

# Inferring socioeconomic attributes of public transit passengers using classifiers

Hamed Farooqi<sup>1</sup>, Mahmoud Mesbah<sup>2,3</sup>, Jiwon Kim<sup>4</sup>

- 1- PhD candidate, School of Civil Engineering, The University of Queensland, Australia. Tel: +61-481247314
- 2- Faculty Member, Department of Civil and Environmental Engineering, Amirkabir University of Technology, Iran
- 3- Honorary senior lecturer, School of Civil Engineering, The University of Queensland, Australia. Tel: +61-733651569
- 4- Lecturer, School of Civil Engineering, The University of Queensland, Australia. Tel: +61-733463008

Email for correspondence: h.farooqi@uq.edu.au

**Abstract.** Emerging new datasets in the public transit network, such as smart card datasets, are changing the transport research area to a data-rich one. While these datasets include a large number of passenger trips in the network, they miss the socioeconomic attributes of passengers. Estimating the socioeconomic attributes from trip attributes is a way to enrich these datasets. To do so, Household Travel Survey (HTS) can be used to learn the relation between trips and socioeconomic attributes of passengers. This paper compares the performance of three well-known classifiers of Naïve Bayes, Random Forest, and Support Vector Machine for estimating age and income of passengers in the public transit network using the HTS. The explanatory variables are considered as the start time of the trip, activity duration (time period between two consecutive trips), land use around the origin, and land use around the destination of trips. The target variables are age and income of passengers. HTS data, which include both explanatory and target variables, are used to train and validate the classifiers. Three measures including Accuracy, F-score, and Informedness are used to evaluate the performance of the classifiers. Also, the three classifiers are compared to a random classifier. The case study is the HTS from the South East Queensland (SEQ) between 2009 and 2012. Results show that the Naïve Bayes classifier is a better classifier for estimating the age and income of passengers from trip attributes.

**Keywords:** Household travel survey; data mining; machine learning; public transport system

## 1. Introduction

Emerging datasets in the public transit network are changing the transport research area to a data-rich one. These datasets are large in size and include data on public transit passenger trips in the network. Big datasets, such as smart card data in the public transit network, are collected continuously, passively and usually as a by-product of other applications, such as fare gathering. However, these big datasets miss the socioeconomic attributes of the passengers (Farooqi et al., 2018b). Thus, inferring socioeconomic attributes from trip attributes of passengers can enrich the available datasets. For instance, elderly people usually start their trip after the morning peak and take short activities, such as shopping; but, students usually take trips in the morning peak and stay at school for around 6 hours. Therefore, available big datasets in the public transit network can be enriched by inferring socioeconomic attributes of passengers from the trip attributes.

Enriching the big datasets in the public transit network with socioeconomic attributes of passengers would extend the current applications of these datasets. First, emerging big datasets in the transportation systems have the potential to support advanced transport models, which require trip demand and socioeconomic attributes of individuals. Socioeconomic attributes are an inevitable part of these models. Household travel surveys are used to build the transport models; while these surveys include both trip demand and socioeconomic attributes of individuals, they suffer from couple of issues such as sampling rate, high cost, and inaccurate information (Liang et al., 2007; Chen et al., 2010; Bethlehem et al., 2011; Faroqi et al., 2017). Hence, the enriched big datasets with the socioeconomic attributes could be considered as an alternative data source for the transport models. Second, enriched big datasets with socioeconomic datasets can trigger a new generation of targeted applications in the public transit network (Faroqi et al., 2018a). These applications help to discover groups of passengers with similar attributes such as age or place of work. Especially, these targeting applications would be interesting in the field of marketing. Consequently, the enriched datasets can extend current applications in the public transit network.

Classifiers as a subset of machine learning and data mining techniques can learn the relations and features of objects in a dataset. Then, the classifiers can predict desired labels for a new set of objects based on the learnt relations. Classifiers are basically trained on a training dataset and then assign each of the records in the new datasets to one of the identified classes from the training dataset. In the case of learning socioeconomic attributes from trip attributes, classifiers can learn from a dataset that includes both socioeconomic (target) and trip (explanatory) attributes, such as HTS; then, the trained classifiers can assign socioeconomic attributes to any set of trip attributes that have similar trip attributes to the HTS (this process is also called as data fusion) (El Faouzi et al., 2011). Trip attributes can include, but not limited to, the start time of the trip, activity duration, land use around the origin, and land use around the destination of trips. Also, socioeconomic attributes can vary from age and income to car ownership and household size. Therefore, classifiers can learn from HTS and assign socioeconomic attributes to the trip attributes of individuals in the smart card datasets.

The effect of passenger socioeconomic attributes on travel behaviour has widely been studied in the literature (Recker et al., 1985; Giuliano and Dargay, 2006; Limtanakool et al., 2006; Pasha et al., 2016). However, estimating socioeconomic attributes of passengers from the HTS in a big dataset of the public transit has attracted little attention. An example is the data fusion technique explained in El Faouzi et al., 2011. More recently, two research papers in the literature estimated the trip purpose in a smart card dataset by fusing it with the HTS using a Naïve Bayes classifier (Kusakabe and Asakura, 2014; Alsger et al., 2018).

This paper compares the performance of three well-known classifiers (Naïve Bayes, Random Forest, and Support Vector Machine) for estimating age and income of passengers in the public transit network using the HTS. The explanatory variables are the start time of the trip, activity duration, land use around the origin, and land use around the destination of trips. Age and income are the target variables. HTS data, which include both explanatory and target variables, are used to train and validate the classifiers. Three measures of Accuracy, F-score, and Informedness are used to evaluate the performance of the classifiers. The case study is the HTS from the SEQ between 2009 and 2012.

The paper is structured as follows. Upcoming is a section for explaining the classifiers, explanatory and target variables. Next, a case study and analysis are discussed in the Results section. Finally, the findings of the paper are summarized and discussed in the Conclusions section.

## 2. Methodology

This section explains the explanatory and target variables, and classifiers. The purpose of the classifiers in this study is to estimate age and income of the passengers based on the start time of the trip, activity duration, land use around the origin, and land use around destination of trips. The explanatory and target variables are explained first in the next subsection and the classifiers are briefly described after the variables.

### 2.1. Explanatory and target variables

Explanatory variables explain the target variables. In other words, target variables are dependent variables that can be inferred from the explanatory variables. In this study, the target variables are passengers' age and income, which are expected to be inferred from the explanatory variables. Each of the explanatory and target variables can be categorised considering specific applications, which is described in the Results section.

Explanatory variables:

The start time of the trip: the boarding time of the trip.

Activity duration: the time spent at the destination to perform an activity.

Origin land use: The dominant land use around the boarding location.

Destination land use: The dominant land use around the alighting location.

Target variables:

Age: a passenger age.

Income: a passenger's household income per week.

Figure 1 presents the explanatory and target variables in one frame. Age and income are two target variables, which are separately inferred from the explanatory variables. Classifiers learn the relation between the explanatory and target variables. In simple words, classifiers differ in learning the relation among the variables; different methods in learning the relations lead to different classifiers.

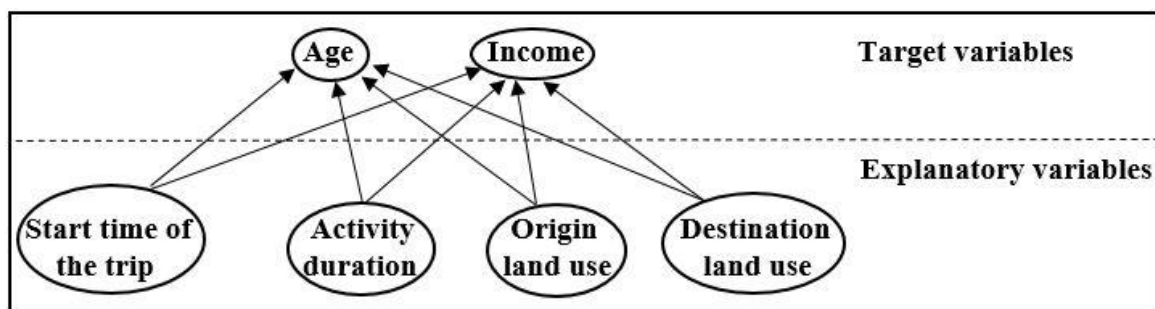


Fig. 1. Explanatory and target variables

## 2.2. Classifiers

The classification task is to predict the label or class for a given unlabelled point. Formally, a classifier is a model or function  $M$  that predicts the class label  $\hat{y}$  for a given input example  $x$ , that is,  $\hat{y} = M(x)$ , where  $\hat{y} \in \{c_1, c_2, \dots, c_k\}$  is the predicted class label (a categorical attribute value). To build the model a set of points with their correct class labels is required (training set). After training the model, the class of any new data point can be predicted (Zaki and Meira, 2014).

### 2.2.1. Naïve Bayes

It is a classification technique based on Bayes' theorem with an assumption of independence between explanatory variables (predictors). In simple terms, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Equations 1 and 2 presents how a Naïve Bayes classifier calculates the posterior probability. In equations 1 and 2,  $P(c|x)$  is the posterior probability of class (target) given an explanatory variable (predictor);  $P(c)$  is the prior probability of class;  $P(x|c)$  is the likelihood which is the probability of the explanatory variable given the class;  $P(x)$  is the prior probability of the explanatory variable (Murphy, 2006).

$$P(c|x) = \frac{P(x|c) P(c)}{P(x)} \quad (\text{eq. 1})$$

$$P(c|x) = P(x_1|c) * P(x_2|c) * \dots * P(x_n|c) \quad (\text{eq. 2})$$

### 2.2.2. Random Forest

Random Forest stands for a group of decision trees. In Random Forest, a set of decision trees (so known as "Forest") together decide for the label of the class. To classify a new data point on attributes, every tree gives a "votes" for that class. The Forest chooses the label that has the most votes (among all the trees). A decision tree in the Forest draws an axis-parallel hyperplane to divide the data space  $R$  into two resulting regions, which also create a partition of the input data points to two spaces. All these regions are recursively divided via the axis-parallel hyperplanes till all the data points within a created partition mostly have the same class label (Zaki and Meira, 2014).

One advantage of using decision trees is that they produce models that are relatively easy to interpret. In particular, a tree can be read as a set of decision rules, with each rule's antecedent comprising the decisions on the internal nodes along a path to a leaf, and its consequent being the label of the leaf node. Furthermore, since the partitions are all separated and cover the whole space, the set of rules can be inferred as a set of disjunctions (Zaki and Meira, 2014).

### 2.2.3. Support Vector Machine

Support Vector Machine (SVM) is a classification method derived from maximum margin linear discriminants concept. SVM tries to find the optimal hyperplane that maximizes the margin between the classes. This classifier plots each data item as a point in  $n$ -dimensional space (where  $n$  is the number of available explanatory variables) with the value of each feature being the value of a particular coordinate. SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane

that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier (Zaki and Meira, 2014).

Furthermore, the linear SVM approach can be used for datasets with a non-linear decision boundary. Conceptually, the idea is to map the original  $d$ -dimensional points  $x_i$  in the input space to points  $\phi(x_i)$  in a high-dimensional feature space via some non-linear transformation  $\phi$ . Given the extra flexibility, it is more likely that the points  $\phi(x_i)$  might be linearly separable in the feature space. Note however that the linear decision surface in the feature space actually corresponds to a non-linear decision surface in the input space (Steinwart and Christmann, 2008; Zaki and Meira, 2014).

### 2.3. Evaluation measures

Outcomes of each classifier can be summarized in a confusion table that is presented by four values including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which are defined as follows. In the case of having more than two classes, each of these values is calculated by considering one class against the other classes.

TP: number of truly predicted objects in the class.

TN: number of truly predicted objects in the other class(es).

FP: number of falsely predicted objects in the other class(es).

FN: number of falsely predicted objects in the class.

Three common measures to evaluate the performance of the classifiers are Accuracy, F-Score, and Informedness. Accuracy evaluates the number of truly predicted objects in all classes in relation to the population. F-Score evaluates the number of truly predicted objects in the class in relation to the population. Informedness considers all true and false positive and negative values; and is considered as a balanced measure. Informedness can be used even in the case of having classes with very different sizes. Each of the mentioned measures is formulated in equations 3 to 5. Accuracy and F-Score vary between 0 and 1. Informedness varies between -1 and +1. In all three measures, values closer to +1 means a better estimation (Powers, 2011).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (\text{eq. 3})$$

$$F - score = \frac{2*TP}{2*TP+FP+FN} \quad (\text{eq. 4})$$

$$Informedness = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1 \quad (\text{eq. 5})$$

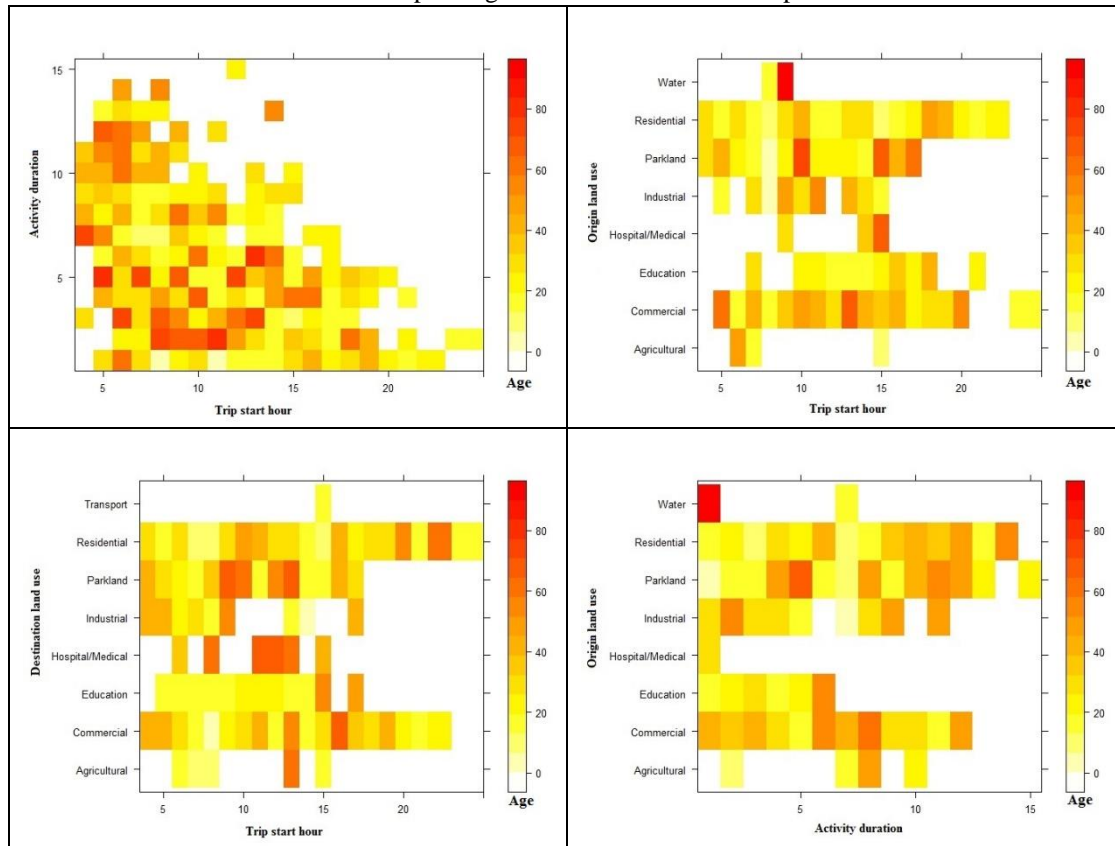
## 3. Results

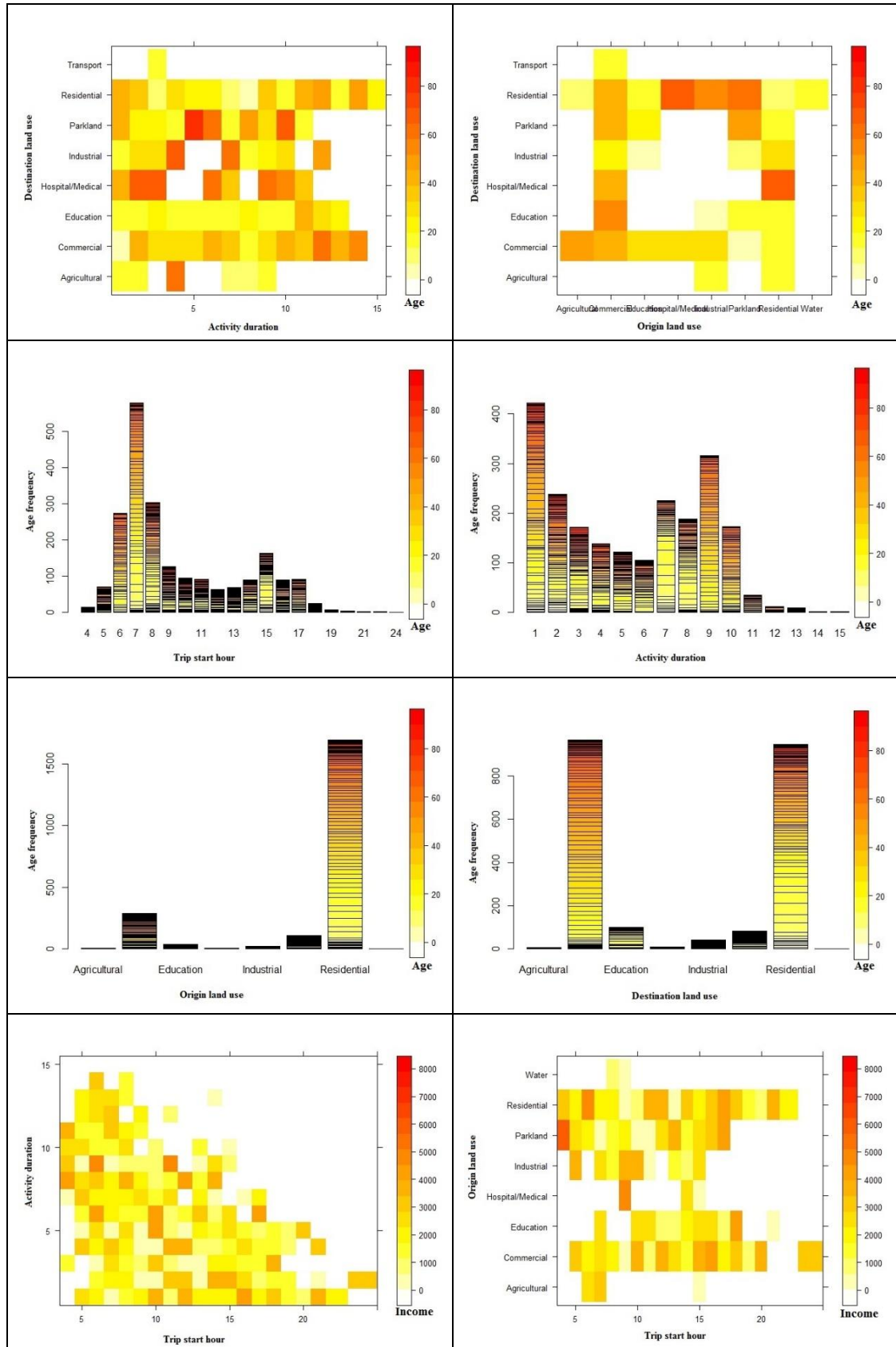
The HTS was undertaken across the SEQ from 2009 through 2012. The survey includes about 2,500 public transit trip records, which is used in this study. The dataset is divided into 80% for training and 20% for validating the classifiers. This means 2,000 records are used to train the classifiers and the rest to validate them. Information from HTS includes origin time and location, destination time and location, activity duration, age and income of passengers, which are used in this study. Furthermore, the land use data from the Australian Bureau of Statistics (ABS) at the mesh block level are used (Alsger et al., 2018). All the computation process is undertaken in R version 3.2.2 using the “e1071” library.

Details of the trip and passenger attributes in the dataset are as follows. Start hour of the trip (origin time) is in between 4 and 24. The activity duration is between less than 1 hour and more than 15 hours. The land use around origin and destination is one of Agriculture, Commercial, Education, Hospital/Medical, Industrial, Parkland, Residential, and Water types. Also, the age and income of the passengers continuously vary between 0 and a maximum value in the database. Age and income are then categorised. Age of the passengers is discretised into four groups: younger than 21 years old (known as generation z), between 21 and 41 years old (known as generation y), between 41 and 61 years old (known as generation x), older than 61 years (known as baby boomers). Income of passengers is discretised into four groups: less than 625 A\$ per week (low income), between 625 and 1230 A\$ per week (medium-low income), between 1230 and 2200 A\$ per week (medium-high income), more than 2200 A\$ per week (high income).

Table 1 presents the heat maps for each of the age and income attributes across the trip attributes. Colours at each diagram present the frequency of the age or income, which are determined at the bottom right side of the diagram. For instance, the heat map in the first row and the second column presents the distribution of age across start hour of the trip and duration of the activity; for example, passengers older than 60 years old prefer to have their trip after 9 am with an activity less than 3 hours. Also, considering the heat map in the seventh row and first column, people with a low income start their trip before 6 am and head to industrial areas. Diagrams at rows four, five, nine and ten presents the age and income distributions across one trip attribute. For example, the diagram at the fourth row and first column presents the frequency of age attribute across start hours of the trip.

Table 1. Heat maps of age and income across the trip attributes





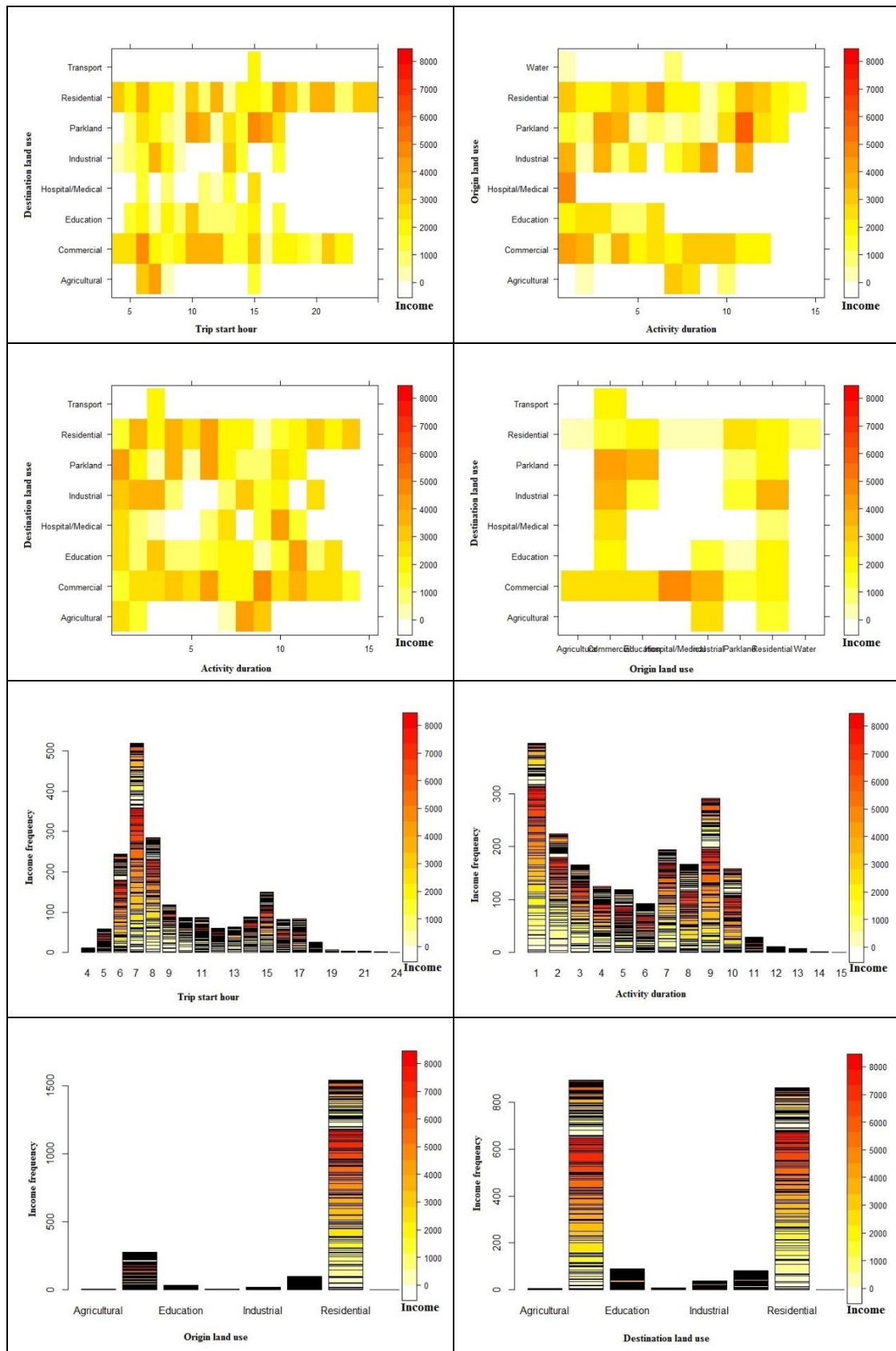


Figure 2 presents the accuracy of each classifier for the age and income groups. Overall, classifiers have the highest accuracy in estimating members of the age group 4 and income group 2. Also, it is seen that all three classifiers have close values of Accuracy for the age and income groups.

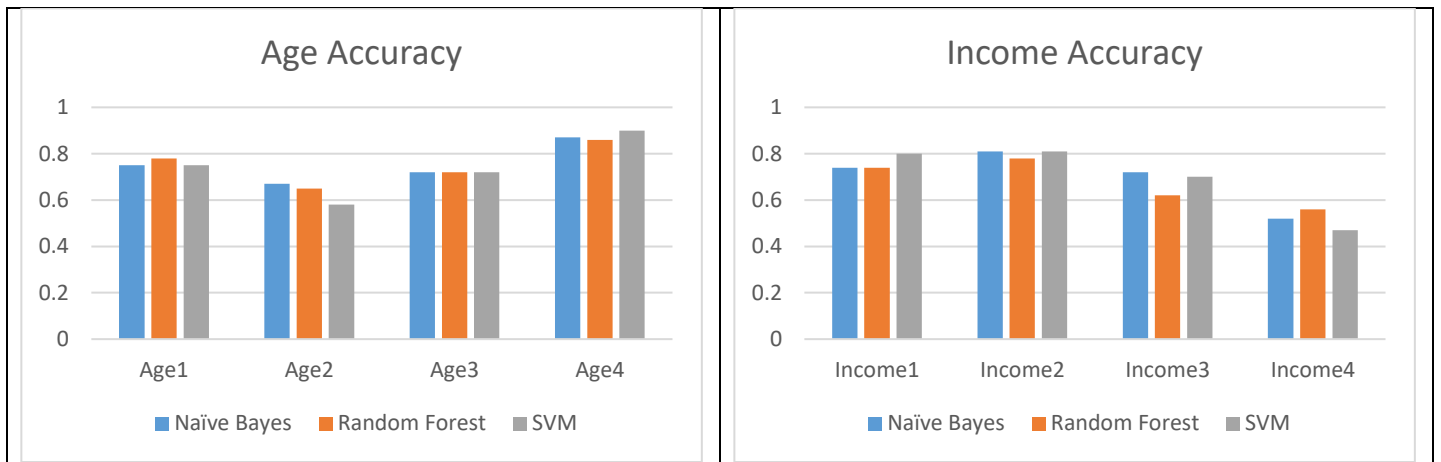


Fig. 2. Accuracy

Figure 3 shows the F-score values for the age and income groups. Totally, the age group 1 and income group 4 have the highest F-score among all other groups. Also, it can be inferred that the classifiers have relatively close F-score values for some of the age and income groups. However, SVM has relatively lower values for the age group 3 and 4, and income group 2.

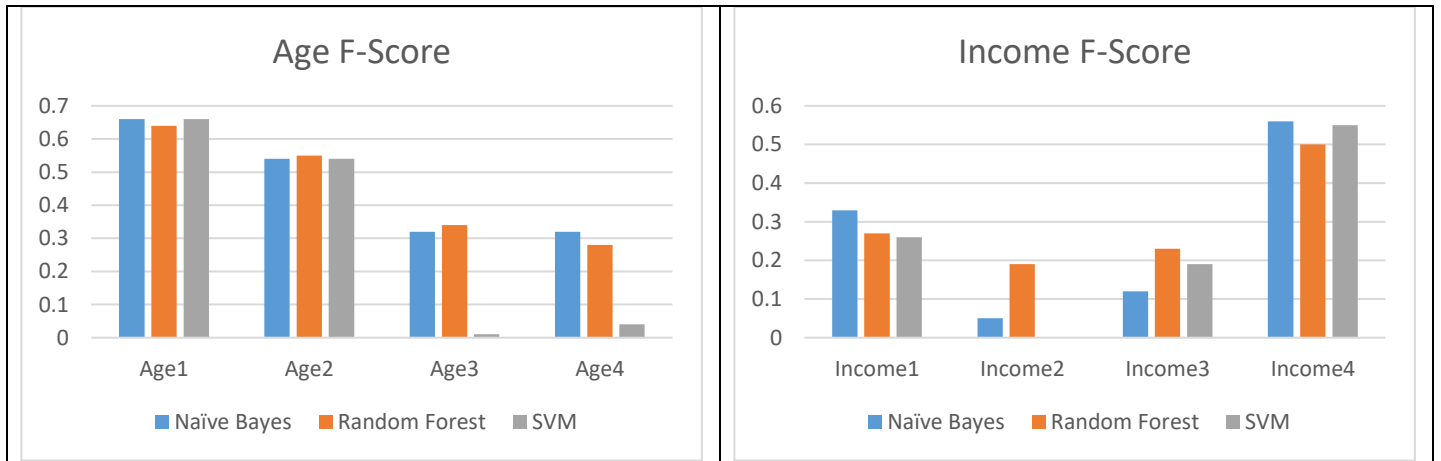


Fig. 3. F-score

Figure 4 demonstrates the Informedness values. Overall, classifiers have the highest Informedness in predicting members of age group 1 and income group 1. Also, the three classifiers have close values of Informedness for some of the age and income groups. On another hand, SVM has relatively lower values for the age group 3 and 4, and income group 2.

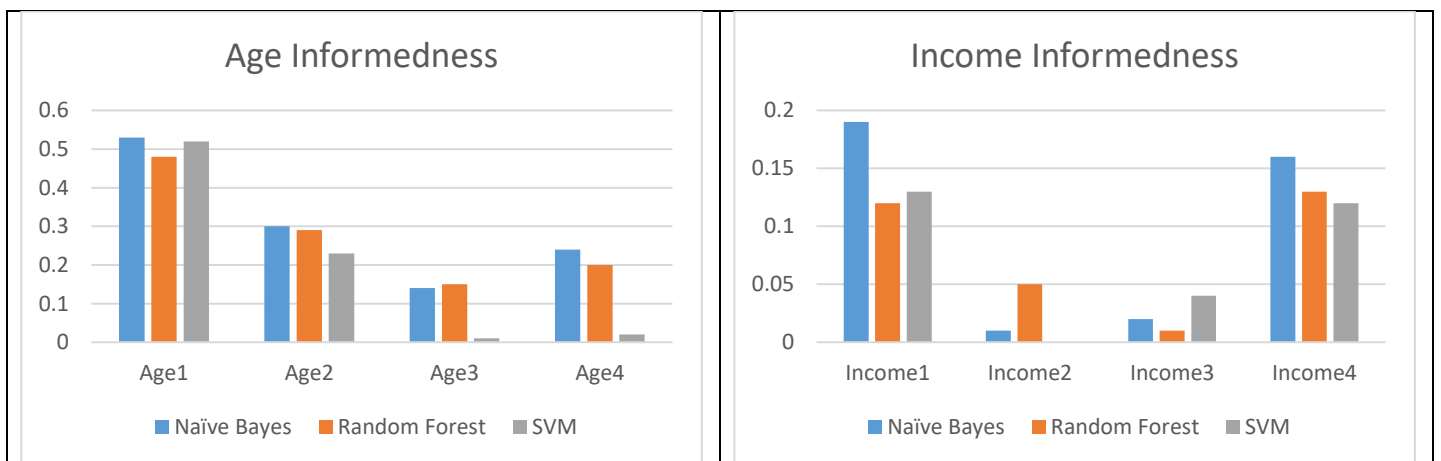


Fig. 4. Informedness

To compare the performance of the classifiers, a random (considering age and income distribution in the training set) classifier is applied. Portions of age classes 1 to 4 respectively are 31%, 34%, 25%, and 10%. Portions of income classes 1 to 4 respectively are 19%, 17%, 27%, and 37%. This means randomly generated classes for age and income are compared with Naïve Bayes, Random Forest, and SVM classifiers in terms of accuracy, F-score, and Informedness. According to Table 2 that represents the results for the random classification, the random classifier is rejected by all three classifiers regarding the accuracy in all classes, but age 4 class; the random classifier is rejected by all three classifiers regarding the F-score in all classes, but age 3 and age 4 classes for SVM classifier. However, both accuracy and F-score can be biased (especially in the case of different sample sizes) because they do not consider all four TP, TN, FP, and FN (Powers, 2011). On the other hand, none of the classifiers is rejected by the random classifier in regard to Informedness, which considers all four TP, TN, FP, and FN. Therefore, all three classifiers have a better performance than the random classification given the Informedness evaluation measure.

Table 2. Random classifier

	Accuracy	F-score	Informedness
Age1	0.58	0.28	-0.02
Age2	0.49	0.34	-0.07
Age3	0.59	0.23	-0.04
Age4	0.90	0.10	0
Income1	0.67	0.22	0.03
Income2	0.72	0.06	-0.07
Income3	0.58	0.12	-0.06
Income4	0.55	0.23	-0.07

Table 3 summarizes the best performance of the classifiers across each evaluation measure for each age and income group. Overall, Naïve Bayes (NB) has a better performance than Random Forest (RF) and SVM because it appears 13 times (out of 24) in Table 3. RF has the second best performance by appearing 8 times in table 3. In the end, SMV just appeared 3 times in the table. In addition, considering the Informedness column, NB dominates the RF and SVM because NB appears 5 times (out of 8) in the fourth column. Therefore, it can be concluded that Naïve Bayes has a better performance than the two other classifiers.

Table 3. Comparing the classifiers

	Accuracy	F-score	Informedness
Age1	RF	NB	NB
Age2	NB	RF	NB
Age3	NB	RF	RF
Age4	SVM	NB	NB
Income1	SVM	NB	NB
Income2	NB	RF	RF
Income3	NB	RF	SVM
Income4	RF	NB	NB

Each of the age and income groups represents a specific passenger type. Age group 1 represents young people who use the public transit system mostly to go to the schools or universities. Age group 2 represents young professionals who use the public transit system mostly to go to the

work. Age group 3 represents older professionals who travel mostly due to work. Age group 4 can represent both older professionals and retired passengers who travel by the public transit mostly to work or recreational purposes. As expected, age group 1 has higher values of Informedness compare to other groups because education trips are more regulated than work, shopping, or recreational trips; diversity of location and time of education groups are more regulated than the other trip purposes. In addition, each of the income groups represents a specific group of passengers. Income group 1 represents low-income people who usually work long hours in the day. Income group 2 and 3 represent medium income people who might have a wide diversity in choosing their trips. Income group 4 represents high-income people who usually work in a more disciplined way. Therefore, as expected, income groups 1 and 4 have more regular trips than income group 2 and 3.

## 4. Conclusions

This research paper compares the performance of three well-known classifiers for estimating age and income of passengers in the public transit network using HTS data. The considered explanatory variables are the start time of a trip, activity duration, land use around the origin, and land use around the destination. Age and income are the target variables. A Household travel survey dataset, which includes both explanatory and target variables, is used to train and validate the classifiers. Three measures including Accuracy, F-score, and Informedness are used to evaluate the performance of the classifiers. The case study is based on HTS data from SEQ between 2009 and 2012. Results show that the Naïve Bayes classifier is a better classifier for estimating the age and income of passengers based on the trip attributes.

Overall, the Naïve Bayes classifier has a better performance in all three evaluation measures than Random Forest and Support Vector Machine. Also, all three classifiers are better than a random (considering age and income distribution in the training set) classifier regarding Informedness. In addition, some of the age and income groups have been classified better than the others, e.g. age group 1 and income groups 3 and 4 have the highest Informedness values, which can be reasoned to more regular trips by the members of those groups.

The main application of inferring the socioeconomic attributes of the passengers in the public transit network is in enriching the emerging big datasets such as smart card datasets. Enriched big datasets can run two streams of novel research and applications. First, the enriched datasets can replace the conventional travel surveys in developing the transport models. Second, developing targeting applications in the public transit network, which focus on discovering groups of passengers with the similar trip and socioeconomic attributes.

Future works can include three main areas. First, evaluating other socioeconomic attributes of passengers such as gender or car ownership. Second, comparing other methods such as linear regression or Neural Network with the classifiers in this study. Third, investigating the potential of replacing conventional travel surveys with the enriched big datasets focusing on the passive and dynamic nature.

## References

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L., and Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87, 123-137.

- Bethlehem, J. G., Cobben, F., and Schouten, B. (2011). The Nonresponse Problem. In *Handbook of Nonresponse in Household Surveys* (pp. 1–25). Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/9780470891056.ch1>
- Chen, C., Gong, H., Lawson, C., and Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice*, 44(10), 830–840. <https://doi.org/10.1016/j.tra.2010.08.004>
- El Faouzi, N. E., Leung, H., and Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges—A survey. *Information Fusion*, 12(1), 4-10.
- Faroqi, H., Mesbah, M., and Kim, J. (2017). Spatial-temporal similarity correlation between public transit passengers using smart card data. *Journal of Advanced Transportation*.
- Faroqi, H., Mesbah, M., and Kim, J. (2018a). A model for measuring activity similarity between public transit passengers using smart card data. *Journal of Travel Behaviour & Society*.
- Faroqi, H., Mesbah, M., and Kim, J. (2018b). Applications of transit smart cards beyond a fare collection tool: A literature review. *Advances in Transportation Studies*.
- Giuliano, G., and Dargay, J. (2006). Car ownership, travel and land use: a comparison of the US and Great Britain. *Transportation Research Part A: Policy and Practice*, 40(2), 106-124.
- Kusakabe, T., and Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, 46, 179-191.
- Liang, L. (2007). Model-Based Synthesis of Geographically Variable Household Travel Data in Small-or Mid-Size Areas, PhD thesis. University of Illinois at Chicago.
- Limtanakool, N., Dijst, M., and Schwanen, T. (2006). The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium-and longer-distance trips. *Journal of transport geography*, 14(5), 327-341.
- Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, 18.
- Pasha, M., Rifaat, S. M., Tay, R., and De Barros, A. (2016). Effects of street pattern, traffic, road infrastructure, socioeconomic and demographic characteristics on public transit ridership. *KSCE Journal of Civil Engineering*, 20(3), 1017-1022.
- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
- Recker, W. W., McNally, M. G., and Root, G. S. (1985). Travel/activity analysis: pattern recognition, classification and interpretation. *Transportation Research Part A: General*, 19(4), 279-296.
- Steinwart, I., and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.
- Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.