02}$	-0.434	-1.472	-0.535
$\tilde{\beta}_{03}$	-1.187	-1.267	-0.496
$\tilde{\beta}_{04}$	1.874	1.32	2.439
$\tilde{\beta}_{05}$	-3.952	-4.537	-3.383
$\tilde{\beta}_{06}$	-2.605	-3.618	-2.227
$\tilde{\beta}_{10}$	0.144	-0.802	-0.245
$\tilde{\beta}_{11}$	1.601	2.513	2.969
$\tilde{\beta}_{12}$	-1.828	-0.578	-0.935
$\tilde{\beta}_{13}$	0.792	0.324	0.065
$\tilde{\beta}_{14}$	2.389	2.154	2.127
$\tilde{\beta}_{15}$	2.646	2.892	2.407
$\tilde{\beta}_{16}$	-1.252	0.628	7.624e-3
RSS	213900	205800	218400
*RSS**	20230	30410	13640
RMSE	**41.366**	**40.572**	**41.802**
*RMSE**	**28.448**	**34.876**	**23.354**

* - out-of-sample (testing sample)

All iterations showed not so high out-of-sample prediction power of the model, but RMSE was smaller for all cases.

5.3 Experiment 3

Data on coaches' delay times for the month of October from 2013 to 2017 was used for experiment 2.

Transition rates from state i to state j are given by the transition matrix:

$$\lambda := \begin{pmatrix} 0 & 0.0427 \\ 0.0698 & 0 \end{pmatrix}.$$

Stationary probabilities of states are as follows:

$$\pi = (0.62 \quad 0.38)^T$$

105 observations were used as a training set and the rest of 50 observations as a validation set.
Dimensions of given matrices and vectors: $\boldsymbol{X}_{105x7}$, $\boldsymbol{Y}_{105x1}$, $\boldsymbol{\tau}_{105x1}$, $\mathbf{I}_{105x1}$.

In the Table 8 comparison of the expectation of the responses (*E(Y)*) and actual responses (*Y*) for validation set is shown.

Table 8. Comparison of actual and expected responses Y

n	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**…**
Y	0	12	10	40	3	0	5	**…**
E(Y)	21.7	16.2	28	54.3	13.5	13.4	48	**…**
n	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**…**
Y	5	3	15	44	22	54	0	**…**
E(Y)	16.2	28	33.8	11.5	54.3	15.7	13.4	**…**

Table 9 contains estimated model quality criteria such as RSS and RMSE. Results within the validation set showed satisfactory results.

Table 9. Model quality criteria

Type of set	**RSS**	**RMSE**
Training	114500	33.017
Validation	20660	20.328

5.4 Experiment 4

Given experiment combines data of two autumn months: October and November, thus expanding the sample twice.

Transition rates from state i to state j are given by the transition matrix:

$$\lambda := \begin{pmatrix} 0 & 0.05 \\ 0.064507198 & 0 \end{pmatrix}.$$

Stationary probabilities of states are as follows:

$$\pi = (0.563 \quad 0.437)^T.$$

Sample size is equal to 305. Estimation gives the following results: RSS = 241400, RMSE = 28.18 and R-squared = 0.173. If we compare with all the experiments, the last model showed the best results based on RMSE criterion.

6 Conclusion

This paper considers application of Markov-modulated linear regression model into practice. Preliminary data preparation was carried out both for external environment description and regression model development itself.

Data on weather conditions in the Ventspils city provided by the Latvian Environment, Geology and Meteorology Centre database is used for the external environment description: two states are assumed: "no precipitation" and "precipitation". The average sojourn time in the state "No precipitation" decreases from the first autumn month to the last. The model of the external environment is tested for the markovian properties. Two criteria were selected to check markovian properties of the described external environment: distribution of the sojourn time in each state (which is supposed to be exponential) and independence of the observations' pairs. Despite the fact that the results obtained were partially negative, it was decided that the data is relevant for describing the external environment in the context of this experiment. Actual data on coaches' trip times is provided by the Riga International Coach Terminal. Actual data on coaches' trip times is provided by the Riga International Coach Terminal.

The application of Markov-modulated linear regression model on this sample did show quite adequate results. The low accuracy of the prediction is not related to the incorrectness of the proposed model, rather it is due to the low quality of the model for describing the delays of coaches: supposedly the day of the week is not the only factor determining the size of the delay. One more reason is connected with the quality and amount of data: matrices $\mathbf{X}$ and vector $\mathbf{I}$ are sparse matrix and vector, also vector $\boldsymbol{\tau}$ contains mostly repeating elements, that could cause unreliable results. The following tasks are requested for further investigation:

- Try different k-folds within cross-validation technique obtaining more reliable results.
- To analyse the remaining months of the year.
- To consider inclusion of other factors in the model (as independent variables).

Markov-modulated linear regression model has recommended itself positively and requires further development.

Acknowledgement

This work was financially supported by the specific support objective activity 1.1.1.2. "Post-doctoral Research Aid" (Project id. N. 1.1.1.2/16/I/001) of the Republic of Latvia, funded by the European Regional Development Fund.

Nadezda Spiridovska research project No. 1.1.1.2/VIAA/1/16/075 "Non-traditional regression models in transport modelling".

The author also appreciates the management of the Riga International Coach Terminal for provided data.

References:

[1] Andronov, A., Spiridovska, N. Markov-Modulated Linear Regression. *In proceedings' book: International conference on Statistical Models and Methods for Reliability and Survival Analysis and Their Validation (S2MRSA)*, 2012, pp.24–28. Bordeaux, France

[2] Andronov A. Parameter statistical estimates of Markov-modulated linear regression, in: *Statistical Methods of Parameter Estimation and Hypothesis Testing 24*, Perm State University, Perm, Russia, 2012, pp. 163–180. (Russian).

[3] Spiridovska N. Markov-Modulated Linear Regression: Tasks and Challenges in Transport Modelling. In: Kabashkin I., Yatskiv I., Prentkovskis O. (eds) Reliability and Statistics in Transportation and Communication. RelStat 2017. Lecture Notes in Networks and Systems, vol 36. Springer, Cham, 2018.

[4] I. Yatskiv. (Jackiva), E. Budilovich. (Budiloviča), and V. Gromule, "Accessibility to Riga Public Transport Services for Transit Passengers," *Procedia Eng.*, vol. 187, pp. 82–88, 2017.

[5] E. Steiger, B. Resch, J. P. de Albuquerque, and A. Zipf, "Mining and correlating traffic events from human sensor observations with official transport data using self-organizing-maps," *Transp. Res. Part C Emerg. Technol.*, vol. 73, pp. 91–104, Dec. 2016.

[6] N. Earl, I. Simmonds, and N. Tapper, "Weekly cycles in peak time temperatures and urban heat island intensity," *Environ. Res. Lett.*, vol. 11, pp. 074003-, Jul. 2016.

[7] M. A. Rotondi, "To Ski or Not to Ski: Estimating Transition Matrices to Predict Tomorrow's Snowfall Using Real Data," *J. Stat. Educ.*, vol. 18, no. 3, 2010.

[8] P. Jordan and P. Talkner, "A seasonal Markov chain model for the weather in the central Alps," *Tellus Dyn. Meteorol. Oceanogr.*, vol. 52, no. 4, pp. 455–469, 2000.