How European Travelers Make Decisions Regarding Airport Access Mode Choice

Slavica Dožić, Katarina Kukić, Milica Kalić, Danica Babić, Tatjana Krstić Simić
University of Belgrade – Faculty of Transport and Traffic Engineering
Belgrade, Serbia
s.dozic@sf.bg.ac.rs; k.mijailovic@sf.bg.ac.rs; m.kalic@sf.bg.ac.rs; d.babic@sf.bg.ac.rs; t.krstic@sf.bg.ac.rs

Aleksandra Colovic, Michele Ottomanelli
Dept. of Civil, Environmental and Building Engineering and Chemistry (DICATECh),
Polytechnic University of Bari
Bari, Italy
aleksandra.colovic@poliba.it; michele.ottomanelli@poliba.it

***Abstract*— Different socio-economic characteristics and factors related to transport system influence travellers' behaviour regarding transport mode choice for the travel to/from the airport. In this paper, characteristics that influence travel mode pattern in multimodal door-to-door service, with air transport mode as a main leg, are analysed. The Decision Tree (DT) models are built, based on the data obtained from the survey, conducted as a part of the SYN+AIR[1] project. Datasets, derived from the answers of the respondents from Italy, Spain, Greece, and Serbia are analysed separately. The results highlight the travellers' attitude and importance of factors that influence their travel mode choice, and show that travellers from different countries differently valued proposed factors. The reliability of the transport was often chosen as the most important factor by the respondents (regardless their country of origin). However, the proposed model showed that different factors for different markets also influenced travel mode choice. Obtained results confirm heterogeneity in different European air transport markets. Moreover, it is highlighted that different aspects of future service should be prioritized in order to implement a new multimodal door-to-door service.**

Keywords – classification model; decision tree analysis; multimodal transport; transport mode choice; airport access

I. Introduction

In order to provide seamless door-to-door (D2D) air passenger transport, it is necessary to implement fully integrated multimodal transport system, which will further lead to advanced transport system and new concepts (urban air mobility, autonomous vehicle, etc.). Fully integrated multimodal transport system will allow airport to evolve into multimodal nodes, where different modes of transport interact and cooperate. However, on their path to the full integration, due to the rapid urbanization of the cities, transport systems are facing many issues (such as traffic congestions, infrastructure accessibility, parking places, etc.) in providing efficient and coordinated multimodal transport service. Therefore, the analysis of travellers' mode choice to/from the airport is essential for having better understanding of travel demand, passengers' behaviour, as well as the factors that influence the passengers' mode choice. Through the last years, several papers focused on the characteristics that influence travel mode pattern, e.g., sociodemographic characteristics, age, income, residence, etc., [1]. The investigation of these factors is fundamental for establishing the relationship between transport supply and demand, and individuating the directions for providing cost-effective service by taking into account e.g., reliability, reduced waiting time, comfort, etc. Furthermore, this is the first step for providing coordinated multimodal service, where the main objective would be to shift the travellers' choice from private to public travel modes, [2]. This would contribute not only to the seamless multimodal D2D service, but also to the sustainability and environmental improvement of the cities.

In this paper, all of these aspects are considered in providing efficient multimodal D2D service, by having the main focus on the air transport mode, and including different transport modes (bus, train/metro, car, taxi) for arriving to/from the airport. For analysing the travellers' behaviour related to mode choice (for arriving to/from the airport), the Decision Tree (DT) model is built, based on the collected data (socio-economic characteristics, travel habits and factors that influence the travellers' mode choice) from the survey conducted as a part of the SYN+AIR project. Based on the answers collected from four countries that participate in the project (Italy, Spain, Serbia and Greece), corresponding datasets were prepared and analysed. The aim of this research is to explore the differences among travellers' attitudes and importance they give to the factors that influence their travel mode choice (for arriving to/from the airport), which has been scarcely investigated in the literature. Those differences are highlighted in obtained results of the DT models that also indicate pronounced heterogeneity of different air transport markets within Europe, caused by different levels of transport and infrastructure developments, but also by some cultural differences. Different approaches in decision making while

[1] Synergies between transport modes and Air transportation (SYN+AIR) project under grant agreement No. 894116, funded by SESAR Joint Undertaking.

choosing airport access/egress mode are also influenced by passengers' experience related to the most common purpose of travels. DT models are chosen because of their simple graphical representation and possibility to quantify importance of factors. Policy makers and transport service providers can easily interpret obtained results and use them to prioritize their decisions.

The structure of the paper is organized as follows. The literature review related to the different methods that are used for building DT, as well as their application for analysing travel mode choice behaviour, are reported in the Section 2. Section 3 provides the methodology and the concept of the DT model, as well as the description of the data that are used for building the corresponding model. In the fourth Section obtained results are reported, while the last section is devoted to the conclusions and further development.

II. Literature Review

There are certain number of papers that applied different decision tree methods for estimating and predicting the travellers' behaviour and mode choice, as well as the factors that influence the mode choice. For example, in [3] Random Forrest Decision Tree (RFDT) method is used for modelling the mode choice behaviour in Delhi, by considering eight different mode choices (public and private). The DT model has been compared to traditional Multinomial logit showing superiority of Random Forrest based DT model. Also, in [4] RFDT method is used for analysing travel mode choice behaviour considering household and individual attributes, built environment, and travel information as explanatory variables for travel mode choice. Differently, in [5] a hierarchical tree-based regression model for investigating the students' travel habits and mode frequency considering walking, cycling and public transit mode choice is used. The obtained results indicated the influence of distance, bicycle ownership, and school location for choosing the mode choice. Recently, two structured models have been developed for investigating the decision process of travellers' mode choice and the trip pattern based on the multi-day GPS data collected in Shanghai, where the car and public transit modes choice were chosen by 79.46% of respondents, while the rest use bicycle or walk. The results of the proposed models showed that the trip chain pattern decision precedes travel mode choice, as well as the tendency to shift to public transit when private cars are unavailable, [6]. Also, in [7] a DT is used to predict travel mode switching in Transjakarta, where the results showed that 57.1% of private vehicle users tend to shift from their mode choice. In recent paper [8], an application of DT, as tool to analyse satisfaction of highly educated people with airlines' services is presented.

Travel model choice is considered as an integral process of urban transportation planning in [9]. A systematic machine learning (ML) framework is proposed for a better understanding of traveller's mode choice decisions. To model the travel mode choices of travellers five different ML models are developed: Logistic Regression, Random Forests, Decision Tree, Multilayer Perceptron, and Light Gradient Boosting Decision Tree (LightGBDT), and tested on Dutch National Travel Survey data. The results showed that LightGBDT outperformed other models for both under and oversampling strategies. Additionally, it is revealed that trip distance, travellers' age and annual income, number of cars/bicycles owned, and trip density, significantly influence the travel mode decisions.

One of the most popular DT algorithms is Classification and Regression Trees (CART). For example, in [10] CART method is used for investigating the quality of metropolitan transit system based on five market segmentations (gender, age, frequency, travel habit and type of ticket), where punctuality and information resulted to be the most important variables. Also, in [11] CART is used for investigating the travel behaviour considering socio-economic, land use, and activity participation as variables, where the high-income travellers tend to use private modes. Hence, according to authors' (of this paper) knowledge, there are no papers in the literature that used CART for analysing the airport travellers' mode choice behaviour. Furthermore, the evaluation of passengers' mode choice regarding the airport access is based on the collected data related to the socio-economic and travel characteristics, which are obtained from the questionnaire considering four different European air transport markets (Italy, Spain, Serbia and Greece).

III. Methodology

This section provides basic concepts of decision tree model and gives insights in data used to develop the model.

A. *The decision tree model*

A DT is a statistical model widely used in many areas for predicting an outcome of targeted or dependent variable based on selected independent variables (also called attributes, predictors, features or input variables). One of the most popular methods for building DT is CART algorithm, which was introduced in [12], with the idea of representing data as a tree, based on the set of if-then rules. Starting from root node (known as the first parent node), which contains whole dataset used for training the model, DT continues to grow through internal (decision or child) nodes, which denote a tests or conditions with branches defining disjoint subsets of the data defined by the outcome of the test (True/False). Splitting continues up to leaf or terminal nodes that hold class label, and in that way a disjoint partition of the original sample is created. There are several different techniques to decide how to split the given data: Gini index, Information Gain, Information Gain Ratio, Entropy, etc. In this paper, Gini index for impurity (GI) is used for making splits of the dataset. Gini impurity represents the probability that randomly chosen data would be wrongly classified by a certain node, and can be calculated for each node by (1):

$$GI(node) = 1 - \sum_{i=1}^{N} \left(\frac{number\ of\ class\ i\ cases}{number\ of\ all\ cases\ in\ the\ node}\right)^2 = 1 - \sum_{i=1}^{N} (\mathrm{p}_i)^2 \quad (1)$$

where N is the number of classes of the independent variable and p_i is the probability of a particular element belonging to a specified class i. Further, weighted Gini indexes (WGI) for left and right sub-nodes are calculated. The weight of a sub-node is the number of samples in that sub-node divided by the total number of samples in its parent node. Weighted GI for entire split is obtained as sum of products of weight and GI for left and right sub-nodes. In order to clarify this explanation, an example of WGI calculation is provided with data from the root node on Fig.1 consisting of totally 310 samples divided in 113 in left and 197 in right sub-node. Thus, the weight of its' left sub-node is 113/310 and of the right 197/310. Further, GI of left sub-node (GI_L) is obtained from:

$$GI_L = 1 - \left(\frac{49}{113}\right)^2 - \left(\frac{64}{113}\right)^2 = 0.4912 \quad (2)$$

It is analogous for the right sub-node:

$$GI_R = 1 - \left(\frac{114}{197}\right)^2 - \left(\frac{83}{197}\right)^2 = 0.4876 \quad (3)$$

Hence, from (2) and (3) and calculated weight of sub-nodes, weighted GI for factor *Familiarity of the city* is calculated by (4):

$$WGI = \frac{113}{310} \cdot GI_L + \frac{197}{310} \cdot GI_R = 0.4889 \approx 0.5 \quad (4)$$

This procedure is performed for all attributes, and one with the smallest Weighted GI is selected for splitting. Finally, the most common label in leaf nodes is taken as a prediction.

Additionally, DT models provide opportunity to obtain feature importance that measures how GI decrease due to splits over a given feature. Briefly, importance for every node is calculated by subtracting sum of products *GI(sub-node)* and percentage of sample in left and right sub-nodes from percentage of sample in a node multiplied by *GI(node)*, by (5):

$$Importance(node) =$$

$$= \frac{(n_l+n_r)}{N} \cdot GI(node) - \left(\frac{n_l}{N} GI(left) + \frac{n_r}{N} GI(right)\right) \quad (5)$$

After that, importance for each feature is obtained as a ratio of sum of importance of nodes splitting on that feature and sum of all nodes' importance.

Standard metrics for evaluating the classification model are based on numbers of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) predictions. To describe the performance of a model on the test dataset for which the true values are known, confusion matrix is commonly used. Further, the accuracy is proportion of true predictions (TP+TN) among the total number of cases (TP+TN+FP+FN). Also, precision, recall (sensitivity) and F1 score are widely used and defined by equations (6) - (8):

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1score = \frac{2 \cdot Precision \cdot Recall}{Precision+Rec} \quad (8)$$

DT models become quite simple to understand when they are visualized. They can be used both for classification and regression problems. Additionally, scaling and normalization of data are not required for this technique and different types of independent variables can be used (continuous, ordinal and nominal). However, DTs have some disadvantages, such as probability of overfitting and the fact that small change in data can cause instability of the model.

Four different DT models were built, for four examined European air transport markets (Italy, Spain, Serbia and Greece). In order to make models comparable, the same set of independent variables is used in all DT models, with the same depth for all trees (set to three). Other hyper parameter tunings of the DT models were omitted, since the goal of this paper is to compare reasons for passengers' choices regarding airport access mode choice for different markets. Introduced assumptions result in models which are not the most accurate possible ones, compared to the models which would be obtained for each sample separately. Models of higher accuracy might be constructed with different sets of independent variables (e.g., in [13] where another DT model for Serbian market is built based on responses from Serbian sample).

B. Data

The data used to build models are obtained from the survey conducted in April and May 2021, for the purpose of the SYN+AIR project. The questionnaire consisted of 25 questions divided in three sections: the socio-economic characteristics of respondents, their travel related habits, and the factors that affect the choice of travel mode. The English version of the questionnaire, as well as descriptive statistics of the survey, may be seen in [14].

Due to Covid-19 restrictions, survey was conducted online, which resulted with certain limitations, common for online surveys. For example, difficulties to reach certain types of participants, such as those who do not have internet access or older respondents, possibilities to have too many unemployed respondents, chances of survey fraud etc. In order to make it easier for respondents to fill out the questionnaire, it was disseminated in five languages (Italian, Greek, Serbian, Spanish, and English). The planed size of the sample was 1200 respondents – at least 300 respondents from each of four countries. However, some administrative obstacles appeared with distribution of questionnaire for one of the projects' partners, which resulted by a lower sample size in Spain. The questionnaire was disseminated across different social networks, websites, air travel organizations, passengers' forums, by sending emails to different universities etc.

The total sample was quite well balanced related to socio-economic and travel related characteristic, which was accomplished by constant monitoring on the data collection. In total sample of 2199 valid responses, there were 54.43% females and 44.52% males (the rest of respondents declare either as other or rather not say), average age of respondents was 39 years. Regarding income, 61.1% of respondents have average household income, 20.6% high, 9.9% low and the rest choose not to say that information. The most of respondents (51.7%) are employed in private sector, 27.5% in public, 10.5% are students and the rest are retired, unemployed or other. When it comes to sample statistics related to travel profile: 44.8% of respondents travel often by plane, 32.8% rarely, 15.6% frequently and the rest 6.7% almost never. Mostly for leisure travel 42% of respondents, followed by 28.1% of respondents who travel mostly for business, 26.7% only for leisure and 3.1% of respondents travel only for business.

During the survey process, constant monitoring on the sample was performed in order to react and try to correct some observed irregularities in the total sample by choosing proper distribution channels. For example, after noticing that in sample from Greece there were far more females (which resulted in gender skewness of total sample), additional effort was put to reach more male respondents. It was balanced with a particularly good response from males in Serbia, and resulted by correction of gender distribution in total sample, but unfortunately also by gender skewness in samples from Greece and Serbia. Beside survey results for total sample, obtained results allowed comparing passenger behaviour in four countries to some degree, keeping in mind limitations of online surveys. Based on data from total sample, the influence of the main factors that have an impact on the non-coordination in the multimodal travel chain was examined using model based on Multinomial Logistics regression regarding airport access mode choice in [15]. The insight into the heterogeneities related to passengers' attitudes in observed four transport markets came as an added value of this research.

The fact that survey was conducted at the end of the first year of emergence of Covid-19 pandemic caused a problem in a sense that for vast majority of respondents it was difficult to focus on a specific, recent journey. Through the questionnaire, respondents were reminded to give answers related to regular travel conditions, before Covid-19 pandemic. Additionally, with the aim to quantify the trade-offs that users consider when selecting travel alternatives, the questions were hypothetical in an attempt to catch general emotions and attitudes of respondents, and not to reflect their experience with airport and public transport infrastructure only in their own country. Based on the obtained answers, one can say that this attempt was partially successful, since that, despite reminders in the text of the questions not to do so, part of the respondents replied based on their most common experience with the local airport and corresponding transport system.

In this paper, data from Italy (444 responses), Spain (194 responses), Greece (719 responses), and Serbia (562 responses), as countries where project partners originated from, were separately analysed, while totally 280 responses from other European countries were excluded, since the information about exact residence for those respondents was unavailable. Significantly lower number of respondents from Spain than from the other countries was not hindrance for implementing DT, since obtained sample size was adequate for this model. The selected socio-economic and travel related characteristics of respondents, which are used as variables in model are presented in Table I. In order to build DT based on GI as splitting criterion in Python's scikit-learn module, categorical data are recoded, and corresponding numerical values are given in parenthesis.

TABLE I. SELECTED SOCIO-ECONOMIC AND TRAVEL RELATED CHARACTERISTICS

Socio-economic and travel related characteristics	**Number of respondents**			
	Italy	**Spain**	**Greece**	**Serbia**
Gender				
Female (1)	243	82	581	188
Male (2)	197	104	135	371
Rather not say (3)	4	2	1	1
Other (4)	0	6	2	2
Average age of respondents (standard deviation) - shortly Age				
	40.6 (13.5)	41.3 (14.0)	34.0 (9.1)	42.3 (11)
Average household income – shortly Income				
Low (1)	46	17	139	7
Average (2)	324	140	479	264
High (3)	34	30	57	217
Rather not say (4)	40	7	44	74
Most common purpose of travel by plane – shortly Common trip purpose				
Business (1)	146	64	101	269
Leisure (2)	298	130	618	293
Frequency of travel by plane – shortly TravelFreq				
Almost never (1)	33	9	43	59
Rarely (2)	182	45	293	171
Often (3)	185	75	322	245
Frequently (4)	44	65	61	87
Type of luggage you usually have when travelling – shortly Luggage				
Carrie-on luggage (1)	371	126	488	281
Large baggage (2)	42	33	221	262
Small bag, backpack (3)	31	35	10	19
Mode choice for travelling to/from the airport (all modes available)				
Bus	16	18	11	7
Car (park at/near the airport)	50	47	182	37
Car (someone drops me off/picks me up)	149	30	382	216
Combination of modes	45	4	27	14
Metro	94	46	64	119
Other	2	0	0	2
Taxi (or ridesharing)	34	35	45	124
Train	54	14	8	43

Two main questions from the questionnaire for analysis are: a) question related to the mode choice to/from the airport, if all modes are available (distribution of answers per countries and per mode presented at Table I), and b) question regarding importance of factors which influence such choice. Namely, answers of the question "If all of the following transport modes are available, which one would you choose to travel to/from the airport?" were grouped into two new categories: PT (joined Bus, Metro, Train, Other, Combination of modes) – coded as 1 and car or taxi (joined Car (park at/near the airport), Car (someone drops me off/picks me up) and Taxi (or ridesharing services like Uber or Lyft)) – coded as 0. This binary variable is taken as dependent variable. Further, another question considered the importance of selected factors in mode choice when travelling to/from the airport, and it is taken as a set of nine independent variables, one for each factor. Answers were converted from five-point verbal scale (from not important to most important) to five-point numerical scale (respectively 1 to 5). Considered factors, their mean values regarding importance and the standard deviations (SD), for four considered samples, are presented in Table II.

From valuation of these factors, independently of residence, *Reliability* appears to be the most important for all respondents when choosing mode of transport to/from the airport (red coloured in Table II). The influence of the *Reliability* as a factor in the multimodal travel chain is analysed in detail in [13].

TABLE II. HOW MUCH THE FOLLOWING FACTORS INFLUENCE MODE CHOICE WHEN TRAVELLING TO/FROM THE AIRPORT

Factors	**Italy** Mean (SD)	**Spain** Mean (SD)	**Greece** Mean (SD)	**Serbia** Mean (SD)
Reliability (e.g., whether your bus may be delayed or stuck in traffic)	**3.98** (0.85)	**3.85** (0.90)	**4.08** (0.97)	**3.89** (0.86)
Security (e.g., the possibility of getting mugged)	**3.61** (0.97)	3.52 (1.02)	**3.61** (1.19)	**3.48** (1.14)
Crowdedness (a crowded bus or crowded train platform)	3.12 (0.95)	3.35 (0.89)	**3.54** (1.13)	**3.28** (0.93)
Travel Time (e.g., time spend in the bus)	**3.25** (0.89)	**3.70** (0.91)	3.41 (1.09)	3.18 (0.95)
Waiting Time (e.g., waiting for the train at the platform)	3.10 (0.91)	3.28 (0.90)	3.27 (1.15)	3.07 (0.93)
Cost (e.g., total cost of a bus ticket)	3.24 (0.91)	3.65 (0.95)	3.26 (1.16)	2.95 (0.98)
Familiarity of the city (e.g., whether it is your first time visiting a location, or travelling within your own city)	2.82 (1.00)	3.39 (0.90)	3.36 (1.16)	2.86 (1.03)
Trip purpose (e.g., leisure or business)	2.64 (1.07)	**3.66** (1.06)	2.74 (1.26)	2.59 (1.10)
Weather (e.g., rainy or cold weather conditions)	2.60 (1.05)	2.64 (0.88)	3.20 (1.18)	2.51 (1.02)

The second most important (blue coloured in Table II) is *Security*, and for respondents from Spain it is *Travel time*. But, starting from the third most important factor (green coloured in Table II), differences regarding residence, start to be more visible. That was the trigger to make DT models for each sample separately and to see how respondents make decisions whether to take public or private transport modes to/from the airport.

IV. RESULTS

For building DT models, Python's scikit-learn module was used. Separate datasets for each country were divided into test and train data, and the percentage split of 70% was used to build DT model, while the remaining 30% are used for validation. The tree depth was specified as three in all models, since those trees are quite easy interpretable, and all have satisfactory accuracy. Note that for example, in cases of Italy and Spain, tree depth of four would provide better accuracy. However, since the main idea is to compare the trees for different European markets, the parameters are set to be the same (in order to make meaningful comparison). The performances of the models are presented with the confusion matrix (Table III), which shows the correctly and incorrectly classified instances for each class, and from numbers of TN, TP, FN and FP other evaluation metrics may easy be calculated. Recall here that sizes of samples for Italy, Spain, Greece and Serbia are 444, 194, 719 and 562 respectively (of which 70% have been used for training the models and presented at DT models on Fig.1-4).

The fact that accuracies of obtained models vary between 56% for Italy and 87% for Greece is not surprising since the vast majority of respondents from Greece choose Car/Taxi, as transport mode choice to/from the airport. This is followed by slightly less percentage of respondents from Serbia, while in Italy and Spain numbers of respondents who chose Car/Taxi and those who chose PT were more balanced. Additionally, each model outer performs null accuracy (accuracy that could be achieved by always predicting the most frequent class). Only model for Serbia has almost the same accuracy as null accuracy (67%), while for Italy, Spain and Greece null accuracies are 52.5%, 57.7% and 84.7% respectively. As stated before, it is possible to obtain models for each country with higher accuracy by hyper-parameter tuning, but that assumes different combination of attributes per countries and comparison would not be meaningful.

TABLE III. CONFUSION MATRIX

		Predicted values							
		Italy		**Spain**		**Greece**		**Serbia**	
		0	1	0	1	0	1	0	1
Actual values	0	40	30	30	4	187	4	104	8
	1	29	35	18	7	24	1	49	8
	Accuracy	56%		62.7%		87%		66.3%	

Four decision trees are built, and graphically presented on Fig. 1-4, based on 15 input variables: importance of nine factors given in Table II and six socio-economic and travel related characteristics from Table I. Nodes coloured in blue predict PT, as a mode choice to/from the airport, while nodes in different tones of peach colour predict use of car or taxi (darker tone of both colours indicates higher data purity or greater homogeneity of nodes). Hence, darker nuances of blue denote nodes with greater share of respondents who chose public transport to/from the airport. Nodes painted in shades close to white colour are quite heterogeneous and share of respondents in such nodes is close to fifty-fifty, while darker peach nuances of the nodes mainly describe respondents who chose car or taxi.

From Fig. 1, one can see that respondents from Italy first split according to how important for them is familiarity of the city in which they are going to/from the airport. Those who find that factor more important, choose car or taxi, in majority. Male respondents who value *Reliability* as more or most important are more likely to choose PT. Finally, important factors for making decision about mode choice in Italian sample are also *Age* and *Gender*, and it could be concluded that passengers of age between 36 and 54 years more likely will choose PT, while older than 54.5, as well as younger than 36.5 more likely will choose car or taxi. Female respondents also prefer car or taxi choice, especially older female respondents (older than 57.5).

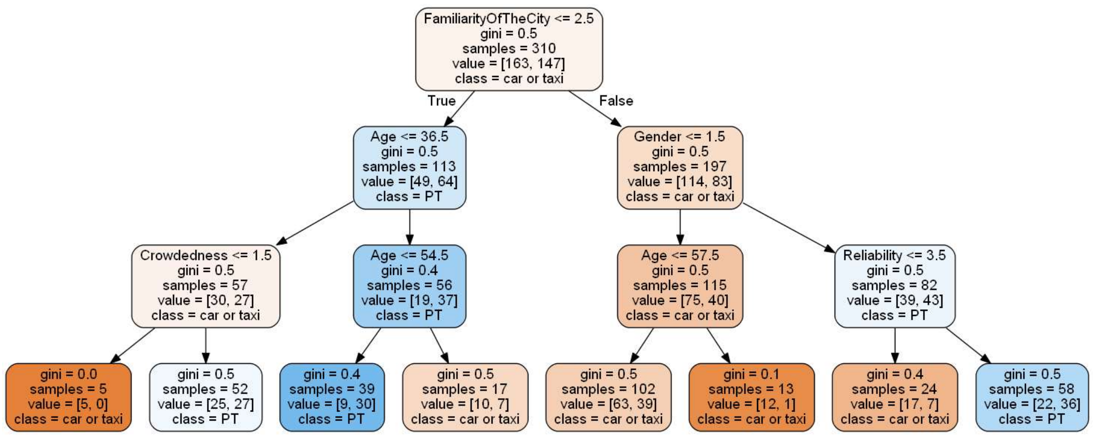

Figure 1. Decision tree model for respondents from Italy

Fig. 2 which presents DT model based on sample from Spain, shows that business travellers from Spain choose car or taxi in greater percentage than leisure travellers. Maybe not as expected, Spain respondents with high income or those who rather not say their approximate income are more likely to take PT. This might be related to better level of public transport infrastructure development in Spain compared to other countries involved in this research Also, this might be related to the residence location (in the major cities, the central areas are more expensive and attract residents with higher income levels, meaning that they have better accessibility to PT). However, since we do not have data about city of residence for respondents, this cannot be confirmed. Respondents who find *Travel time* as more or most important factor for transport mode choice will choose car or taxi in significantly higher percentage.

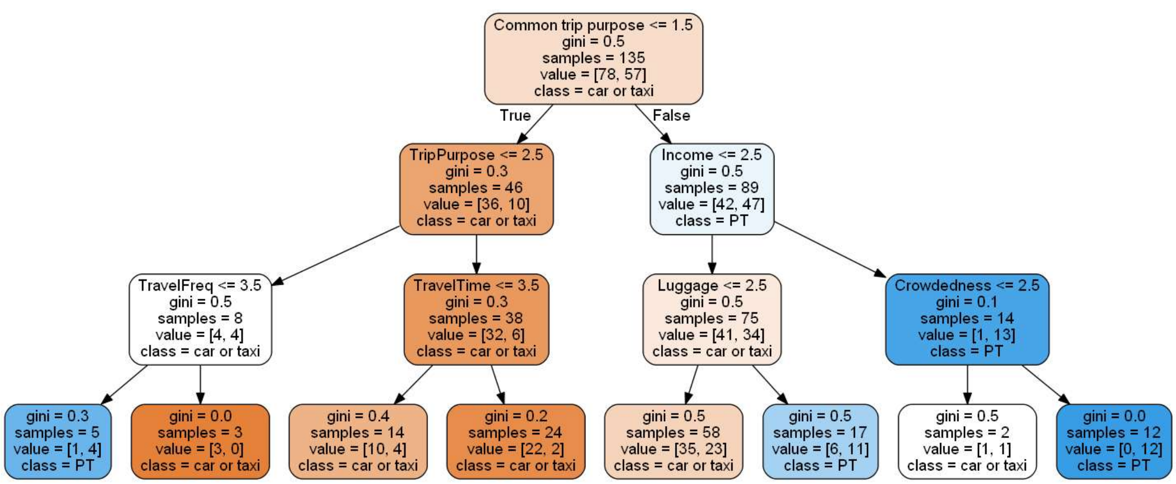

Figure 2. Decision tree model for respondents from Spain

Regarding respondents from Greece, vast majority of them opted for car or taxi (84.7%). Based on such sample, DT model also predicts class car or taxi for majority. It can be seen from Fig. 3 that travellers at age 20.5 or younger are more likely to take PT, as well as travellers without checked baggage and those who do not value *Travel time* as very important.

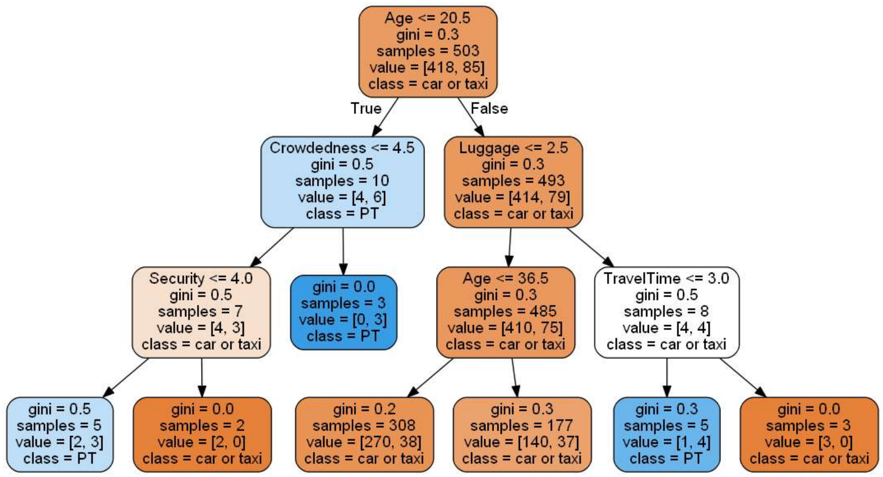

Figure 3. Decision tree model for respondents from Greece

Also, respondents from Serbia mainly chose car or taxi as a mode choice. Model presented on Fig. 4 shows that travellers from Serbia with carry-on luggage will take car or taxi in less percentage (60% of them) then those with large baggage (74% of them). Further, those who find *Security* as not or less important will take PT in higher percentage, as well as travellers who find *Cost* as most important factor. Note here, that *Cost* as a splitting factor does not appear in DT models for respondents from European Union, which might indicate lower income of Serbian respondents compared to respondents from other countries. *Familiarity of the city* is also splitting factor for Serbian respondents, and those who find it important, very important or most important will choose PT. Finally, respondents from Serbia who valued *Reliability* as most important factor that influence their mode choice to/from the airport will take car or taxi in smaller proportion than other respondents. Based on the obtained results, respondents older than 59 years for whom *Reliability* is most important factor, prefer PT when travelling to/from the airport.

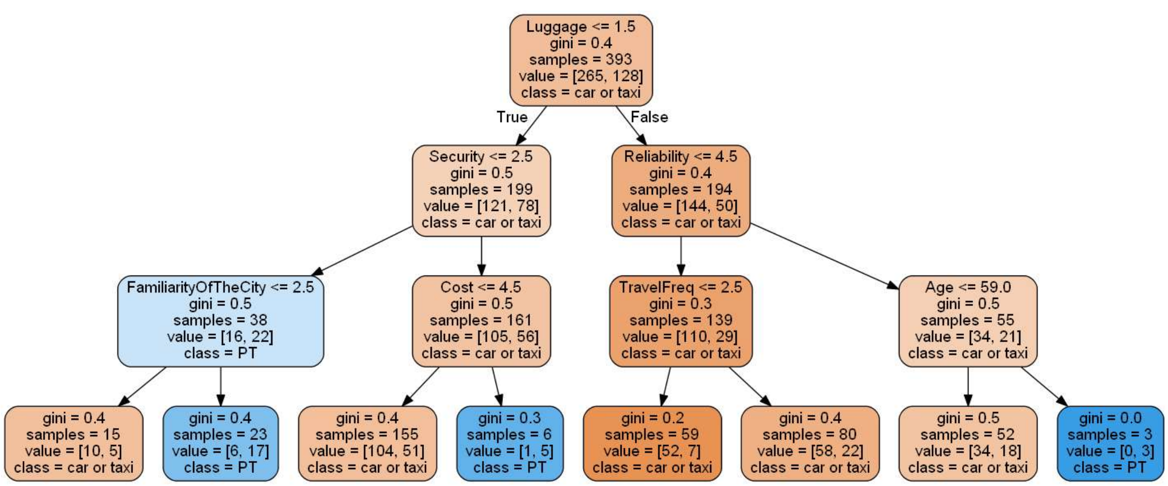

Figure 4. Decision tree model for respondents from Serbia

Based on developed DT models, importance of input variables for each sample is obtained, and presented in Fig. 5. It can be noticed that Spain and Serbia have seven different factors which influence decision making, while respondents from Greece and Italy use five different factors to make their decisions.

It clearly shows that travellers from different countries differently value selected factors when making decision about mode choice to/from the airport. It can be seen that the factor *Age* has the greatest importance in the case of Greece and Italy, while in the case of Spain the *Common trip purpose* and *Income* are the most influencing factors. In the case of Serbian respondents, factors *Luggage*, *Security*, *Familiarity of the city* and *Cost* are almost equally distributed, followed by *Age, Reliability* and *Travel frequency*. *Reliability* appears to be important only in DT for Italian and Serbian respondents, indicating that respondents from these two countries have less trust in public transport systems. *Security* as a splitting factor can be observed in the case of DT for Greek and Serbian respondents. *Crowdedness* appears to have influence while making decisions in all countries except in Serbia and it might be reflected through the fact that possibilities to reach airports in Serbia by public transport are very limited. *Gender* is a splitting feature only in DT obtained for Italian sample, where females appear to be less prone to use public transport services.

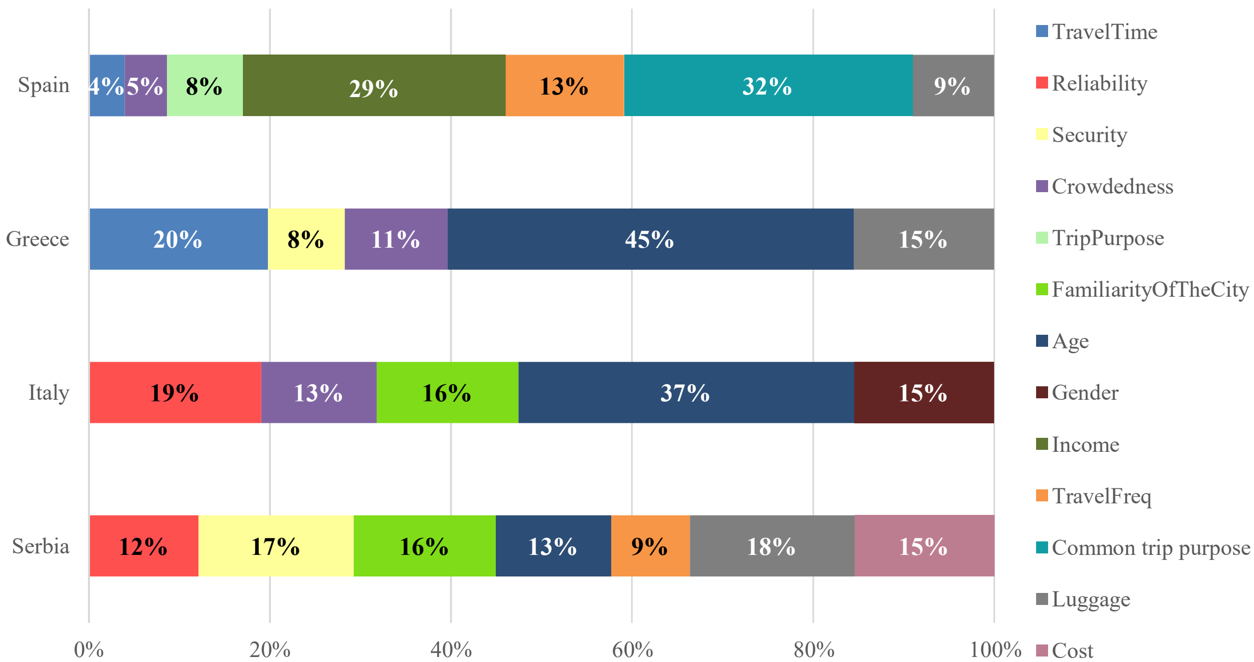

Figure 5. Importance of features in Decision trees

Despite the fact that the question regarding mode choice was hypothetical and asked in the form "if all modes are available what would you choose" (because the aim was to cover both departing and arriving airport and general attitudes of respondents), it is evident that respondents gave answers based on their experience, mainly related to their hometown. Generally, when respondents rated the importance of selected nine factors from Table II, *Reliability* appeared to be the most important factor (regardless the respondent's origin country). However, according to DT models, respondents do not make decision whether to take public or private transport to/from the airport mainly based on that factor. Proposed models showed that the given mode choice depends more on, for example, type of luggage travellers have, purpose of the trip, their age, gender etc.

V. CONCLUSION

Transport systems that incorporate efficient public transport are one of the most important elements for sustainable cities. Providing coordinated multimodal service, with the main objective to shift the travellers from private to public travel modes, is one of the essentials to support such system, and to provide certain environmental benefits, as well.

Different socio-economic characteristics (e.g., age, income, residence, etc.) and factors related to transport system (e.g., reliability, reduced waiting time, comfort, etc.) could influence different travellers' behaviour regarding travel mode choice.

Therefore, the analysis of travellers' mode choice is essential for having better understanding of travel demand.

The objective was to compare different transport markets (Italian, Spanish, Greek, and Serbian) and to investigate what are the most important factors for European air passengers when choosing whether to use public or private transport, to reach the airport. Decision tree algorithm was chosen, since the outcomes of these models are easily understandable and suitable for visual representation.

It is revealed that respondents from different countries differently valued proposed factors, when choosing travel mode to/from the airport. *Reliability* of the transport was valued as more or most important factor for vast majority of respondents, when they assessed importance of selected factors one by one. On the other hand, according to DT models, respondents make decision whether to take public or private transport to/from the airport mainly based on other factors (their age, gender, type of their luggage, purpose of the trip, etc.).

Developed DT models point out that importance of different factors which influence travel mode choice to/from the airport is different for analysed air transport markets in Europe. Thus, transport service providers, public transport managers and operators, as well as policy makers, should take into consideration heterogeneity of European air transport markets even when developing a new service on the European level. The results of DT analysis may provide useful information, highlighting measures which should be prioritized in order to introduce a new multimodal D2D service and helping to define directions and paths for implementation of such a new service regarding the continental, as well as regional and local levels, and with steps adjusted to considered markets.

The authors believe that presented models offer an easy interpretable and transparent method for identification of the most important decision variables. Additional value of DT is in their graphical presentation which would help those who will make decisions or create future plans.

However, some limitations of DT algorithm can be observed. Namely, these models are instable, in a sense that small changes in input training dataset may cause significant changes in output classification rules. Also, limitations of dataset obtained from online survey must be kept in mind. But, despite perceived shortcomings with data (obtained from online survey), such as the number of respondents from Spain and gender skewness in Greece and Serbia, obtained data provided solid foundation to highlight differences among respondents' attitudes based on their country and related transport markets.

In future work, the results of this research could be compared to actual/quantified provisions of public transport, to assess if the increased reliability of public transport would influence passenger shift from private to public transport (for arriving to/from the airport).

ACKNOWLEDGMENT

The authors are grateful to anonymous reviewers for their helpful and constructive comments, questions and suggestions that contribute to the paper improvement.

Parts of this study has been supported by SESAR Joint Undertaking within the project "Synergies between transport modes and Air transportation" (SYN+AIR) under grant agreement No. 894116 and the Ministry of Education, Science and Technological Development, Republic of Serbia, as part of projects TR36033 and TR36002.

REFERENCES

[1] Q. Wan, Z. Li, Y. Qi, J. Yu, Z. Pu, G. Peng and Q. Liu, "Comparing uncertainties in travel mode choice decisions for various trip chains," Adv. Mech. Eng. Vol. 11, pp. 1-11, 2019.

[2] D. Babić, M. Kalić, M. Janic, S. Dožić, and K. Kukić, "Integrated Door-to-Door Transport Services for Air Passengers: From Intermodality to Multimodality," Sustainability, vol. 14, 6503. 2022.

[3] C.R. Sekhar, Minal, and E. Madhu, "Mode Choice Analysis Using Random Forrest Decision Trees," Transp. Res. Procedia, vol. 17, pp. 644–652, 2016.

[4] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a Random Forest Method Approach to Model Travel Mode Choice Behavior," Travel Behav. Soc. Vol. 14, pp. 1–10, 2019.

[5] G. Zhan, X. Yan, S. Zhu, and Y. Wang, "Using Hierarchical Tree-Based Regression Model to Examine University Student Travel Frequency and Mode Choice Patterns in China," Transp. Policy, vol. 45, pp. 55–65, 2016.

[6] Y. Huang, L. Gao, A. Ni, and X. Liu, "Analysis of Travel Mode Choice and Trip Chain Pattern Relationships Based on Multi-Day GPS Data: A Case Study in Shanghai, China," J. Transp. Geogr. Vol. 93, 103070, 2021.

[7] D.M. Safitri, and I. Surjandri, "Travel mode Switching Prediction using Decision Tree in Jakarta Greater Area," International Conference on Information Technology Systems and Innovations, 2017.

[8] M.G. Bellizzi, L. Eboli, G. Mazzulla, and M.N. Postorino, "Classification trees for analysing highly educated people satisfaction with airlines' services", Transport Policy, Vol. 116, 199-211, 2022.

[9] M.T., Kashifi, A. Jamal, M.S. Kashefi, M. Almoshaogeh, and S.M. Rahman, "Predicting the travel mode choice with interpretable machine learning techniques: A comparative study," Travel Behaviour and Society, Vol. 29, pp. 279-296, 2022.

[10] R. De Oña, and J. de Oña, "Analysis of transit quality of service through Segmentation and Classification Tree Techniques," Transp. A Transp. Sci. vol. 11, pp. 365–387, 2015.

[11] C. Pitombo, A.J. Sousa, and L.N. Filipe, "Classification and Regression Tree, Principal Components Analysis and Multiple Linear Regression to summarize data and understand travel behavior," Transp. Lett. vol. 1, pp. 295–308, 2009.

[12] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and regression trees. Boca Raton: CRC Press; 1984.

[13] K. Kukić, S. Dožić, D. Babić, and M. Kalić, "A decision tree model for airport access mode choice: serbian air transport market," In Proceedings of the XLIX International Symposium on Operational Research (SYM-OP-IS 2022), Vrnjačka Banja, Serbia, 19–22 September 2022, pp. 557-563.

[14] K. Mavromatis, M. Kalić, K. Kukić, M. Ottomanelli, and A. Colovic, SYN+AIR: D3.1 Report on Customer Journeys, H2020-SESAR-2019-2. (Available online: https://syn-air.eu/).

[15] A. Colovic, S.G. Pilone, K. Kukić, M. Kalić, S. Dožić, D. Babić, and M. Ottomanelli, "Airport Access Mode Choice: Analysis of Passengers' Behavior in European Countries," Sustainability, vol. 14, 9267, 2022.