UHASSELT

KNOWLEDGE IN ACTION

School of Transportation Sciences

Master of Transportation Sciences

Master's thesis

Mode-Choice Model for High-Speed Railway in Hungary Using Stated Preference Data

Márton Hegedüs

Thesis presented in fulfillment of the requirements for the degree of Master of Transportation Sciences

SUPERVISOR :

Prof. dr. ir. Tom BELLEMANS

CO-SUPERVISOR :

dr. Muhammad ADNAN

CO-SUPERVISOR :

N/A N/A

2021

2022

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

School of Transportation Sciences

Master of Transportation Sciences

Master's thesis

Mode-Choice Model for High-Speed Railway in Hungary Using Stated Preference Data

Márton Hegedüs

Thesis presented in fulfillment of the requirements for the degree of Master of Transportation Sciences

SUPERVISOR :

Prof. dr. ir. Tom BELLEMANS

CO-SUPERVISOR :

dr. Muhammad ADNAN

CO-SUPERVISOR :

N/A N/A

Mode-choice model for high-speed railway in Hungary using stated preference data

Author: **Márton Hegedüs**

Hasselt University, Belgium

Promotors: dr. M. Adnan
Prof. dr. ir. T. Bellemans

Submission: 2022.01.21.

Word count: 6457
Table count: 5
Figure count: 9

Abstract

The high-speed railway (HSR) has an increasing importance in the future of transportation in Europe. Hungary is still in the planning phase of implementing a HSR service, which requires the determination of the potential demand for this service. To estimate this demand a stated preference (SP) survey was conducted. This paper shows how the stated preference survey was used to determine a mode-choice model which includes the HSR service as a new mode. The analysis of the survey provided information to establish that the classical multinominal logit model is not sufficient to handle a new transport mode like the HSR. Therefore, other model structures were investigated to find the one that fits the data best. The best fitting model was the mixed logit model, which is not a commonly used model structure, however, it provided the necessary flexibility to analyse the potential HSR service. The mixed logit model resulted in a parameter set which side by side with other HSR mode-choice models a valid solution to be used in transport planning practice.

Highlights

- SP data are indispensable, when examining potential demand for new transport modes.
- The best fit model is the mixed logit model for predicting HSR demand in Hungary
- The VoT parameter set is consistent with other models, so it can be used for further research.

Keywords

Stated preference, High-speed railway, Mode choice, Mixed Logit Model

1. INTRODUCTION

High-speed railway (HSR) is one of the most recent focuses of European transport developments, even more so since the passing of the European Green Deal in 2019. (EU Commission, 2019) although there is no clear definition on what is considered as high-speed rail, the common agreement is that the HSR is a special railway service that runs on a special, usually separate infrastructure, with a speed of at least 250 km/h or above. The first HSR line in Europe was the TGV in France that was opened in 1981. Since then, however, several other lines were introduced, sometimes with mixed results. (Arduin, 2005) The expending infrastructure resulted in an increase of demand. In the past 10 years the volume of distance travelled by HSR increased by almost 500%.

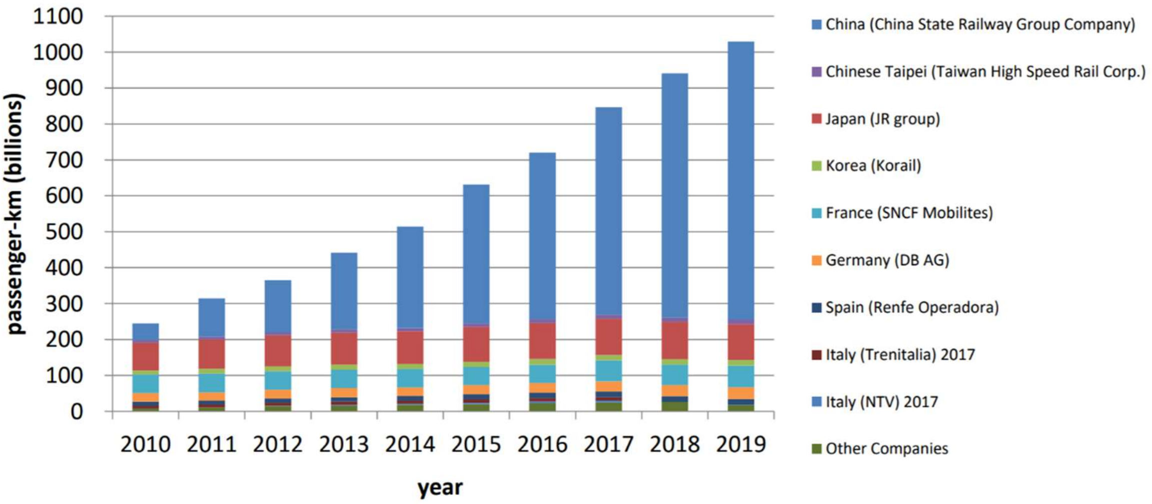


FIGURE 1: Passenger-km travel by high-speed railway
(source: https://uic.org/IMG/pdf/20210201_high_speed_passenger_km.pdf)

Evaluating the introduction of a new transport mode like high-speed rail is always challenging. One of these challenges is the prediction of future demand, where the mode-choice modelling is a key tool. As several examples show, the HSR service causes a significant decrease in the modal share of air travel, but this shift is way less in the case of passenger car, regional bus, and railway demand. (Givoni, 2016) As a result, most of the researchers focused on how the HSR service competes with air travel, but not so much on traditional transport modes. Hungary, with a population of 9.8 million can be seen as a little market for national HSR service to be operated, especially as there is no domestic air service. So far, there have been no comprehensive studies conducted in Hungary that could provide information on the potential modal split of a HSR service. However, as part of the recently planned international HSR lines there has been a necessity to assess the potential demand of HSR in Hungary.

2. METHODOLOGY

The steps towards creating a mode-choice demand model for the high-speed railway has four major steps. The data collection, the data analysis, the construction of the mode-choice model and the validation.

1.The **data collection** is the first step to create a mode-choice model. The data collection has a more important part than many would think. The data collection has a high determining factor on the result of the demand model. The data collection process has almost infinite possibility on how, where, what and when it should be conducted. (TAG Unit 1.2, 2020) The analysis of a new mode, like HSR, however, sets the limit and goals of the data collection. To assess a new mode properly, collecting the current behaviour and travel patterns are not enough, a stated preference (SP) survey is required, to determine who the potential users of the new transport mode will be and under what circumstances. A badly planned and executed data collection can hinder the whole mode-choice modelling attempt and result in a false, unrealistic result.

2. The **data analysis** follows the data collection as the second step of the process. The data analysis consists of cleaning the data and then creating the proper format that can be used for the creation of the model. The initial analysis of the model can tell the transport planners if the collected data is appropriate for further use. Can the surveyed participants represent the population? Did the participants answer correctly and realistically?

3. After the initial analysis of the data, the **construction of the model** can be done. This is the main step of the research process. The first step of the modelling process is to determine the utility functions of the different transport modes. The elements of the utility function are limited by the structure of the stated preference survey, as only those variables can be included in the function that had been examined in the SP questionnaire. The next step is to determine the type of mode-choice model that can fit the collected data the best.

4. The **validation** of the model is the last step, where it is determined if the mode-choice model can be used in practice. Even if the created model fits the collected data, it is important to validate the results. The validation in a case where a new transport mode is introduced can be problematic, as there are no actual revealed preference data to compare the results to. The only option is to compare the results to other similar mode-choice model results.

3. DATA COLLECTION

The data collection for this research has been carried out by an online survey due to the extensive area and the COVID-19 restrictions. The survey was conducted with **4012 participants** spread across the whole country. The participants were selected from a panel provided by Impetus Research Ltd. The only restriction we determined is that all participants must have had a trip in Hungary, above 80 km, in the past 2 years (also taking into consideration the Covid restriction prior to the survey). The distance restriction was to ensure that the described trips could be potential HSR trips. The survey had two main parts. The first part consisted of a household and a trip diary questionnaire, and the second part was a stated preference questionnaire.

Survey

The 4012 participants were asked about their social and economic attributes and about the last trip they had, that matched the previously mentioned criteria. This part of the survey provided a control over the validity of the survey over the whole population and a possible segmentation criteria for the analysis of the data. The description of each trip of the participants provides a baseline revealed preference data for the transport modelling, of which the calculated mode-choice model can be validated against.

Questions about the social and economic stand of the participants:

- Place of living
- Age
- Sex
- Household income
- Activity group
- Access to a private vehicle

Questions about the trip:

- Main original mode – Car, bus or rail
- Purpose of trip
- Was it a one-day trip?
- Luggage
- Frequency of the travel

Stated Preference Questionnaire

Kroes and Sheldon in 1988 defined the stated preference (SP) as a “family of techniques which use individual respondents’ statements about their preferences in a set of transport options to estimate utility functions”. The use of SP was first used in marketing, but soon after transport researcher found that it can solve to limitations of the revealed preference (RP) data. The RP data often cannot provide sufficient data to carry out a detailed examination of all relevant variables. The SP provides a more flexible option as a design of the SP questionnaire determines the number of variables to be examined. The SP analysis can also help to counter the possible correlation between variables. The SP design makes it considerably easier to evaluate variables that are difficult to measure such as the comfort of vehicle. The most important difference between RP and SP analysis is that the RP can't handle any evaluation of a situation, or transport mode that does not yet exist. (Adamowicz, Deshazo, 2006) These attributes of the SP make it obvious that it is preferable to use the SP questionnaire in the case of examining the HSR service in Hungary.

The SP can have different designs from which three are used in transport related studies. (Willumsen, 2011) Out of the possible *contingent valuation*, *conjoint analysis* and *stated choice* methods in this case the stated choice method (SC) was used. The SC presents a set of alternatives to the respondents from among which they must choose their preferred one.

In the second part of the questionnaire two stated preference surveys were conducted. The two SPs differed by the presented transport modes. In the first part of the SP **8** questions were asked, in the second part **5** additional questions were presented to the participants. In the first part of the questions only the transport modes that are currently available were presented, and in the second part the HSR was also introduced as an alternative transport mode. Altogether **52 156** answers were given. The following transport modes were determined in the SP questionnaire (the availability of the personal car was dependent on answers in the first part of the survey):

- Car
- Train
- Bus
- Highspeed rail (HSR)
- No trip

The variables used in the SP are all quantifiable and measurable variables. The aim of the research, as it was detailed in Chapter 1, was to determine a generic mode-choice model that only requires quantifiable measurements. These measurements serve as the basis of the utility functions for each transport mode. The presented values for the questions in the stated preference survey were based on the revealed trips of the participants and altered as it is shown in TABLE 1. The following variables were used in the SP design.

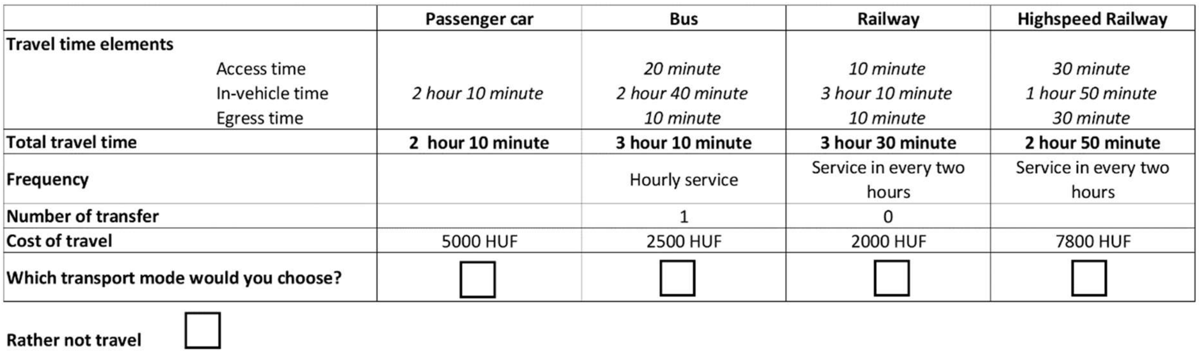
- In-vehicle time
- Access and egress time
- Travel cost
- Service frequency
- Number of transfers

TABLE 1: Used levels for each attribute in the stated preference survey

Level		1	2	3	4	5
Access/Egress time		0.5*Baseline value	0.75*Baseline value	1*Baseline value	1.25*Baseline value	1.5*Baseline value
Travel time	Car	0.9*Baseline value	1*Baseline value	1.1*Baseline value	1.25*Baseline value	1.5*Baseline value
Cost	Rail/Bus	0.5*Baseline value	0.75*Baseline value	1*Baseline value	1.25*Baseline value	1.5*Baseline value
	HSR	0.7*Baseline value	0.75*Baseline value	0.8*Baseline value	0.9*Baseline value	1*Baseline value
	Car/Rail/ Bus	0.5*Baseline value	0.75*Baseline value	1*Baseline value	1.25*Baseline value	1.5*Baseline value
	HSR	1*Baseline value	1.2*Baseline value	1.5*Baseline value	2*Baseline value	3*Baseline value

Level	1	2	3	4	5
Frequency	Every hour	Every two hours	Less than 2 hours		
Number of transfers	0 transfer	1 transfer	2 transfers		

Based on the number of attributes and possible levels for these attributes, with the use of the Ngene software 40 different choice scenarios were created to be presented to the survey participants. Ngene is a software is a capable tool to create experimental designs for stated preference surveys. (http://www.choice-metrics.com/features.html)



	Passenger car	Bus	Railway	Highspeed Railway
Travel time elements				
Access time		*20 minute*	*10 minute*	*30 minute*
In-vehicle time	*2 hour 10 minute*	*2 hour 40 minute*	*3 hour 10 minute*	*1 hour 50 minute*
Egress time		*10 minute*	*10 minute*	*30 minute*
Total travel time	**2 hour 10 minute**	**3 hour 10 minute**	**3 hour 30 minute**	**2 hour 50 minute**
Frequency		Hourly service	Service in every two hours	Service in every two hours
Number of transfer		1	0	
Cost of travel	5000 HUF	2500 HUF	2000 HUF	7800 HUF
Which transport mode would you choose?	□	□	□	□

Rather not travel □

FIGURE 2: Example of SP survey question card

Evaluation tools for the stated preference analysis

The evaluation of the data was carried out in the opensource software of Biogeme - Version 3.2.8. (Bierlaire, 2020) The software was designed to calculate discrete choice model parameters in a Python environment using maximum likelihood estimation. By applying the software, it is possible to calculate several different mode-choice model types. The output file contains not only the calculated values of the parameters, but several key statistical information that can be used to determine the goodness of fit for the model.

4. RESULTS

4.1. Data analysis

Data cleaning

The collected data should always be considered suspect. Any type of data collection may be affected by some type of observation bias and errors occurring during the collection process. In this particular survey, as the chosen population was selected from a predefined pool of participants (described at Chapter 3) there were only a few answers that were needed to be discarded. During the examination the following criteria were checked:

- Plausible trip description: The given parameters of the described trip – distance, time etc. – are in the range of the survey requirement and are logical.
- Origin and destination check
- Time taken to fill out the questionnaire and stated preference survey
- Consistent use of answer "No trip" during stated preference questions

Based on these criteria the answers of 13 people – i.e., **169 observations** – were discarded during the analysis of the data based on the criteria listed above.

Initial analysis

The socio-economic composition of the 4012 participants is essential part of interpreting the results. 49% of the participants were male and 51% female. The age distribution of the participants is even, no age groups are underrepresented or overrepresented. 84% of the partakers have access either to a personal car or a company car for their trip and only 16% have access solely to public transportation.

This high percentage of access translates to a 65% modal split for the car. The 35% of public transport mode share is close to the current modal-split for long distance trips in Hungary.

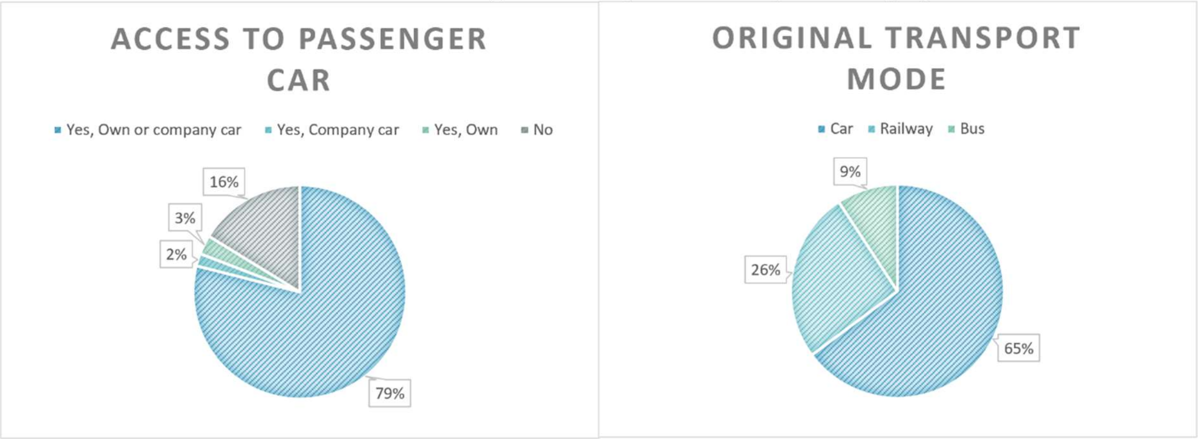


FIGURE 3: Access to passenger car and original transport mode of participants

The trip length distribution of the participant's revealed trips also shows that the participants of the survey provided a good representation of the target population. Among these trips the most frequently stated trip purpose was family visit, with a percentage of 41%. The trip purpose leisure was attributed by 33% of the trips and trips related with work or school was 16%. The low percentage of trips related to work or school can be explained by the restrictions determined by the survey, as only trips above 80 km were asked, which is commonly above the normal commuting distance.

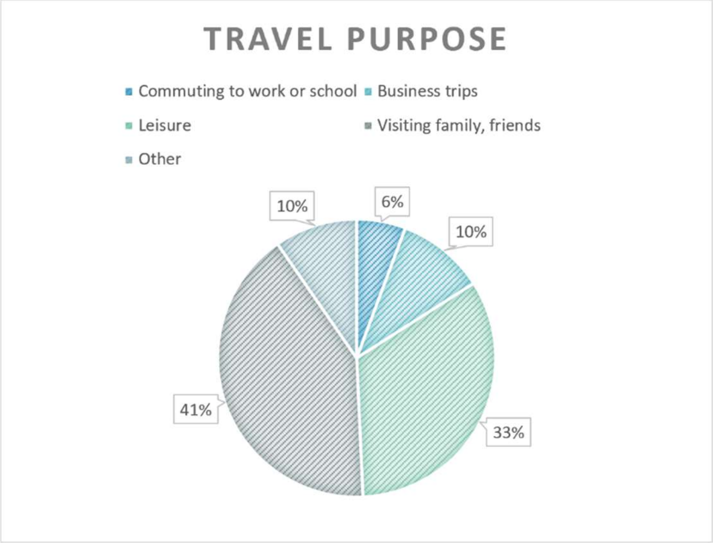


FIGURE 4: Distribution of travel purpose in the survey

The distribution of the stated preference questionnaire is shown in FIGURE 5. Only 5% of the answers fell under the 'No trips' alternative. I In all other cases the participants found at least one other transport mode desirable. In both parts of the SP choices the most dominant transport mode was the car with 55% and 45% respectively. The introduction of the HSR transport mode influenced the participant to choose it over 24% of the possibilities. Less than 20% of the participants chose only one transport mode for all stated choice questions, and out of these participants 90% chose their original transport mode as the sole answer.

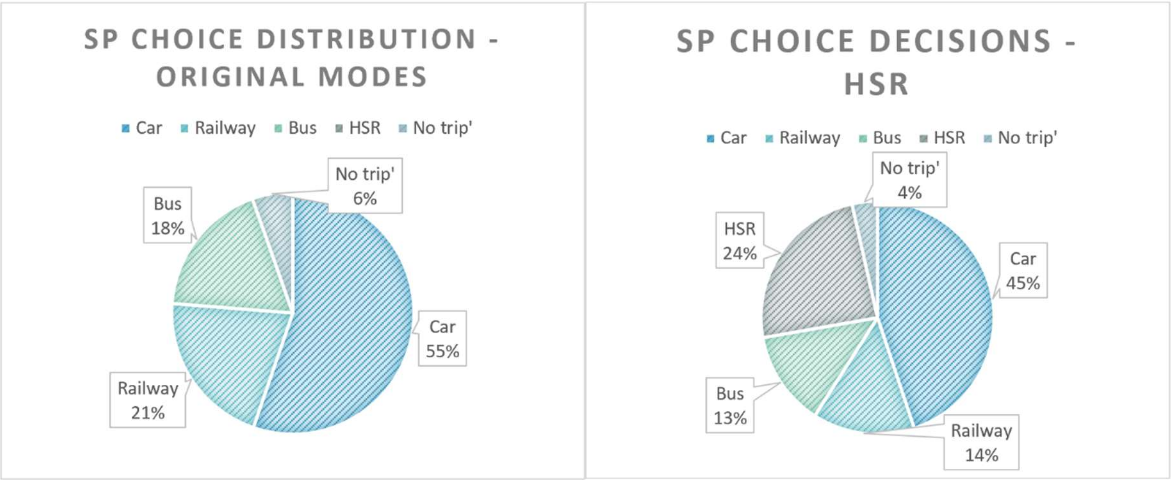


FIGURE 5: Distribution of SP choices

4.2. Results of the stated preference survey

Utility functions

Defining the utility functions is key to specify a mode choice model. The utility describes the set of parameters considered during the mode choice for each mode. The presumption is that for all users and modes a utility function has two parts, a measurable and a random part. The random part equates to the possibility that two individuals – in a presumable homogenous population – can choose different options with the same attributes. However, this random part of the utility is considered to have a mean equal to 0 with a distribution. (Willumsen, 2011)

The utility functions are highly determined by the information that the mode-choice model is based on, in this case the stated preference questionnaire. Even though the maximum number of attributes are fixed by the survey there are several considerations that had to be made. The main consideration is whether the used attributes are continuous or discrete parameters. In this case it is easy to determine that the cost, time and transfer parameters can be considered as continuous values, only the frequency can be considered as either a continuous or a discrete attribute. After the investigation of prior mode-choice models in Hungary and the evaluation of the data it was determined that the frequency should be used in the utility function as a continuous parameter. The following utility functions were used by all examined mode-choice models.

1. Utility function for car (U_{Car}):

$$ASC_{Car} + B_{TIME_{IVT}} * Time_{car} + B_{Cost} * Cost_{Car} \tag{4.1}$$

2. Utility function for rail (U_{Rail}):

$$ASC_{Rail} + B_{Time_{IVT}} * Time_{Rail_{IVT}} + B_{Time_{AccEgg}} * Time_{Rail_{AccEgg}} + B_{Cos} * Cost_{Rail} + B_{Transfer} * NoOfTransfers_{Rail} + B_{FRQ} * (Frequency_{Rail}) \tag{4.2}$$

3. Utility function for bus (U_{Bus}):

$$ASC_{Bus} + B_{Time_{IVT}} * Time_{BUS_{IVT}} + B_{Time_{AccEgg}} * Time_{BUS_{AccEgg}} + B_{Cos} * Cost_{Bus} + B_{Transfer} * NoOfTransfers_{Bus} + B_{FRQ} * (Frequency_{Bus}) \tag{4.3}$$

4. Utility function for high-speed rail (U_{HSR}):

$$ASC_{HSR} + B_{Time_{IVT}} * Time_{HSR_{IVT}} + B_{Time_{AccEgg}} * Time_{HSRV_{AccEgg}} + B_{Cost} * Cost_{HSR} + B_{FRQ} * (Frequency_{HSR}) \tag{4.4}$$

5. Utility function for "No Trip" choice ($U_{"NoTrip"}$)

$$ASC_{NO} = 0 \tag{4.5}$$

Multinominal Logit Model

The Multinominal Logit Model (MNL) is the simplest and most popular choice for a mode-choice evaluation. (Willumsen, 2011) The MNL model is based on a decision-making process of maximizing the utility. The model has three main underlying assumptions, a) the variables are assumed as random; b) equal variability through the cases; c) follows a Gumble (also called Weibull) distribution. (Li, 2011) The MNL model calculates the choice probability as the following equation:

$$P_{ijm} = \frac{e^{U_{ijm}}}{\sum_k^m e^{U_{ijk}}} \tag{4.6}$$

where U_{ijm} is the utility function of mode m.

One of the main advantages of the MNL model is that it allows handling new independent alternatives in the model as it satisfies the independence of irrelevant alternatives, which is described by Luce and Suppes (1965) “Where any two alternatives have a non-zero probability of being chosen, the ratio of one probability over the other is unaffected by the presence or absence of any additional alternative in the choice set.”. However, in the case of correlating alternatives, this attribute of the MNL model can prove to be disadvantageous. As the HSR is a new alternative the basic assumption is that the MNL model can provide a valid mode-choice model parameter set. Although, it is possible that HSR is seen as correlating alternative to the railway mode.

The result of the MNL model calculated by the Biogeme software (shown in TABLE 2) shows a high constant value for all transport modes. The rank of the constant values follows the previously assumed order, as the private car has the highest value, and the bus has the lowest. The constants also show, compared to the time and cost parameter values, that the initial affection to each mode has a high determining factor. The in-vehicle time has a beta value that corresponds to a value of time of 5.88 EUR/hour. The value of time for the access-, and egress time is 8.69 EUR/hour, which is almost double the other time component. The MNL model also shows a significant value for the transfers, as one transfer has almost the same value as one hour of in-vehicle time. The service frequency of public transport modes is less valuable, than any other parameter in research.

TABLE 2: Result of the MNL model

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
Likelihood ratio test for the init. model	45661.45								
Rho-square-bar for the init. model	0.307								
Akaike Information Criterion	102983.1								
Bayesian Information Criterion	103062.9								
ASC_BUS	4	0.0591	67.7	0	0.0649	61.7	0	**52.16**	*[EUR]*
ASC_HSR	4.74	0.0692	68.5	0	0.0733	64.7	0	**61.82**	*[EUR]*
ASC_CAR	5.45	0.042	130	0	0.0503	108	0	**71.07**	*[EUR]*
ASC_RAIL	4.25	0.0665	64	0	0.071	59.9	0	**55.43**	*[EUR]*
B_COST	-0.000213	3.20E-06	-66.7	0	3.47E-06	-61.6	0	**-**	-
B_FRQ	-0.00133	0.00022	-6.06	1.34E-09	0.000213	-6.27	3.64E-10	**1.04**	*[EUR/ hour]*
B_TIME_accegg	-0.0111	0.000679	-16.4	0	0.000658	-16.9	0	**8.69**	*[EUR/ hour]*
B_TIME_ivt	-0.00752	0.00015	-50	0	0.000151	-49.8	0	**5.88**	*[EUR/ hour]*
B_TRANSFER	-0.353	0.0117	-30.1	0	0.0114	-31.1	0	**4.60**	*[EUR]*

Nested Logit Model

The HSR transport mode can be easily perceived similar to an already existing transport mode. As it was stated previously, the MNL model has a high vulnerability in terms of calculating the parameters for multiple correlating transport modes. The nested logit model (NLM) can provide a reasonable alternative model to evaluate the HSR mode-choice parameters. The NLM model is declared by the grouping of the correlating transport modes, called nests, where each nest is represented by a so-called composite alternative which competes with other alternatives. (Willumsen, 2011) The utility (U_j) is calculated by a function of the utility of its alternatives. (4.7) The calculation of choice probability is the same as in the MNL model.

$$U_j = \phi_j * \left(log \sum_k \exp\,(\frac{U_k}{\phi_j} \right) \tag{4.7}$$

where U_j is the utility of the nest
U_k is the utility of the corresponding transport mode
ϕ_j is the structural parameter for the nest

There are also limitations to the use of the NLM model. The NLM model similarly to the MLM model cannot calculate with variation in the choice of the individuals. Also, as there are many possible nest structures available for a model with five transport modes, either prior knowledge or an extensive testing is needed to determine the correlating transport modes. (Wilumsen, 2011) In this research, with prior knowledge of other HSR mode-choice models, two different nesting structures were examined in detail.

Nest Structure 1.

The first examined NLM model presumes that the choice of "No trip" is included in one nest and all other transport modes are in the second nest. This structure is built on the presumption that the first choice of the individuals in the survey is whether they would like to make the trip or not. After that decision all the other available transport modes can be considered individual modes.

The results of the first NLM is shown in TABLE 3. The t-test shows that all the parameters calculated in the model are significant. The results show that the base perception of the transport modes is the same as in the MLM model, but closer to each other. The other significant difference between this model and the MLM model is that the value of the time spent accessing or egressing the particular transport mode is relatively lower compared to the time spent in the vehicle. In this model the value of the access and egress time is 7.8 EUR/hour and the value of the in-vehicle time is 6.6 EUR/hour. The MU structural parameter has a value different to 1, which provides a valid argument to use the NLM model over the standard MLM model.

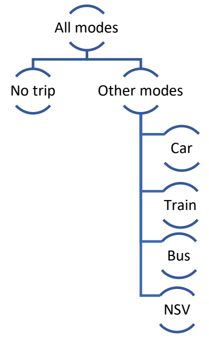


FIGURE 6: Nest structure for the NLM 1. model

TABLE 3: Result of the NLM 1. Model

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter
Likelihood ratio test for the init. model	47311.43							
Rho-square-bar for the init. model	0.318							
Akaike Information Criterion	101335.2							

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
Bayesian Information Criterion	101423.7								
ASC_BUS	3.44	0.0349	98.5	0	0.0344	100	0	**137.10**	[EUR]
ASC_HSR	3.61	0.0442	81.7	0	0.0436	82.8	0	**143.87**	[EUR]
ASC_CAR	3.81	0.0526	72.5	0	0.0517	73.7	0	**151.84**	[EUR]
ASC_RAIL	3.49	0.0382	91.5	0	0.0376	93	0	**139.09**	[EUR]
B_COST	-0.0000697	4.36E-06	-16	0	4.31E-06	-16.2	0	**-**	
B_FRQ	-0.000462	6.35E-05	-7.28	3.29E-13	6.22E-05	-7.43	1.08E-13	**1.10**	[EUR/hour]
B_TIME_accegg	-0.00326	0.000269	-12.1	0	0.000266	-12.2	0	**7.80**	[EUR/hour]
B_TIME_ivt	-0.00276	0.000172	-16	0	0.000168	-16.5	0	**6.60**	[EUR/hour]
B_TRANSFER	-0.0941	0.00667	-14.1	0	0.00663	-14.2	0	**3.75**	[EUR]
MU	4.02	0.255	15.7	0	0.255	15.8	0	**-**	

Nest Structure 2

The second examined NLM model assessed a situation where the “No trip” choice was one nest, the HSR was the second nest, as it is a currently non-existing mode in Hungary and all other transport modes were in the third nest. The result of this model structure is shown in TABLE 4.

The t-test revealed that with this structure the calculated parameters are also statistically significant. The separation of the HSR transport mode resulted in lower constant values across the model and as the HSR was separated it has a lower value than the other transport modes. The values of other parameters are close to the results given by the first NLM model structure, where the value of time for the access and egress time is 8.01 EUR/hour and value for the in-vehicle time is 6.23 EUR/hour. Similarly to the other nest structure, the result of the structural parameters shows that this classification of the nests is reasonable and cannot be calculated as an MNL model.

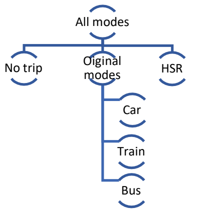


FIGURE 7: Nest structure for the NLM model 2.

TABLE 4: Result of the NLM 2. Model

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
Likelihood ratio test for the init. model	46302.01								
Rho-square-bar for the init. model	0.312								
Akaike Information Criterion	102244.6								
Bayesian Information Criterion	102333.2								
ASC_BUS	3.81	0.0432	88	0	0.0464	81.9	0	**71.03**	*[EUR]*
ASC_HSR	3.78	0.063	60.1	0	0.0682	55.4	0	**70.47**	*[EUR]*

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
ASC_CAR	4.68	0.0466	100	0	0.0502	93.2	0	**87.25**	*[EUR]*
ASC_RAIL	3.94	0.0493	79.8	0	0.0525	75	0	**73.45**	*[EUR]*
B_COST	-0.000149	3.49E-06	-42.8	0	3.55E-06	-42.1	0	**-**	
B_FRQ	-0.000689	0.000118	-5.84	5.31E-09	0.000115	-6	1.95E-09	**0.77**	*[EUR/ hour]*
B_TIME_accegg	-0.00716	0.000497	-14.4	0	0.000493	-14.5	0	**8.01**	*[EUR/ hour]*
B_TIME_ivt	-0.00557	0.000135	-41.2	0	0.000125	-44.4	0	**6.23**	*[EUR/ hour]*
B_TRANSFER	-0.207	0.00852	-24.3	0	0.00846	-24.4	0	**3.86**	*[EUR]*
MU	1.8	0.0437	41.1	0	0.0436	41.2	0	**-**	

Mixed Logit Model

The mixed logit model (ML) is one of the model structures that has been introduced in mode-choice modelling recently. The ML model is flexible and can counteract the main restrictions of the standard logit models by considering the random taste variation, unrestricted substitution pattern and correlation in unobserved factors over time. (Berkeley, 2002) The ML model probability is calculated by taking the integral of the standard logit probability over the density of a parameter. (4.8)

$$P_{ni} = \int \frac{e^{\beta x_{mi}}}{\sum e^{\beta x_{mi}}} f(\beta) d\beta \tag{4.8}$$

Hensher and Greene (2003) defined several crucial issues in the framework of a mixed logit model. The key issues regarding this research are the following:

- selection of the random parameter,
- selecting the distribution function of the random parameter,
- number of points in the distribution.

In the course of this research several options were considered for the questions above. After careful consideration and research on other HSR mode-choice models the parameter for in-vehicle time and cost were analysed in detail for the random parameter. In the initial analysis phase, the *normal* and the *lognormal* distributions were also considered for both random parameter distribution functions. The lognormal distribution function provided no valid results, therefore only the result of the normal distribution is presented in this research.

MLM model – random parameter: time

The MLM model, where the chosen random parameter was the in-vehicle time with normal distribution, provided a similar result to the MNL model. The output of the model displays that all parameters are statistically significant in this modelling construction. In terms of constant parameters for each transport mode there is a maximum of 1% difference compared to the MNL model. The biggest variation compared to the MNL model is the value of the service frequency.

TABLE 5: Result of the ML-time model

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter
Likelihood ratio test for the init. model	328009.8							

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
Rho-square-bar for the init. model	0.761								
Akaike Information Criterion	102989.1								
Bayesian Information Criterion	103077.7								
ASC_BUS	3.98	0.0578	68.8	0	0.0637	62.4	0	**51.90**	*[EUR]*
ASC_HSR	4.71	0.0678	69.5	0	0.072	65.4	0	**61.42**	*[EUR]*
ASC_CAR	5.45	0.042	130	0	0.0503	108	0	**71.07**	*[EUR]*
ASC_RAIL	4.23	0.0653	64.8	0	0.0699	60.5	0	**55.16**	*[EUR]*
B_COST	-0.000213	3.20E-06	-66.7	0	3.47E-06	-61.5	0	**-**	
B_FRQ	-0.00106	0.000185	-5.73	1.03E-08	0.000179	-5.93	3.11E-09	**0.83**	*[EUR/ hour]*
B_TIME_accegg	-0.0111	0.000688	-16.1	0	0.000664	-16.7	0	**8.69**	*[EUR/ hour]*
B_TIME_ivt	-0.00752	0.000151	-50	0	0.000151	-49.8	0	**5.88**	*[EUR/ hour]*
TIME_random	1.31E-06	0.000201	0.00652	0.995	1.28E-05	0.102	0.918	**-**	
B_TRANSFER	-0.353	0.0118	-29.8	0	0.0115	-30.8	0	**4.60**	*[EUR]*

MLM model – random parameter: cost

The model, where the cost was chosen as the random parameter, also has similar values to the ML model. By choosing cost as the random parameter all the calculated parameter values became lower (5-10%) compared to the MNL and to the previous ML model. This means that the cost proves to be a factor of higher significance in mode-choice than in the previously presented models. The values for the constant parameter range from 49.36 EUR to 67.48 EUR with a value of time of 8.23 EUR and 5.52 EUR.

TABLE 6: Result of the ML-cost model

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
Likelihood ratio test for the init. model	440854.4								
Rho-square-bar for the init. model	0.811								
Akaike Information Criterion	102918.6								
Bayesian Information Criterion	103007.2								
ASC_BUS	4.14	0.0624	66.3	0	0.0718	4.14	0.0624	**49.36**	*[EUR]*
ASC_HSR	4.92	0.0742	66.3	0	0.0825	4.92	0.0742	**58.66**	*[EUR]*
ASC_CAR	5.66	0.0505	112	0	0.0631	5.66	0.0505	**67.48**	*[EUR]*

Name	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value	Value of parameter	
ASC_RAIL	4.41	0.0704	62.6	0	0.0786	4.41	0.0704	**52.58**	*[EUR]*
B_COST	-0.000233	4.24E-06	-55	0	4.82E-06	-0.000233	4.24E-06	**-**	
COST_random	9.19E-05	7.04E-06	13.1	0	7.44E-06	12.4	0	**-**	
B_FRQ	-0.00117	0.000188	-6.21	5.37E-10	0.000182	-6.42	1.34E-10	**0.84**	*[EUR/hour]*
B_TIME_accegg	-0.0115	0.000699	-16.5	0	0.000676	-17	0	**8.23**	*[EUR/hour]*
B_TIME_ivt	-0.00772	0.000157	-49.1	0	0.000162	-47.7	0	**5.52**	*[EUR/hour]*
B_TRANSFER	-0.356	0.012	-29.8	0	0.0116	-30.8	0	**4.24**	*[EUR]*

4.3. Summary of the results

The five presented models have all shown to result in a parameter set that are statistically significant and show a reasonable value considering prior knowledge. However, there are statistical measurements presented in this chapters, that adhere to an objective decision on which model fits the dataset best. The models were ranked by these goodness to fit measurements and the best mode was chosen based on these ranks.

The first measurement is the *likelihood ratio test (LLR)*. The LLR provides a better understanding of the fitness of the model than the final log likelihood measurements. The log likelihood is highly affected by the number of calculated parameters, as any new parameter increases the degree of freedom and thus the fit, even though it might not explain the data better. (Bierlaire, 2020) The LLR can be calculated by using the initial log likelihood ($\mathcal{L}^i$) and the final log likelihood ($\mathcal{L}^*$)values. (4.9)

$$LLR = -2 * \left(\mathcal{L}^i - \mathcal{L}^*\right) \tag{4.9}$$

Among the five models the ML model, which considered cost as the random variable, proved to be the best. The other ML model also turned out to be also superior to the other three models. The worst performing model regarding the LLR measurement is the MNL model.

The second measurement that was applied is the *rho-square-bar* value. The ρ^2 provides an alternative measurement that takes the number of estimated parameters (K) to account. (4.10) The ρ^2 as value cannot be used as a measure of goodness, it’s only sufficient for comparing different model scenarios. (Bierlaire, 2020) The higher the ρ^2 value the model proves to be a better fit.

$$\rho^2 = 1 - \frac{\mathcal{L}^i - K}{\mathcal{L}^i} \tag{4.10}$$

Similarly to the LLR measurement the ML models also proved to be the best performing models, while the MNL model is the least performing model. The ML mode which considers the cost as random parameter is just slightly better than the ML model calculating with the time parameter.

The third measurement is *Akaike Information Criterion (AIC)*. The AIC similarly to the ρ^2 takes the number of estimated parameters into the calculation of the value. (4.11) In the case of the AIC measurement the lower the value the better the model fits.

$$AIC = 2K - 2\mathcal{L}^* \tag{4.11}$$

The AIC measurement shows different results than the previous two measurements. In the case of the AIC the two NL models provided the best results.

The fourth measurement is *Bayesian Information Criterion (BIC)*. The BIC goes a step further and takes not only the number of estimated parameters, but also the number of observations (N) to provide a statistical value that can describe the fitness of the model to the observed data. (4.12)

$$BIC = -2\mathcal{L}^* + \left(K * ln(N)\right) \tag{4.12}$$

The BIC measurement shows a similar result as the AIC measurement. The NL models show the best fitness to the observed data and the ML model with time as a random parameter is the worth.

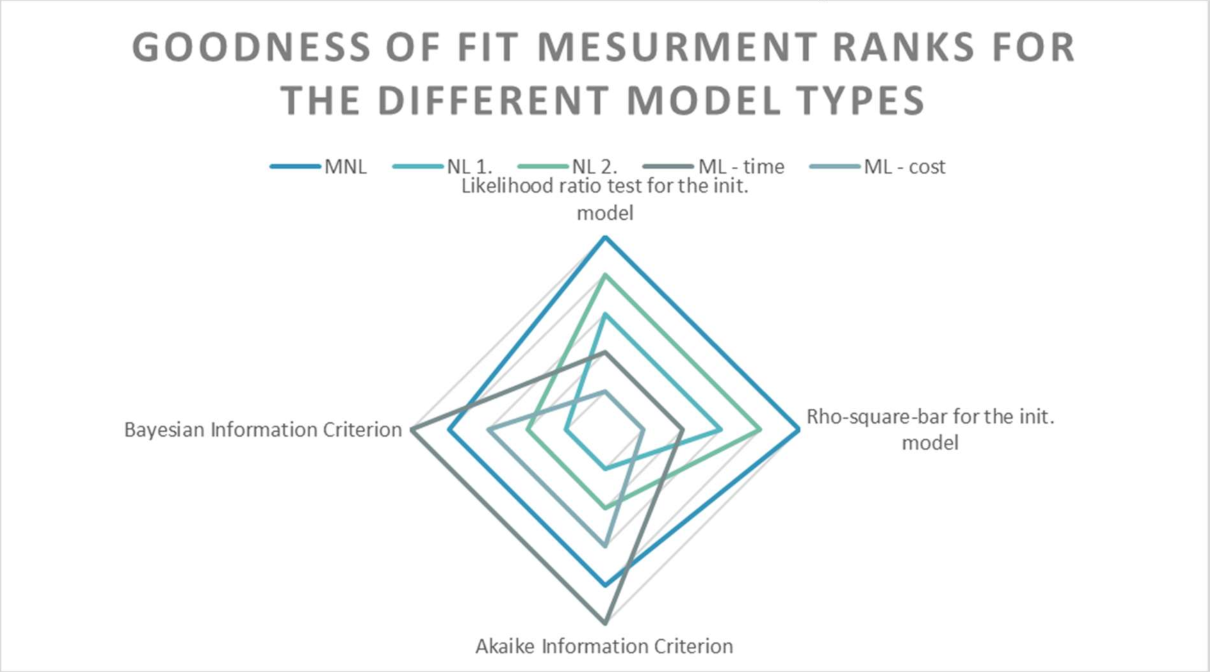


FIGURE 8: Goodness of fit measurement ranks for all models

Considering all the statistical measurements for the goodness of fit a ranking order was determined for the examined models. Based on the rank each model got a point measurement for each goodness of fit measurement. The model with the lowest value provides the best solution for the mode-choice model. Based on this calculation the first NL model and the ML – cost model had the same result. To determine which of these two models is the better one, an approach, similar to a Multi Criteria Analysis was used, which is common in transport development appraisals. The methodology differs from the previous calculation, as here a point rate between 0 and 10 was assigned for for each measurement of each model based on where the value of the measurement is situated between the minimum and maximum of all values. Based on this calculation the overall order didn't change from the third and the fifth place. Out of the best two models the NL model reached 20.2 points and the MNL model got 20.8 points. Based on this methodology the **Multinominal Logit model proved to be the best fit for the observed data**, although the nested logit model shows that it fits the data almost as well as the ML model.

The ML model that fit the data best has higher constant values than all other models. Even though the constants are high indeed, the difference between the best and worst transport mode is only 18.12 EUR. This means that the potential to change mode is quite low in Hungary, which is consistent with professional opinion.

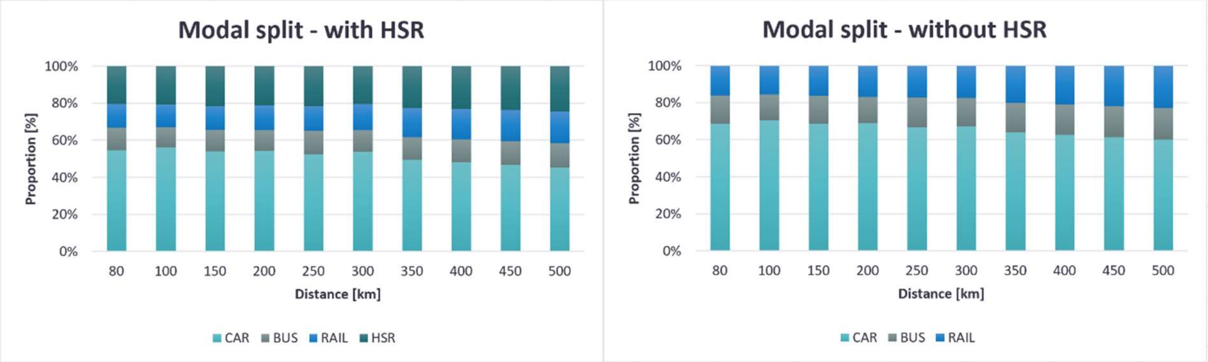


FIGURE 9: Modal split for the Mixed Logit Model – cost

The example modal split figure clearly shows that the introduction of the high-speed railway as a new transport mode decreases the modal split of all modes. The modal split of the HSR service is between 20-25%, dependant on the distance. The modal split of the passenger car decreases by 16-19%, the bus mode by 3-4% and the traditional railway mode by 0-2%. This means that based on this mode-choice model the introduction of a new HSR transport system in Hungary decreases the modal split of the private transport mode by a much higher percent than public transport modes.

4.4. Comparing the results with previous studies

The lack of Hungarian practice makes it essential to compare the results to previous studies. Although, there are several high-speed railways all over the world including Europe, there are only a few recent studies published on the methodology and results of the mode-choice demand that use stated preference data. Also, it is difficult to compare different study results is that the structure of the model is never the same, however, there are studies that have certain attributes that makes them comparable to the findings of this paper.

The most recent study on potential HSR demand was published in the Czech Republic for the planned HSR line between Praha and Brno. (Sudop Praha, 2020) In the Czech Republic there are no HSR lines established, similarly to the infrastructure of Hungary, so they also needed to conduct a stated preference survey to determine the future demand of the planned HSR lines. This research can provide a good comparison to my research as the Czech Republic has a similar size, population, and demand structure to Hungary. The survey had 3170 participants and had a similar structure as the Hungarian survey. In the first part the respondent's socio-economic status were asked and in the second part an SP questionnaire was presented. However, they chose to conduct the SP structure differently, using a hierarchical demand model structure. The respondents had chosen between public transport and private car in the first step and then between different public transport modes, including HSR in a separate question. This model structure is similar to a nested logit mode-choice model but the values of the variables were calculated separately. The model calculated values for cost, access and egress time, travel time, service frequency. The number of transfers in the first level and on the second level the parameters for in-vehicle time were separately calculated for each public transport mode. The used methodology has a very similar structure even though they used a different model structure. The result of the survey calculated 8.8 EUR/hour for the value of time, with 11.32 EUR/hour for the value of the access and egress time for public transport. These values are higher than the values calculated in Hungary, but consistent with the assumption that the access and egress time has a higher value than the in-vehicle time. The base perception of the different transport modes is the same as the conclusion of this paper, as the best transport mode is the private car followed by the HSR, the railway and the bus respectively.

A different model structure was presented by Wen, Weng and Fu in 2012 for a mode-choice model for a high-speed railway in Taiwan. The basic assumption of that paper was that a standard MNL model structure is not good enough to measure the demand of the HSR service and predict whatthe key variables are in the mode-choice decision. The analyzed data included 1200 respondents collected at the public transport terminals. The research evaluated several different model structures and determined that the latent class nested logit model proved the best fit for their mode-choice model. The final model has four different segments with public transport modes that were grouped in a nest. Even though this model has a significantly different structure, there are parts of the result that are similar to what was found in Hungary. The most obvious similarity is that the value of time for the access and egress times are higher than the in-vehicle time and to increase the potential demand of the HSR service it is important to consider the access to the service. Also, a very important conclusion of the paper is that different segments can have significantly different values, as the access time is more important for the lower income segment of the population and to trips related to leisure than for other types of trips. The use of different segments can significantly improve the goodness of fit of the model.

5. LIMITATION AND FUTURE RESEARCH

The research is focused on creating a universal mode-choice model to determine a potential transport demand for a high-speed railway service. The research is based on the assumption that a detailed socio-economic data, where the generated and distributed transport demand can be calculated for all segments is not available to all transport models. The collected data set is capable of estimating further models with the inclusion of socio-economic data as car-ownership, financial status, activity, or trip purpose, which can open the door for a segmented mode-choice model analysis.

Even though the result of this research is suitable to use in cases where there is no HSR service available, but as the data collection was limited to long distance national journeys, above 80 km (as it was stated in Chapter 3,I. This means that while investigating other transport development projects, these calculated mode-choice parameters cannot be used for trips under 80 km, which is a significant portion of the total national demand. In the future the research can be extended to short distance journeys with additional data collection.

6. CONCLUSION

The research had reaffirmed the initial hypothesis that the most widely used multinominal logit model cannot provide the desired accuracy for the mode-choice model, because it is lack of flexibility when introducing a new transport mode. The best fit mode-choice model was the mixed logit model out of the five examined scenarios, but only with a small margin to the nested logit model. The ML mode-choice model showed that the access and egress time for long-distance trips have a higher value than the in-vehicle time, which highlighted the importance of accessibility of the HSR services.

Altogether the analysis of the stated preference data proved that there can be a relevant domestic demand for the high-speed railway in Hungary. The mode-choice model had a similar result as the other research in this topic, even if the methodology was different. This proves the robustness of the calculated model. The approximately 20% modal-split of the HSR service reinforces the fact that under the right circumstances the HRS service can be desirable. The potential demand change from passenger cars is around 15-16%. This model can be used in the transport modelling practice as it is easy to import different transport modelling tools with minimal input data required. At the same time the collected data is suitable to create a sub-model for segment groups of the population.

7. REFERENCES

Adamowicz, W., Deshazo, J.R. Frontiers in Stated Preferences Methods: An Introduction. Environ Resource Econ 34, 1–6 (2006). https://doi.org/10.1007/s10640-005-4818-z

Arduin, J.P.; Ni, J.C. French TGV network development. Jpn. Railw. Trans. Rev. 2005, 40, 22–28.

Baibing Li, The multinomial logit model revisited: A semi-parametric approach in discrete choice analysis, Transportation Research Part B: Methodological, Volume 45, Issue 3, 2011, Pages 461-473,

Berkeley (2002), Choice models (https://eml.berkeley.edu/choice2/ch6.pdf)

Bierlaire, M. (2020). A short introduction to PandasBiogeme. Technical report TRANSP-OR 200605. Transport and Mobility Laboratory, ENAC, EPFL.

Chien-Hua Wen, Wei-Chung Wang, Chiang Fu, (2012) Latent class nested logit model for analyzing high-speed rail access mode choice, Transport Research Paper Part E 48(2012) 545-554

Department of transport, TAG Unit M1.2 (2020) https://www.gov.uk/guidance/transport-analysis-guidance-tag

M. Givoni, Development and impact of the modern high-speed train: A review, Transport Reviews 26(5) (2006) 593–611.

Hensher, D.A., Greene, W.H. The Mixed Logit model: The state of practice. Transportation 30, 133–176 (2003). https://doi.org/10.1023/A:1022558715350

European Union Commision (2019) European Green Deal (https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_hu)

Kroes, E. P., & Sheldon, R. J. (1988). Stated Preference Methods: An Introduction. Journal of Transport Economics and Policy, 22(1), 11–25. http://www.jstor.org/stable/20052832

Luce, R.D. and Suppes, P. (1965) Preference, utility and subjective probability. In R.D. Luce, R.R. Bush and E. Galanter (eds.), Handbook of Mathematical Psychology. John Wiley & Sons, Inc. New York.

UIC High Speed Department, High speed lines in the World. UIC High Speed Department, 2020, Available from Internet: http://www.uic.org/highspeed

Sudop Praha (2020), Summary of FS HSR Praha – Bruno SP survey

Willumsen (2011) Modelling transport, 4th edition