XLIII CILAMCE
21-25, November, 2022
FOZ DO IGUAÇU - PR - BRAZIL

# Infering the passenger's trip purpose in the *Smart card* data using data mining techniques, an case study of the Belo Horizonte Brazilian city

M. G. O. Pinheiro [1], G. F. Moita [2], A. L. Guerra [3], R. G. Ribeiro [4], I. M. Silva [5],

[1],[3] ,[4] ,[5]*Dept. of Transportation Engineerinc, Federal Center for Technological Education of Minas Gerais*
*5253 Amazonas Avenue, Nova Suíça, Zip-Code, Minas Gerais/Belo Horizonte, Brazil*
*mirian.greiner@cefetmg.br, andreguerra@cefetmg.br, renato.ribeiro@cefetmg.br, iagomanancezzi7@gmail.com*
[2]*Mathematical and Computational Modeling, Federal Center of Technological Education of Minas Gerais*
*5253 Amazonas Avenue, Nova Suíça, Zip-Code, Minas Gerais/Belo Horizonte, Brazil*
*gray@cefetmg.br*

**Abstract.** Planning a quality public transport system starts with collecting data about passenger demand. Traditional data collection forms are expensive and do not precisely represent the travel demand. On the other hand, secondary data collected passively, for long and continuous periods, and at low cost is emerging as a new opportunity, such as Smart-cards data. However, these data are also limited and miss important information, such as trip purpose. Knowing the trip purpose of public transport passengers is essential to ensure integrated planning between transport and land use and to guarantee higher quality of the public transport service and attract more users, thus contributing to the mitigation of excessive use of the car and its impacts. In this context, the process of Knowledge Discovered in Databases can be used to extract knowledge from these secondary data, improving their applicability in travel demand models. Under this perspective, this study contributes to the enrichment of Smart cards data from the public transportation system of the Metropolitan Region of Belo Horizonte by infering the passengers' trip purposes using data mining techniques.

**Keywords:** Smart card data, Trip purpose inference, Data minning, Transport planning.

## 1 Introduction

For many years, the transportation planning was focused on the short term, with decisions based on the planner's professional experience or on small sample data sets collected by active methodologies. These long periods of weak and limited planning brought to light the realization that old transportation problems don't disappear over time, but rather resurface in more vigorous, extensive, complex and difficult to deal with forms [1].

Congestion, pollution, socio-spatial exclusion, traffic accident and insufficient road infrastructure are some of the challenges caused by a badly skewed planned transportation system. However, the advance of information technology, the availability of large data sets and the expansion of computationall power have encouraged a new scenario of transportation planning with most trusted technical solutions [1].

In general, the objective of transportation planning is to satisfy a highly qualitative, dynamic and differentiated demand for movement, whether by trip purpose, time of day, the mode of transport used or the spatial characteristics of the movement. In this context, the basic structure for developing transportation solutions its starting with the collection of data to identify and characterize the current transport demand [1–3]. According [1], the Origin-Destination (OD) matrices are essential to describe and summarize the transportation demand, providing information about the number of trips made between an area of origin and an area of destination for a in a specific period of time [2]. For this reason, the OD matrix is the core component of the transportation planning process. The information brought by it makes it possible to plan a new transportation system, restructure existing systems and make operational improvements [4].

For decades, transportation planners and researchers have used active data collection methods to identify transportation demand and obtain OD matrices. Active methods are based on sample surveys conduced by observation, interviews, or self-completion forms. This type of data collection has several advantages, but collecting data by these methods is expensive and therefore these surveys are infrequent and applied to a small sample.

Over time, due to the advancement of technology applied to the transportation sector, Big Data sources has become a new possibilities for studying urban mobility patterns and determining the origin and destination of the movements [5]. Smartphones data, social media data or Intelligent Transport Systems data, such as vehicle tracking by Global Position System (GPS), Automatic Number Plate Recognition (ANPR), Smart cards from the public transportation system, among others, are being used by several researchers as possible solutions to the problems brought by traditional data collection methodologies [6]. However, although these Big data sources are valuable for analyzing travel patterns on an ongoing, long-term and at a low cost, these datasets don't collect some important information about the trip and the individuals who are commuting, and this is a disadvantage of contemporary data collection methodologies compared to traditional surveys.

In light of the foregoing, this study seeks to apply data mining techniques in the inference of missing trip information in big data sources, thus contributing to improve the input of travel demand models and to improve the decision making process of planners and managers of public transport service.

As a specific goal, we intend to estimate the primary trips purpose on smart card data using data mining techniques. In this context, there are two explanations: the first is related to the database chosen, and the second to the type of attribute that will be inferred.

We chose to use smart card data because it is the only source of passively and continuously collected data that is readily available to planners and managers of public transportation services in cities. The trip purpose attribute was considered because transportation is a derived demand which only occurs because of the utility of the destination activity.

In light of this, this study will contribute to improve the interpretability of Smart card data, qualifying them with an missed attribute that until has been collected only by household surveys. This will make it possible to continuously monitoring the demand for public transport by trip purpose, contributing to ensure integrated planning between public transport and land use and to provide a quality public transport service, attracting more users, and mitigating excessive private car use and its resulting impacts.

## 2 Literature Review

### 2.1 Travel surveys data collection

Among the traditional data collection methods for obtaining OD matrices, household travel surveys are the most complete. In this survey the respondent answers questions about your travel behavior and providing detailed information about all trips made on the previous the survey day. The same survey includes questions about the economic and social characteristics of the respondent such as income, age, education, vehicle ownership, among others [1].

Travel surveys has several advantages: includes all transportation modes and detailed information about the trips and socioeconomic information of the respondents. However, despite the advantages collecting data by these methods is expensive and therefore these surveys are infrequent and applied to a small sample, both in terms of observations and spatio-temporal coverage [2, 7]. In addition, these surveys match the sociodemographic and travel characteristics of a specific time period, because is difficult collect day over day data for extended periods using active methods due to the cost and burden of the processing, accuracy, and protection of respondents' privacy [8, 9]. Thus, these conventional survey do not capture the fast changes in individuals' travel patterns and are not effective to supporting the dynamic planning transportation

### 2.2 Smart cards data collection

Two decades ago, when automatic fare collection (AFC) systems were implemented in public transportation, smart card data have been the focus of several studies that seek to understand the travel pattern of users. Most of these studies focus on using these data to estimate OD matrices that can be obtained quickly and at lower cost than traditional methods [6, 10–14].

According [15], the main method of obtaining OD matrices from Smart cards data is trip chaining. Some basic premises of this method are:

- One card is linked to a single user;
- a passenger will not walk long distances to access his next boarding point;
- a passenger does not use another mode of transport in the interval of two consecutive boardings on public transport;
- a passenger must have at least two legs of travel for the origin and destination can be identified

From the established assumptions, the chaining method develops a list of passenger trips by linking the corresponding commuting legs for each smartcard. The destination of each trip is considering as the boarding stop of the next trip [15].

Some of the advantages of using smart card data in the OD matrices estimation are: the greater spatiotemporal data collected coverage, the possibility of collecting data on continuously, long-term and at a low cost and the minimal interaction required from users. On the other hand, the disadvantages are: missing information about the destination of the trip, the reason of the trip and the socioeconomic characteristics of the users. These information are not directly collected, and it is necessary to apply algorithms and inference techniques to deduce them [13].

### 2.3 Important attributes of OD matrices

The presented methodologies to obtain OD matrices have advantages and disadvantages regarding how to obtain the data, cost of the data collect, processing load, user interaction, spatial and temporal resolution, sample size, among others. However, an important comparison to be made between these methods is related to the travel attributes available in each of them.

Among the attributes that characterize a trip are: the trip purpose, the time of day when the trip was made, the mode of transport used, and the type of person who is moving. In this context the Table 1 presents a comparison of the availability of these attributes in each type of OD matrix

Table 1. Availability of attributes by matrix OD type

| Attribute | Travel surveys | *Smart cards* |
|---|---|---|
| Trip purpose | ✓ | X |
| Mode of transport | ✓ | ✓ |
| Type of person | ✓ | X |
| Time attributes | ✓ | ✓ |

As show in Table 1 the travel surveys data are the most complete, containing information on all the attributes described. This is because these surveys are designed exclusively to collect travel data and therefore all the information needed to characterize a trip is part of the scope of the survey. However, the disadvantages of this collection method, already discussed in this study, at times outweigh the advantage of completeness of information. On the other hand, the secondary data collected from Smartcards do not have as their main objective the estimation of OD matrices. For this reason important travel attributes are missing which need to be inferred.

### 2.4 Trip purpose importance

One of the main premises of conventional transportation is that travel is a derived demand from the activity that will be accessed at the destination. In this context, the travel occurs only because the benefits obtained at the destination exceed the financial and time costs to get there [16]. Therefore, understanding how individuals access distributed activities in urban space and how they behave during their trips is essential for decision makers, planners, and government agencies to manage the distribution of urban and transportation resources.

Knowing the longitudinally trip purpose also makes it possible to bring valuable information to commercial establishments or for public health decision-making, as was recently observed given the spread of COVID-19 [17]

Given the absence of this relevant attribute in Smart cards data, the inference of trip purpose, has been strongly discussed in the literature. The inference algorithms varying in complexity, input data requirements, accuracy, performance, and most importantly by the method [6]. Several studies have used rule-based methods whose heuristics for categorically determining activities are predefined and previously specified by researchers [8, 17–22]. In general, rule-based methods use the trip start time and the duration of the activity to infering the most likely purpose of a trip.

Rule-based algorithms, despite serving as a basis for more complex methods, are often criticized for their lack of generalizability, which makes it difficult to replicate the method in other contexts [22]. Model-based methods have emerged as a possible solution to this problem. Model-based methods use machine learning models such as decision tree [15, 23], random forest [24, Kim et al.], artificial neural networks [17] Bayesian neural networks [9, akahiko Kusakabe and Asakura] and clustering techniques [21, 23, 27, 28] to infer travel purpose.

### 2.5 Knowledge-discovered in databases Process

Data are collected facts that, when processed for a specific purpose, are transformed into information. This information, when interpreted and applied to a specific purpose, are transformed in knowledge that will be the basis for decision making. One of the ways to find hidden patterns in data, i.e. patterns that cannot be observed explicitly, is the application of a process called Knowledge-discovered in databases (KDD).

The KDD process consists of five phases: Business or problem understanding, where seeks to understand the context in which data mining will be applied; Pre-processing data where the structure of the data, its quality, ease of access and relationship with other data sources will be evaluated; Data transformation where objective is to organize, clean and select the data that will be used in the modeling process; Data mining, which consists of building a model using machine learning techniques with the objective of extracting knowledge from the database; and model evaluation and interpretation of results phase where the model will have its performance evaluated and in case of success will be implemented to make predictions on new collected data.

Among the phases of the KDD process, Data Mining is the most important, and although it is not the only way to transform data into knowledge, it is undoubtedly the most sophisticated way, capable of bringing insights into the problem that no other technique could produce. [29].

## 3 Methodologies

### 3.1 Study Area and Data bases

The Metropolitan Region of Belo Horizonte (RMBH) was established in 1973 by the same law that created the Metropolitan Regions of São Paulo, Porto Alegre, Recife, Salvador, Curitiba, Belém and Fortaleza. Located in the state of Minas Gerais, southeastern Brazil, the RMBH is formed by 34 municipalities. The RMBH population, according to IBGE estimates for 2020 was 5.4 million inhabitants and its economic development is in the central area of the metropolis, which reinforces the dependence that other municipalities have on this centrality.

In the RMBH, traditionally, the data used to formulate OD matrices are collected by household surveys. The first survey was conducted in 1972 and the others were done with 10-year intervals between them, i.e., in the years 1982, 1992, 2002, and 2012, the latter being the most recent edition of this type of collection.

The 10-year interval between editions of the household OD survey is due to all the costs involved in this type of data collection, and although 10 years have passed since the last edition, the OD Matrix conducted in 2012 is still the most recent one done by direct data survey methods, since the subsequent OD matrices prepared in 2019 and 2021, respectively, used passively collected mobile phone data to quantify passenger movement between OD pairs in RMBH.

For this study, the table of trips from the household OD survey was used. This table contains 140,540 records with the complete information of a trip such as origin, destination, trip purpose and mode of transport. This database will serve as the labeled database for training the machine learning model in the data mining step.

The second database used in this study (OD_SBE matrix) is the OD matrix made for RMBH from the smart card data, contains 1,042,347 records. In this base will be made the prediction of the passengers' travel motive, after the model is trained and validated in the historical database (OD_2012).

### 3.2 Preprocessing and data transformation

In the data pre-processing stage, it was defined which predictor attributes would be considered in the machine learning model. In addition, there was also a selection of records from the historical database, keeping only the records referring to trips made by public transport, because the data of the OD_SBE matrix contemplates only the trips of this mode of transport. The variables from the labeled database used in the transformation and modeling steps were: ID of household surveyed (id_h), ID of the individual interviewed (id_i), sequential ID of the trips of each individual (id_t), start time of trip (st_trip), end time of trip (end_trip), place of origin of trip (o_trip), Place of destination of trip (d_trip), sample expansion factor (exp_fac), Origin trip purpose (o_purpose) and destination trip purpose (d_purpose).

The first transformation performed on the OD_2012 matrix data was the creation of an attribute called "KEY", which corresponds to the concatenation of the variables ID_DOM and ID_IND. This is the unique identification key for a person. After creating the key, this column was used to sort the data set and identify all the travels made by the same person. An example of this transformation step in the data is shown in Table 2.

Table 2. creating and checking the key field

| ID_H | ID_I | ID_T | KEY |
|---|---|---|---|
| 44122 | 3 | 1 | 44122_3 |
| 44122 | 3 | 4 | 44122_3 |
| 44122 | 4 | 1 | 44122_4 |
| 44122 | 4 | 2 | 44122_4 |

Next, two checks were performed in relation to the trip reasons. The first, to check if the reason for destination of the trip is equal to the reason for origin of the person's subsequent trip and the second to check if the trips present in the database are complete trips, removing the trips whose reason is stopover or integration. An example of these checks is shown in Table Table 3, in this example it can be seen that the destination reason of a trip corresponds to the origin reason of the following trip for the same individual. In the case of the individual with the key "44122_3", we notice that his trip is divided into two trips, the first one leaving his home to a stopover or transfer point and the second one leaving this stopover or transfer point and going to work. The correction of this inconsistency occurred with the suppression of this transfer, remaining only the information of the complete trip of that user, in this case, leaving his home to his workplace.

Table 3. evaluation purpose of origin and destination of trips

| KEY | O_PURPOSE | D_PURPOSE |
|---|---|---|
| 44122_3 | Residence | stopover |
| 44122_3 | stopover | Work |
| 44122_4 | Residence | School |
| 44122_4 | School | Residence |

After checking the trip reasons for each key and correcting the inconsistencies in the database, a new variable was created, called EXIT_DESTINATION. This variable is equivalent to the time when a certain individual finished his activity at the destination and will be used to calculate the duration of this activity. The time of departure from the destination considered was based on the time at which the analyzed individual had his or her next boarding. In this context, the duration of the activity can also be understood as the time elapsed between two consecutive departures, as shown in Table 4.

Table 4. Calculation of activity duration

| KEY | ST_TRIP | END_TRIP | EXIT_DESTINATION | DURATION |
|---|---|---|---|---|
| 44122_3 | 05:30 | 07:22 | 17:30 | 10:08 |
| 44122_3 | 17:30 | 19:30 | 05:30 | 10:00 |
| 44122_4 | 12:30 | 13:55 | 17:30 | 03:35 |
| 44122_3 | 17:30 | 18:00 | 12:30 | 18:30 |

Regarding the OD_SBE matrix, because it is a base composed of data already processed previously, with the objective of converting them into an OD matrix, the inconsistencies of records have already been discarded. In addition, this is a database with few attributes, and for this reason no selection was applied to it. Regarding data transformation, it was not necessary to create a key attribute, as was done in the OD_2012 database, because in the case of OD_SBE this is already a collected attribute and refers to the individual's card number. However, the key verification step was undertaken in the same way as previously presented.

Since the OD_SBE database does not have the trip reason attribute, the verification step of equality between the destination reason of the trip and the origin reason of the subsequent trip was not performed in this case.

The steps for calculating the departure time from the destination and the elapsed time between two subsequent departures were done in the same way as discussed above.

### 3.3 Data Minning

Data mining is the main step in the KDD process and its execution involves applying algorithms on the data that are responsible for exploring it in order to produce models that will effectively provide the intended knowledge.

As already described, the entire KDD process is guided by the goal initially defined, including the data mining step. In this context, as the goal of this study is to find, for each trip record, to which class of travel reasons it belongs, the data mining task employed will be classification, a typical supervised learning task.

Supervised learning seeks to abstract a knowledge model from data presented in the form of ordered pairs (input, desired output). The input, refers to the predictor attributes that will feed the algorithm and the desired output, corresponds to the value of a variable called the target variable that is expected to be obtained whenever the algorithm receives the specified input values [30]. For this study, we considered as input variables, or predictor variables, the travel start time and the passenger's length of stay at the destination, calculated by the time difference between two consecutive boardings. As a target variable, the first model considered all the reasons for travel, primary and secondary, divided into: work, school, residence, health, shopping and leisure, and the second considered only the primary, or mandatory, work, school and residence trips.

All modeling was done on the historical data base, with a part of this base set aside to train and adjust the model and another part destined to test its performance. The proportion of training and testing data used was 80% and 20%, respectively. The data mining algorithm used was Random Forest. This choice was made because this algorithm reduces the risk of overfitting and is a popular classification method capable of inferring labels with high accuracy.

## 4 Results and Conclusions

To interpret the results of the applied model, the confusion matrices were used. These matrices allow us to identify if the classification occurred correctly based on the metrics of true positives, true negatives, false positives and false negatives. To do this evaluation, the confusion matrix presents three metrics, namely: precision, recall and F1_SCORE.

Recall corresponds to the proportion of positives that was correctly identified. This metric is defined as the ratio between true positives and the sum of true positives and false negatives. The precision metric, on the other hand, seeks to communicate what proportion of positive identifications was actually correct, in other words, this metric tells us how well the model worked. Finally, the F1_SCORE metric presents the balance between precision and recall metrics. In addition to these metrics, accuracy was also used, which is the percentage of correct hits over all the algorithm's bets.

As shown in Table 5, the evaluation metrics proved satisfactory for the modeling performed. This indicates that the proposed model performed well in predicting travel motives from smart card data in the application context. As a continuation of this study, we intend to disaggregate the ”other” reason to predict secondary travel motives such as leisure, health, and shopping. For this purpose, it will be necessary to include other predictor variables in the model, since secondary activities do not generally have regular time patterns like primary activities. We also intend to evaluate the same case study using an unsupervised machine learning model and compare the results obtained.

Table 5. Model evaluation metrics

| | PRECISION | RECALL | F1_SCORE |
|---|---|---|---|
| School | 0.83 | 0.83 | 0.83 |
| Others | 0.73 | 0.72 | 0.73 |
| Residence | 0.97 | 0.96 | 0.96 |
| Work | 0.86 | 0.88 | 0.87 |

## References

[1] J. d. D. Ortúzar and L. G. Willumsen. *Modelling Transport*. John Wiley and Sons, 4 edition, 2011.

[2] A. L. Guerra, H. M. Barbosa, and de L. K. Oliveira. Estimativa de matriz origem/destino utilizando dados do sistema de bilhetagem eletrônica: proposta metodológica. vol. 22, pp. 26–38, 2014.

[3] M. J. Bruton. *Introdução ao planejamento dos transportes*. Interciência, 1 edition, 1979.

[4] M. J. Kang, S. Ataeian, and S. M. M. Amiripour. A procedure for public transit od matrix generation using smart card transaction data. vol. 13, pp. 81–100, 2020.

[5] P. García-Albertos, M. Picornell, M. H. Salas-Olmedo, and J. Gutiérrez. Exploring the potential of mobile phone records and online route planners for dynamic accessibility analysis. vol. 125, pp. 294–07, 2019.

[6] E. Hussain, A. Bhaskar, and E. ching. Transit od matrix estimation using smartcard data: Recent developments and future research challenges. vol. 125, 2021.

[7] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. vol. 68, pp. 285–299, 2016.

[8] F. Devillaine, M. A. Munizaga, and M. Trepanier. Detection of activities of public transport users by analyzing smart card data. vol. 2276, n. 1, pp. 48–55, 2012.

[9] T. Kusakabe and Y. Asakura. Behavioural data mining of transit smart card data: A data fusion approach. vol. 46, pp. 179–191, 2014.

[10] J. J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda. Origin and destination estimation in new york city with automated fare system data. vol. 1817, n. 1, pp. 183–187, 2002.

[11] J. Zhao, A. Rahbee, and N. H. Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. vol. 22, n. 5, pp. 376–387, 2007.

[12] M. Trepanier, N. Tranchant, and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. vol. 11, n. 1, pp. 1–14, 2007.

[13] M. P. Pelletier and C. Morency. Smart card data use in public transit. vol. 19, n. 4, pp. 557–568, 2011.

[14] M. A. Munizaga and C. Palma. Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. vol. 24, pp. 9–18, 2012.

[15] A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman. Public transport trip purpose inference using smart card fare data. vol. 87, pp. 123–137, 2018.

[16] D. Banister. *Inequality in transport*. Marcham, Oxfordshire : Alexandrine Press, 1 edition, 2018.

[17] N. S. Aslam, D. Zhu, T. Cheng, M. R. Ibrahim, and Y. Zhang. Semantic enrichment of secondary activities using smart card data and point of interests: a case study in london. vol. 27, 2020.

[18] K. K. A. Chu and R. Chapleau. Enriching archived smart card transaction data for transit demand modeling. vol. 2063, pp. 63–72, 2008.

[19] Q. Zou, X. Yao, P. Zhao, H. Wei, and H. Ren. Detecting home location and trip purpose for cardholders by minning smart card transaction data in beijing subway. vol. 45, pp. 919–944, 2018.

[20] Z. Zhao, H. N. Koutsopoulos, and J. Zhao. Discovering latent activity patterns from transit smart card data: a spatiotemporal topic model. vol. 116, 2020.

[21] H. Faroqui and M. Mesbah. Inferring trip purpose by clustering sequences of smart card records. vol. 127, pp. 103–131, 2021.

[22] sanjana Hossain and K. N. Habib. Inferring the purposes of using ride-hailing services through data fusion of trip trajectories, secondary travel surveys, and land use data. vol. 2675, n. 9, 2021.

[23] S. G. Lee and M. Hickman. Trip purpose inference using automated fare collection data. vol. 6, pp. 1–20, 2014.

[24] Y. Zhu. Inference of activity patterns from urban sensing data using conditional random fields. vol. 49, n. 2, 2021.

[Kim et al.] E. J. Kim, Y. Kim, and D. K. Kim. Interpretable machine-learning models for estimating trip purpose in smart card data. In *Proceedings of the Institution of Civil Engineers: Municipal Engineer*.

[akahiko Kusakabe and Asakura] akahiko Kusakabe and Y. Asakura. Combination of smart card data with person trip survey data. In *Public Transport Planning with Smart Card Data*.

[27] K. Lu, A. Khani, and B. Han. A trip purpose-based data-driven alighting station choice model using transit smart card data. vol. 2018, 2018.

[28] C. Pieroni, M. Giannotti, B. Alves, and R. Arbex. Big data for big issues: Revealing travel patterns of low-income population based on smart card data mining in a global south unequal city. vol. 96, 2021.

[29] F. Amaral. *Aprenda Mineração de dados, teoria e prática*. Alta Books, 1 edition, 2016.

[30] R. Goldschmidt, E. Passos, and E. Bezerra. *Data minning: conceitos, tecnicas, algoritmos, orientações e aplicações*. Elsevier, 2 edition, 2015.