Transport and Telecommunication, 2011, Volume 12, No 1, 40–49
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

APPLICATION OF A DISCRETE CHOICE MODEL TO ANALYSIS OF PREFERRED TRANSPORT MODE FOR RIGA–DAUGAVPILS ROUTE

[1] Dmitry Pavlyuk, [2] Vaira Gromule

[1] Candidate of Economic Science, Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019, Latvia
Phone: (+371)29958338. E-mail: Dmitry.Pavlyuk@tsi.lv

[2] „Rīgas Starptautiskā Autoosta" JSC
Prague 1б, Riga, LV-1050 Latvia
Phone: (371) 67503646. Fax: (+371) 67507009. E-mail: autoosta@autoosta.lv

This paper is devoted to discrete choice analysis of behaviour of bus and train passengers and their choice between alternative transportation modes. We develop a nested discrete choice model and estimate model parameters on the base of data sample collected in our own survey.

We consider Riga-Daugavpils two-way journeys with three possible transportation options – a car, a coach, and a train. Using our survey data we analyse factors, which influence passengers' choices. A set of factors includes travel specific factors (departure time), factors, which describe passengers' age, income, etc. and factors, which describe their behaviour (time of arrival to a station before departure).

The resulting conclusions of the research have a significant practical utility and have been used to improve operations of Riga International Coach Terminal.

Keywords: discrete choice model, transportation, coaches, railways

Introduction

The entry of Latvia into the single European market and political integration of Latvia into the EU have brought with them qualitatively new requirements for passengers transportation; a necessity for higher levels of mobility, intermodality, passenger's comfort and support for passenger's rights, as well as new environmental requirements to transport. Riga International Coach Terminal ("Rīgas Starptautiskā Autoosta") is a leading provider of passenger's bus transportation services in Latvia. It provides international, intercity and regional coach services. Recent studies of a role of buses and coaches confirm excellent safety, environmental and social record of bus and coach transport [1]. In Latvia this mode of transportation competes with railways and also with private cars [2].

Discrete choice modelling is a modern econometric approach to analysis of selection between a set of predefined alternatives. In this research we constructed two discrete choice models for predicting a preferred transportation mode for Riga–Daugavpils two-way journeys. The first model deals with predicting a choice between using a car and using public transport, while the second model is focused on a choice between using a bus and using a train.

A discrete choice model predicts a decision made by a person (such as a mode or a route chosen) as a function explanatory variable set. In the research we investigated an influence of a wide range of factors affecting passenger's choice and estimated their marginal effects. The key factors of passenger's choice were revealed. Direction of these factors' influence on passenger's choice can be used to improve services of bus and railway carriers and stations.

Theoretical Background

There are many examples of economic research where an outcome is considered as a qualitative decision, or more exactly a discrete choice from a set of alternatives. In this work we consider a choice of a transportation mode from three alternatives; a car, a train, and a bus.

For choice modelling it is impossible to use a simple regression (OLS), because an outcome is not a continuous variable. Also there are some more problems with using of a simple regression in this case (related to violations of Gauss–Markov conditions) [3].

In this research a discrete choice model [3, 4] was used for predicting a preferred transportation mode. The mathematical formalization of the model can be presented as:

$$P\left(y_i = 1 \middle| x_i = X_i\right) = F\left(X_i^T \beta\right), \tag{1}$$

where y – a discrete variable, which equals to 1 if a passenger accepts an alternative and 0 if he/she declines it (a binary choice);
X – a set of explanatory variables;
β – a vector of unknown coefficients to be estimated;
F – a function, transforming a set of real numbers into [0, 1].

An estimation base is different from a simple regression in this case – it is necessary to estimate a conditional probability of a particular choice alternative as a function of a set of explanatory variables.

A selection of a form of F function is an important decision. Usually F is a standardised normal cumulative distribution function (a model is called *probit* in this case) or a cumulative logistic distribution function (a *logit* model). There are no exact practical rules for this selection (some recommendations can be found at [5]). Both logit and probit models give similar results for intermediate values of $X^T\beta$ (a functional form makes a difference for larger deviations where the logistic function has fatter tails). In this research there is no systematic difference in results of logit and probit model estimations, so we include estimation results for logit models only.

Usually the maximum likelihood estimator is used for estimating the coefficients of unknown discrete choice models. This provides asymptotically efficient and consistent estimates. For hypothesis testing it is possible to use a Wald test, a likelihood ratio test, or a Lagrange multipliers test. We use the Wald procedure to test linear hypotheses about coefficients as we need to compare models estimated on different samples due to having a non-balanced data set.

There is a complication with analysis of estimated coefficients $\hat{\beta}$. We can calculate an influence of each explanatory variable x_i as:

$$\frac{\partial P\left(y = 1 \middle| X^T \beta\right)}{\partial x_i} = \frac{\partial F\left(X^T \beta\right)}{\partial x_i} = f\left(X^T \beta\right) \cdot \beta_i, \tag{2}$$

where f is an appropriate probability density function. So a level of influence of each variable depends on the full set of values of explanatory variables. Usually marginal effects are calculated for each variable using sample mean values for all other explanatory variables.

In our case the majority of explanatory variables are discrete (qualitative), so it is not possible to use derivatives for calculation of marginal effects. Therefore we use a discrete change formula instead:

$$P\left(y = 1 \middle| \bigcup_{j \neq i} x_j = \bar{x}_j, x_i = 1\right) - P\left(y = 1 \middle| \bigcup_{j \neq i} x_j = \bar{x}_j, x_i = 0\right). \tag{3}$$

The discrete choice model can be extended for more than two alternatives. In our case there are three possible modes of transportation – a bus, a car, and a train. The set of alternatives is unordered, so an appropriate model in this case is a multinomial discrete choice model. A multinomial logit model for K alternatives can be expressed [3] as:

$$P\left(y = j\right) = \frac{e^{X^T \beta_j}}{\sum_{k=1}^{K} e^{X^T \beta_k}}.$$

Using of this model is related with some interpretation difficulties – coefficients of the model are not tied to marginal effects directly. This general variant of the model can be split into a set of separate binary choice models, if model's alternatives are irrelevant. Hausman and McFadden [7] suggested that if an alternative is irrelevant to other, then omitting it from the model will not bias parameter estimates.

To test this irrelevance Hausman and McFadden developed a procedure, usually called a test of independence from irrelevant alternatives.

The procedure is based on a common Hausman specification test:

$$\chi^2_{obs} = \left(\hat{\beta}_R - \hat{\beta}_U\right)^T \left(\hat{V}_R - \hat{V}_U\right)^{-1} \left(\hat{\beta}_R - \hat{\beta}_U\right), \tag{4}$$

where $\hat{\beta}$ and $\hat{V}$ are estimates of model's unknown coefficients and covariance matrices respectively. The U index means that a full sample is used for estimation (an unrestricted model), and the R index means that irrelevant alternatives are excluded from the sample (a restricted model). A null hypothesis is an absence of systematic differences between these two models, and the test statistics will have a chi-square distribution under the null hypothesis (a number of degrees of freedom is a number of restrictions).

To measure a goodness of fit we use an analogue of a determination coefficient R^2 – McFadden's likelihood ratio index [6]:

$$LRI = 1 - \frac{Ln(L)}{Ln(L_0)}, \tag{5}$$

where

$Ln(L)$ – a value of a log-likelihood function for a full model,

$Ln(L_0)$ – a value of a log-likelihood function for a model with a constant only.

There is another way to measure a goodness of fit, based on comparison of model predictions and observed choices (using a '2x2' table for binary models – a table with values for all 4 combinations of observed and predicted successes/failures).

Outputs of a discrete choice model are probabilities $\hat{p}$ of a particular alternative for all sample cases. So there is a need for a predicted probability threshold value p^* to make a choice prediction $\hat{y}$:

$$\hat{y} = \begin{cases} 1, \hat{p} \geq p^* \\ 0, \hat{p} < p^* \end{cases}. \tag{6}$$

The natural threshold value is 0.5 (if a predicted probability value for an alternative is more than 0.5 we predict that this alternative will be accepted, otherwise we predict that the alternative will be rejected). However, because in our case the relative frequencies of selection of different transport modes vary considerably, we will use smaller threshold values for predicting rarely selected alternatives.

Data

We planned and organised a survey of passengers to collect information about their preferred transportation mode. A questionnaire contains 20 questions; main 15 of them are presented to respondents, and other 5 are filled by a pollster (respondent's sex, a language, a place, and date and time of an interview). The survey included 177 respondents.

All questionnaires were collected during a week, so we consider our data as related to one time point (cross-sectional). It's planned to repeat the survey later, to compare results and reveal influence of a season on passengers' choice.

Respondents for this survey were chosen randomly from people visited Riga coach terminal. Riga coach terminal is located aside of a railway station, so it means that respondents are bus passengers only. This definition of a statistical population and a sample has a critical influence on final results' interpretation.

Model Specification

The first practical task is to define variables describing passenger's choice. In this research we use passengers' answers to the question: "Which transportation mode do you usually choose for travelling to Riga/Daugavspils? (If this journey is not a unique event)" with three possible answers: a car, a train, or a bus. So the resulting model is based on two general methods of discrete choice modelling – Revealed

and Stated Preference. Revealed preference method utilises information about real behaviour of respondents, while Stated Preference is based on answers about hypothetical situations.

Note that the sample comprises bus passengers, so a discrete choice can be formulated as: 'a usual transportation mode of bus passengers'.

Passengers who identified this trip as a unique event (31 passengers, 17% of the sample) were excluded from our consideration. Therefore the modelling sample comprised 146 passengers. A distribution of answers to the goal question is presented in Table 1 and on Figure 1.

Table 1. The distribution of preferred transportation modes

Answer	*Number of passengers*	*Percent*
Usually use a bus	111	76.03%
Usually use a train	22	15.07%
Usually use a car	13	8.90%
Total	*146*	*100.00%*

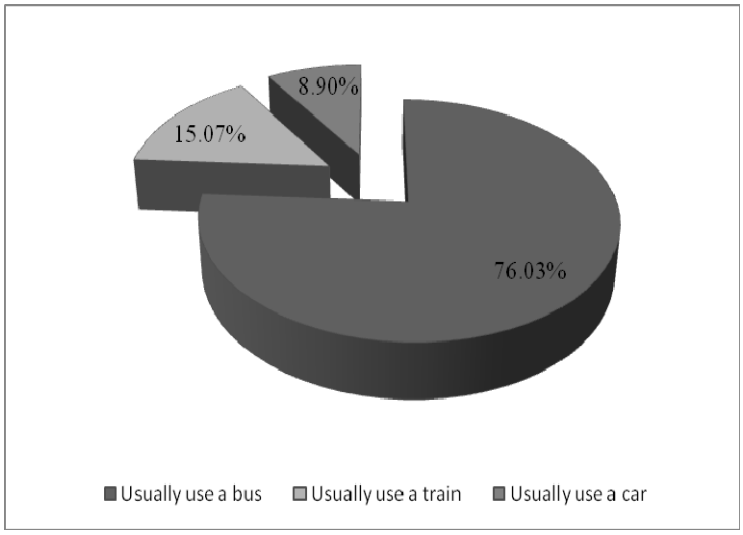

Figure 1. The distribution of preferred transportation modes

It would seem logical that bus passengers in 76.03% of cases identified a bus as their usual mode of transport.

Consequently, there are three possible outputs for our discrete choice model (such models are usually denoted as *multinomial discrete choice* models).

It is presumed that passengers do not choose their usual means of transport from all three possible alternatives directly, but that their decisions include two steps. The first step is to decide to use or not to use their own car (the choice is not applicable when the passenger does not have access to a car). The second stage, if the passenger has decided to use public transport, is to choose between a bus and a train. We used Hausman's and McFadden's test of independence from irrelevant alternatives to test this hypothesis.

In our case we consider the 'use a car' alternative as irrelevant for the choice between a train and a bus. The full model for the 'use a train' outcome includes the full sample, and the restricted model includes the 'use a train' and 'use a bus' answers only (the 'use a car' answer is excluded). The observed value for the Hausman's test and the p-level is as follows:

$$\begin{aligned} &\chi^2_{obs} = 1.77, \\ &p - value = 0.9946, \end{aligned} \tag{7}$$

Consequently, we can definitely accept the null hypothesis that the choice 'use a car' is irrelevant for the choice between a train and a bus.

Usually such conclusions mean that a nested discrete choice model should be used, but in our case it is lessened to a sequential discrete choice model [8] (Figure 2).

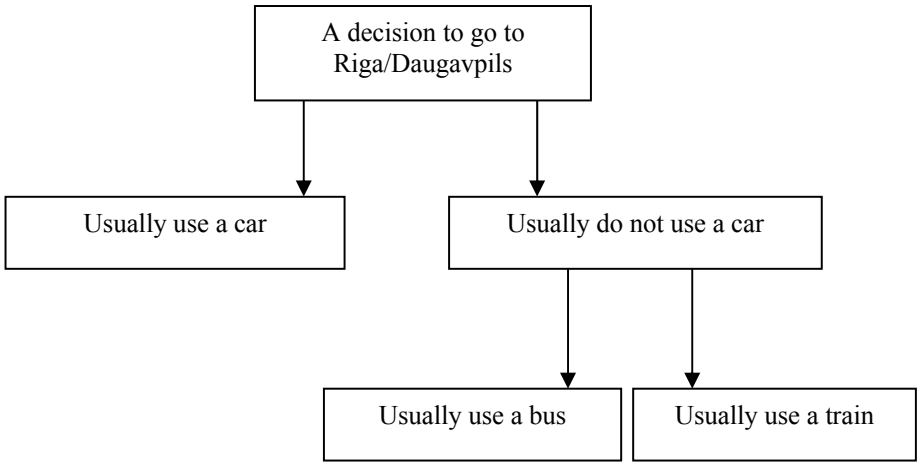

Figure 2. Hierarchy of passenger's decisions

The first stage choice ('use a car' vs. 'don't use a car') is represented by an usual discrete choice model (the Model I) estimated on the basis of the full sample, and the second stage ('use a train' vs. 'use a bus') is represented by a conditional discrete choice model (the Model II) estimated on the basis of the restricted sample.

Estimation Results

Model I is a discrete choice model for passengers' decisions to use a car or not to use a car for regular journeys to Riga/Daugavpils. The original sample contained passengers who stated that this trip was a unique event. Such passengers did not answer the question about their *usual* mode of transport and were excluded from the data set for the Model I. Questionnaires with missing answers were also excluded from the sample.

Some characteristics of Model I and an estimation of its results are presented in Table 2.

Table 2. Model I estimation results

Sample		The full sample with single trips excluded		
Sample size		126		
Dependent variable		'use a car'		
Dependent variable values		1, a person usually uses a car (10 cases, 7.94%) 0, a person usually uses another means of transport (116 cases, 92.06%)		
Log likelihood function value		– 21.448048		
LR test for goodness of fit, χ^2_{obs}		26.96		
LR test for goodness of fit, p-value		0.0007		
LRI value		0.3860		
Explanatory variable	*Description*	*Coefficient*	*p-value*	*Marginal effects*
Why-comfort	A person states a comfort as a reason for bus selection	-1.654	0.151	0.381
Income 500	Person's family income is more than 500 Ls per member	2.400	0.036	0.167

The continuation of Table 2

Age 4060	A person is from 40 to 60 years old	-2.025	0.059	0.333
Time6-12	Departure time is from 6 till 12 AM	-1.630	0.165	0.333
Time15	A person arrived at a station in less than 15 minutes before a departure	-0.468	0.635	0.333
Time1530	The person arrived at a station from 15 to 30 minutes before a departure	-3.540	0.007	0.524
Destination-final	The person's destination is the terminal point	-1.537	0.136	0.690
Language-lat	The person preferred to use the Latvian language	-3.399	0.008	0.519
Constant		1.872	0.163	

The value of McFadden's likelihood ratio index (formula 5, $LRI = 0.3860$) is significantly different from 0. The p-value for the likelihood ratio test (0.007) allows us to accept the hypothesis about the goodness of fit for Model I. Both facts identify Model I as a significant model which can be used for analysis and forecasting.

All significant coefficients in the model have expected signs (correct influence directions). In Table 2 we presented coefficient values, p-values for testing the insignificance hypothesis, and marginal effects for each variable. Marginal effects were calculated as discrete changes (formula 3) as all explanatory variables were discrete.

The income per family member variable had a significant influence on the car usage decision. The positive coefficient ($\hat{\beta}_{income500} = 2.400$) indicates that passengers with a higher level of income used a car for travelling to Riga/Daugavpils more frequently.

Passenger's age was also a significant factor in the model. According to the negative coefficient for the *age4060* variable we conclude that older people (from 40 to 60 years old) use a car rarely for this long trip. As the influence of the income level was included in the model separately, we suppose that older people preferred a worry-free trip on public transport to a long drive.

There was a negative sign for the *why-comfort* variable coefficient. This meant that passengers who liked the level of comfort in buses usually did not use a car for this reason. This fact can be explained in two ways. The first explanation (an obvious one) is that a bus is less comfortable (or equal) compared with a car and people who use cars do not consider the level of comfort of a bus one of its advantages. The second option is that passengers who like the level of comfort in a bus usually prefer this mode of transport to a car. We think both factors could be important.

Also we can note that passengers who began their journeys in the morning (from 6 till 12 AM) usually did not use a car. This fact could be useful for bus carriers and railways, as they could pay special attention to morning trips. The level of competition between trains and buses is most intense at this time.

The time of the passengers' arrival at a station is also an important consideration for transport services management. According to the model, the majority of passengers who usually prefer public transport arrived at a station between 15 and 30 minutes before departure time (the *time1530* variable). This information can be used to inform policies related to advertising or improving station services. The 15–30 minutes time range was compared with a long wait (more than 30 minutes) and last moment arrival (less than 15 minutes). No significant difference between two latter ranges was discovered.

We should also note the high marginal effect value for the arrival variable, which proves its importance. The marginal effect value means that if a passenger arrived at a station between 15 and 30 minutes before departure, there was a higher probability value (0.524 or 52.4%) that he would prefer public transport (other conditions being equal).

There was a higher probability that passengers travelling to a terminal point (Riga/Daugavpils) would usually use public transport (the coefficient for the *destination_final* variable is negative). This conclusion could be expected, and that people prefer to use cars for shorter trips (to intermediate stops).

From a practical point of view it is useful to separate the explanatory variables into two groups: observable and unobservable. Values of observable variables (age or destination point, for example) can be obtained without direct contact with a person, but unobservable values (passengers' opinions about the service, for example) cannot be obtained without asking questions.

Observable factors are easier to control in a particular case (by observing potential passengers). In addition, the average values can be obtained from statistical reports for the majority of observable factors. For example, the percentage of people who are from a particular age group in a geographical area can be collected from government statistical reports, or a destination point's distribution can be collected from ticket offices).

We included both observable and unobservable variables in the model, but after the stepwise reduction procedure there were only two unobservable variables that remained significant. The first variable was positive passenger's opinion regarding the level of comfort of buses (*why_comfort*), and the second was high level of passenger's income (more than 500 Ls, *income500*).

We tested the hypothesis concerning the joint insignificance of coefficients for these two variables using the Wald test:

$$\hat{\beta}_{why\_comfort} = \hat{\beta}_{income500} = 0$$

$$\chi^2_{obs} = 6.04$$

$$p - value = 0.0488$$

Therefore we should reject the null hypothesis regarding the joint insignificance at the 95.12% significance level. Consequently, neither variable can be excluded from the model.

The next step was to estimate the forecasting power of Model I. Using the model we predicted the probability values for all sample records. The distribution of these probability levels is presented on Figure 3.

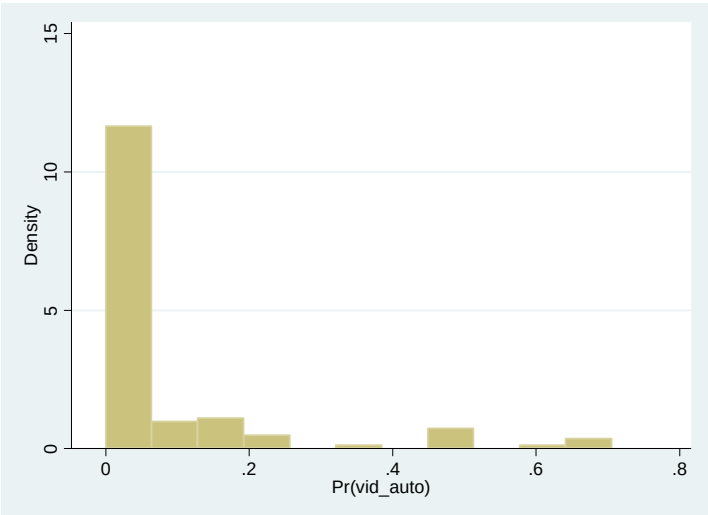

Figure 3. The distribution of predicted probability values for Model I

There are many instances of low probability values (representing a small possibility of using a car). This distribution matches our expectations as only 7.94% of the sample used a car as their usual means of transport (see the dependent variable description in Table 2).

Due to this fact we used a lower threshold value for forecasting. The selected value was 0.45, so all records with lower predicted probability values were classified as 'don't use a car' and all records with higher predicted probability values were classified as 'use a car'. The threshold value was selected to maximise the percentage of correctly classified cases.

We used a 2x2 table for comparing forecasted and observed values (Table 3).

Table 3. Model I forecasting power

The threshold = 0.45	Observed 'use a car'	Observed 'don't use a car'	*Total*
Predicted 'use a car'	5	5	*10*
Predicted 'don't use a car'	5	111	*116*
Total	*10*	*116*	*126*
Correctly classified			92.06%

The percentage of correctly classified cases was high (92.06%). Also a large percentage of persons were correctly classified as 'don't use a car' (111 out of 116). The forecasting of usual car users was not so good (5 from 10) as an absolute value, but nonetheless, 5 usual car users were correctly identified from 126 persons. This shortcoming can be explained by the small size of the usual car user subsample.

Model I is a discrete choice model for passengers' decisions to use a train or use a bus for regular journeys to Riga/Daugavpils. We used a Hausman's test to test the null hypothesis about the irrelativeness of the car alternative to a bus/train choice (formula 7). The null hypothesis was accepted; therefore passengers who usually use a car were excluded from the sample for Model II estimating.

Model II estimation results are presented in Table 4.

Table 4. Model II estimation results

Sample		The sample with 'use a train' and 'use a bus' answers for the goal question only		
Sample size		102		
Dependent variable		"use a train"		
Dependent variable values		1, a person usually uses a train (18 cases, 17.65%) 0, a person usually uses a bus (84 cases, 82.35%)		
Log likelihood function value		– 31.69122		
LR test for goodness of fit, χ^2_{obs}		31.68		
LR test for goodness of fit, p-value		0.0002		
LRI value		0.3333		
Explanatory variable	*Description*	*Coefficient*	*p-value*	*Marginal effects*
Time12–18	Departure time is from 12:00 till 18:00	-1.785	0.116	-0.084
Riga	A person travels from Riga	1.976	0.051	0.106
Why-habit	A person states a habit as a reason for bus selection	-2.079	0.088	-0.892
Why-price	A person states a price as a reason for bus selection	2.657	0.005	0.399
age4060	A person is from 40 to 60 years old	-1.988	0.029	-0.115
direct	A trip is a direct one (vs. a return one)	-1.867	0.023	-0.193
Freq-year	A person travels this way once a year or more rare	1.014	0.183	0.084
Destination-final	Person's destination is the terminal point	1.178	0.131	0.069
Alt-cheaper	A person thinks that a train is cheaper than a bus	1.167	0.150	0.100
Constant		-2.919	0.049	

The value of McFadden's likelihood ratio index (*LRI* = 0.3333) is significantly different from 0. We also accept the hypothesis about the goodness of fit for Model II on the basis of the p-value for a likelihood ratio test (0.002).

The majority of variables have the expected direction of influence.

Departure time had a significant influence on the bus/train choice. Evening bus trip passengers (leaving between 12:00 and 18:00, the *time12–18* variable) use a bus as a rule with higher probability. This fact can be used by bus carrier managements to introduce additional features for regular evening bus passengers, and to attract non-regular passengers at other times of day.

The *Riga* discrete variable increases the probability of train selection, so we can conclude that there are more passengers who leave Riga by bus, rather than by train (their preferred means of transport).

This can be explained by a higher level of mobility for passengers leaving Riga, and by the possibility that it is more difficult to switch modes of transport from a train to a bus in Daugavplis.

There are two significant variables related to the reasons for bus selection, and both have obvious meanings. Firstly, passengers who state that habit is their reason for choosing the bus use the bus more frequently than the train. The highest marginal effect value (-0.892) matches our expectations. Also, passengers who choose price as a key factor in their selection prefer to use the train (a train ticket is slightly cheaper than a bus ticket).

Passengers from 40 to 60 years old use the bus more frequently than the train for regular trips. This fact can be useful for the marketing campaigns of railways and bus carriers.

The negative value of a coefficient for the *direct* variable suggests that people prefer their usual means of transport for their direct trips. In other words, the fact that a person chooses a bus for his direct trip indicates that he uses a bus as a rule.

Passengers, who usually travel by bus very seldom (once a year or less), choose to use the train. This could indicate that they have the habit of using trains mainly for longer trips or when there is more information available about railways. The terminal point as a destination predictably increases the probability of train selection.

In order to estimate the model forecasting power we used the same method as for Model I.

The distribution of the predicted values is presented on Figure 4. Again, one of the alternatives is chosen rarely (only 17.65% of respondents use a train as their usual means of transport. This figure may possibly be due to limitations in the sample), and this could explain why there are many records with a low level of forecast probability. Using the same principle we defined the threshold as 0.35 for forecasting (all records with a predicted probability value less than 0.35 were classified as 'use a bus' and all records with a predicted probability value more than 0.35 were classified as use a train).

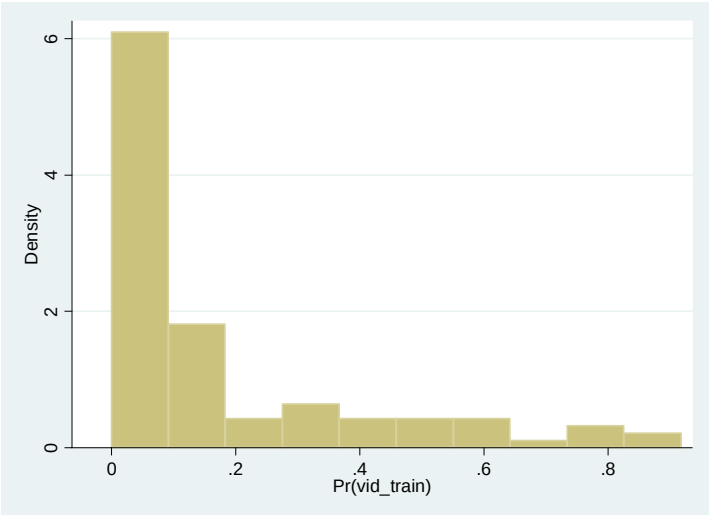

Figure 4. The distribution of forecast probability values for the Model II

The Table 5 contains the comparing of forecast and observed choices.

Table 5. The Model II forecasting power

The threshold = 0.35	Observed "use a train"	Observed "use a bus"	*Total*
Predicted "use a train"	13	9	*22*
Predicted "use a bus"	5	75	*80*
Total	*18*	*84*	*102*
Correctly classified			86.27%

86.27% of cases are classified correctly. Also we note a high level of coincidence for the unusual alternative (13 of 18 passengers who usually use a train are classified correctly).

Conclusions

We developed two discrete choice models for a choice of transportation mode for Riga–Daugavpils two-ways journeys. The first model allows the prediction of the choice between using a car and using public transport, and the second model focuses on the choice between using a bus and using a train. Both models have a high percentage of correctly classified cases (92.06% and 86.27% respectively).

We investigated an influence of a wide range of factors, which affected passenger's choice, and estimated their marginal effects. Some key factors were discovered. Directions of these factors' influence could be used by bus and railway carriers and stations to improve their services. A set of most important significant factors includes departure time, ticket price information, and passenger's personal characteristics like age and income. Also a significant influence of a destination point is discovered.

We are planning to extend this research with additional surveys, executed at different time points (seasons of a year). We assume that passengers' preferences and explaining factors can change depending on seasonal variables like weather and workload.

References

1. Gromule, V., Yatskiv, I. Coach Terminal as an important element of transport infrastructure. In: *Proceedings of the International Scientific Conference "Transbaltica 2007", Vilnius 11–12 April 2007, Transport.* Vilnius: Technika, 2007. Vol. 22, No 3, pp. 200–206.
2. Gromule, V., Jackiva, I. Information System Development for Riga Coach Terminal. In: *Conference on System Science and Simulation in Engineering, Venice, Italy, November 21–23, 2007*, pp. 170–175.
3. Greene, W. H., *Econometric Analysis, fifth edition.* NJ: Prentice Hall, Upper Saddle River, 2003. 1024 p.
4. Ben-Akiva, M., Lerman, S. *Discrete Choice Analysis: Theory and Application to Travel Demand (Transportation Studies)*. Massachusetts: MIT Press, 1985. 384 p.
5. Amemiya, T. Qualitative Response Models: A Survey, *Journal of Economic Literature,* Vol. 19, No 4, 1981, pp. 481–536.
6. McFadden, D. The Measurement of Urban Travel Demand, *Journal of Public Economics,* Vol. 3, 1974, pp. 303–328.
7. Hausman, J., McFadden, D. A Specification Test for the Multinomial Logit Model, *Econometrica,* Vol. 52, 1984, pp. 1219–1240.
8. McFadden, D. *Econometric Analysis of Qualitative Response Models: Handbook of Econometrics,* Amsterdam: North Holland, 1984. Vol. 2, pp. 1395–145.