Enhancing public transportation planning through travel data analysis: a data mining application in the inference of passenger trip purpose in the Metropolitan Region of Belo Horizonte, Brazil

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

[1], [2]*Mathematical and Computational Modeling, Federal Center of Technological Education of Minas Gerais*
5253 Amazonas Avenue, Nova Suiça, 30.421-169, Minas Gerais/Belo Horizonte, Brazil
mirian.greiner18@gmail.com, gray@cefetmg.br
[34]*Dept. of Transportation Engineering, Federal Center for Technological Education of Minas Gerais*
5253 Amazonas Avenue, Nova Suiça, 30.421-169, Minas Gerais/Belo Horizonte, Brazil
renato.ribeiro@cefetmg.br, andreguerra@cefetmg.br

Abstract. Planning an efficient transport system begins with the collection of demand data. Traditionally, this data has been gathered through travel surveys conducted by interviews. These surveys involve questioning people about their trip's starting and ending points, purpose, mode of transportation, travel duration, and other relevant details. However, this method is costly and can be somewhat inaccurate since it relies on respondents' ability to accurately describe their journeys. With technological advancements in the transportation sector, Big Data sources have emerged as a new possibility to studying urban mobility patterns. In the public transport sector, since the 2000s the data collected by automatic fare collection systems provide a quick, accurate, and cost-effective means to estimate Origin and Destination (OD) matrices. However, a significant challenge arises when using this data source for OD matrix estimation—the lack of trip purpose information. To address this, researchers have turned to data mining techniques to inference this trip atribute. This paper contributes to the field by applying these emerging data mining approaches to infer the trip purposes of public transport passengers in the Metropolitan Region of Belo Horizonte (RMBH).

Keywords: Trip purpose inference, Transport planning, Data minning, Random Forest, Smart card data

1 Introduction

One of the main assumptions of conventional transport thinking is that traveling it is a derived demand from the activity to be carried at the destination place. This argument comes from the utilitarian perspective of the economic rationality who considers that travel occurs only because the benefits obtained at the destination place outweigh the financial and time costs spends for the people to arrive there [1]. Therefore, understanding how people access activities distributed in urban space and how they behave during your trip is essential for that decision makers, planners and government agencies can manage the distribution of urban and transport resources, to attend the diverse needs of the population [2]. In general, the trip purposes are categorized into: work, home, school, shopping, leisure, health, business and other. The first three are known as primary or mandatory trips and all others are called discretionary or secondary trips [3].

Nowadays, the only source data that collecting the purpose and other relevant details about the trip are the origin-destination surveys, also known household surveys. However, this data collection method has major disadvantages, like:

1. The inaccuracy of some information, especially temporal and spatial attributes, due to people's lack of ability to accurately report the requested information [4].
2. The high human and financial costs involved in this type of data collection, which impacts on the decrease of the frequency conducting the research and on the reduction of its sample size, both in terms of volume and space-time coverage [5, 6],
3. The static characteristic of this data collection type, because they corresponds to the sociodemographic and

travel reality of a specific time period (research datetime). Thus, travel surveys do not capture the rapid changes of the individuals' trip patterns and are not effective to understand the dynamism of the mobility in the urban centers [7, 8].

With technological advancements in the transportation sector, Big Data sources have emerged as a new possibility to studying urban mobility patterns and overcome the household travel surveys limitations [9]. In the public transport sector, since the 2000s the data collected by automatic fare collection systems provide a quick, accurate, and cost-effective means to estimate Origin and Destination (OD) matrices [5, 9]. However, a significant challenge arises when using this data source for OD matrix estimation are the lack of trip purpose information. To solve this, researchers have turned to data mining techniques to inference this trip atribute using algorithms that vary in complexity, input data requirements, accuracy, performance and mostly by the method, which can be rule-based or model-based [10]. As a contribution to this search topic, this paper applying these emerging data mining approaches to infer the trip purposes of public transport passengers in the Metropolitan Region of Belo Horizonte, Minas Gerais, Brazil.

2 Inference of trip purpose: a literature review

The trip purpose inference studies differ mainly by the methodological approach used, which can be supervised or unsupervised. Unsupervised learning methods group trips with similar characteristics and then assign a trip purpose through a specific criterion or a decision rule. The advantage of this method is that the trip purpose label is not required to train the model, however, so that the results are satisfactory, there needs to be prior specialized knowledge to assign the trip purpose.This can become especially difficult for secondary trip purposes such as leisure, health or shopping because in general, these trips do not have a well characterised pattern of behavior as the mandatory trips. On the other hand, the supervised learning methods use historical data to train a model that will later be applied to another dataset. In this approach, the model will find the patterns in the data, without the need for prior knowledge of specialists, however, as already described, for this methodology to be applied, travel data labeled with the trip purpose need to be available [11].

2.1 Studies using supervised methodologies

Among the studies that apply supervised methods for the trip purpose inference [12] Used smart card data, land use data, home travel survey data, the southeast transport strategic model of Queensland (Australia) and General Transit Power Specification (GTFS) data to train a probabilistic choice model that classifies trips into five purposes : work, education, shopping, home and recreation. The results of this study demonstrated an improved accuracy after applying temporal attributes jointly spatial attributes in model training. However, different trip purposes presented different sensitivities to the spatial and temporal attributes applied, with home and work trips being those that presented the best correct inference results, with 92% and 96% respectively, already for secondary trips the sensitivity was significantly lower , around 46%.

[13]Presents a method to identify weekly temporal patterns of primary activities made by public transport users in Singapore. In this study, a household trip survey data were used to train two logit models of discrete choice: one for workers and one for students. Using start time and duration of an activity the models were trained to infer the trip purpose in home, work, study or others. [14] Inferred the trip purpose for public transport users from South Korea using smart card data, a travel search data, the population census data and national transport infrastructure data. In this study, the authors compared the performance of four models: Random Forest, Gradient Boosting Machine, Naive Bayes and Multinomial Logit. All algorithms had their hyperparameters adjusted based on a Grid Search with quadruple cross validation, to classify a trip purpose in four categories: work, business, leisure and return home. Among the various models used, the Random Forest performance was the best. However, the results of this model were different for each class, with 91% of balanced accuracy for the work purpose and 67% for the business purpose, for example.

[2] Proposed a neural network-based framework to classify the trip purposes of London public transport passengers using smart card data, travel search data and Points of Interest data collected using an Twitter API and Foursquare location API. In this study, the trip purposes were classified into two categories, namely primary (work and home) and secondary (entertainment, meal, shopping, delivery/pick-up of children and part-time work activities). The framework for prediction of trip purposes was called ActivityNET and was divided into two phases: the first phase was focused on extracting spatiotemporal activity patterns in the smartcard data and later in the combination of these patterns with points of interest (POIs) using an confirmation algorithm. The second phase of the study used input features such as start and end time of the trip and the duration of activity to predict trip purposes using a neural artificial network-based approach. Due to the need to process large datasets, this study

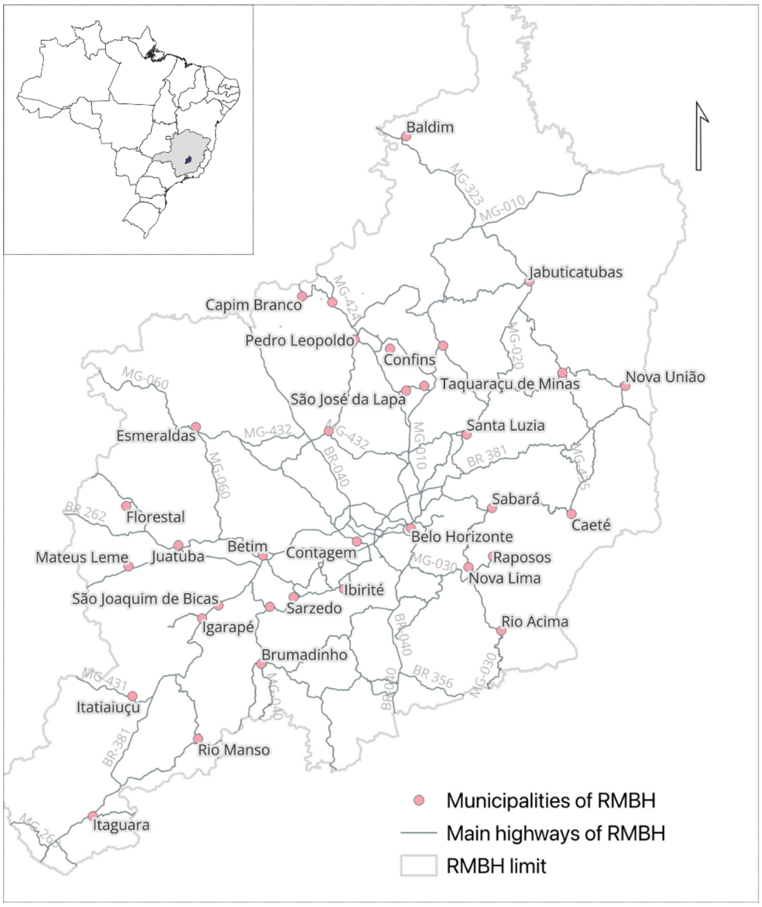
used PySpark to enable the execution of the proposed methodology. The F1-score results showed significantly higher for primary activities in compared to secondary activities, with values greater than 90% for the work and residence and about 68% for the shopping, for example.

Finally, in one of the most recent studies on travel motive inference, [11], estimated public transport trip purposes in Santiago (Chile) using a supervised machine learning methodology, specifically the XGBoost algorithm. In this study, several data sources were used, such as origin and destination research, land use data from the Federal Revenue Service, smart cards data and social priority data for the region generated by the Ministry of Social Development and Family . The model proposed in this study is a concatenation of three XGBoost models, the first was trained to discretize trips into primary and secondary purposes with an F1-score of 88%. The second model was used to refine the primary purposes into Work, Home, and School with an F1-score of 86% and, finally, the third model was used to characterize the secondary purposes in health, leisure and others with an F1-score of 72%. The variables used in this study are the number of stages of a trip, the travel time, the duration of the activity, the start time of the trip, the arrival time, the distance between the embarkation and disembarkation point in meters, two boolean variables that indicate whether the trip is the first or last trip of the day, the social priority index associated with the embarkation and disembarkation zones, the percentages of Y-type land use within a 500m radius of the embarkation and disembarkation points and the number of X-type points of interest within a 500m radius of the departure and arrival points. The individual F1-score results for each of trip purpose discretized in the models 2 and 3 were not presented.

3 Methodology

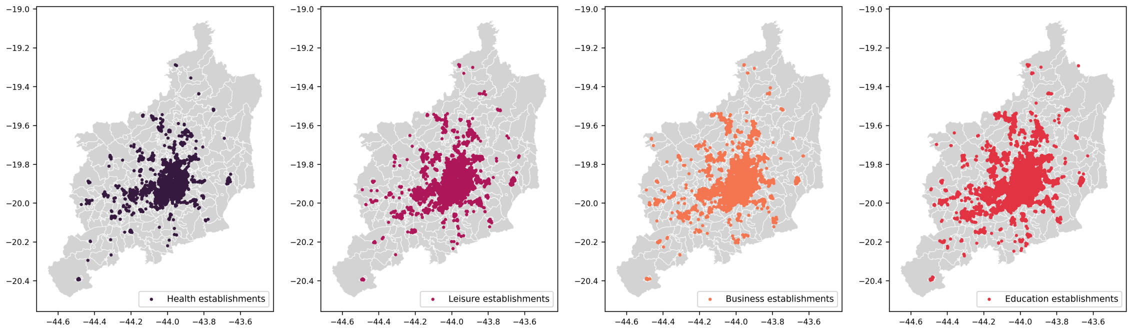
The Metropolitan Region of Belo Horizonte (RMBH) is an administrative political sector created to enable the management of the metropolitan agglomeration resulting from the expansion of the Belo Horizonte metropolis. Currently, the RMBH is made of 34 municipalities of the state of Minas Gerais, as shown in Figure 1, that together have just over 5 million inhabitants, being considered the third largest urban agglomeration in Brazil.

Figure 1. Graphic representation of the RMBH with its municipalities and main highways



To infer the trip purposes of public transport passengers in the Metropolitan Region of Belo Horizonte, two main data sources are used: the public transport travel records from the last edition of the Origin and Destination Survey carried out in 2012 in the RMBH (called in this study ”OD_2012”) and the data of National Register of Legal Entities provided by the Federal Revenue Service of Brazil, which contains informations of all companies registered in the country. This last database was be used to calculate the number of health, leisure, business and education Points of Interest (POIs) available in each destination of trips places. These POIs spatial distribution in the territory of the RMBH is presented in Figure 2 and this dataset was called in this study ”POIS”.

Figure 2. Distribution of Points of Interest in the territory of the RMBH



The set of variables that will be used in the methodological stages of this study it’s composed by the predictor variables: trip start time, activity duration, number of trips made each person, trip sequential identifier and number of health, education, leisure and business POIs available in each destination of trips places. The first four variables are available in the data source (OD_2012) and the others in the data source (POIS). As it is a supervised learning application, in addition to the predictor variables, the trip purpose label also available in OD_2012.

The trip purpose inference process in this study will be conducted in three stages. In the first, only the travel variables available in the OD_2012 data source will be used and the travel records will be classified into seven purposes: Home, Work, School, Health, Leisure, Business and Others. One of the problems identified in this stage was the trip purpose class are very imbalance. The reason for this imbalance is mainly related to the way the household trip research was conducted, with interviews carried out on a typical working day where mandatory trips prevail, to the detriment of secondary trips. The distribution of the trips records by purpose are: residence (47%), work (32%), school (7%), other (5%), health (4%), business (3%) and leisure (2%). In this context, because it is about a unbalanced multiclass classification problem, three algorithms were tested in this step, with the aim of identifying the one that best adapts to the studied problem. The tested algorithms were Random Forest (RF), Support Vector Machines (SVM), and XGBoost (XGB).

To class balancing in this step, the random oversampling technique was used and a grid search with 5-fold cross-validation was used to optimize the hyperparameters of each algorithm. The parameter grid used in the search was:

- **Random Forest**:
  n_estimators: [100, 500, 1000]
  criterion: [’gini’, ’entropy’]
  max_depth: [5, 10]
  min_samples_leaf: [1,2]
  max_features: [’sqrt’, ’ log2’]

- **SVM with OneVsOneClassifier strategy**:
  C: [0.1, 1, 10]
  Kernel: [’linear’, ’rbf’]
  gamma: [0.1, 1, 10]

- **XGB**:
  n_estimators: [100, 500]
  learning_rate: [0.1, 0.01]
  max_depth: [3,5,7]
  subsample: [0.8, 1.0]

At the end of the first stage training process, the metrics of accuracy, Matthews Correlation Coefficient (MCC), precision, recall, and F1-score were calculed and a ANOVA statistical test was conducted to evaluate whether there was a statistically significant difference between the performance of the three algorithms.

Finally, in the third stage, the spatial variables available in the data source (POIs) were included as predictor variables set in the two submodels trained in the previous stage. The objective of this stage was to understand the contribution brought by the joint use of spatial and temporal variables in improving the quality of inferences.

4 Discussion of results

As demonstrated in Table 1, The results of the three implemented algorithms were very similar. The ANOVA statistical test confirmed that there are no statistically significant differences between the performance of the three algorithms, with a significance level of 99%, statistics of 0.009 and p_value of 0.99. So that, it was decided to continue with the Random Forest algorithm for the next methodological steps. This decision was based on the interpretability and moderate/low training and prediction costs of this algorithm. Furthermore, this algorithm has frequently used in studies with objectives similar to the one being worked on in this paper. The grid search found, for the Random Forest algorithm, the following values in hyperparameter optimization: criterion (gini), max_depth (10), max_features (sqrt), min_samples_leaf (1), n_estimators (1000). These hyperparameters will be used in the next modeling steps described below.

Table 1. comparison between algorithms for the first methodological step

	Random Forest	**Support Vector Machine**	**XGBoost**
Accuracy	*0.84*	*0.83*	*0.83*
MCC	*0.76*	*0.75*	*0.74*
Precision	*0.53*	*0.53*	*0.51*
Recall	*0.53*	*0.54*	*0.52*
F1-Score	*0.52*	*0.52*	*0.52*

The results presented in Table 2 demonstrate that the Random Forest algorithm has a good performance to predicting the mandatory trip purposes, especially residence and work. However, there is a significant worsening in the performance of this model to predicting the secondary trip purposes, in particular the leisure purpose. In addition, in Table 2 it is also possible to notice the substantial difference between the Macro average and the Weighted average for the Precision, Recall and F1-score measures. This is because in addition to the primary trip purposes having presented the best results for these metrics, they are also the classes that have the most data, due to the data collection procedure already discussed earlier. Therefore, as in this classification problem all classes have equal importance, it was decided to use the macro average of the Precision, Recall and F1-score to evaluate these metrics

Table 2. Random Forest Algorithm - Classification Report

Classe	**Precision**	**Recall**	**F1-score**	**Support**
School	*0.66*	*0.64*	*0.65*	*295*
Leisure	*0.25*	*0.15*	*0.18*	*88*
Business	*0.33*	*0.32*	*0.33*	*137*
Other	*0.36*	*0.31*	*0.34*	*200*
Residence	*0.96*	*0.98*	*0.97*	*1993*
Health	*0.40*	*0.36*	*0.38*	*153*
Work	*0.89*	*0.92*	*0.91*	*1366*
Accuracy			*0.84*	*4232*
Macro avg	*0.55*	*0.53*	*0.54*	*4232*
Weighted avg	*0.83*	*0.84*	*0.84*	*4232*

That said, the second methodological step was applied: splitting the classification model, so far unique, into two. The model trained to classify mandatory trips showed very good results, with an accuracy of 92%, Matthews coefficient of 87%, macro-precision of 86%, macro-recall of 84% and macro F1-score of 85%. For the model trained to classify secondary trips, although the results remained poor, with accuracy of 42%, Matthews coefficient of 21%, Precision of 41%, Recall of 41% and F1-score of 41%, the individual performance per class, shown in the Table 2, got improvements, especially for the leisure purpose, whose gain in the result of the F1-score was 22 percentage points. For health and ”other” purposes the gains in the F1-score result were more subtle, 8 and 11 percentage points, respectively. Finally, for the business motive, there was no change in the result of the F1-score, remaining at 33%.

Finally, the spatial variables were included in the set of predictors of the two models described above (the one that classified primary trips and the one that classified secondary trips). The spatial variables considered were the number of Points of Interest (POIs) of leisure, health, education and business available at the destination of the trip. As a result of the including spatial variables along with temporal variables in the set of predictors, the model trained to classify mandatory trips showed an accuracy of 92%, MCC 88%, macro-precision 87%, macro-recall 85% macro F1-score 86%. The main improvements brought to this model were observed in the classification of school trips, whose the hit rate rose from 66% to 70%. As for the secondary trip classification model, the contributions arising from the inclusion of spatial variables in the set of predictors was more expressive. The accuracy of the secondary trips model went from 42% to 46%, MCC went from 21% to 27%, macro-precision from 41% to 46%, macro-recall from 41% to 46% and macro F1-score from 41% to 46%.

5 Conclusions

This study applied emerging data mining approaches to infer the trip purposes of public transport passengers in the Metropolitan Region of Belo Horizonte (RMBH) on smart card data using a machine learning supervised approach. This approach proved to be solid for predicting the trip purposes in passively collected datasets, getting rich this type of data and, consequently reducing dependence that transport planners have on households travel surveys. The model applyed in this study uses two databases to infer the trip purpose, namely: the RMBH Origin and Destination Survey and the CNPJ registration database of the Federal Revenue Service of Brazil. After applying, calibrating and validating different trip purpose inference models of public transport passengers, it is concluded that Residence, Work and School are trip purposes can be classified with a high level of precision, while secondary purposes show good results when they are classified all together in a single category called ”Secondaries”. When trying to drill down to the secondary trip purposes class in the Leisure, Health, Business and Others, the method using in this study did not show good results, having presented an accuracy of 46%. In this regard, it is important to highlight that, in most cases, the households travel survey are carried out on typical weekdays that the main travel made by people are the mandatory. For this reason, the data generated by these surveys are significantly unbalanced. In the Metropolitan Region of Belo Horizonte, the area studied in this paper, more than 80% of the trips collected by the OD 2012 survey correspond to mandatory trips, and this significant imbalance compromised the performance of the secondary trips purpose classification model.

Hardly the inference methodologies will fully replace travel data surveys, that’s why, one of the most important conclusions of this paper is that the planning of these surveys needs to be reviewed so that relevant, comprehensive and high-quality data on secondary trips be collected.This is because the mandatory travel pattern is already well defined, as demonstrated in this study with the high inference accuracy of the Home, Work and School trip purposes. Furthermore, this pattern does not tend to change significantly. On the other hand, the way data is currently collected makes it difficult to clearly understand individuals’ secondary travel patterns and this significantly compromised the inference model for these type of trips, whose precision reached only 46%.

Finally, even though the model for inferring the secondary purpose of the trip did not perform satisfactorily, the importance of continuing this study is highlighted in an attempt to obtain better results for these cases. Currently, transport planning in the Metropolitan Region of Belo Horizonte is guided by mandatory trips, but planning an efficient public transport system means ensuring that the population has good access conditions also for leisure trips or for health promotion, for example . In this context, the availability of quality transport, which meets the population’s needs beyond routine requirements, can be a decisive factor for urban development, for promoting accessibility and to encourage the emergence of other centralities. It should be noted here that this conclusion is especially relevant for low-income people, who depend exclusively on public transport for their transportation and for older people, whose trips for secondary reasons are even more frequent than trips for primary reasons. This reinforces the importance of understanding the pattern of secondary travel, as only then will it be possible to understand and longitudinally monitor the trip journeys of the population as it ages.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

The source code and datasets used in this paper are available in: GitHub repository

References

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

[2] N. S. Aslam, M. R. Ibrahim, T. Cheng, H. Chen, and Y. Zhang. Activitynet: Neural networks to predict public transport trip purposes from individual smart card data and pois. vol. 24, 2021.

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

[4] P. R. Stopher, C. FitzGerald, and M. Xu. Assessing the accuracy of the sydney household travel survey with gps. vol. 34, pp. 723–741, 2007.

[5] M. Bagchi and P. R. White. The potential of public transport smart card data. vol. 12, n. 5, pp. 464–474, 2005.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] R. Pezoa, F. Basso, P. Quilodrán, and M. Varas. Estimation of trip purposes in public transport during the covid-19 pandemic: The case of santiago, chile. vol. 109, 2023.

[12] 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.

[13] S. A. O. Medina. Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. vol. 12, pp. 91–101, 2018.

[14] 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*, pp. 108–117, 2021.