Prediction of Journey Destination for Travelers of Urban Public Transport: A Comparison Model Study

Vera Costa[1,2(✉)], Tânia Fontes[1], José Luís Borges[1,2], and Teresa Galvão Dias[1,2]

[1] INESC TEC - INESC Technology and Science, Porto, Portugal
veracosta@fe.up.pt
[2] FEUP - Faculty of Engineering, University of Porto, Porto, Portugal

Abstract. In public transport, smart card-based ticketing system allows to redesign the UPT network, by providing customized transport services, or incentivize travelers to change specific patterns. However, in open systems, to develop personalized connections the journey destination must be known before the end of the travel. Thus, to obtain that knowledge, in this study three models (Top-K, NB, and J48) were applied using different groups of travelers of an urban public transport network located in a medium-sized European metropolitan area (Porto, Portugal). Typical travelers were selected from the segmentation of transportation card signatures, and groups were defined based on the traveler age or economic conditions. The results show that is possible to predict the journey's destination based on the past with an accuracy rate that varies, on average, from 20% in the worst scenarios to 65% in the best.

Keywords: Urban public transportation · Travel patterns · Journey destination · Prediction models

1 Introduction

Enlargement of urban public transport (UPT) is essential to promote sustainable development of cities [1]. Nevertheless, the use of such systems by its users is not always easy, due its complexity and inflexibility [2]. To improve the efficiency of existing transportation networks, in recent years, UPT systems have adopted sophisticated Information and Communication Technologies (ICT). The use of such technologies allows the possibility to provide information to travelers using innovative ways [3] rather than expanding infrastructures [4]. Two main factors have contributed to this: (i) the adoption of smart cards in UPT; and (ii) the significant increase in the usage of mobile devices.

Adoption of smart cards has provided several benefits to the UPT management. To monitoring the network, surveys and other less reliable methods were replaced by these cards [5]. Therefore, UPT systems can provide real-time access to public transport data, which could be used for estimating the arrival times of buses, incidents, or delays [6]. Data provided by smart cards enable access to detailed information on the use, travel

J. C. Ferreira et al. (Eds.): INTSYS 2018, LNICST 267, pp. 113–132, 2019.
https://doi.org/10.1007/978-3-030-14757-0_9

patterns and demand. Also, the exploration of this data allows deriving useful information about transit passenger behavior, such as travel purpose or activity [7].

Availability of descriptive data about service usage will allow UPT providers to optimize the transport network and manage their resources more efficiently [8]. To do this, some information, as the main factors that influence travel's occurrence, should be well known. Having this knowledge, it is possible to define measures at two levels: (i) redesign the network, for example, by providing customized transport systems (e.g., DRT); or (ii) creating incentives to change certain travel patterns. Incentives may be, for example, pricing policies to restrict travels to some specific locations (e.g., access to monuments, cable cars), information management policies during the occurrence of critical events (e.g., floods, popular manifestations), or commercial policies to influence consumers to explore/visit other locations or at different times of the day (e.g., through the attribution of offers or discount vouchers). Taking in mind the development of efficient tools to implement these and other possible measures, the knowledge of journey destination when the user validates his ticket at the entrance of public transport, is of great importance. Thus, the central questions of this research are:

- How to predict the journey destination of a traveler from a UPT system? How can past data be efficiently used to improve such prediction?
- Are there any significant differences between the predictions of journey destinations for different traveler's groups? What are those differences and main reasons for this occurs?
- Is there any significant difference in the prediction of a journey destination for different time periods? How model these periods to improve such predictions?

To answer these questions, data from travel validations of a multimodal network of a UPT system, collected over a year, was explored. Two factors that may influence the travel purpose were considered: travel day and traveler type.

The paper is structured as follows: Sect. 2 presents the related work. Section 3 describes the methodology used while the presentation and discussion of the most significant results are present in Sect. 4. Some policy implications are presented in Sect. 5. Finally, the conclusions and answers to the previous questions are given in Sect. 6.

2 Related Work

The implementation of smart card-based ticketing occurred in many public transport systems around the world with different characteristics. Closed gate system allows an explicit recognition of patterns of mobility. It is possible to identify the origin and destination, time and duration travel, but there no transshipments information when multiple alternatives are provided. To avoid alighting delays, several bus services around the world use open gate systems. This system collects the origin of each journey without identifying the exact destination [9] and uses flat fare structures. However, some public transport providers are driven to adjust the price, based on the travel distance [10]. In this context, an algorithm of destination's inference based on the past usage of travelers is crucial.

Journey prediction is a central component that supports the development and delivery of personalized information services in UPT. Destination's inference provides relevant information to UPT providers, identifying behavior patterns, namely the traveler entrances and exits. The trend towards personalized Traveler Information Systems (TIS) supports the development of services capable of assessing and delivering contextually relevant information. The vast amount of data requires efficient processing and storage methods.

The latest developments in ICT have paved the way for the emergence of ubiquitous environments and ambient intelligence in UPT, mostly supported by miniaturized computer devices and pervasive communication networks. Such environments have been simplifying the collection and distribution of detailed real-time data, allowing the access to a rich information and support the development of next-generation TIS [11]. Some research has used these technologies to produce large matrices of origin-destination from smart card data [12]. These approaches focus on the destination's inference after the trip ended, allowing the identification of behavior patterns [13, 14], traveler segmentation [15, 16] and provision of information services [17]. Another group of studies was developed to understand travel patterns. Table 1 shows a summary of some studies formulated the last years.

Several authors studied urban traffic in different cities [18–21]. In these studies, some places in the city were identified as more popular. The conclusions allowed optimizing the public transport demand. On the other hand, the knowledge about travel time distribution along weekdays seemed to be another critical factor to optimize the system. Additionally, improving the knowledge of demand for public transport and identify travel peaks can allow transport providers to adjust the availability of vehicles [22].

The occurrence of individual travel and their primary purpose (work, school, for example) was also predicted [7, 23–25]. Thus, travelers with different mobility patterns were found. The acknowledging of their main patterns was one of the goals of a vast number of transport providers. In that way, two conditions are analyzed: if regular transit travelers tend to maintain their patterns and if it is easier to predict the future travels [26–28].

Although the analyzed studies focused on understanding the usage of public transportation and in the optimization of their network, a lack of knowledge about the best way of how a model can predict mobility patterns can be highlighted. Only a restrict number of works studied the prediction of journeys of travelers from public transport [23, 26]. In these studies, some of them distinguished travel patterns between different groups of passengers (with fixed routines and almost random routines) [20], which does not allow to understand if some models can be more suitable for specific traveler groups. Last, to our best knowledge an inter-comparison study to analyze different models and data configurations to predict the journey destination using UPT is missing, particular by exploring the potential of an extensive dataset.

Table 1. Summary of studies developed in the last years to know travel patterns using of urban public transport.

Author	City (Country)	Input variables	Methods or analysis	Period	Number of trips or users	Main conclusions
[18]	London (UK)	Origin, destination, individual trajectories without their history	Clustering (stations)	7 days	11.22 million trips	Heterogeneous patterns of intra-urban movement Large flows around a limited number of activity centers
[13]	Beijing (China)	Distance between stops, card ID, route number, driver ID, transaction time, remaining balance, transaction amount	Markov chain (prediction of origin)	1 day	36,246 validations	The method is effective in extracting transit passengers' origin information from transactions with relatively high accuracy (90%)
[22]	Shenzhen (China)	Card ID, action type, station ID, time of the action, check-in and check-out records	Spatial/temporal analysis (day's peak)	6 days	2.5 million trips	The intra-urban trips: - have two significant peak hours over a day - are different between weekday and weekend - have significant periodicity
[19]	London (UK)	Boarding the bus, entering into or exiting	Probabilities (visited locations)	3 months	626 users	Two most frequent locations can be modelled with fixed probabilities Other destinations (not the two most visited) are popular places in the city
[7]	Minneapolis/St. Paul (USA)	Date/time, route number, card type, is initial boarding or transfer, GPS location	Inference of trip purpose (work or school)	1 week	3,687 validations	Different groups of users have different routines The return trip time in the Post Meridiem (PM) peak is the primary determining factor of whether an activity is work-related
[26]	Beijing (China)	Card ID, route number, driver ID, transaction time, remaining balance, transaction amount, boarding and alighting stop	Clustering (mobility patterns)	5 days	3.8 million users	Most regular transit riders are commuters who do not own private cars and thus tend to be very sensitive to service reliability
[23]	Lisbon (Portugal)	Card ID, bus boarding time	Prediction of travel (travel occurs/not occurs)	61 days	24 million trips	Longitudinal data from automated fare collection (AFC) systems can be mined to uncover characteristic patterns of temporal regularities in accessing transport system

(*continued*)

Table 1. (*continued*)

Author	City (Country)	Input variables	Methods or analysis	Period	Number of trips or users	Main conclusions
[20]	Brisbane (Australia)	Boarding and alighting stop, boarding and alighting times, route ID, direction, card ID, card type, trip ID	GIS techniques (travel patterns)	1 day	5 million trips	Identification of traffic/users in the city's zones for different groups of people
[21]	Beijing (China)	Deal time and status, entry time, line and station, exit line and station	Analysis of spatial relationships (location, times)	1 day	8.7 million users	The urban development is increasingly concentrated near subway lines and transit stations The people flow in the morning peak shows that the construction of the new cities in Beijing's surrounding area is reinforced
[24]	Lisbon (Portugal)	Bus stops, geographic locations, bus line id, direction, stops on the route, card ID, time of bus boarding, id of the bus boarded	Personal+, Network+, (mobility patterns of urban bus riders)	61 days	24 million trips	Prior knowledge of the user's behaviour can improve the prediction For active users, the rider's own history covers a large portion of the future stop usage. For low demand riders, there is a high degree of uncertainty involved resulting in inaccurate prediction
[25]	All country (Netherlands)	Card ID, date, check-in time and location, check-out time and location	Route deduction	5 days	500,000 journeys	Found the route deduction to perform an accuracy of over 90% for the best selection rule, STA (Selected Least Transfers Last Arrival)
[27]	Oporto (Portugal)	Origin, destination (inferred), date, card ID, line, direction	Probabilities (destination)	2 months	5,000 journeys	Depending the probability stabilizes around two months of data (about 120 travels) or near three months of data (around 351–400 travels)
[28]	Oporto (Portugal)	Origin, destination (inferred), date, card ID, line, direction	Top-K, NB, J48 (destination)	2 months	800 users	The performance of journey predictions seems to be directly related to the mobility patterns
[9]	Oporto (Portugal)	Origin, destination (inferred), date, card ID, line, direction	Top-K, J48 (destination)	2 months	803,892 trips	Similar accuracy in the two methods. The Top-K is several orders of magnitude less memory demanding and much faster, showing great promise for large-scale systems

3 Materials and Methods

To predict the journey destination of a traveler using urban public transport three main steps were considered: (i) firstly, travel data was collected from different travel providers (first subsection); (ii) secondly, journey destination's inference was performed (second subsection); and (iii) last, journey destination's prediction was made (third subsection). Such methodology was applied to a European medium size Metropolitan Area. Figure 1 presents an overview of the methodology followed.

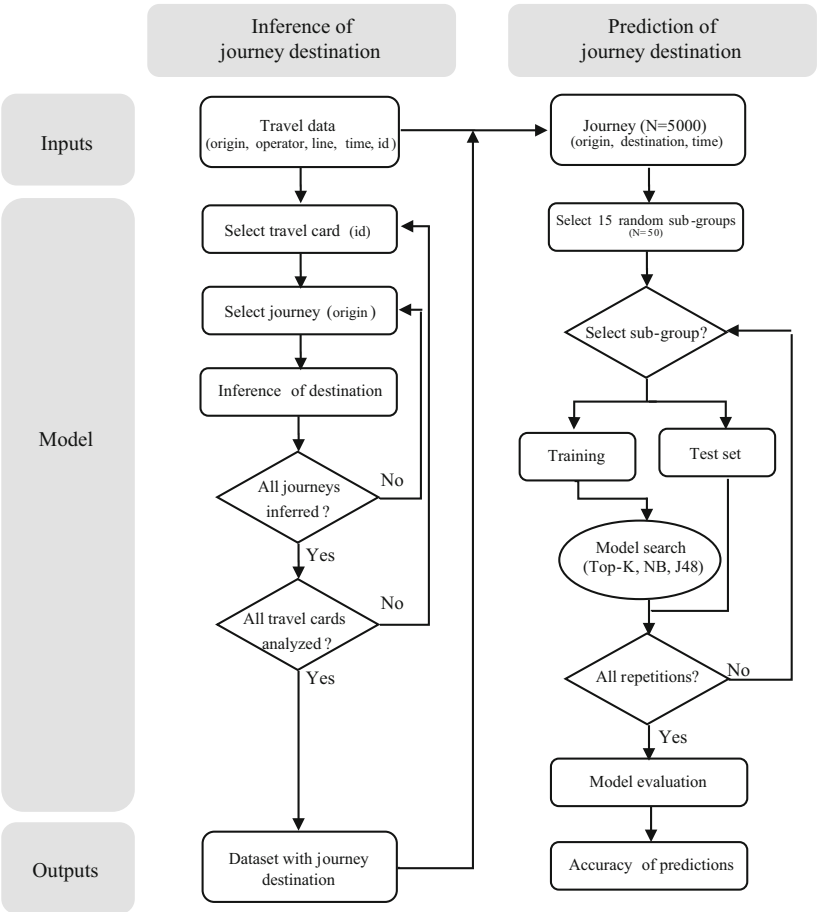

Fig. 1. Methodology overview.

3.1 Data Collection

One year of travel data obtained from the UPT of the Metropolitan Area of Oporto was collected (January 2013 to December 2013). Such network covers an area of 1,575 km^2

and serves 1.75 million of inhabitants. It is composed of 126 buses lines (urban and regional), six metro lines, one cable line, three tram lines and three train lines. This system is operated by 11 transport providers, which Metro do Porto (metro system) and STCP (bus system) are the largest [29].

Oporto network is based on an open and intermodal zonal system. The payment uses a rechargeable intermodal smart card called Andante. There are two types of Andante transport tickets: Signature Titles and Occasional Titles. Signature Titles have different groups of users where the charge depends, besides the journey length, also on the traveler age or economic conditions. While signatures cards can only be used to the cardholder, Occasional Titles can be used by different travelers (it has no personal information). Both cards are valid for a set of adjacent areas previously chosen by the passenger. Signature Titles are valid for the charged month while Occasional Titles are valid within the limit ring acquired during a particular period, currently 1 h for the minimum 2-zone ticket and longer as the number of valid zones increases. Thus, one journey may have one or more stages (validations), depending on the journey's period and the number of zones included in that journey.

For each traveler (i.e., for each Andante smart card), the information related to the boarding time (first boarding on the route), the line (or lines for each journey) and the stop (or stops for each journey) is available. Table 2 shows an extracted trip chain information for an individual traveler during a week of January 2013. The first row shows a journey with two stages. First, the traveler uses stop 2716 and line 303 at 9:14 a.m. followed the stop 3175 and line 302. The second row shows a journey with only one stage.

Table 2. Extracted trip chain information for a traveler during a week of January 2013.

Journey ID	Date	First boarding time of the route	Route sequence (Line ID)	Stop sequence (Stop ID)
1036866	02/01/2013	09:14 a.m.	303 → 302	2716 → 3175
1036867	02/01/2013	06:27 p.m.	203	1822
1036868	03/01/2013	11:24 p.m.	200	1035
1036869	04/01/2013	09:02 a.m.	402	2632
1036870	04/01/2013	10:29 p.m.	400 → 400	1675 → 1689
1036871	05/01/2013	09:09 a.m.	402	2632
1036872	05/01/2013	07:11 p.m.	206	1338
1036873	06/01/2013	08:45 a.m.	302 → 500	2632 → 1390
1036874	06/01/2013	10:11 p.m.	303	1338

To model urban travel patterns, this study used data from the two most significant urban public transport providers operating in Oporto's city: STCP and Metro do Porto, which corresponds to 135 million of annual validations distributed by different types of cards. Traveler routines were identified for different traveler profiles using Andante signature. Andante cards represent 67% of the total number of validations. In this work, six datasets of typical travelers were selected from the segmentation of Andante signatures based on the traveler age or economic conditions, as follow:

- G1 (4–12 students): includes students from 4 to 12 years old;
- G2 (13–18 students): includes students from 13 to 18 years old;
- G3 (sub23-superior students): contains students of higher education, public or private, with less than 23 years;
- G4 (normal): in general, this group is composed of regular workers. Any discount is provided for this group;
- G5 (social): includes people with a low monthly income (gross monthly income per household member smaller than 1.2 times of IAS - Social Support Index);
- G6 (seniors): includes retired people or people with 65 years old or more.

3.2 Inference of Journey Destination

UPT of the Metropolitan Area of Oporto is an open system which means there is tickets validation at the entrance only to each stage/journey. Thus, to know each traveler's destination, the application of an inference algorithm is required [30].

For this purpose, an updated version of an algorithm proposed by [10] was used in this work. Such algorithm is supported by the following assumptions:

1. "The most likely destination of a journey stage is the route stop located downstream from its own origin that is nearest to the origin of the next journey stage from that passenger";
2. "The most likely destination of the last journey stage of a day is the route stop located downstream from its own origin that is nearest to the origin of the first journey of the day from that passenger."

After setting candidate destinations, spatial validation rules were used to ascertain whether these assumptions are likely to hold for each transaction record. Some additional spatial validation rules were included in the proposed algorithm [10]. The rules are:

1. Origin and candidate destination of a journey stage are the same;
2. Candidate destination of a journey stage is beyond a set Euclidean distance from the next journey origin (or from daily origin if the stage is last) for the passenger;
3. Number of travel zones is exceeded for the passenger to reach the candidate destination.

Before applying this algorithm, the data was first pre-processed namely to remove: (i) validation records with missing data; and (ii) repeated validation records (or spaced by seconds or few minutes, but insufficient time for to go, to return and to go again). Also, a maximum walkable distance of 640 m was considered which corresponds to 8 min on foot at 4.8 km h^{-1}. This distance is recognized as the maximum walking distance for bus stops in Great London (TfL, 2010).

The algorithm used in this work inferred 85.9% of journey destinations. The destinations' percentage not inferred in the presented work could be related to the use of only 2 of the 11 transport providers operating in the network. Still, such result not affect

the further development of the work since traveler samples will be used to evaluate the capability of a model to predict the journey destination. Other values were obtained in similar studies: 71% in estimating alighting stations for rail boarding [31], 66% using a bus-only system [30]; and 80% in a multimodal public transport system [32].

3.3 Models of Prediction of Journey Destination

In this section, it is described models, models' optimization and results' evaluation used. All implementations and computations were performed using R software.

Models

To estimate the traveler destination, three models were applied: Top-K, Naïve Bayes (NB) and Decision Trees (J48).

Top-K. Top-K model is focused on the demand for more numerous elements (or item sets) based on an increment counter [33]. Two different techniques of Top-K are available: (i) Counter-based techniques, that keep an individual counter for a subset of the elements in the dataset, guaranteeing their frequency; and (ii) Sketch-based techniques, that provide an estimation of all elements, with a less stringent guarantee of frequency.

To optimize the performance and efficiency of predictions, required in the context of UPT, an update of Top-K model based in the Space-Saving technique that targets performance and efficiency for large-scale datasets was used in this work [33, 34]. The Space-Saving maintains partial information of interest, with accurate estimates of significant elements supported by a lightweight data structure, resulting in memory saving and efficient processing.

In this work, Top-K Space-Saving model was updated to account for the specificities of transportation networks, where a journey is considered to be an edge, i.e., a connection between any node A and B. This method showed to be several orders of magnitude less memory demanding and much faster [9]. Therefore, to identify the journey destination, only the origin of the journey (O) and the past origins of that traveler (OP) were considered in Top-K.

Top-K model applied for the first three days of the test set is shown in Fig. 2. Note that on day 1, all journeys would be correctly predicted based on origin, except for journey AC. In this case, journey AB is more frequent. However, on day 2, all counters are incremented based on the previous day, and journey CA would become increasingly relevant. Also, the new journey CD is added to the list. On day 2, the traveler always starts on C stop. As CA is the most frequent journey, the predicted destination failed for the three journeys (stop G, stop G and stop I). Finally, on day 3, no new journeys occur, but the counters are updated accordingly.

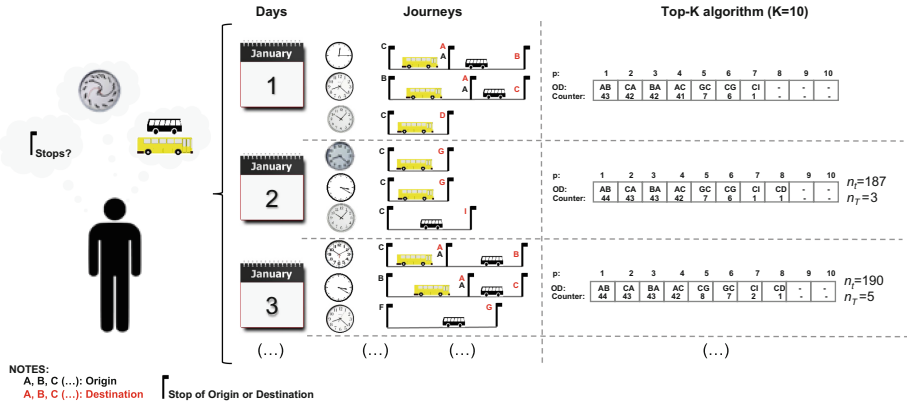

Fig. 2. Application of the Top-K model for the first three days of the test set.

Naïve Bayes. Naïve Bayes model is one of the most well-known classification techniques. This technique is based on statistical data and uses Bayes' Theorem proposed by Thomas Bayes to compute unknown conditional probabilities [35], assuming all attributes are independent given the class value, that is,

$$P(X|C) = \prod_{i=1}^{n} P(X_i|C) \tag{1}$$

where $X_i = (X_1, \ldots, X_n)$ is the feature vector and C is a class. A feature's probability in data appears as a member of the probabilities' set and is calculated by the frequency of each feature value within a class of a training dataset. Training dataset is a subset, used to train a classifier algorithm by using known values to predict unknown values [36].

Naïve Bayes is a very efficient model which have simplicity and unrealistic independence assumption. However, the Naïve Bayes classifier's performance is remarkably successful in practice [37].

Naïve Bayes classifier is also well known as very sensitive to the presence of redundant and/or irrelevant attributes. Redundant (highly correlated) attributes can bias the decision taken by this classifier [38]. Thus, only relevant attributes should be considered in this model.

In this work, the e1071's package of R software was used.

Decision Tree. Decision Tree model is one of the most widely used techniques for text-based automatic classification [39]. It is a tree-based knowledge representation methodology, which is used to represent classification rules in a simple structure. Tree' non-terminal nodes represent tests on one or more attributes, and terminal nodes reflect decision outcomes [40].

Decision Tree has several advantages over traditional supervised classification algorithms [41]. In particular, it is strictly nonparametric and does not require assumptions regarding input data distributions. Also, for missing values, it accepts nonlinear relations between features and classes and can receive both numeric and categorical inputs naturally [42].

To generate a decision tree model to classify the destination based on available training data's attribute values, J48 was performed. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data-mining tool.

In this work, RWeka's package of R software was used.

Model Search

To obtain a traveler's behavior representative sample throughout a year, they were selected 5,000 travelers with at least 300 validations inferred with success for each traveler group previously defined in Sect. 3.1 (G1–6). Student's groups (G1, G2 and, G3) do not have enough travelers in these conditions. So, in these cases, a reduced number of travelers were used (NG1 = 870, NG2 = 4,337 and NG3 = 3,707).

For each traveler group (G1–6) and model (Top-K, NB, and J48) 15 repetitions were computed. Each repetition included 50 travelers randomly selected. Predictions started on the 2nd day (January 2nd) by using the 1st day (January 1st) as a training set. Also, the training set was updated with previous journeys for each iteration. Consequently, for each traveler on an nth day, the model was trained with the corresponding training set, up to the (n − 1)th day, and predicted the journeys' destination for the nth day. After performing predictions, all journeys are added to the training set, and the iteration moves on (n + 1)th day. Thus, the test set is always composed of journeys of one day while training set continuously grows.

For each simulation, two different approaches were considered. Firstly, a simulation was applied to all days of the week (Sunday to Saturday), i.e., not distinguish different patterns of weekdays and weekends. Then, to consider such differences, two another simulation was performed, one for weekdays (Monday to Friday) and another for weekend days (Saturday and Sunday).

Model Evaluation

Accuracy measure (2) was applied to evaluate the algorithm's performance. This measure uses the confusion matrix, a two-way table, which summarizes the classifier's performance to represents the proportion of correctly identified results. Considering one class as positive (P) class and other as negative (N) class, four quantities may be defined: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). So, the accuracy (A) is given by:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

4 Results

Results are represented in two main sections. Firstly, a sensibility analysis of Top-K model is presented. Here, it is discussed the average accuracy variability for different K values (1st section). Secondly, the accuracy of the destination's prediction is analyzed taking into account the various traveler group routines (G1–6), the influence of different periods (weekdays and weekends) and the models' sensibility (Top-K, NB, J48). In this study, frequency and the average number of daily travels are analyzed (2nd section).

4.1 Sensibility Analysis

Figure 3 shows the annual average accuracy on weekdays and weekends for different K values (2–16) and traveler groups (G1–6). As it can be observed, the highest values are always obtained for travelers from group G4, both on weekdays and on weekends (65% and 45%, respectively). Travelers from group G6 and G2 has, respectively, the lowest average annual accuracy on weekdays (48%) and weekends (28%). Nonetheless, as it is possible to observe when K value is higher than ten, the accuracy keeps approximately a constant value for all traveler groups. Thus, a K value of 10 was adopted to retain more information.

4.2 Prediction Analysis

Figure 4 shows the monthly accuracy of destination's prediction for each traveler group (G1–6) on different time periods (weekdays and weekends) using three different models (Top-K, NB, J48). The variation represents the standard deviation using 15 repetitions.

In general, the maximum levels of accuracy to predict the destination of a traveler using UPT are reached after the model learn two months of travel patterns. For travelers from groups with high variance of daily travel patterns, such optimal is only achieved three to four months after the model starts the learning process. This usually happens, for example, for elderly and retired people (G6), since these travelers tend to frequent a big list of different places. Thus, a long-time period is required to identify all locations used by these travelers. This suggests the model's efficiency is directly affected by the number of different destinations used by a traveler and historical data available.

An in-depth analysis of average monthly accuracy variability shows different levels between different traveler groups, which is in line with the above findings. G4 and G5 are the groups with highest average accuracy. During weekdays, values are around 65% and 55%, respectively, for almost every month. Standard deviation has also lower than the remaining analyzed groups which suggest a higher confidence level.

In fact, G4 and G5 are mostly workers. Travelers from these groups have no fixed period for vacations. However, many of them keep routines in summer which may explain the higher accuracy values obtained. As opposite, particular groups of students (G1 and G2) need to restart frequently the learning process due holiday or exam periods, which affect the average capability of models to predict destination of these travelers. In these cases, the average accuracy range, during weekdays, between 25% and 55%. Lower accuracy values are obtained during periods of routine change.

For each traveler group, similar model performances were achieved for each analyzed model (Top-K, NB, J48). Accuracy monthly differences between each model are constant ranging between 1% and 5% (Fig. 4). A daily analysis of this accuracy shows higher differences.

Figure 5 shows an example for traveler group G4 distinguishing (i) weekdays and (ii) weekends. Note, however, in both graphs of Fig. 5. There is an evident similarity between the results from different models (Top-K, NB, and J48). This suggests, regarding accuracy, all models can be used equivalently.

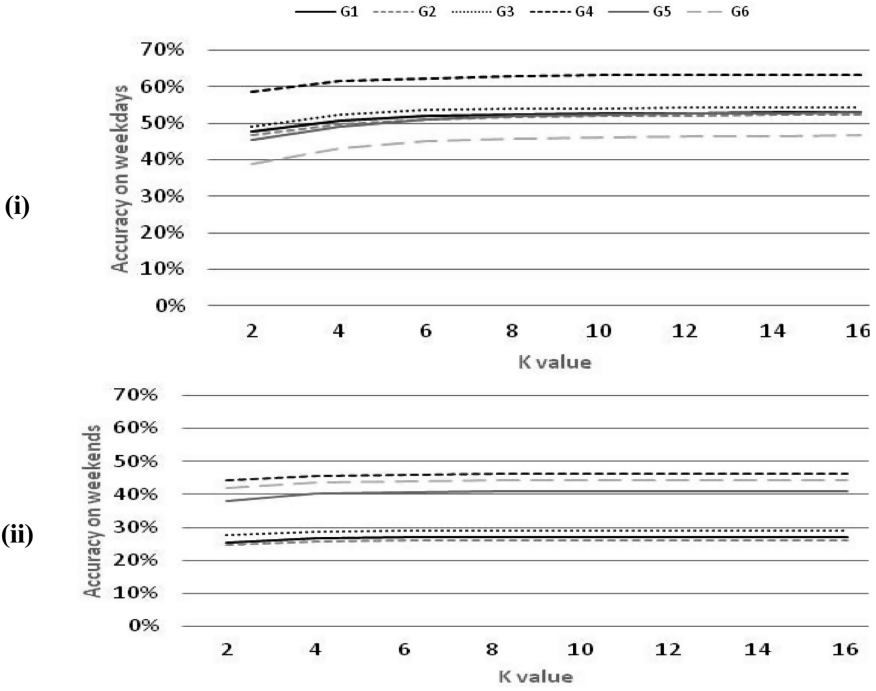

Fig. 3. Annual average accuracy of the prediction of journey destination for (i) weekdays and (ii) weekends of 2013. Predictions were made for different K values (2 to 16) and traveler groups (G1–6).

Although no variations of ranking accuracy performance between the models are observed, differences between weekdays and weekends can be identified. During weekdays J48 presents the highest values of accuracy (AG1 = 49.4%, AG2 = 47.3%, AG3 = 51.4%, AG4 = 63.2%, AG5 = 53.8% and AG6 = 47.2%), while during weekends the best model is NB (AG1 = 29.9%, AG2 = 29.0%, AG3 = 33.5%, AG4 = 48.9%, AG5 = 43.3% and AG6 = 46.5%). Additionally, besides accuracy other factors as the computing time must be considered by decision makers to select the most efficient model.

Analysis of accuracy's performance in predict a journey destination by traveler group allow to conclude the model ranking is not affected by short routines interruption. Short routines could be public holidays with one or two days, medium routines interruptions (as occurred during public holidays Easter and Christmas with one or two weeks), or even long routines interruption as summer vacations (July and August). Still, especially during long break travel patterns, the average capability of models to predict the journey destination decreases (Figs. 4 and 5). Such changes are not usually observed for short and medium interruptions. An exception to these patterns is the perceived for elderly and retired people (G6). For this traveler group, no significant variations of accuracy are observed across the year. This happens for this group because travel patterns seem not change significantly over the year. For long interruption periods, a high variance of the standard deviation is also observed.

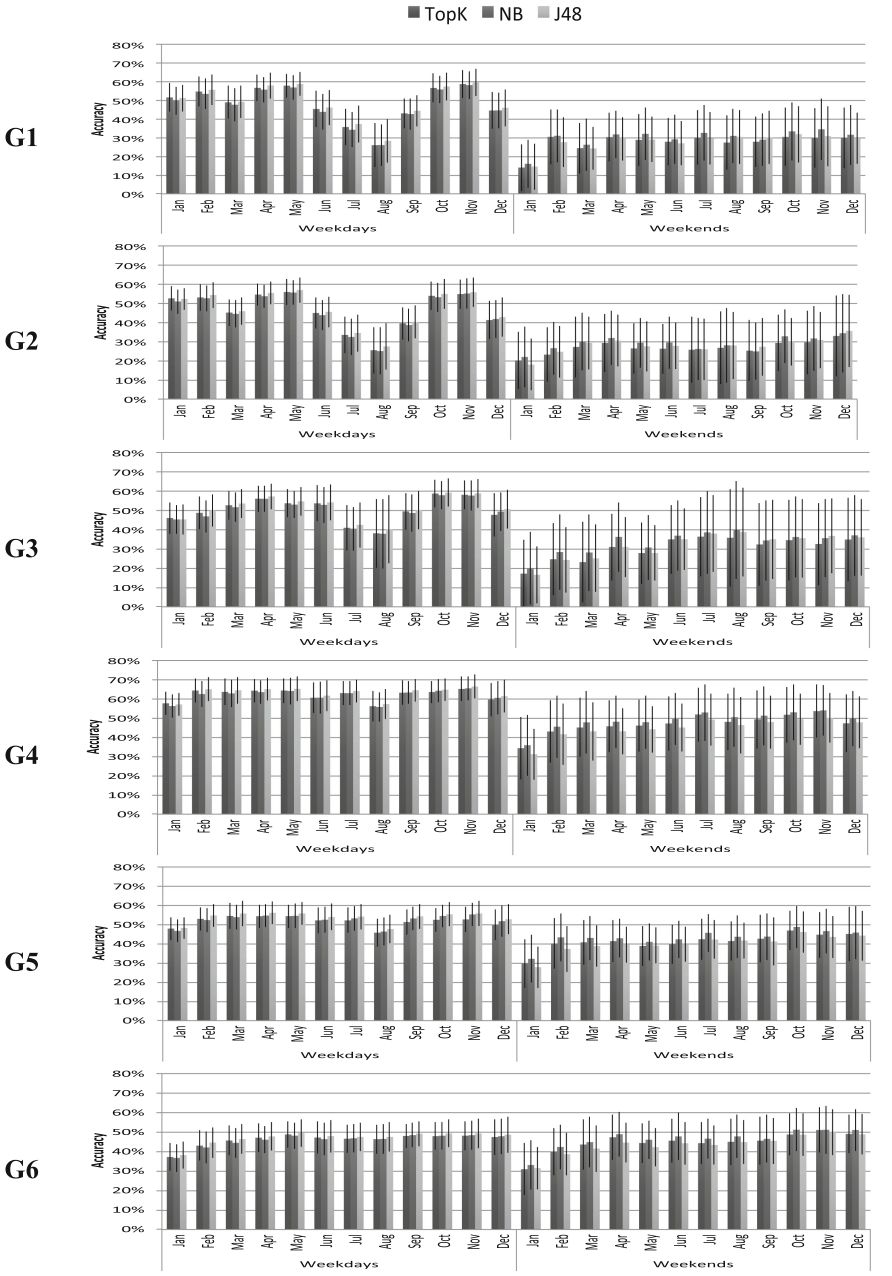

Fig. 4. Monthly accuracy of destination's prediction recorded for each model (Top-K, NB, J48), traveler group (G1–6) and time period (weekdays and weekends).

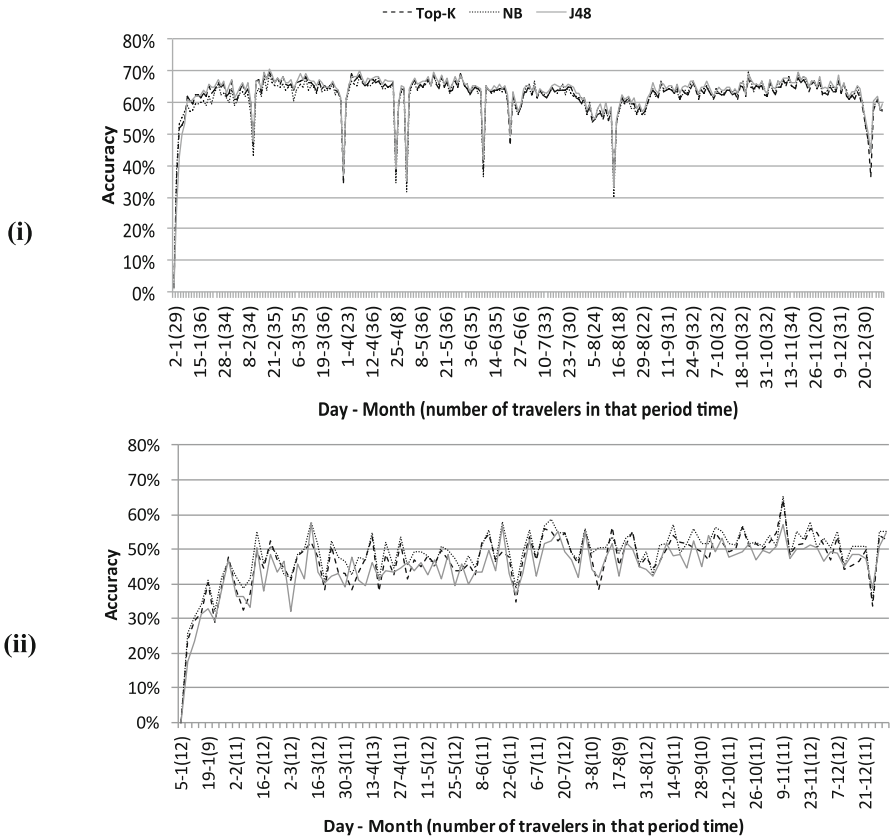

Fig. 5. Daily accuracy of the prediction of journey destination recorded during the (i) weekdays and (ii) weekends of 2013 for the traveler group G4 and applying different models (Top-K, NB and J48).

Since all models have similar values of average monthly accuracy for prediction of journey destination, Fig. 6 shows the daily accuracy (grey line) and the average number of daily travels (black line) for G4 using Top-K algorithm. The variation represents the standard deviation using 15 repetitions.

The average number of daily travels displayed in Fig. 6 was obtained by dividing the number of journeys in each group by the number of signature cards used. During weekdays and weekend days, the average number of trips is approximately constant along the year. The main exceptions are observed during strikes (February 1st, March 5th, June 27th, November 7th, November 26th) or public holidays (January 1st, February 12th, March 29th, April 25th, May 1st, June 10th, June 24th, August 15th, December 25th). During these periods, both on weekdays and weekends, a smaller number of signature travel cards are recorded (approximately 20%).

The average number of daily travels for G4 individuals are around 2.2 travels daily, which suggest that, on weekdays, journeys are mostly home-to-work and work-to-

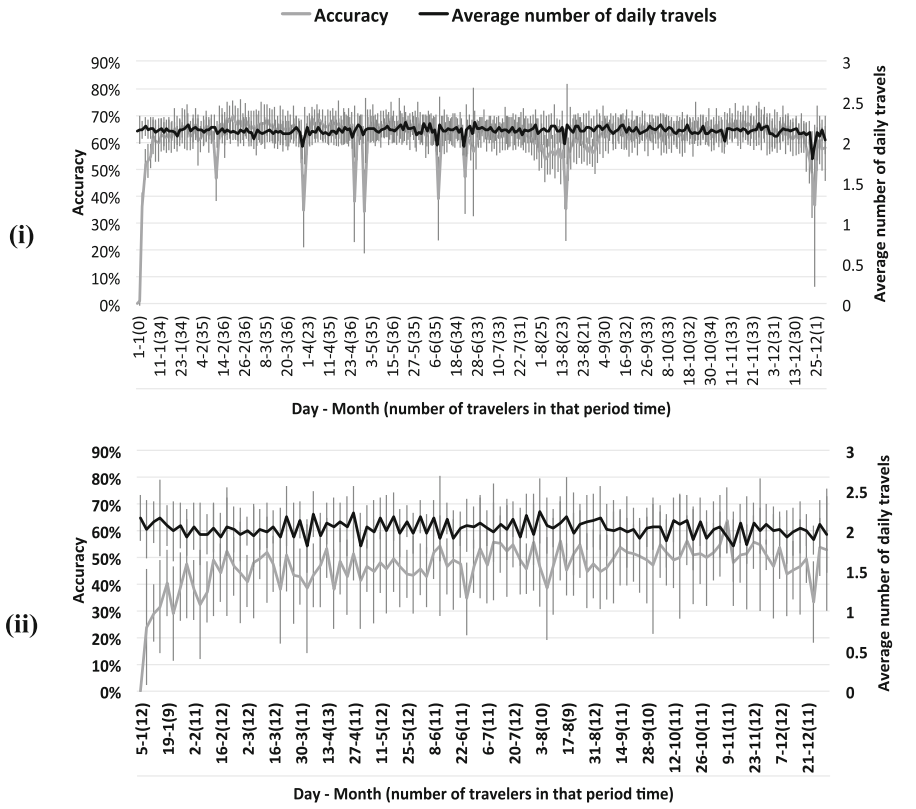

Fig. 6. Daily accuracy of destination's prediction (grey line) and average number of daily travels (black line), recorded during (i) weekdays and (ii) weekend days of 2013, for the traveler group G4 using the Top-K algorithm.

home. On weekends, this average has approximately the same values, which indicate the travelers go to a place and return.

Regarding the accuracy of destination's prediction, differences between weekdays and weekends were obtained. On weekdays, the average accuracy of travelers is around 65%. If holidays and strikes are excluded, it is possible to verify the very regular patterns for travelers from group G4, throughout the year, either during weekdays either weekend. In this group, deviations are lower and the accuracy only down about 10% in the summer months. In summer, the excellent weather usually invites travelers to visit different places in the city.

5 Policy Implications

Personalized information provision to travelers is a new research topic. While in closed urban networks the prediction of traveler destination is irrelevant, the opposite not happens with an open system as the implemented in Metropolitan Area of Oporto.

As explained previously in Sect. 1, such personalized information can be used to redesign the UPT network, by providing customized transport services, or incentivize travelers to change specific patterns. The redesign of UPT network can bring several benefits for travelers, namely: (i) to optimize travelers waiting time; (ii) allow to use public transport by a lower price by choosing a different transport route or time; (iii) benefit from offers or discount vouchers; and (iv) explore unknown parts of the city for travels with similar purpose (e.g. shopping, leisure). For transport providers, such redesign may: (i) improve levels of service appraisal; (ii) optimize the efficiency of services provided; and (iii) develop innovative and competitive services.

In this work, we obtained a maximum average accuracy of models evaluated ranging between 55% and 65%. Although these values are quite low, the results' deep analysis suggests the acceptance of many wrong predictions as correct by travelers. It happens because in Oporto's city the density of UPT stops is very high: 3,959 stops distributed by 1,575 km^2. Thus, for time-saving, travelers often use close stops to their primary destination.

Such happens especially during journeys affected by traffic congestion (e.g., accidents) or trips with double purpose (e.g., take a cafe before work).

For some traveler groups (e.g., G1 and G2) and particular days (e.g., weekend and public holidays) low levels of accuracy in journey destination's prediction are achieved. In these cases, personalized information cannot be provided with confidence. It seems not be critical to the overall system since during days with higher congestion (mostly weekdays) an acceptable level of confidence was achieved to the majority of travelers' population (G4 = 47.6%, G5 = 19.8% and G6 = 23.9% of total travelers using Andante).

6 Conclusions

Three models (Top-K, NB, and J48) were used to predict journeys destination for travelers of urban public transport of Oporto (Portugal). More than 90 million of trips recorded from signatures cards in two main transport providers of the city along one year were considered. Travelers were targeted in different groups, and different periods were analyzed. The results obtained show no differences statistically significant between the three prediction methods studied. Also, they provide answers to the initial questions:

- It is possible to predict the journey destination based on traveler's past. Such predictions are improved when past travel data are used. Additionally, some differences in accuracy in prediction of journey destination are observed between weekdays and weekends. While on weekdays, the higher average accuracy is reached between the second (February) and the fourth (April) months, on weekends, although little pronounced, the accuracy continuously grows over the year;
- Several differences were found between the predictions of journey destination for different traveler groups. The degree of success is widely affected by travel patterns. Since group G4 (normal) is the most regular, the high accuracy is found for this group. On the opposite, the worst values are found for students (G1, G2, and G3), the most irregular. In general, groups with stable work (G4) have established

routines while groups without work (G1–3 and G6) are very flexible regarding journey destinations;
- Significant differences in predictions of journey destination were found between (i) weekdays and weekends, and (ii) regular and irregular days. On weekends, the average accuracy is 20% to 30% lower than weekdays for all travelers. When a model is applied to all days of the week (i.e., without distinction between weekdays and weekend) 10% lower accuracy is obtained during weekends, which highlights the importance to predict weekdays and weekends separately. The learning process is also affected by the interruption of routines during irregular days (strikes, public holidays or school holidays). During these periods accuracy decreases 20–30%.

All this knowledge allows transport providers to detailed know their customers and to adjust the network or service. Future work expects to study the variation of travel patterns during irregular days, the model sensibility to sample size and to the use of past data. To improve the results for weekends and irregular days, an analysis of more efficient models using these datasets must be explored. Other types of travelers (as tourists) must also be investigated. Additionally, since the high density of stops in Oporto's city may be a reason for the low accuracy values, prediction of the user destination within a certain radius should be analyzed.

Acknowledgments. The Portuguese Science and Technology Foundation (FCT) funded the Doctoral scholarship of V. Costa (Ref. PD/BD/128065/2016) and the Post-Doctoral scholarship of T. Fontes (Ref. SFRH/BPD/109426/2015). The authors also acknowledge to the transport providers of Oporto, TIP, STCP, Metro do Porto and Transdev which provide travel data for the project, and also to our partner in the project, OPT company.

References

1. Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., Bento, C.: Predictability of public transport usage: a study of bus rides in Lisbon, Portugal. IEEE Trans. Intell. Transp. Syst. **16**, 2955–2960 (2015)
2. Gardner, B., Abraham, C.: What drives car use? A grounded theory analysis of commuters' reasons for driving. Transp. Res. Part F Traffic Psychol. Behav. **10**, 187–200 (2007)
3. Zimmerman, J., et al.: Field trial of Tiramisu: crowd-sourcing bus arrival times to spur co-design. In: Proceedings of 2011 Annual Conference Human Factors Computing System - CHI 2011, pp. 1677–1686 (2011)
4. Caragliu, A., del Bo, C., Nijkamp, P.: Smart cities in Europe. J. Urban Technol. **18**, 65–82 (2011)
5. Gordon, J.J., Koutsopoulos, H., Wilson, N., Attanucci, J.: Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. Transp. Res. Rec. J. Transp. Res. Board. **2343**, 17–24 (2013)
6. Weigang, L., Koendjbiharie, M.W., De Jucá, R.C.M., Yamasiiita, Y., Maciver, A.: Algorithms for estimating bus arrival times using GPS data. In: IEEE Conference on Intelligent Transportation Systems Proceedings, ITSC, pp. 868–873 (2002)
7. Lee, S.G., Hickman, M.: Trip purpose inference using automated fare collection data. Publ. Transp. **6**, 1–20 (2014)

8. Giannopoulos, G.A.: The application of information and communication technologies in transport. Eur. J. Oper. Res. **152**, 302–320 (2004)
9. Costa, V., Fontes, T., Costa, P.M., Galvao, T.: How to predict journey destination for supporting contextual intelligent information services? IEEE Conference on Intelligent Transportation Systems Proceedings, ITSC, pp. 2959–2964 (2015)
10. Nunes, A.A., Dias, T.G., Cunha, J.F.: Passenger journey destination estimation from automated fare collection system data using spatial validation. IEEE Trans. Intell. Transp. **17**, 133–142 (2016)
11. Patterson, D.J., et al.: Opportunity knocks: a system to provide cognitive assistance with transportation services. International Conference on Ubiquitous Computing (UbiComp), pp. 433–450 (2004)
12. Bera, S., Rao, K.V.K.: Estimation of origin-destination matrix from traffic counts: the state of the art. Eur. Transp. - Trasp. Eur. **49**, 3–23 (2011)
13. Ma, X., Wang, Y., Chen, F., Liu, J.: Transit smart card data mining for passenger origin information extraction. J. Zhejiang Univ. Sci. C. **13**, 750–760 (2012)
14. Kusakabe, T., Asakura, Y.: Behavioural data mining of transit smart card data: a data fusion approach. Transp. Res. Part C Emerg. Technol. **46**, 179–191 (2014)
15. Krizek, K., El-Geneidy, A.: Segmenting preferences and habits of transit users and non-users. J. Publ. Transp. **10**, 71–94 (2007)
16. Kieu, L.M., Bhaskar, A., Chung, E.: Transit passenger segmentation using travel regularity mined from smart card transactions data. In: Transportation Research Board, 93rd Annual Meeting Washington, D.C. (2014)
17. Foth, M., Schroeter, R., Ti, J.: Opportunities of public transport experience enhancements with mobile services and urban screens. Int. J. Ambient Comput. Intell. **5**, 1–18 (2013)
18. Roth, C., Kang, S.M., Batty, M., Barthélemy, M.: Structure of urban movements: polycentric activity and entangled hierarchical flows. PLoS ONE **6**, e15923 (2011)
19. Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C.: Spatiotemporal patterns of urban human mobility. J. Stat. Phys. **151**, 304–318 (2013)
20. Tao, S., Rohde, D., Corcoran, J.: Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. J. Transp. Geogr. **41**, 21–36 (2014)
21. Yu, W., Mao, M., Wang, B., Liu, X.: Implementation evaluation of Beijing urban master plan based on subway transit smart card data. In: Proceedings of 2014 22nd International Conference on Geoinformatics, Geoinformatics 2014, pp. 1–6 (2014)
22. Gong, Y., Liu, Y., Lin, Y., Yang, J., Duan, Z., Li, G.: Exploring spatiotemporal characteristics of intra-urban trips using metro smartcard records. In: Proceedings of 2012 20th International Conference on Geoinformatics, Geoinformatics 2012, pp. 1–7 (2012)
23. Foell, S., Kortuem, G., Rawassizadeh, R., Phithakkitnukoon, S., Veloso, M., Bento, C.: Mining temporal patterns of transport behaviour for predicting future transport usage conference item mining temporal patterns of transport behaviour for predicting future transport usage. In: ACM Conference on Pervasive Ubiquitous Computing Adjunct Publication, pp. 1239–1248 (2013)
24. Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., Bento, C.: Catch me if you can: predicting mobility patterns of public transport users conference item predicting mobility patterns of public transport users. In: IEEE 17th International Conference. Intelligent Transportation Systems, pp. 1995–2002 (2014)
25. Van Der Hurk, E., Kroon, L., Maróti, G., Vervest, P.: Deduction of passengers' route choices from smart card data. IEEE Trans. Intell. Transp. Syst. **16**, 430–440 (2015)
26. Ma, X., Wu, Y.J., Wang, Y., Chen, F., Liu, J.: Mining smart card data for transit riders' travel patterns. Transp. Res. Part C Emerg. Technol. **36**, 1–12 (2013)

27. Fontes, T., Costa, V., Costa, P.M., Dias, T.G.: Analysis of urban mobility patterns using data from public transport ticketing system: implications for developing autonomic systems. In: Autonomous Road Transport Support Systems Early Career Research Conference, Greece, pp. 1–6 (2015)
28. Costa, V., Fontes, T., Costa, P.M., Dias, T.G.: Prediction of journey destination in urban public transport. In: Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, pp. 169–180 (2015)
29. TIP: Relatório e Contas. Transportes Intermodais do Porto (2014)
30. Trépanier, M., Tranchant, N., Chapleau, R.: Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. Syst. Technol. Plann. Oper. **11**, 1–14 (2007)
31. Zhao, J., Rahbee, A., Wilson, N.H.M.: Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput. Civ. Infrastruct. Eng. **22**, 376–387 (2007)
32. Munizaga, M.A., Palma, C.: Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C Emerg. Technol. **24**, 9–18 (2012)
33. Metwally, A., Agrawal, D., El Abbadi, A.: Efficient computation of frequent and top-k elements in data streams. In: Eiter, T., Libkin, L. (eds.) ICDT 2005. LNCS, vol. 3363, pp. 398–412. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30570-5_27
34. Sarmento, R., Cordeiro, M., Gama, J.: Streaming networks sampling using top-K networks. In: ICEIS 2015 - 17th International Conference on Enterprise Information Systems Proceedings, vol. 1, pp. 228–234 (2015)
35. Yildirim, P., Birant, D.: Naive Bayes classifier for continuous variables using novel method (NBC4D) and distributions. In: 2014 IEEE International Symposium Innovations in Intelligent Systems and Applications Proceedings, pp. 110–115 (2014)
36. Patil, T.R.: Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**, 256–261 (2013). ISSN 0974–1011
37. Hand, D.J., Yu, K.: Idiot's Bayes-not so stupid after all? Int. Stat. Rev. **69**, 385–398 (2001)
38. Inza, I., Larrañaga, P., Etxeberria, R., Sierra, B.: Feature subset selection by Bayesian networks based optimization. Artif. Intell. **123**, 157–184 (2000)
39. Peng, Y., Ye, Y., Yin, J.: Decision tree construction algorithm based on association rules. In: Proceedings 2nd International Conference on Computer Science and Application System Model, pp. 754–756 (2012)
40. Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H.: A comparative study of reduced error pruning method in decision tree algorithms. In: Proceedings 2012 IEEE International Conference on Control System Computing and Engineering, ICCSCE 2012, pp. 392–397 (2012)
41. Friedl, M.A., Brodley, C.E.: Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. **61**, 399–409 (1997)
42. Fayyad, M.U., Irani, K.B.: The attribute selection problem in decision tree generation. In: Proceeding of AAAI 1992 Proceedings of tenth National Conference on Artificial intelligence, pp. 104–110 (1992)