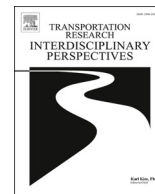


Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Transportation Research Interdisciplinary Perspectives

journal homepage: www.sciencedirect.com/journal/transportation-research-interdisciplinary-perspectives



Blue collar laborers' travel pattern recognition: Machine learning classifier approach

Aya Hasan Alkhereibi^{a,b,*}, Shahram Tahmasseby^a, Semira Mohammed^a, Deepti Muley^{a,c}

^a Qatar Transportation and Traffic Safety Center, College of Engineering, Qatar University, Doha, Qatar

^b Department of Mechanical and Industrial Engineering, College of Engineering, Qatar University, Doha, Qatar

^c Department of Civil Engineering, College of Engineering, Qatar University, Doha, Qatar

ARTICLE INFO

Keywords:

Travel behavior
Transportation planning
Activity-based model
Machine learning
Blue collar travel diary
Travel Pattern

ABSTRACT

This paper proposes a pattern recognition model to develop clusters of homogenous activities for blue-collar workers in the State of Qatar. The activity-based data from the travel diary of 1051 blue-collar workers collected by the Ministry of Transportation and Communication (MoTC) in Qatar was used for analysis. A pattern recognition model is applied to a revealed preference (RP) survey obtained from the Ministry of Transportation and Communication (MoTC) in Qatar for the travel diary for blue-collar workers. Raw data preprocessing and outliers detection and filtering algorithms were applied at the first stage of the analysis, and consequently, an activity-based travel matrix was developed for each household. The research methodology undertaken in this paper comprises a combination of different machine learning techniques, predominantly by applying clustering and classification methods. A bagged Clustering algorithm was employed to identify the number of clusters, then the C-Means algorithm and the Pamk algorithm were implemented to validate the results. Meanwhile, the interdependencies between the resulted clusters and the socio-demographic attributes for the households were examined using crosstabulation analysis. The study results show significant diversity amongst the clusters in terms of trip purpose, modal split, destination choice, and occupation. Furthermore, whilst the Bagged Clusters and Pamk Clusters techniques on the three attributes yielded similar results, the Cmeans Clusters differed significantly in a number of the clusters. Applying such pattern recognition models on big and complex activity datasets could assist transport planners to understand the travel needs of segments of the population well and formulating better-informed strategies.

Introduction

Transport modeling is applied as an effective tool to manage sustainable development in most developed countries. Considerable investments have been made in transportation planning and policy development to observe travel behaviors and forecast future travel demand. These forecasting tools must integrate the design of transport systems, based on global infrastructures and the inclusion of the travel behavior of the inhabitants of the study area. The intention is to develop a system with the capability of accommodating travel demands for the future. Moreover, the rapid global urban development raised concern about the level of traffic congestion, which promotes the need for activity-based travel behavior modeling that leads to better decision-making in urban transportation planning. The activity-based approach is mainly derived from the analysis of activity sequences that are

distributed in different areas and places. It also highlights the effect of demographic attributes such as gender, income, and occupation, as well as spatial characteristics on individual travel behavior. As such the activity-based travel analysis can give an index on the reliability of the urban network and transportation infrastructure since it evaluates the travel sequences for a defined purpose of activities, which is influenced by place distributions, and travel modes.

Qatar's population is made up of 8.641% locals or Qatari nationals and 91.358% non-Qataris, predominantly expatriates ([Planning and Statistics Authority, 2020](#)). The population growth rate has been rising since 2003, and especially since the announcement of the awarding of the FIFA World Cup 2022 bid to Qatar in 2010. The massive investment in developments witnessed over the years in Qatar increased the number of foreign workers, especially male workers significantly. Due to the nature of work and immigration regulations, demographically most of

* Corresponding author at: Qatar Transportation and Traffic Safety Center, College of Engineering, Qatar University, Doha, Qatar.

E-mail address: aalkhereibi@qu.edu.qa (A.H. Alkhereibi).

<https://doi.org/10.1016/j.trip.2021.100506>

Received 6 July 2021; Received in revised form 4 October 2021; Accepted 21 November 2021

2590-1982/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

the non-Qatari population is heavily distorted towards males (1:7) and working-age groups 18–64 (60.72%). Qatar also has a large proportion of its foreign males (>70% of all resident men) who dwell in single household labor camp quarters (Planning and Statistics Authority, 2020). Overall, this segment of the population is young and often illiterate. In this paper, the aforementioned segment will be referred to as “blue-collar” workers. Given the unique demographic characteristics of the blue-collar worker population in Qatar, the objective of this research is to contribute toward advancing and operationalizing the activity-based approach for the aforementioned group by;

- i. Apprehending the time-based behavior for blue-collar workers,
- ii. Developing a comprehensive representation of attributes characterizing working and nonworking daily activity-travel patterns for blue-collar workers, and
- iii. Proposing models of clustering model for analyzing the activity-based travel behavior for blue-collar, given socio-demographic attributes, travel diary, and Traffic Analysis Zones (TAZs).

Literature review

This section highlights the latest studies and research related to the analysis of activity-based travel behavior. In addition, the applicability of machine learning classifiers for the analysis of activity-based travel behavior is investigated. Activity-based models have been widely used in practice; for their ability to define the following four components; i) identification of activities, ii) recognition of the travel destination, iii) determination of the mode of travel, iv) and allocation of the transport network. However, activity-based models incorporate other aspects into the models, such as time, space, and the relationship between activities and trips made by a single household or by several people in the household.

The more realistic representation echoes the effect of travel structures on the generation of activities and choice of family trips. Additionally, activity-based models allow capturing more detail at the personal level of the family, which can lead to more precise analysis and modeling such as disaggregating attributes at the personal level. There are behavioral aspects associated with travel patterns and characteristics, such as where to travel (choice of destination), when to travel (choice of departure time), and how to get to these activities (choice of mode). This realistic representation of behavior makes activity-based models more representative of how changes in investments and policies might affect people’s travel behavior. Castiglione et al. (2014) defined activity-based travel models as models that describe the movement and mobility of people based on their needs and desire to move from one point to another point.

Cui et al. (2018) investigated a more microscopic behavioral analysis of individual travel decision-making, conducting an experimental method using a process-tracing technique to acquire travel data based on repetitive activities. In their study, they built a dynamic stochastic model based on decision field theory. The results of the study demonstrated the model’s ability to reproduce and analyze travel decision-making. Bcikova (2014) indicates four distinct groups of young tourists in the UK student travel market. These groups were based on the motivation, behavior, and age of the students. The study expands the know-how of travel motivations and behaviors among young people and aims to go further in implementing the segmentation of UK university students based on their travel habits during the holidays.

In recent years, there has been a growing interest in incorporating machine learning techniques into the activity generation and activity planning phases in activity-based travel demand models (Allahviranloo and Aissaoui, 2019; Li, 2017). Machine learning methods are reproducible techniques that are commonly used to recognize complex patterns without being programmed (Kubat, 2015). Machine learning is a well-known technique among computer science scholars but has very limited applications in transportation planning. In the context of this

research study, Joh et al. (2002) showed that machine learning can be applied for dealing with the activity/trip sequence alignment problems. Given the lack of best fitted explanatory variables to represent the process of population grouping with the nearest characteristics, Delhoum et al. (2020), applying the machine learning technique is recommended for activity-based travel behavior analysis. The aforementioned technique demands employing explanatory variables for population clustering, which originate from differentiated groups and their associated probability distributions. This hurdle is overcome in this research study by considering only one segment of the population, i.e., is the blue-collar workers, having convergent characteristics.

Lately, machine learning techniques inspired scholars of transportation planning to incorporate new insights into activity-based modeling. Allahviranloo and Aissaoui (2019) applied k-means clustering algorithm to identify unique trip clusters. They applied the AdaBoost algorithm to predict the start time and the duration of activity-based trips based on the socio-demographic characteristics of individuals as well as the daily schedule of individuals (Li, 2017). The interpersonal variability detected in cross-sectional data for the same day of a group of individuals was utilized to generate daily intrapersonal variability. Drchal et al. (2019), introduced a data-driven activity planning approach for the next activity planning component in agent-based mobility models. The new technique exchanges several expertly designed mechanisms and their complex engineering interactions with a handful of machine learning techniques. Despite advancements in the generation of timing information for the planning engine in activity-based travel demand models, there are significant room and opportunities to improve the performance of those models in terms of estimation accuracy, the efficiency of calculation, and their implication for practical cases.

Hafezi et al. (2021) proposed an algorithm that uses the random forest approach to model the temporal dimensions associated with the traveler’s daily activity patterns in their study. The authors indicate Random Forrest (RF) models using the Classifier Classification and Regression Tree (CART) technique along with the methods for measuring the importance of variables, that have not previously been used to forecast the start time and the duration of travel in the analysis of travel behavior. Nonetheless, the RF method is applied in other areas of transport such as the detection of traffic accidents and the recognition of the mode of transport (Awais and Hato, 2014; You et al., 2017).

By employing a time-use and activity model, Hafezi et al. (2021) improved the current activity-based models by utilizing data daily and pattern recognition models to identify the groups of individuals having consistent demands for their activities outside their residence. And, by extension, similar travel patterns e.g. the type of activities, the start time, end time, and the duration of activities alongside the probability distribution. The Scheduler for Activities, Locations, and Travel (SALT) framework uses a sequence of advanced rule-based econometric models to predict the specifications of daily trip demand and route generation. Identifying homogeneous population clusters can greatly improve the estimation accuracy of activity-based travel demand models.

Sun and Huang (2006) explained the Support Vector Machine (SVM) as a machine learning technique that applies kernel for data transformation into a high-dimensional space. Its algorithm attempts to find the optimal linear decision surface in the resulted space of data transformation by classifying data into two categories. Multiple kernels could be applied for data transformation such as; linear, polynomial, Sigmoid, Gaussian, and Radial Basis Function (RBF) (Niculescu-Mizil, and Caruana, 2005). In the case of linearly separable data, the optimal hyperplane perfectly divides the data into two classes with a minimum misclassification and a maximum feasible margin (distance between the surface and nearest data points – support vectors). SVMs are integrally binary classifiers. They could be used as a multiclass classifier, which results in a continuous score for each prediction output. The scores represent the confidence of the classification, but don’t correspond to the class probabilities (Niculescu-Mizil, and Caruana, 2005; Bar-Gera

and Boyce, 2003).

Several scholars stressed that the activity-based models demand excessive efforts in the data preparation stage because the data are constructed in a sequence of activities and travel (Joubert and de Waal, 2020; Chen et al., 2020; Koushik et al., 2020). These characteristics could result in a gap or inconsistencies among the travel diaries which may not be addressed in the traditional four-stage models since it concentrates on the individuals' trips, not on the sequences of activities.

Consequently, this study addresses the limitations presented in the above-mentioned literature review. The paper implements machine learning classifiers to recognize activity-based travel behavior patterns for the blue-collar laborer segment of the Qatari society. The subject is tackled from a different perspective by developing a classification activity-based model that groups the homogeneous daily activity data into clusters. Each cluster has its essential characteristics such as the type of activity, sequences, start time, activity duration, etc.

Methods and materials

Many modules were applied to recognize and comprehend activity-based travel patterns (Cui et al., 2018; Allahviranloo and Aissaoui, 2019; Chen et al., 2020; Hafezi et al., 2021). The bagged clustering (B-Cluster) algorithm was employed to initialize and recognize the number of clusters, as suggested by Leisch (1999). Subsequently, the C-Means and Pamk algorithms were applied to optimize the performance of the B-Cluster algorithm as recommended by Hung and Yang (2001). In this study, the socio-demographic attributes were then examined with defined clusters and using the CROSSTABS test to deal with interdependencies. Finally, the CART classifier algorithm was applied to align the person's socio-demographic with the predefined clusters. The theoretical model proposed in this study has the capability of learning and predicting the temporal information of the patterns of the daily activities of individuals concerning the characteristics of heterogeneity. Not only, does applying the aforementioned framework to activity-based models reveal the effectiveness of machine learning techniques in distinguishing the temporal attributes of travelers' daily activities patterns, but also helps better understand the relationship between travel planning and the pattern of activities with respect to the characteristics of heterogeneity. A flowchart of the research methodology followed in this study is displayed in Fig. 1.

Data acquisition, processing, and transformation

The State of Qatar has experienced population growth since 1961 according to the WorldBank.¹ The growth rate ranges from 8.19 in 1961 to 1.17 in 2020. The country's maximum growth rate reached 17.15 % in 2007. The population of the country at the beginning of August 2021 was estimated at approximately 2,380,011 persons constituting 1,756,026 males and 623,985 females, according to the 91st issue of the Qatar Monthly Statistics Bulletin, published by the Planning and Statistics Authority (PSA) for July.² The nature of the population of Qatar is unique since it consists of a vast and diverse segment of different nationalities. The majority of the population originates from blue-collar workers which comprise 79.7–89.1% of the total population since 2006 (Planning and Statistics Authority, 2020). For this paper, the blue-collar segment of the population is male, single households who reside in the labor camp and are semi-skilled to unskilled, less-educated workers engaged in different economic sectors.

The primary data used for this research study was obtained from the Ministry of Transportation and Communication's (MOTC) household

¹ Population Growth Annual Percentage Qatar- <https://data.worldbank.org/indicator/SP.POP.GROW?locations=QA>.

² Qatar population drops to 2.38mn in July, lowest since July 2016, 2021, Gulf Times.

survey for blue-collar workers conducted in 2018 by using a travel diary. As a part of this survey, 1051 blue-collar workers were interviewed. The laborers' survey program aimed to collect their socio-economic characteristics, activity patterns, and the modes of travel used by laborers who are engaged in different sectors of the economy. It was expected that the activity patterns and travel; characteristics of laborers are predominantly in the fields of construction, industry, and services. Tables 1 and 2 show the sample and population distribution of laborers by their economic sector and municipality they dwell in, respectively.

Fig. 2 depicts the 2010 labor buildings overlaid on the 2010 Census Block boundaries. The 2010 data is color graded according to the number of residents in each camp in each Block, and the labor buildings are represented by points reflecting each building location. There are 1,442 labor buildings defined in the building's database of MoTC.

The labor buildings were selected randomly and confined to the study areas. Within the buildings, a random selection technique was applied to identify laborers based on sampling done by floor numbers as well as room numbers. The sample returned from the interviews was constantly monitored against the proportion of the occupation sector. Where it became evident that there was over-sampling in a specific category, then screening questions were being raised to enable the sample frame to be adjusted. The survey was conducted for one month between the 9th of April 2018 and the 10th of May 2018. After data cleansing and filtering outliers, 1051 datasets of individuals were considered for further analysis with 4990 travel records. Seventy-eight percent of the dataset contained a response to all the survey questions. Table 3 shows the general characteristics, in terms of the age, income, education level, and occupation, of the respondents. As shown in Table 3, more than 93% of the respondents were from the category of economically active ages ranging from 19 to 54 years. Similarly, the overwhelming majority of the respondents (93%) earn QR 3,000 (\cong USD 820) or less per month. The education level of respondents is mostly secondary and high school educators with a total percent of 70.2.

The activities conducted by laborers were mainly personal activities including personal care, leisure, watching TV, shopping, etc with a relative frequency of 39%. Work duties-related activities had a relative frequency of 31%, and preparing meals with a relative frequency of 30% were two other major activities of laborers. Transportation and study activities had a minor proportion in this case study, with less than 1% relative frequency. Table 4 and Fig. 3 show activities descriptions and frequencies respectively.

Initialization of clusters numbers and characteristics

Bagged clustering algorithm

The partitioning and hierarchical methods were combined to introduce a new clustering method called the "bagged clustering algorithm" (Leisch, 1999). Since partitioning methods are known as unstable methods, Leisch (1999) attempted to stabilize them by repeatedly running the algorithm of the cluster.

The bagged clustering algorithm initiates from constructing B bootstrap training sample X_1^N, \dots, X_N^B , through drawing in respect to a replacement from the original sample X_N . At the second step, the algorithm follows running the cluster method technique e.g. K-means, on each set of the trained data. This will result in $B \times K$ centers ($c_{11}, c_{12}, c_{13}, c_{14}, \dots, c_{1K}, c_{21}, \dots, c_{BK}$) where K represents the number of centers used in the base method, and c_{ij} is the jth center that was found using X_i^N . Then, it prunes the data set C^B by computing a portion of X_N concerning C^B and removing all centers where the corresponding cluster is either empty or below the predefined threshold θ , as per Eq. (1). The outcome is a new set,

$$C^B(k-\theta) = \{ C \in C^3(k) \forall (x : C = C(x)) \geq \theta \} \quad (1)$$

C-means clustering algorithm

This section briefly describes the Fuzzy C-means, considering a set of

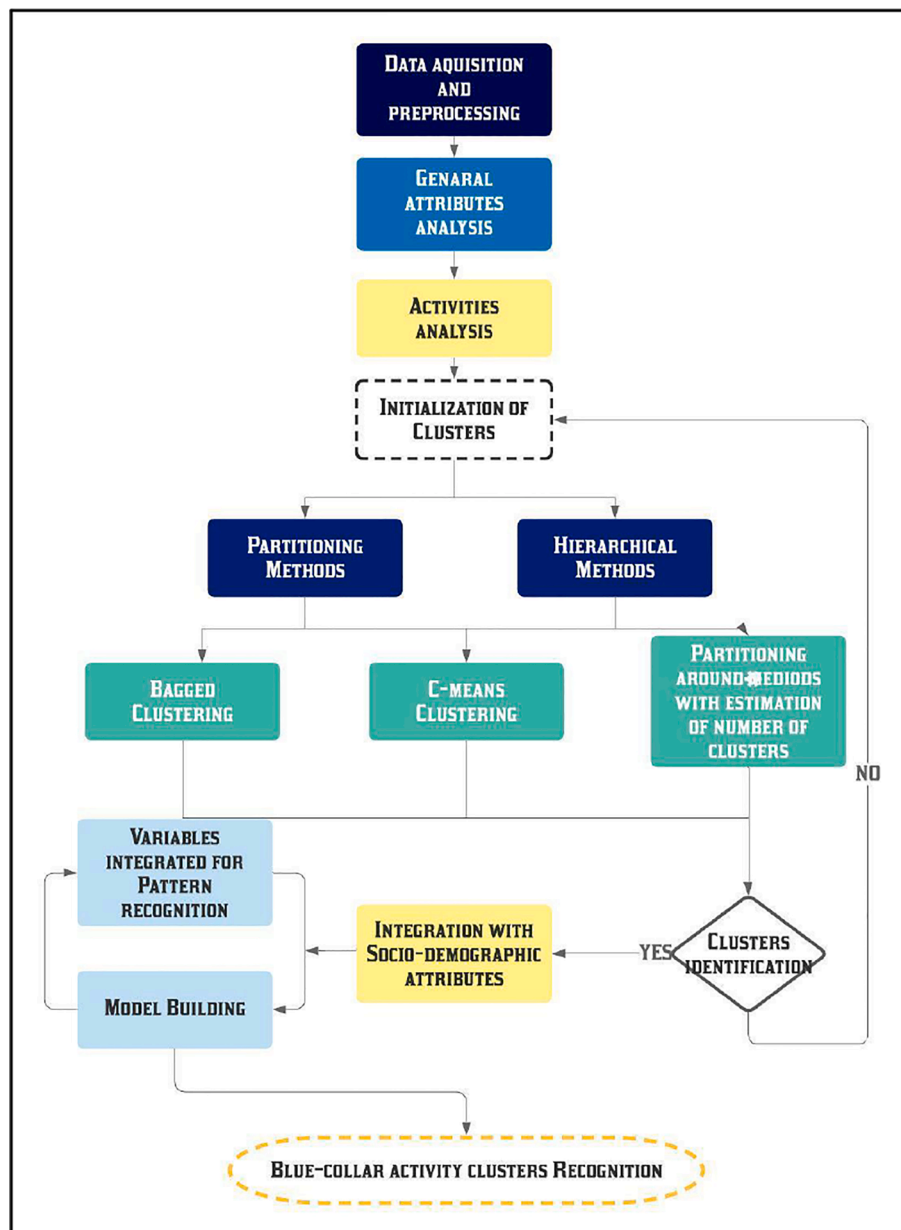


Fig. 1. Graphical illustration of the methodology applied in this study.

Table 1
Distribution of Laborers by Economic Sector.

Economic Sector	Percentage of Total Laborers in Qatar (%)	Percentage of Samples' sectors (%)
Accommodation and food services	2.90	5.30
Agriculture	1.90	2.70
Construction	56.60	46.60
Manufacturing	11.20	13.20
Mining/Quarrying	8.90	8.60
Transportation	3.30	3.70
Wholesale and retail, and vehicle repair	15.20	11.70

Table 2
Spatial distribution of labor camps.

Area	Labor Population	Percent of Total
Al Khor	61,150	7
Al Wakra	28,250	3
Central Doha	120,000	13
Industrial areas in Al Rayyan	265,000	29
Mesaieed	32,000	3
Other areas	83,000	9
Ras Laffan	115,000	12
Rest of Doha (excluding Al Rayyan industrial areas)	210,000	23
West Coast	15,000	2

unrecognized patterns $X = \{x_1, x_2, x_3, \dots, x_N\}$, where $x_1 \in R^f$, N is the number of patterns and f is the dimension pd pattern vectors. The main focus of the FCM algorithm is to minimize the value of an objective function, which is measuring the partitioning quality the divides the

dataset into C clusters (Hung and Yang, 2001).

The main concept of improving the quality of the clusters is by measuring the distance from pattern x_i to the candidate cluster center w_j with the distance from pattern x_i to any other candidate cluster. As

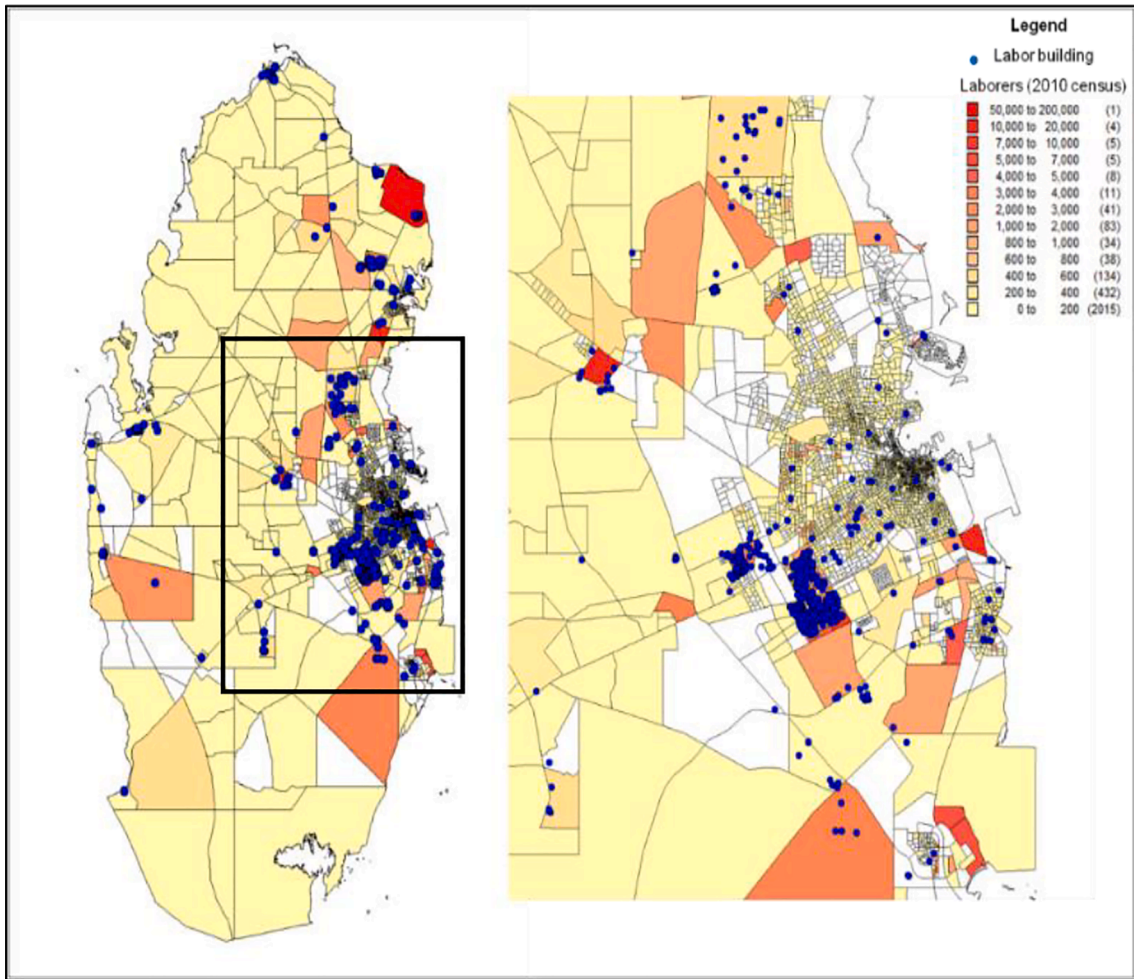


Fig. 2. Distribution of Laborers and Labor Buildings (source: MoTC).

previously mentioned the FCM is mainly an optimization function that aims to minimize the weight within Groups of Squared Errors (WGSS) as follows:

$$J_m(U, W) = \sum_{j=1}^C \sum_{i=1}^N (\mu_{ij})^m d_{ij}^2 \quad (2)$$

where:

N: the number of patterns in X; C: the number of clusters; U: the membership function matrix; the elements of U are μ_{ij} ; μ : the value of the membership function of ith pattern belongs to jth cluster; d_{ij} : the distance from x_i to w_j ; w : cluster center vector; m : the exponent on μ_{ij} to control fuzziness or the amount of clusters overlap.

The minimization objective function of the FCM algorithm is subject to the following constraints on U:

$$\mu_{ij} \in [0,1], i=1, \dots, N, j=1, \dots, C \quad (3)$$

$$\sum_{j=1}^c \mu_{ji} = 1, i=1, \dots, N \quad (4)$$

$$0 < \sum_{i=1}^d \mu < N \quad (5)$$

Partitioning around medoids with an estimation of the number of clusters (Pamk)

The partitioning-based algorithms were found to be the best method to discover the geo-location that is hidden in the data (Ibrahim et al., 2012). The core concept of the partitioning clustering methods originates from K-medoids and K-means. The mutual function between these two methods is that it partitions the dataset randomly into K subsets, then it refines the centers of the clusters repeatedly which will result in a reduction in the cost function. The cost function represents the spatial domain in the total summation of distance error between all data objects and their assigned centers. Where there is no center for a data point, it is assigned to the nearest center. The K-means clustering algorithm is well known and commonly used since it is easy to understand and be implemented. Besides, the K-means algorithm builds clusters centers that act like gravity points for all dataset points, which easily build the clusters. By determining these gravity points, the method guarantees minimum error distances. However, Nanopoulos (2014) in their research suggested that the characteristics of the gravity points may differ; therefore, instead of representing the clusters by their gravity points, it is preferred to apply the K-medoids technique by which actual points are implemented to represent the cluster as a medoid. This feature of using a real point of data as a representative can diminish the sensitivity to outliers. The Partitioning Around Medoids (PAM) is a K-medoids-based algorithm. It initializes the number of desired clusters and accordingly selects the k medoids randomly. Later, each point of the dataset is tested to figure out whether it should be a medoid or not. Finally, the algorithm makes a comparison between the pairs of medoids

Table 3
Sample general characteristics.

Variable	Category	Frequency	Relative Frequency
Age	19_24 Years old	164	15.2
	25_34 Years old	472	43.9
	35_44 Years old	299	27.8
	45_54 Years old	115	10.7
	55_74 Years old	21	2.1
Income	Confidential	2	0.2
	Less than or equal to QAR 1,000/month	352	32.7
	QAR 1,001_3,000/month	656	61.0
	QAR 3,001_6,000/month	31	2.9
	QAR 6,001_10,000/month	5	0.5
Education attained	QAR 10,001_20000/month	3	0.3
	Nursery	66	6.1
	Primary	152	14.1
	High School	756	70.2
	Graduate	83	7.7
Occupation Sector	Masters Degree (M.Sc.)	1	0.1
	Other	10	0.9
	Construction Buildings	112	10.4
	Construction Infrastructure	390	36.2
	Offices Government Administration	4	0.4
	Offices Private	42	3.9
	Agriculture and fishing	29	2.7
	Electricity, gas, water, and waste management	4	0.4
	Hotels and Food Service Activities	56	5.3
	Industrial Mechanic shops, Factory, Logistic (Manufacturing)	142	13.2
Occupation Sector	Mining, Quarrying, oil and gas extraction	92	8.6
	Transport and storage	40	3.7
	Wholesale and retail trade/ repair of motor vehicles	126	11.7
	Other	29	2.7
	Clerks, Documenters, and office workers	12	1.1
	Clerks, Documenters, and office workers	12	1.1
	Craft and related trades workers	14	1.3
	Elementary Occupations (Drivers, labourers, e.t.c.)	833	77.4
	High-Level Directors, Managers, and Legislators	3	0.3
	Industrial workers (Plant/ Machine Operators & Assemblers)	94	8.7
Occupation Sector	Shop employees, and service workers	4	0.4
	Skilled agriculture and fishing workers	18	1.7
	Specialties/Experts (Professionals)	5	0.5
	Technicians and Associate Professionals	31	2.9
	Professionals	53	4.9
	Other	53	4.9

and non-medoids objects and decides which pair of medoids are contributing to the overall quality of the clusters. The quality here is measured by the overall distances between the medoids and medoids objects in the cluster. In this research study, since the authors implemented several clustering techniques, it is hard to distinguish the obvious (correct) results. Nonetheless, the implementation of different techniques will validate the clustering quality.

Results and discussion

By applying various clustering techniques discussed in Section “Initialization of clusters numbers and characteristics”, different

Table 4
Frequencies of observed activities.

Activity	Activity Description	Frequency	Relative Frequency
1	Transportation	27	0.541%
2	Personal activities	1924	38.565%
3	Preparing meals/eating	1487	29.806%
4	Study/school	2	0.040%
5	Work/job duties	1542	30.908%
6	Other	7	0.140%

structures are observed. Table 5 outlines the results of these clustering techniques. Figs. 4–6 show the dot plots for the three clustering techniques. As outlined, it appears that CMeans and Pamk bring about an even distribution compared to the Bagged Clustering technique. The latter returns negligible records for clusters 5, 6, 7, 9, and 10, and nil records for cluster 8.

Based on the C-means clustering technique, it is observed that (a) work and job duties, (b) personal activities, and (c) preparing meals for a group of co-workers are the main activities of the people with the share of 21%, 43%, and 31% respectively. Engagement in other activities such as study is negligible. For destination choice, returning home and commuting to the workplace, and traveling to other places are the three main destinations. By the way, Al Rayyan, Al Sheehaniya, and Doha are respectively three main municipalities where trips were made. In the context of the modal split, private shuttle (i.e. 40%), as well as car/van/truck (40%), are the main transport modes whilst the share of walking is about 16%. Only 3% ride public transport in cluster 1. Regarding the income, the majority (almost more than 63%) are earning between QR1,001 and QR 3,000 per month. Also, 33% earned less than QR 1,000 monthly. Around 2.5% earned between QR 3,001 and QR 6,000 per month and below 1% earned more Qatari Rails monthly.

Regarding nationality, the majority (around 43%) were Indian. Workers from Nepal (23.773%), Bangladesh, and the Philippines are the other main nationals respectively. The other demographic attributes indicate young adults between the ages of 25 to 34 years by almost 40% followed by 33% of mid-age adults ranging from 35 to 44 years old. Moreover, almost 99.9% of the adults are male and are employed full-time. Around 84% of the occupation refers to elementary occupations such as drivers, laborers, while almost 9% are industrial workers such as plant/machine operators, and assemblers. Approximately 3% in this cluster are technicians and associate professionals. (a) construction infrastructure, (b) industrial mechanic shops, factories, logistics (manufacturing), and (c) mining, quarrying, oil and gas extraction, and (d) motor mechanics are the main domain of occupations.

The above-mentioned trends are more or less valid for the other nine (9) remaining clusters. Table 6 outlines the figures for all ten (10) clusters predominantly based on the Pamk clustering technique. Comparing the trend of the attributes in the above-mentioned clusters, it is evident that (1) trip purpose, (2) modal split, and (3) destination choice from trip attributes and occupation from socio-economic attributes indicate significant alterations amongst the clusters. For instance, while “personal activities” is a dominant trip purpose for clusters 2, 7, 9,

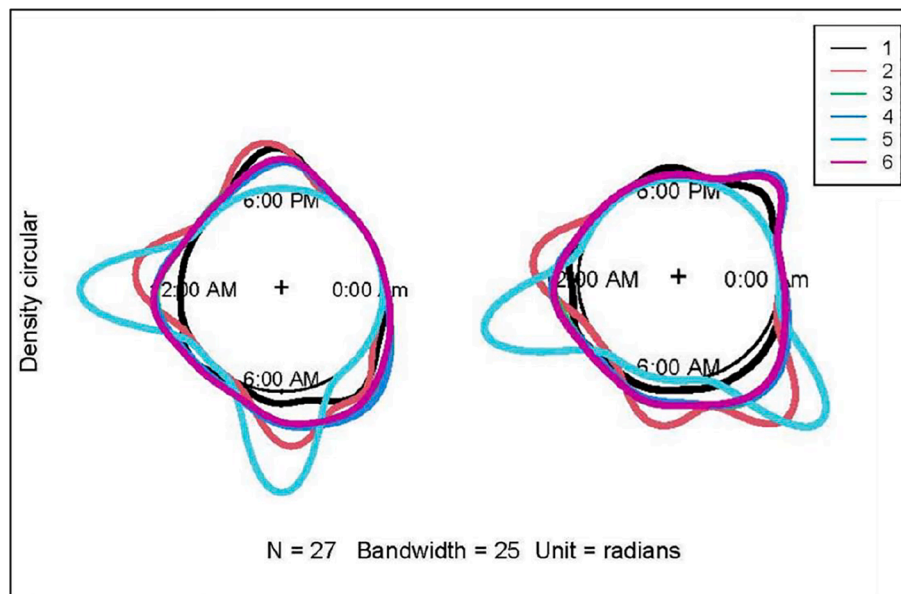


Fig. 3. Frequency of blue-collar labor activities over the day. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Number of clusters per method.

Method	Cluster number										Total
	1	2	3	4	5	6	7	8	9	10	
Bclust	2150	2405	28	356	32	3	4	0	8	4	4990
Cmeans	358	583	43	658	353	612	404	1369	572	38	4990
Pumk	604	1692	693	443	591	529	387	38	8	5	4990

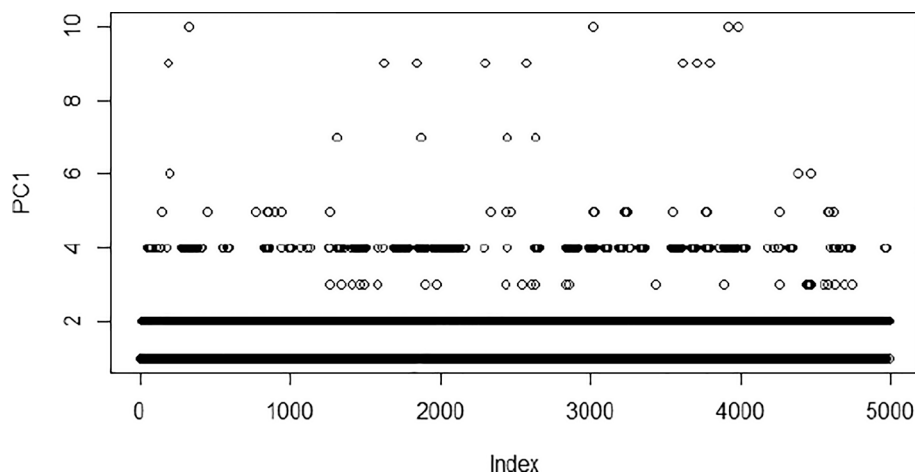


Fig. 4. Dot plot of clusters generated using Bagged clustering method.

and 10; “work & job duties” is the foremost purpose for clusters 3, 5, and 9. Besides, “preparing meals” for co-workers is a major activity for clusters 4 and 8. Regarding the destination choice, travel to accommodation/home is pivotal for clusters 1, 3, 5, 6, 7, 9, and 10. Furthermore, the travel to the workplace is not considered for cluster 9 and was not applicable for cluster 10.

The Modal split, traveling by “private shuttle” is dominant for all the ten clusters. The results also show that Car/van/truck are major trip modes for clusters 1, 4, and 10, while walking is a key mode for clusters 2, 3, 5, 6, and 9. The results also indicate that almost the majority of laborers had income between QR1,001 and QR 3,000. A significant

proportion of the respondents evidently earn less than QR 1,000 per month. Between 40% and 50% of the respondents’ age were between 25 and 34. Similarly, between 23% and 33% of the respondents’ age in the clusters were between 35 and 44, which means that the majority of respondents are between 25 and 44 years old in all the clusters shown in pamk clustering.

More than 75% of the respondents in all the clusters are engaged in elementary occupations. In cluster 10, 100% of the respondents indicate elementary occupation. Industrial workers and technicians are the other types of occupation indicated by the respondents significantly. Between 9% and 13% of the respondents in clusters 1, 2, 3, 4, 6, 7, and 8

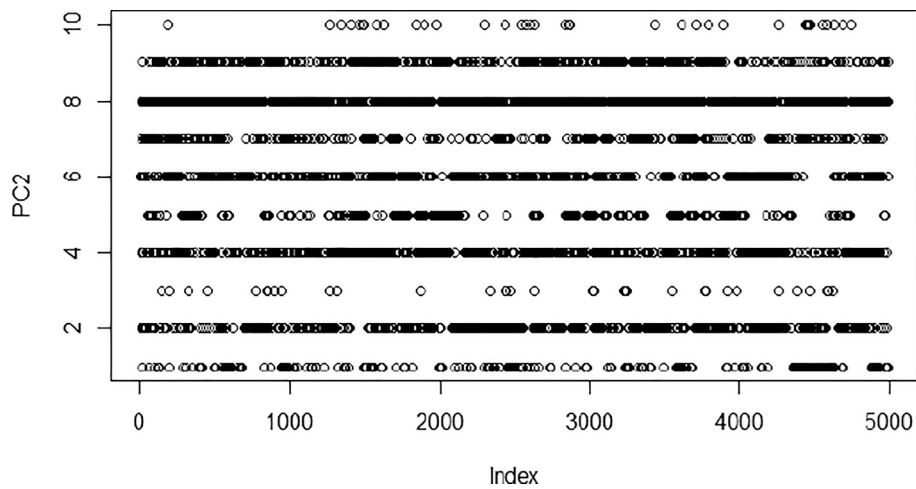


Fig. 5. Dot plot of clusters generated using C-means clustering method.

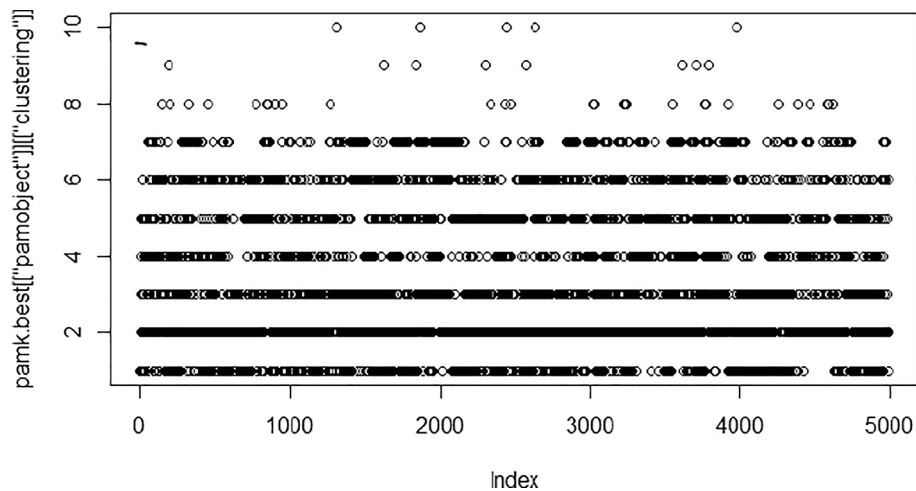


Fig. 6. Dot plot of clusters generated using Pamk clustering method.

introduced themselves as “Industrial workers”. Around 10% indicated they are “Technicians” in clusters 3, and 8. In cluster 9, neither the industrial worker nor technician is the case as the “other” type of occupation contributes 13% in this cluster. Construction is a dominant occupation in clusters 4 to 10. Industrial mechanic jobs are one of the main occupations amongst clusters 3, and 6 to 9. In addition, mining, quarrying, oil, and gas extraction are the other important profession in clusters 2, 4, and 6 to 8.

Comparison of clusters interpretation of results

Clusters integration using different three techniques of cross-tabulation i.e. Bagged Clusters, Cmeans Clusters, and Pamk clusters may show different outcomes in a sample (Hafezi et al., 2019). This was observed in this study as well. Tables 7–9 outline the impacts of the aforementioned techniques on (1) trip purpose, (2) modal split, and (3) destination choice from trip attributes for pivotal clusters 1, 2, and 8. Applying the three clustering techniques resulted in completely different outcomes for Cluster 8 and that is why it has been chosen for inclusion in the analysis. It should be noted that

- The Bagged Clusters and Pamk clustering techniques yielded similar results for Cluster 1.
- The three above-mentioned clustering techniques yielded different results for Cluster 2.

- While the Bagged Clusters technique was not applicable for Cluster 8, applying Cmeans and Pamk Clustering techniques yielded completely different outcomes for Cluster 8.

Applying clustering techniques demonstrate accuracy and efficiency in representing trip purpose. The outcomes for cluster 1 and cluster 2 obtained by CMeans and Pamk, and Bagged clustering techniques are demonstrating the efficiency of those techniques for foreseeing trip purposes. Furthermore, both CMeans and Pamk demonstrate almost identical results for cluster 8.

Interpretation:

- The Bagged Clusters and Pamk clusters techniques yielded almost the same result (with some marginal difference) for Cluster 1 and Cluster 2. The Cmeans Clusters technique also showed a similar outcome for Cluster 2.
- While Bagged Clusters was not applicable for Cluster 8, the Cmeans and Pamk Clusters techniques yielded different results for the aforementioned cluster.

Applying clustering techniques demonstrate accuracy and efficiency in forecasting mode choice. This is the case particularly for cluster 2 as the results of CMeans and Pamk, and Bagged are quite close to each other.

Interpretation:

- The Bagged Clusters and Pamk clusters techniques yielded similar results for Cluster 1 and Cluster 2.
- The Cmeans Clusters and Pamk clusters techniques yielded almost the same result (with some marginal difference) for Cluster 8.

Applying clustering techniques demonstrate accuracy and efficiency in forecasting destination choice. All three clustering techniques have brought about similar effects for clusters 1 and 2 in terms of destination choice. Both CMeans and Pamk demonstrate very close results for cluster 8.

The above-mentioned results are somehow in line with that of Ding and Zhang study in the city of Nanjing in China (Ding and Zhang, 2016). They applied a Multinomial Logit (MNL) model for the travel mode choice analysis. The study demonstrated that individual grouping based on cluster analysis benefits the mode choice estimation and indicates the accuracy rates of 89.8%, 85.6%, and 78.2% for a three-group clustering for 524 respondents, while 65.5% accuracy anticipated without grouping.

Comparison of the study findings with those of other models

The outcomes of the models investigated in this research were compared to similar models indicated in case studies in three different domains;

- The impact of clustering technique on trip purpose
- The impact of clustering technique on mode choice
- The impact of clustering technique on trip destination

In the context of trip purpose, the results are compared with an Australian study by Faroqi and Mesbah (2021) where they used spatial and temporal attributes to infer trip purpose. Our study, and Faroqi and Mesbah (2021) both use clustering methods to identify travel patterns, considering mainly travel purposes. Adding to Faroqi and Mesbah (2021) key finding indicating the incorporation of the trip sequence in choice behavior modeling which was more significant than considering land-use variables, we hereby conclude that trip purpose could also substantially impact travel pattern recognition.

In the context of mode choice, a Chinese study by Ding and Zhang (2016) proposed a forecasting model to estimate the choice of travel mode with individual grouping based on cluster analysis. They used socio-economic information to classify all travelers into three groups. The estimate of the choice of travel modality was presented and compared to survey data. The results showed that the accuracy rates of the estimates for the three groups were 89.8%, 85.6%, and 78.2%, respectively. They are significantly higher than those without clustering, which was 65.5%. This agrees with our findings indicates that individual grouping based on cluster analysis is beneficial in travel pattern analysis and modal choice estimation. Yet, the findings of Ding and Zhang (2016) showed that the statistical cluster analysis method is an efficient mathematical method for dividing individuals into groups.

In this study, the machine learning techniques when used for the individuals' travel behavior, demonstrated a better accuracy of modeling. Nonetheless, both studies showed that the clustering techniques can also potentially be adopted to analyze the travel behavior of residents and refine the necessary strategies.

In the context of destination choice, another Chinese study by Sun et al. (2021) applied the Kmeans clustering method to identify specific segments of tourists and locals' travel behavior in choosing a holiday destination. Their results indicate that the proposed method can identify the most popular viewpoints and the main transport corridors for tourists. The prove of the efficiency of K-means clustering of Sun et al. (2021) comes in compliance with our research finding that Pamk clustering technique is the most significant clustering method.

Contribution to policy implications

This study brings insights for road transport authorities, public transit operators, and city planners by demonstrating the feasibility of forecasting blue collars mode and destination choice with individual grouping based on the revealed preference survey. This research study outcome could also be valuable for decision-makers to decide the priorities of providing public transport services to blue-collar workers as well as devising travel demand management strategies considering blue-collar workers' demographic distribution and activity classifications. The study demonstrated that a transport policy for the blue-collar segment of Qatar needs to be developed by focusing mainly on commuting trips between the home and workplace of blue-collar workers.

Conclusion and way forward

This article presented the applicability of machine learning classifiers models for an activity-based model for blue-collar workers in the city of Doha, Qatar. The proposed model applied clustering techniques to consolidate the activity data of blue collars into a homogenous activity pattern. A revealed preference (RP) survey results obtained from Qatar's Ministry of Transport and Communications (MOTC) were used to extract blue collars' social-economic information and travel information. Three clustering techniques (Bagged Clustering, C-means Clustering, and Pamk Clustering) were applied separately to find out the activity patterns of blue-collar workers particularly in the contexts of trip purpose, mode choice, and destination choice. The findings demonstrate that all three aforementioned clustering techniques are plausible and could result in providing a better understanding of the blue collars' travel pattern. Nonetheless, the C-Means and Pamk Clustering show better efficiency and representation compared to the Bagged Clustering technique.

The study clustered a large number of activities collected from travel diaries into meaningful groups using machine learning techniques to assist in the further development of prediction models for different elements of the transportation system. Travel patterns have been identified that aids in developing effective decision support systems in transportation planning and assessment of different strategies. The study results showed that the cluster analysis techniques are mathematically efficient and can classify individuals such as blue-collars into groups and consequently analyze travel behavior. By applying clustering techniques, detailed traveler characteristics such as income, gender, age, nationality, and occupation are taken into consideration for activity-based models in addition to mode/destination choice.

Travel patterns identified are valuable knowhow and uptake in developing effective decision support systems in transportation planning and assessment of different strategies. The grouping of the blue-collar segment of Qatar into efficient clusters that represents accurately their travel patterns supports the country's transportation system planning. These results can help devise travel demand management strategies considering blue-collar workers' demographic distribution and activities patterns.

Nonetheless, further research would validate the findings by utilizing different samples, demographics associated with empirical methods. This study can be extended by including the trip patterns of other types of people such as students, and night shift personnel who may have different lifestyles and travel patterns. The inclusion of other types of travelers may result in different findings in terms of the appropriateness and efficiency of Bagged Clustering, C-means Clustering, and Pamk Clustering techniques. Applying such pattern recognition models on big and complex activity datasets could assist transport planners to understand the travel needs of segments of the population well and formulating better-informed strategies.

Table 6
Summary of travel characteristics, and socio-demographic attributes for the ten clusters indicated in this research study (Pamk clustering).

Cat-gory	attribute	Cluster									
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Travel Characteristics	Trip purpose	Personal activities (21%), and work and job duties (43%), Preparing meals (31%)	Personal activities (55%), and work and job duties (38%), Preparing meals (5%)	Personal activities (23%), and work and job duties (76%),	Personal activities (45%), and work and job duties (7%), Preparing meals (47%)	Personal activities (11%), and work and job duties (89%),	Personal activities (43%), and work and job duties (57%),	Personal activities (73%), and work and job duties (27%),	Personal activities (18%), and work and job duties (6%), Preparing meals (75%)	Personal activities (61%), and work and job duties (38%), Preparing meals (1%)	Personal activities (53%), and work and job duties (47%),
	Destination choice	Work(40%), Home (48%), Other (9%)	Work(39%), Home (42%), Other (17%)	Work(40%), Home (47%), Other (10%)	Work(43%), Home (42%), Other (13%)	Work(40%), Home (47%), Other (9%)	Work(35%), Home (49%), Other (12%)	Work(36%), Home (51%), Other (9%)	Work(45%), Home (40%), Other (15%)	Work(12%), Home (62%), Other (12%)_Shop (12%)_	Home (60%), Other (40%)
Demographics	Municipality	Al Rayyan, Al Sheehaniya, and Doha									
	Modal split	Private shuttle main mode (40%), walking (16%), and car/van/truck (40%)	Private shuttle main mode (55%), walking (23%), car/van/truck (19%)	Private shuttle main mode (45%), walking (40%), car/van/truck (12%)	Private shuttle main mode (56%), walking (21%), car/van/truck (20%)	Private shuttle main mode (67%), walking (20%), car/van/truck (10%)	Private shuttle main mode (52%), walking (24%), car/van/truck (18%)	Private shuttle main mode (59%), walking (20%), car/van/truck (18%)	Private shuttle main mode (54%), walking (23%), car/van/truck (19%)	Private shuttle main mode (60%), walking (24%), car/van/truck (12%)	Private shuttle main mode (52%), walking (13%), car/van/truck (26%)
	Income (per month)	63% between QR1,001 and QR 3,000. 33% earned less than QR 1,000.	68% between QR1,001 and QR 3,000. 27% earned less than QR 1,000.	63% between QR1,001 and QR 3,000. 33% earned less than QR 1,000.	63% between QR1,001 and QR 3,000. 33% earned less than QR 1,000.	61% between QR1,001 and QR 3,000. 36% earned less than QR 1,000.	66% between QR1,001 and QR 3,000. 30% earned less than QR 1,000.	57% between QR1,001 and QR 3,000. 38% earned less than QR 1,000.	52% between QR1,001 and QR 3,000. 47% earned less than QR 1,000.	50% between QR1,001 and QR 3,000. 37% earned less than QR 1,000.	60% between QR1,001 and QR 3,000. 40% earned less than QR 1,000.
	Gender	100% male									
Nationality	Indian (43%), Bangladesh and Nepal (40%),, Philippines (9%)	Indian (40%), Bangladesh and Nepal (43%),, Philippines (7%)	Indian (43%), Bangladesh and Nepal (40%),, Philippines (7%)	Indian (43%), Bangladesh and Nepal (46%),, Philippines (8%)	Indian (40%), Bangladesh and Nepal (40%),, Philippines (7%)	Indian (37%), Bangladesh and Nepal (46%),, Philippines (5%)	Indian (43%), Bangladesh and Nepal (42%),, Philippines (8%)	Indian (23%), Bangladesh and Nepal (53%),, Philippines (8%)	Indian (63%), Bangladesh and Nepal (52%),, Philippines (12%)	Indian (60%), Bangladesh and Nepal (40%),,	
Age	40% between 25 and 34, 33% between 35 and 44 years old	46% between 25 and 34, 27% between 35 and 44 years old	42% between 25 and 34, 23% between 35 and 44 years old	46% between 25 and 34, 28% between 35 and 44 years old	46% between 25 and 34, 27% between 35 and 44 years old	46% between 25 and 34, 28% between 35 and 44 years old	42% between 25 and 34, 30% between 35 and 44 years old	46% between 25 and 34, 28% between 35 and 44 years old	49% between 25 and 34, 27% between 35 and 44 years old	47% between 25 and 34, 23% between 35 and 44 years old	
Occupation	Elementary occupations (84%), industrial workers (9%), Technicians (3%)	Elementary occupations (79%), industrial workers (11%), Technicians (1%)	Elementary occupations (79%), industrial workers (9%), Technicians (9%)	Elementary occupations (80%), industrial workers (8%), Technicians (2%)	Elementary occupations (78%), industrial workers (4%), Technicians (2%)	Elementary occupations (76%), industrial workers (11%), Technicians (4%)	Elementary occupations (78%), industrial workers (13%), Technicians (3%)	Elementary occupations (76%), industrial workers (10%), Technicians (10%)	Elementary occupations (87%), industrial workers (0%), others (13%)	Elementary occupations (100%), industrial workers (0%),	
		Construction (20%), Industrial Mechanic (15%), Oil & gas(9%), motor mechanic (11%)	Construction (35%), Industrial Mechanic (13%), Oil & gas(10%), motor mechanic (20%)	Construction (26%), Industrial Mechanic (30%), Oil & gas(5%), motor mechanic (11%)	Construction (41%), Industrial Mechanic (10%), Oil & gas(10%), motor mechanic (15%)	Construction (52%), Industrial Mechanic (9%), Oil & gas(4%), Construction Bldg (11%)	Construction (33%), Industrial Mechanic (14%), Oil & gas(12%), motor mechanic (15%)	Construction (36%), Industrial Mechanic (14%), Oil & gas(11%), motor mechanic (10%)	Construction (33%), Industrial Mechanic (15%), Oil & gas(10%), motor mechanic (10%)	Construction (39%), Industrial Mechanic (17%), Oil & gas(7%), motor mechanic (10%)	Construction (40%), Office Private (8%), Oil & gas(5%), motor mechanic (8%)

Table 7

The impact of clustering technique on trip purpose.

Technique	Clusters		
	Cluster 1	Cluster 2	Cluster 8
<i>Bagged Clustering</i>	Personal activities are the dominant activity (56%). Work/job duties is another major activity (42%)	Preparing meals for co-workers is pivotal (60%). Personal activities are another major trip purpose (27%)	N/A
<i>C-means Clustering</i>	Work/job duties is the major activity (42%), followed by preparing meals for co-workers (31%)	Personal activities are the dominant activity (55%). Work/job duties is another major activity (38%).	Preparing meals for co-workers is pivotal (76%), followed by personal activities (18%)
<i>Pamk Clustering</i>	Work/job duties is the major activity (57%), followed by personal activities (43%)	Preparing meals for co-workers is pivotal (67%). Personal activities are another major trip purpose (18%)	Personal activities are pivotal (76%). Personal activities are another major trip purpose (24%)

Table 8

The impact of clustering technique on Modal Split.

Technique	Clusters		
	Cluster 1	Cluster 2	Cluster 8
<i>Bagged Clustering</i>	Private shuttle is the major travel mode (56%). Walking is another important mode (23%)	Private shuttle is the major travel mode (56%). Walking and traveling by car/van/truck both have almost the same share of modal split (22%)	N/A
<i>C-means Clustering</i>	Private shuttle and traveling by car/van/truck both have almost the same share of modal split (40%)	Private shuttle is the major travel mode (55%). Walking is another important mode (23%). Car/Van/Truck partakes 19%	Private shuttle is the major travel mode (55%). Walking is another important mode (23%). Car/ Van/Truck partakes 19%.
<i>Pamk Clustering</i>	Private shuttle is the major travel mode (52%). Walking is another important mode (25%)	Private shuttle is the major travel mode (52%). Walking and traveling by car/van/truck both have almost the same share of modal split (23%)	Private shuttle is the major travel mode (46%). Walking is another major travel mode (40.5%).

Table 9

The impact of clustering technique on destination choice.

Technique	Cluster		
	Cluster 1	Cluster 2	Cluster 8
<i>Bagged Clustering</i>	Travel to home is the major destination choice (47%), followed by traveling to the workplace (40%)	Travel to home is the major destination choice (43%), followed by traveling to the workplace (39%)	N/A
<i>C-means Clustering</i>	Home travel is the major destination choice (41%). Traveling to the workplace, and traveling to "other" destinations are other major destination choices. both have almost similar percentages i.e. 30% and 27% respectively	Travel to home is the major destination choice (47%), followed by traveling to the workplace (40%)	Travel to home and traveling to the workplace both partake similarly in the respondents' destination choice (42%)
<i>Pamk Clustering</i>	Travel to home is the major destination choice (48%), followed by traveling to the workplace (40%)	Travel to home is the major destination choice (41%), followed by traveling to the workplace (39%)	Traveling to the workplace is the major destination choice (45%), followed by traveling to home (40%)

CRedit authorship contribution statement

Aya Hasan Alkhereibi: Conceptualization, materials acquisition, data acquisition, data analysis, development of models and framework, results development, writing original draft. **Shahram Tahmaseby:** Results development, writing original draft, Editing and Formatting, general improvement. **Semira Mohammed:** Data filtering and preparation, writing original draft, Review, Editing and Formatting, general improvement. **Deepthi Muley:** Results discussion, Review, Editing and Formatting, general improvement.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the Ministry of Transport and Communications in the State of Qatar for providing data for analysis. Open Access funding is provided by the Qatar National Library.

References

- Allahviranloo, M., Aissaoui, L., 2019. A comparison of time-use behavior in metropolitan areas using pattern recognition techniques. *Transp. Res. Part A: Policy Pract.* 129 (September), 271–287. <https://doi.org/10.1016/j.tra.2019.08.007>.
- Awais, M.S., Hato, E., 2014. Use of acceleration data for transportation mode prediction. <https://doi.org/10.1007/s11116-014-9541-6>.
- Bar-Gera, H., Boyce, D., 2003. Origin-based algorithms for combined travel forecasting models. *Transp. Res. Part B: Methodol.* 37 (5), 405–422. [https://doi.org/10.1016/S0191-2615\(02\)00020-6](https://doi.org/10.1016/S0191-2615(02)00020-6).
- Bicikova, K., 2014. Understanding student travel behavior: a segmentation analysis of British University Students. *J. Travel Tour. Market.* 31 (7), 854–867. <https://doi.org/10.1080/10548408.2014.890154>.
- Castiglione, J., Bradley, M., Gliebe, J., 2014. Activity-based travel demand models: a primer. <https://doi.org/10.17226/22357>.
- Chen, S., Prakash, A.A., De Azevedo, C.L., Ben-Akiva, M., 2020. Formulation and solution approach for calibrating activity-based travel demand model-system via microsimulation. *Transp. Res. Part C: Emerg. Technol.* 119 (August 2019), 102650. <https://doi.org/10.1016/j.trc.2020.102650>.
- Cui, Y., He, Q., Khani, A., 2018. Travel behavior classification: an approach with social network and deep learning. *Transp. Res. Rec.* 2672 (47), 68–80. <https://doi.org/10.1177/0361198118772723>.
- Delhoum, Y., Belaroussi, R., Dupin, F., Zargayouna, M., 2020. Activity-based demand modeling for a future urban district. *Sustainability (Switzerland)* 12 (14), 5821. <https://doi.org/10.3390/su12145821>.
- Ding, L., Zhang, N., 2016. A travel mode choice model using individual grouping based on cluster analysis. *Procedia Eng.* 137, 786–795. <https://doi.org/10.1016/j.proeng.2016.01.317>.
- Drchal, J., Certický, M., Jakob, M., 2019. Data-driven activity scheduler for agent-based mobility models. *Transp. Res. Part C Emerg. Technol.* 98, 370–390. <https://doi.org/10.1016/j.trc.2018.12.002>.

- Faroqi, H., Mesbah, M., 2021. Inferring trip purpose by clustering sequences of smart card records. *Transp. Res. Part C Emerg. Technol.* 127 (April), 103131 <https://doi.org/10.1016/j.trc.2021.103131>.
- Hafezi, M.H., Daisy, N.S., Millward, H., Liu, L., 2021. Ensemble learning activity scheduler for activity based travel demand models. *Transp. Res. Part C: Emerg. Technol.* 123 (September 2019), 102972. <https://doi.org/10.1016/j.trc.2021.102972>.
- Hafezi, M.H., Liu, L., Millward, H., 2019. A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation* 46 (4), 1369–1394. <https://doi.org/10.1007/s11116-017-9840-9>.
- Hung, M., Yang, D., 2001. An Efficient Fuzzy C-Means Clustering Algorithm Ming-Chuan. *IEEE*.
- Ibrahim, L.F., Hamed, M., Harbi, A., 2012. Using modified partitioning around medoids clustering technique in mobile network planning. *Int. J. Comput. Sci. Issues* 9 (6), 299–308.
- Joh, C.H., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity pattern similarity: a multidimensional sequence alignment method. *Transp. Res. Part B: Methodol.* 36 (5), 385–403. [https://doi.org/10.1016/S0191-2615\(01\)00009-1](https://doi.org/10.1016/S0191-2615(01)00009-1).
- Joubert, J.W., de Waal, A., 2020. Activity-based travel demand generation using Bayesian networks. *Transport. Res. Part C: Emerg. Technol.* 120 (September), 102804 <https://doi.org/10.1016/j.trc.2020.102804>.
- Koushik, A.N.P., Manoj, M., Nezamuddin, N., 2020. Machine learning applications in activity-travel behaviour research: a review. *Transp. Rev.* 40 (3), 288–311. <https://doi.org/10.1080/01441647.2019.1704307>.
- Kubat, M., 2015. An introduction to machine learning. In: *An Introduction to Machine Learning*, January 1992, pp. 1–291. <https://doi.org/10.1007/978-3-319-20010-1>.
- Leisch, F., 1999. Bagged clustering. *Adapt. Inf. Syst. Modell. Econ. Manag. Sci.* 51 (51), 11. 10.1.1.33.642.
- Li, S., 2017. Security and vulnerability in the internet of things. In: *Securing the Internet of Things*. Elsevier, pp. 49–68. <https://doi.org/10.1016/B978-0-12-804458-2.00003-2>.
- Nanopoulos, A., 2014. C2P: Clustering based on Closest Pairs C 2 P: Clustering based on Closest Pairs. January 2001.
- Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning, pp. 625–632.
- Planning and Statistics Authority, 2020. Qatar Economic Outlook 2018-2020. 11.
- Sun, H., Chen, Y., Lai, J., Wang, Y., Liu, X., 2021. Identifying tourists and locals by K-means clustering method from mobile phone signaling data. *J. Transp. Eng., Part A: Syst.* 147 (10), 04021070. <https://doi.org/10.1061/jtepbs.0000580>.
- Sun, X.D., Huang, R.B., 2006. Prediction of protein structural classes using support vector machines. *Amino Acids* 30 (4), 469–475. <https://doi.org/10.1007/s00726-005-0239-0>.
- You, J., Wang, J., Guo, J., 2017. Real-time crash prediction on freeways using data mining and emerging techniques. *J. Modern Transp.* 25 (2), 116–123. <https://doi.org/10.1007/s40534-017-0129-7>.