Transport technic and technology
Volume XV Issue 2 Year 2019
DOI: 10.2478/ttt-2019-0005

MODELING COMMUTER'S SOCIODEMOGRAPHIC CHARACTERISTICS TO PREDICT PUBLIC TRANSPORT USAGE FREQUENCY BY APPLYING SUPERVISED MACHINE LEARNING METHOD

Abstract: Predictive modeling is the key fundamental method to study passengers' behavior in transportation research. One of the limited studied topic is modeling of public transport usage frequency, which can be used to estimate present and future demand and users' trend toward public transport services. The artificial intelligence and machine learning methods are promising to be better substitute to statistical techniques. No doubt, traditionally been used econometrics models are better for causal relationship studies among variables, but they made rigid assumptions and unable to recognize the pattern in data. This paper aims to build a predictive model to solve passengers' classification, and public transport usage frequency using socio-demographic survey data. The supervised machine learning algorithm, K-Nearest Neighbor (KNN) applied to build a predictive model, which is the better machine learning method for dealing with small datasets, because of its ability of having less parameter tuning. Survey data has been used to train and validate the model performance, which is able to predict public transport usage frequency of future users of public transport. This model can practically be used by public transport agencies and relevant government organizations to predict the public transport demand for new commuters before introducing any new transportation projects.
Keywords: machine learning; modeling; public transport; socio-demographic status; Hyderabad

Nabeel Shakeel[1, 2]
[1] *MoE Key Laboratory of Complex System Analysis and Management Decision, School of Economics and Management, Beihang University, Haidian District, Beijing 100191, China*
[2] *Department of City and Regional Planning, University of Engineering and Technology, Lahore 54600, Pakistan, plannernabeel@buaa.edu.cn*
Farrukh Baig[3]
[3] *School of Transportation and Logistics, Dalian University of Technology, Dalian 116000, China, farrukhbaig@mail.dlut.edu.cn*
Muhammad Abubakar Saddiq[4]
[4] *School of Computer Sciences and Engineering, Beihang University, Haidian District, Beijing 100191, China, abubakar@buaa.edu.cn*

Introduction

History reveals that in the past, Pakistani cities were characterized by low population, shorter trips length and higher non-motorized transport (Thomson, 1977; Tiwari, 2002; Imran and Low, 2003; Singh, 2005; Imran and Low, 2007). However, these characteristics changed as the spatial structures built with time and that were only accessible by public transport and privately owned vehicles. Therefore, public transport trips share to non-motorized has grown in cities of Pakistan. The expansion of cities has increased the trip lengths for urban residents, which make walking less feasible than before and force residents to shift from non-motorized mobility to motorized mode of transport (Imran, 2009). In these scenarios, public transport including both formal and informal, aims to provide quality services to urban resident's mobility at very lower cost as compare to private mode of transport. Many efforts have been done to improve the public transport system in order to make it more convenient, comfortable and environment friendly (Buehler, Lukacs and Zimmerman, 2015; National Transport Authority, 2016; Transport for London, 2016). However, after these efforts there are many reasons of decline in public transport ridership e.g., higher household income, credit access at low interest rates and low tax ratio, encouraged resident's private vehicle ridership (Bliss, 2017; Levinson, 2017). Urban residents often consider public transport inferior and decline use when their overall household income rises. The dwindling demand aimed at public transport usage is explained by a failure in the quality of service, congested buses and time delays, altogether dishearten residents, push them away from public transport usage and to more reliable replacements i.e., private mode of transport (Orcutt, 2017). Therefore, it is necessary to understand residents' decision making to use public transport in order to learn commuting pattern and public transport perception in the eyes of residents. This helps to ensure that residents continue to use the service of public transport and either existing public transport services attracts new residents (Fujii and Kitamura, 2003; Fellesson and Friman, 2008). Some existing studies have estimated some key effects of public transport service quality based on information collected from users, e.g., (Lei and Mac, 2005; Schiefelbusch and Dienel, 2009; Lai and Chen, 2011; Imaz et al. 2015).

Residents' usage frequency choice of travel mode depends on several socio and demographic factors (Cervero, 2002; Ermagun, Rashidi and Lari, 2015). Being

most important deterministic, these factors have been neglected to estimate public transport usage frequency in cities of Pakistan, which ultimately helps to determine public transport demand. Many existing studies are based on historical perspectives and present situations of public transport in cities of Pakistan but to our knowledge very less attentions has been given to deal the future bottlenecks. This study proposed a deeper understanding of residents' attitude towards public transport usage frequency based on socio and demographic factors. An attempt has been made to build a predictive model using K-Nearest Neighbor (KNN), a machine learning algorithm – an advance artificial intelligence approach, which can predict the public transport usage frequency for future coming users of public transport.

The paper has been arranged as follows: section 2 explains the literature review, which comprises the existing studies used statistical techniques, machine learning and artificial intelligence techniques to study public transportation problems by different means of data. The section 3 explains the data used to train and test the predictive model. The section 4 explains the model formulation, section 5 explains the model evaluation and section 6 explains the conclusion of this study and future directions.

1. Literature Review

Public transport plays a vigorous part in shaping city and has always be the reason for sustainable alternative to private mode choice for travel, because of its capacity to reduce traffic congestion and lessen the environment pollution (Tsai et al., 2008). The association between public transport usage frequency and socio-demographic variables has well understood but empirical evidences on this topic is limited. Only few existing studies, e.g., (Badoe and Yendeti, 2007) investigated the public transport usage behavior and studied the factors influencing the ownership of transit pass and daily number of trips by using binary probit model and count variable regression model. Although, this study described well the role of occupation in holding transit pass but had not reveled other socio-demographic variables influencing the usage frequency of public transport (Habib and Hasnine, 2019). In another existing study, e.g., (Farber et al., 2014) a joint model based on econometrics methods was developed to study public transport trip frequency and travel distance by individuals using household survey data. This model incorporated the endogenous relation of trip frequency and travel distance. Among many socio-demographic variables, gender, age, ethnicity, income, occupation, education level and geographical locations of households was found to be significant to study residents' behavior of public transport usage.

Most of the existing studies used statistical models to estimate the passengers' demand of public transport, e.g., (Vicente and Reis, 2018). In some recent decade, researchers and practitioners of public transportation have used different econometrics techniques to solve public transport problems using smart card data, e.g., (Seaborn et al., 2009; Munzinga et al. 2012; Tao et al., 2014, Tao et al., 2016 and Haibo et al., 2016). The other existing studies found which used smart card data to study public transportation problems, e.g., (Agard et al., 2006) studied the public transport users' behavior and trip habits. (Baghai and White, 2005) studied consistency of passengers' travel behavior over time and proposed approaches to retain users. (Utsunomiya et al., 2006) forecasted the demand and proposed approaches to improve user trust and fare adjustment according to needs of users. (Park and Kim, 2008) used historical data to estimate future trends by creating a future demand matrix and (Trépanier and Morency, 2010) model the loyalty of passengers to use public transport. Unfortunately, in developing countries like Pakistan public transport operators and authorities do not open this kind of rich data for research, which is the limitations to use this kind of data for research.

However, most of these studies used regression models to predict demand, which are more suitable for casual relationships studies. These linear models are not able to take into account the out-of-sample observations which ultimately reduced the predictive performance. Due to these reasons, our aim is to increase predictive capabilities using machine learning technique, which are promising to be the better as compared to traditionally been used econometrics models. In the era of artificial intelligence and machine learning, predictive models have attracted many researchers' attentions but usefulness of these techniques are still largely unexplored in studies of public transportation. The purpose of this paper therefore is clearly defined. We used supervised machine learning algorithm, i.e., KNN to form a predictive model of usage frequency of public transport and trained it using survey data related to socio-demographic attributes of residents. During model training process, it learns the pattern that arose. However, after training the model the testing process was started to check how well our model has trained and able to predict new users' usage frequency of public transport.

2. Methodology

The second largest city of Sindh province, Hyderabad, Pakistan (Baig, Rana and Talpur, 2019) was considered as study area for this study. This emerging metropolitan city consists of 1,732,693 citizens, which includes it among the top 10 most populated cities of Pakistan (Government of Pakistan). All types of city's public transport considered under the umbrella of public transportation in present study i.e., buses, mini buses, mini carriages and auto Rickshaws (Government of Sindh, 2019). In order to collect suitable data, online social media based questionnaire survey was conducted by targeting the selected audience. This technique of getting more data in less period of time has been used in many studies, e.g.,

(Ho, 2015; Talpur et al., 2017). Questionnaire was prepared using Google Forms and link was shared through social media to selected audience. Further, questionnaire was also shared via email to collect the responses. A convenience sampling technique was adopted for questionnaire survey (Ross, 2005). A total of 383 valid responses were collected in return of questionnaire survey.

The study is limited as it includes unequal gender distribution. Males were predominantly participated in survey constituting the 71.8% of total respondents as compare to females which constitutes 28.2%. As, the survey was conducted online, therefore, young people of age 21-30 years were dominant with 78.5% of total respondents. While, 17.5% of respondents belonged to 20 years or below and 4% belonged to 31-40 years' age group. Majority of participants had bachelors or higher degree (73.91%), while the rest had attended high school (2.6%), diploma (1.3%), and college (22.19%). The audience were also predominantly students (73.36%) followed by 18.8% private employees, 3.92% government employees, and 3.92% others (including self-employed and labor). As the audience were predominantly students, therefore, most of the respondent's personal income (51.5%) was less than 10,000 PKR. This may possibly be generated by part time work. Meanwhile, 13.5% of respondent's have monthly income of 11000-20,000PKR, 13% of audience have monthly income of 21000 – 30000 PKR, 10% participants have 31000 - 41000 PKR, and 12% respondents earned more than 40,000 PKR monthly income. The sample is almost equally distributed in respect to car ownership variable as 52.22% of respondents owned a car as compare to rest of 47.78% of respondents. On the other hand, 81.8% of respondents owned a bike, while only 18.2% don't have bike.

Responses related to Public transport usage frequency showed that only 13.6% never used public transport but the rest had experience of traveling public transport. Among the participants, 13.3% said that they use public transport daily, 17.8% use once a week, 28.3% respond in using public transport few days in a month, and 27% respondents told that they used public transport few days in a year.

After getting the data, the modeling approach as shown in **Fig. 1** was applied in order to bring the data, train our model on training data and then validate on testing data to check and evaluate the model performance.

3. The Model

3.1 K-Nearest Neighbor Algorithm

The KNN is a supervised instance-based non-parametric machine learning algorithm which can be used for solving both classification and regression tasks (Hand, Mannila and Smyth, 2001). The KNN belongs to supervised machine learning group of algorithms. Because of this, it always given a labelled dataset consisting of training observations (x, y) and would capture the relationship between x and y. The goal in this is to acquire a function $h: X \rightarrow Y$ so that given an unseen instance x, $h(x)$ can surely predict the corresponding output y.

The KNN works on the concept of majority votes between the k most similar observations to a given unseen observation. The k is the hyper parameter of KNN algorithm which is need to find out during hypermeter tuning of the algorithm (Friedman, Baskett and Shustek, 1975). Similarity in observations is defined according to a distance metric between two data points. A general optimal choice as suggested by many researchers are Euclidean distance which can be calculated using Eq. (1).

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x_i')^2} \quad (1)$$

where, x *and* x' are representing Euclidean vectors, starting from initial point and ending at terminal points respectively.

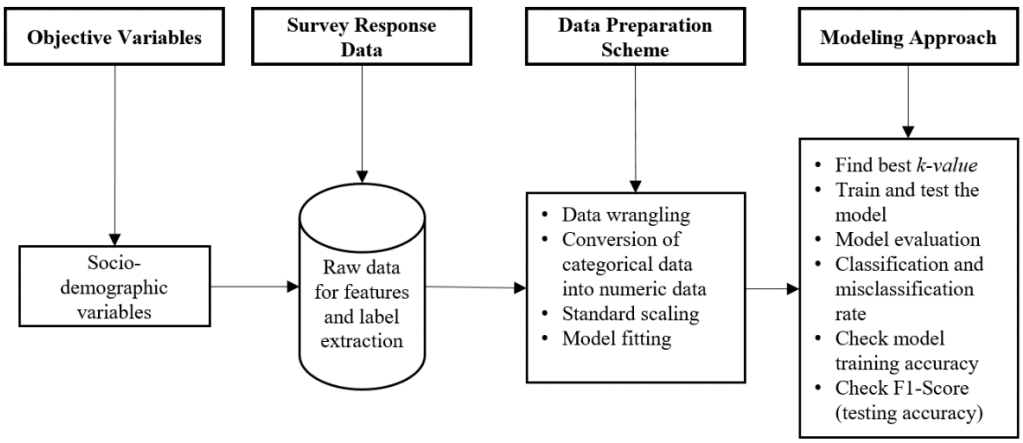

Fig. 1 Modeling approach diagram

Providing numeral k to an invisible instance x and a correspondence metric d, the KNN algorithm runs through the whole dataset for calculating d value between x and every training instance and approximate the conditional probability for each class which can be calculated using Eq. (2). Process diagram for KNN has shown in Fig. 2 indicating the working conceptual method for algorithm.

$$P(y = j \mid X = x) = \frac{1}{k} \sum_{i=N_k} I(y^i = j) \qquad (2)$$

where, $(y = j \mid X = x)$ is representing the corresponding observations, k is hyper parameter $I(y^i = j)$ is the indicator function evaluates to 1 when the argument x is true and 0 otherwise.

Finding the optimal value of k is important tasks during tuning hyper parameter for the model as it required to get the best performance of the model (Jahangiri and Rakha, 2015). One of the frequently used method to find the k value is k-fold cross validation. The k-fold cross validation is a technique to find the prediction error. Subsequently, it is the finest method to define the model parameters.

3.2 Model Formulation

Meanwhile, machine learning techniques are suggested to be used for data having big observations but there are no any restrictions to use these machine learning techniques for data having less observations (Sug, 2012). One problem that may arise for small data is the overfitting and the outliers but experts have proposed different ways to deal with such issues. First suggestion to that is to select ensemble machine learning technique with minimum hyper parameter to tune, which helps to less in complexity in the model (Maheswari, 2018). Among predictive algorithm, the KNN is one of the best algorithm

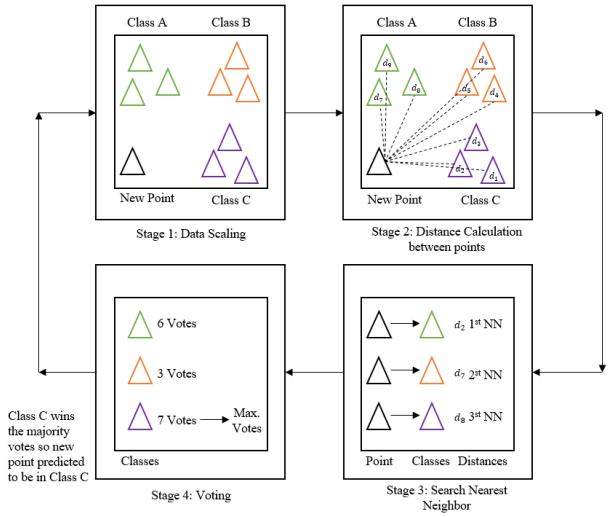

Fig. 2 Conceptual framework of K-Nearest Neigbor Algorithm

to work with small data as this machine learning technique has only one hyper parameter to tune. Second suggestion to small data issue is the increase the training size in order to increase the predictive performance of model (Zhang and Ling, 2018). Data was randomly chosen for training of the model and then validating it using testing data. The 85% observations were selected for training the model and remaining 15% observations were used to validate the model. Analysis has been done using SciKit-Learn library of Python version 3.7 and anaconda framework. The KNN have only one parameter to tune which is number of neighbors named as k value. This number helps in assigning decision boundary to classes. Using k-fold cross validation, in general 5 or 10 folds' cross validation is best for finding optimal value of k (Kohavi et al., 1995; James et al., 2013). The data was examined for different error values as shown in Fig. 3 indicating at 1 neighbors ($k = 1$) have lowest mean error which is not the optimized solution for model because when k-value is very small model shows more blind behavior to overall distribution of classes. On the other hand, for large k value averages extra voters during each prediction and model become more resilient to outliers. It is suggested to choose odd number for k to avoid complexity of class selection so from next error value model at 5 neighbors showing more robustness so $k = 5$ was chosen to train the model.

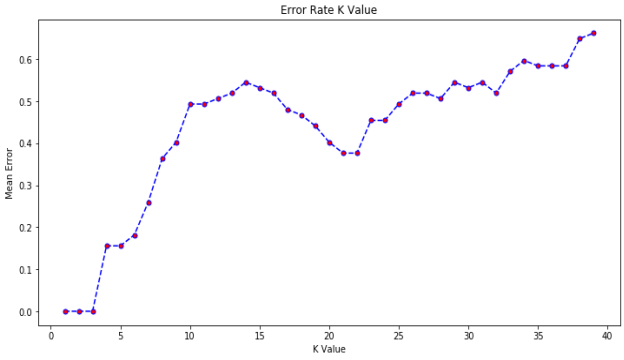

Fig. 3 Mean error value for different k value

4. Model Evaluation

In order to check the predictive performance of model on validation data, confusion matrix (CM) was constructed to estimate the observations of actual and predictive classes. CM is a technique to identify classification and misclassification rate, recall and precision values of the model. The general interpretation of confusion matrix as shown in Eq. (3) can be used to find out recall and precision value.

$$CM = \begin{bmatrix} C_{ij} & \dots & C_{ij} \\ \vdots & \ddots & \vdots \\ C_{ij} & \dots & C_{ij} \end{bmatrix} \qquad (3)$$

where, C_{ij} is the class for ith row and jth column of the confusion matrix. In this matrix all the correct predications are present in the diagonal (correctly classified) and the values outside the diagonal is to estimate misclassification

(wrongly classified) values. Classification and misclassification rate can further be used to calculate precision value as shown in Eq. (4) and recall value as shown in Eq. (5) of the model. Average of recall values can further be used to estimate the training accuracy of the model.

$$\text{Precision} = \frac{M_i}{\sum_j M_j} \quad (4)$$

where, M_i is correctly predicted class i and $\sum_j M_j$ is the sum of both correctly and incorrectly predicted class i.

$$\text{Recall} = \frac{M_i}{\sum_j M_{ij}} \quad (5)$$

where, M_i is correctly predicted and $\sum_j M_{ij}$ is sum of out of all the cases, which labeled as *i*.

CF for model as shown in Eq. (6) indicating the classification and misclassification values observed during testing the model. Rows in the matrix shows the predicted classes and columns shows the actual classes observations. For frequency daily (D), 1 observation has been misclassified as frequency few days in a year (FDY). For frequency few days in a year (FDY), 1 observation have been misclassified as frequency daily (D) and 2 observations as frequency never (N). For frequency few days in a month (FDM), 3 observations have been misclassified as frequency few days in a year (FDY), 4 observations have been misclassified as frequency once a week (OW) and 1 observation has been misclassified as frequency never (N). For frequency once a week (OW), 1 observation have been misclassified as few days in a year (FDY) and 2 observations have been misclassified as frequency never (N). For frequency never (N), 1 observation has been misclassified as frequency once a week (OW). By taking average of the recall values it is estimated that overall training accuracy of the model is 72.4% indicating good predictive performance of the model. For testing accuracy of the model, F1-Score as shown in Fig. 4 indicating testing accuracy of frequency daily, few days in a year, few days in a month, once a week and never is 67%, 76%, 81%, 60% and 57% respectively.

$$CM = \begin{bmatrix} KNN & D & FDY & FDM & OW & N \\ D & 2 & 1 & 0 & 0 & 0 \\ FDY & 0 & 11 & 3 & 1 & 0 \\ FDM & 1 & 0 & 19 & 0 & 0 \\ OW & 0 & 0 & 4 & 6 & 1 \\ N & 0 & 2 & 1 & 2 & 4 \end{bmatrix} \quad (6)$$

Conclusion

This research focused on modeling for predicting public transport usage frequency in the city of Sindh province, Hyderabad, Pakistan. For the purpose of predictive modeling, K-Nearest Neighbor supervised machine learning algorithm is used, which is better for modeling based on observations less in size and small dataset. Looking to the historical studies, we realized that, no attention has been given before to study modeling public transport usage frequency applying machine learning techniques. We proposed a robust model using K-Nearest Neighbor algorithm, which is able to predict the residents' interests toward usage of public transport in future. Such studies can help urban transport planners, policies maker and experts to estimate the future demand before introducing any new transport system. Special attentions should be given to public transport demad when forming and evaluating transport policies in cities of Pakistan. Transportation organizations working in cities of Pakistan do not have the excessive big data but census data and micro-level survey data can revival the socio-demographic status of commuters, which can be further used to study transportation problems applying machine learning methods.

These advance modeling approach needs data relatively big in observations. In future, researchers can bring advance data sets, e.g., GPS location data, smart card data, built environment data and social media data to study transportation problems in cities of Pakistan. The study is limited as it used questionnaire based data, which can be improved in future by using big data. Transport organizations can help to provide the data or open it publically for researchers to better and advance modeling of transport which can ultimately help to improve overall transport system.

Acknowledgment

The authors would like to acknowledge the research grants from the National Natural Science Foundation of China (71771007, 71890971/71890970). The first author also gratefully acknowledges the financial supports from Chinese Government Scholarship (CSC) for graduate studies in China.

References

Agard, B., Morency, C. and Trépanier, M. (2006). Mining public transport user behaviour from smart card data. In: 12th IFAC Symposium on Information Control Problems in Manufacturing – INCOM 2006, Saint-Etienne, France, May 17–19.

Badoe, D. A. and Yendeti, M. K. (2007). Impact of transit-pass ownership on daily number of trips made by urban public transit. J. Urban Plan. Dev., 133(4): 242–249. https://doi.org/10.1061/(ASCE)0733-488(2007)133:4(242)

Bagchi, M. and White, P. R. (2005). The potential of public transport smart card data. Transport Policy 12, pp. 464-474. https://doi.org/10.1016/j.tranpol.2005.06.008

Baig, F., Rana, I. A. and Talpur, M. A. H. (2019). 'Determining Factors Influencing Residents' Satisfaction Regarding Urban Livability in Pakistan', International Journal of Community Well-Being. doi: 10.1007/s42413-019-00026-w.

Bliss, L. (2017). "What's Behind Declining Transit Ridership Nationwide?" CityLab. February 24, 2017. https://www.citylab.com/transportation/2017/02/whats-behind-declining-transit-ridership-nationwide/517701/. Accessed May 17, 2019.

Buehler, R., Lukacs, K. and Zimmerman, M. (2015). Regional Coordination in Public Transportation: Lessons from Germany, Austria, and Switzerland. Final Report VT 2103-04. Virginia Tech, Urban Affairs and Planning. http://www.mautc.psu.edu/docs/VT-2013-04.pdf. Accessed June 8, 2019.

Cervero, R. (2002). Built environments and mode choice: toward a normative framework. Transportation Research Part D, Transport and Environment, (7): 265-284. https://doi.org/10.1016/S1361-9209(01)00024-4

Ermagun, A., Rashidi, T. H. and Lari, Z. A. (2015). Mode Choice for School Trips Long-Term Planning and Impact of Modal Specification on Policy Assessments. Journal of the Transportation Research Board, 97-105. https://doi.org/10.3141/2513-12

Farber, A., Bartholomew, K., Li, X., Paez, A. and Habib, K. M. N. (2014). Social equity in distance based transit fares using a model of travel behavior. Transp. Res. Part A. Policy Pract., 67: 297–303. https://doi.org/10.1016/j.tra.2014.07.013

Fellesson, M. and Friman, M. (2008). "Perceived Satisfaction with Public Transport Service in Nine European Cities." Journal of Transportation Research Forum, 47(3): 93-103. https://doi.org/10.5399/osu/jtrf.47.3.2126

Fujii, S., and Kitamura, R. (2003). "What does a one-month free bus ticket do to habitual drivers? An experimental analysis of habit and attitudes change." Transportation, 30(1): 81-95. https://doi.org/10.1023/A:1021234607980

Friedman, J. H., Baskett, F. and Shustek, L. J. (1975). An algorithm for finding nearest neighbor. IEEE TRANSACTIONS ON COMPUTERS, 1000-1006. https://doi.org/10.1109/T-C.1975.224110

Government of Sindh. (2019). Transport and Mass Transit Department. Available at: https://sindh.gov.pk/dpt/Transport/route.html, Accessed May 17, 2019.

Habib, K. H. and Hasnine, S. (2019). An econometric investigation of the influence of transit passes on transit users' behavior in Toronto, Public Transport 11: 111–133. https://doi.org/10.1007/s12469-019-00195-z

Haibo, L. H. and Chena, X. (2016). Unifying Time Reference of Smart Card Data Using Dynamic Time Warping. Procedia Engineering, 137: 513 – 522. https://doi.org/10.1016/j.proeng.2016.01.287

Hand, D., Mannila, M. and Smyth, P. (2001). Principles of Data Mining. United States of America: The MIT Press.

Ho, J. K. (2015) 'A review of the notions of quality of life (QOL) and livability based on ackovian systems thinking', American Research Thoughts, 1(11), pp. 2513–2532. http://dx.doi.org/10.6084/m9.figshare.1528199

Imaz, A., Habib, K., Shalaby, A. and Idris, A. (2015). "Investigating the factors affecting transit user loyalty." Public Transport, 7(1): 39-60. https://doi.org/10.1007/s12469-014-0088-x

Imran, M. and Low, N. (2003). Time to change the old paradigm: Promoting sustainable urban transport in Lahore, Pakistan. World Transport Policy & Practice, 9(1): 32-39.

Imran, M. and Low, N. (2007). Institutional, technical and discursive path dependence in transport planning in Pakistan. International Development Planning Review, 29(3): 319-352. https://doi.org/10.3828/idpr.29.3.3

Imran, M. (2009). Public Transport in Pakistan: A Critical Overview. Journal of Public Transportation 12(2): 53-83. https://doi.org/10.5038/2375-0901.12.2.4

Jahangiri, A. and Rakha, H. A. (2015). Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, 1-12. https://doi.org/10.1109/TITS.2015.2405759

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning with Application in R. New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kohavi, R. (1995). A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence. Stanford.

Lai, W. and Chen, C. (2011). "Behavioral intentions of public transit passengers — The role of service quality, perceived value, satisfaction and involvement." Transport Policy, 18(2): 318–325. https://doi.org/10.1016/j.tranpol.2010.09.003

Lei, M. and Mac, L. (2005). "Service Quality and Customer Loyalty in a Chinese Context: Does Frequency of Usage Matter?" ANZMAC 2005 Conference: Services Marketing, 138-145.

Levinson, D. (2017). "On the Predictability of the Decline of Transit Ridership in the US." Transportist. March 20, 2017. https://transportist.org/2017/03/20/on-the-predictability-of-the-decline-of-transit-ridership-in-the-us/. Accessed May 17, 2019.

Maheswari, J. P. (2018). Towards Data Science. https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d. Accessed May 17, 2019

Munzinga, M. A. and Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago de Chile, Transportation Research Part C, 24: 9-18. https://doi.org/10.1016/j.trc.2012.01.007

National Transport Authority. (2016). Transport for Dublin: Investment Projects. https://www. nationaltransport.ie/projects-schemes/. Accessed June 4, 2019.

Orcutt, J. (2017). "Why Public Transit Ridership Is Down In Most U.S. Cities." Here & Now, WBUR. March 21, 2017. http://www.wbur.org/hereandnow/2017/03/21/public-transit-ridership-down. May 17, 2019.

Pakistan Bureau of Statistics. Goverment of Pakistan. (2017). Population of major cities census - 2017 population top 10 cities. Available at: http://www.pbscensus.gov.pk/.

Park, J. Y. and Kim, D. J. (2008). The Potential of Using the Smart Card Data to Define the Use of Public Transit in Seoul. Transportation Research Record: Journal of the Transportation Research Board, No. 2063, Transportation Research Board of the National Academies, Washington, DC, pp. 3-9. https://doi.org/10.3141/2063-01

Ross, K. N. (2005) Sample design for educational survey research. Module 3, Quantitative research methods in educational Planning. Module 3. UNESCO International

Institute for Educational Planning. Available at: http://www.unesco.org/iiep/PDF/TR_Mods/Qu_Mod3.pdf.
Schiefelbusch, M. and Dienel (Eds.), H. L. (2009). Public Transport and its Users: The Passenger's Perspective in Planning and Customer Care. London: Routledge.
Seaborn, C., Attanucci, J., Wilson, N. H. M. (2009). Using Smart Card Fare Payment Data to Analyze Multi-Modal Public Transport Journeys in London. Transportation Research Record: Journal of the Transportation Research Board, 2121: 55-62. https://doi.org/10.3141/2121-06
Singh, S. (2005). Review of urban transportation in India. Journal of Public Transportation, 8(1): 79-97. https://doi.org/10.5038/2375-0901.8.1.5
Sug, H. (2012). Applying Randomness Effectively Based on Random Forests for Classification Task of Datasets of Insufficient Information. Journal of Applied Mathematics, 1-13. https://doi.org/10.1155/2012/258054
Talpur, M. A. H. (2017). 'Energy Crisis and Household's Perception about Solar Energy Acceptance: District Hyderabad, Pakistan', SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES), 49(3), pp. 601–604. https://doi.org/10.26692/Surj/2017.09.23
Tao, S., Rohde, D. and Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. J. Transp. Geogr. 41, 21–36. https://doi.org/10.1016/j.jtrangeo.2014.08.006
Tao, S., Corcoran, J., Hickman, M. and Stimson, R., (2016). The influence of weather on local geographical patterns of bus usage. Journal of Transport Geography, 54: 66-80. https://doi.org/10.1016/j.jtrangeo.2014.08.006
Thomson, M. (1977). Great cities and their traffic. Middlesex: Penguin Books Ltd.
Tiwari, G. (2002). Urban transport priorities: Meeting the challenge of socio-economic diversity in cities, a case study of Delhi, India. Cities, 19(2): 95-103. https://doi.org/10.1016/S0264-2751(02)00004-5
Transport for London. (2016). "Improvements & projects." Transport for London. https://tfl.gov.uk/travel-information/improvements-and-projects/. Accessed June 5, 2019.
Trépanier, M., Morency, C., (2010). Assessing transit loyalty with smart card data. In: Presented at the 12th World Conference on Transport Research, Lisbon, Paper No. 2341.
Tsai, T. H., Lee, C. K. & Wei, C. H. (2009). Neural network based temporal feature models for short-term railway passenger demand forecasting. Expert Systems with Applications (36), 3728–3736. https://doi.org/10.1016/j.eswa.2008.02.071
Utsunomiya, M., Attanucci, J. and Wilson, N. (2006). Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning, Transportation Research Record: Journal of the Transportation Research Board, No. 1971, Transportation Research Board of the National Academies, Washington, DC, pp. 119–126. https://doi.org/10.3141/1971-16
Vicente, P. and Reis, E. (2018). Ex-regular Users of Public Transport: Their Reasons for Leaving and Returning. Journal of Public Transportation, 21(2): 101-116. https://doi.org/10.5038/2375-0901.21.2.7
Zhang, Y. and Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. npj Computational Materials, 4:25 https://doi.org/10.1038/s41524-018-0081-z