CLASSIFYING BEHAVIORAL DYNAMICS OF TAXI DRIVERS ROUTE CHOICES USING LONGITUDINAL GPS DATA

Hamzeh Alizadeh
Department of Civil, Geological and Mining Engineering, Polytechnique Montréal,
C.P. 6079, succ. Centre-Ville Montréal (Québec) Canada, H3C 3A7
Tel: (514) 340-4711 ext. 4235 Fax: 5143404191 E-mail: hamzeh.alizadeh-aliabadi@polymtl.ca

Bilal Farooq
Head of Laboratory of Innovations in Transportation (LITrans),
Department of Civil, Geological and Mining Engineering, Polytechnique Montréal,
C.P. 6079, succ. Centre-Ville Montréal (Québec) Canada, H3C 3A7
Tel: (514) 340-4711 ext. 4802, Fax: (514) 340-3981, E-mail: bilal.farooq@polymtl.ca

Catherine Morency
Head of Mobilité Research Chair
Department of Civil, Geological and Mining Engineering, Polytechnique Montréal,
C.P. 6079, succ. Centre-Ville Montréal (Québec) Canada, H3C 3A7
Tel: (514) 340-4711 ext. 4502 Fax: (514) 340-3981 E-mail: cmorency@polymtl.ca

Nicolas Saunier
Department of Civil, Geological and Mining Engineering, Polytechnique Montréal,
C.P. 6079, succ. Centre-Ville Montréal (Québec) Canada, H3C 3A7
Tel: +1 (514) 340-4711 ext. 4962, Fax: (514) 340-3981, E-mail: nicolas.saunier@polymtl.ca

Word count: 6,464 words text + 6 tables/figures x 250 words (each) = 7,946 words

11-November-2016

ABSTRACT

This study aims to capture the behavioral heterogeneity in route choice by identifying subgroups of drivers based on their actual route choices and factors affecting them. We have studied a highly longitudinal GPS dataset, tracking 1,746 taxi drivers over a period of one year, making more than 22,000 trips between the Islands of Montreal and Laval. We opted for a two-step procedure, where in the first step a Principal Component Analysis (PCA) is performed to reduce collinearity among attributes, followed by a Hierarchical Agglomerative Clustering (HAC) to form behavioral clusters in the second step. Results show that four major types of route choice behaviors are observable among taxi drivers. These clusters show significant variations based on the time of day (day/night) and the traveled distance (shorter trips/longer trips) and are labelled: *"Short trips night drivers"*, *"Long trips night drivers"*, *"Short trips day drivers"*, and *"Long trips day drivers"*. Due to the rise of ride-hailing services, the understanding of these patterns are important for city and transportation planners in the context of proposing new laws and policies that safeguard taxi industry as well as encourage sharing economy. The inclusion of similar typologies in route choice models would improve their behavioral aspect as well as their estimation and prediction abilities.

Keywords: Route choice behavior, Principal Component Analysis (PCA), Hierarchical Agglomerative Clustering (HAC), Longitudinal GPS data, Taxi

INTRODUCTION

Individuals have different inclinations toward choosing a route between same Origin-Destination (OD) pairs. These choices form the collective patterns of route choice behaviors, observable in our everyday lives. Understanding these patterns is of crucial importance to the city and transportation planners (*1*). Route choice models are used to capture the complex process of drivers' route selection behavior. These models are used to estimate and predict the probability of a certain route being chosen between a given OD pair. The complexity of this process is mostly due to factors such as the sophisticated nature of human behavior; ambiguity of decision making process; stochasticity of individuals' preferences, as well as the high density of road networks; large number of possible alternatives between OD pairs; and correlation among these alternatives.

Route choices are highly dependent on individuals' characteristics. Therefore, in order to improve the behavioral aspect of route choice models, it is essential to incorporate various sources of heterogeneity influencing drivers' route choice decisions. This heterogeneity mostly comes from having different preferences, experiences, information levels, and attitudes. These sources of heterogeneity are somehow correlated, which makes the understanding of route choice behavior even more complex. Therefore, it might be interesting and necessary to ask "*How factors affecting route choices are correlated?"*, and *"How these factors can be grouped into various categories?"*. Another relevant question could be "*How do these different groups represent different categories of drivers?"*.

Unfortunately, a detailed representative classification of drivers according to their route choice decisions is missing from the existing literature. According to (*2*), dividing drivers into different categories based on their route choice behavior can improve route choice models and can be used in defining different functional forms for traffic assignment models. (*3*) classified drivers according to their age, gender, ethnicity, education, driving experience, and annual driven miles, and found meaningful differences in their route choice behaviors. They have argued that the incorporation of different groups of drivers can improve the accuracy of route choice models. (*4*) illustrated that the incorporation of an activity-based segmentation can improve traffic assignment procedures and proposed stratifying traffic assignment models by incorporating different typologies of attitudes toward route choice.

A further challenge in route choice modeling is the selection of a proper data collection method. Although travel surveys are one of the most adopted methods (*5-8*), it is not easy to capture route choice behavior via diary surveys. With the introduction of data collection through Global Positioning Systems (GPS) in transportation studies, researchers are supplied with unprecedented high-resolution spatial and temporal records on drivers' route choices.

Earlier route choice studies have mostly focused on traditional cross-sectional analysis of drivers' behavior. Therefore, the decision making behavior for a single day has usually been interpreted to derive important attributes affecting drivers' decision making patterns. Studies focusing on driving patterns over an extended period appear to be minimal, which might be mostly due to the challenges of data collection and analysis. In recent years, with the dissemination of GPS technologies and the progressive utilization of in-vehicle or on-person GPS devices, researchers are supplied with longitudinal choice data over an extended period, which provides better insights on drivers' route choice patterns and heterogeneities than cross sectional data.

In this study we aim to extend the understanding of drivers' route choice behaviors by classifying drivers into separate categories, based on their observed route choice decisions, using a longitudinal GPS dataset. Although previous studies have shown that some attributes such as travel time, travel speed, some demographic and socioeconomic attributes, etc., influence drivers' route choices, to the best knowledge of the authors, none of these studies have ever categorized common attributes resulting in the same decision, to identify various classes of route choice behavior. This research relies on a longitudinal GPS dataset, tracking 1,746 taxi drivers making more than 22,000 trips over a period of one year. We focus on trips between the Islands of Montreal and Laval, originating in Montreal.

This work distinguishes itself from the existing literature based on two original contributions. First, to the extent of our knowledge, drivers' route choice behaviors have never been stratified using drivers' actual route choices. Second, route choice preferences have never been looked at through longitudinal GPS data over a one-year period. Previous datasets were not large and/or detailed enough to allow this type of stratification.

The rest of this paper is organized as follows. First, we overview factors affecting drivers' route choice decisions. Second, we present the studied regional context and describe our dataset. Then, we discuss the process of deriving indicators and our clustering methodology. Finally, we present descriptive analysis of studied trips, discuss the clustering results and underline the findings of this study.

STATE-OF-THE-ART

The aim of this study is to classify drivers' route choice behaviors based on their observed choices over a long duration of time. In a perfect world, drivers make route choices by minimizing a certain perceived cost. This cost function is unique for each driver and situation, resulting in different chosen routes, and can be attributed to several factors, such as different levels of information, different capacities to process them, and different computational and prediction abilities (*9*). A large body of literature exists on modelling route choice and incorporating factors affecting them. These factors are correlated and form different types of preferences, which in turn can be used to classify drivers into different categories.

Classifications of driving behaviors have been mostly cited in traffic safety studies to distinguish between conflicts and collisions, risk-taking versus safe drivers and experienced versus novice drivers (*10-13*), as well as car-following research classifying drivers according to their driving behavior (*14*). A classification of driving patterns has been proposed by (*15*), which divides drivers into three categories, namely "*passionate*", "*every day*", and *"leisure time"* drivers, based on their attitudes toward the environment. Behavioral classification studies have also been conducted on bike users. (*16*) studied the causality relation between non-work-related trips and commuters, and classified cyclist into four different clusters, namely *Non-cyclists*, *Non-work cyclists*, *All-around cyclists*, and *Commuter cyclists*. Another major work on cyclists classification is the work by (*17*), in which bike users are classified into four categories: *Strong and Fearless*, *Enthused and Confident*, *Interested but concerned*, and *No Way No How*. The reader is referred to (*18*) for more reflections on this work. Another recent work by (*19*) classified bike-sharing users into two groups of *Commuters* and *Recreational users*. Drivers' perceptions and experiences have been studied by (*20*), using a driving simulator and two initial and final questionnaires. They found observable differences between drivers' route choice behaviors and categorized drivers based on their learning skills into four different categories. In a further study, (*21*) have demonstrated that the inclusion of those learning clusters would improve disaggregated route choice models. In an interesting study by (*22*), cab drivers have been classified into *Top Drivers* and *Ordinary Drivers* based on their incomes, and their trip characteristics have been studied.

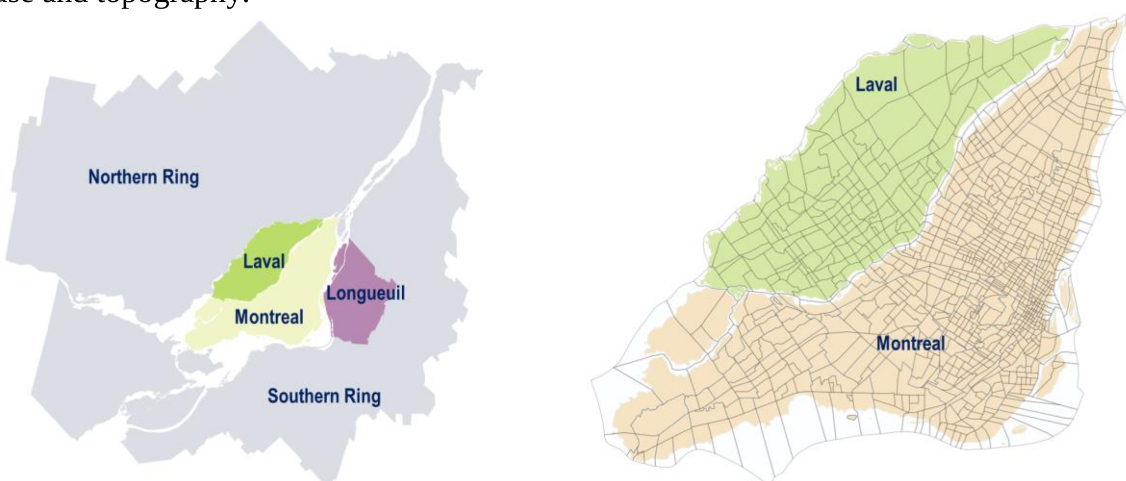
Although previous studies showed that different categories of road users are observable, an explicit classification of drivers based on their actual choices is missing from the literature. Due to the availability of very large and rich datasets from mobility services (e.g. taxi companies), that have detailed GPS trajectories of their fleets for a long duration of time, it has become possible to address this critical research need.

CONTEXT AND DATA DESCRIPTION

Regional Context

This study is performed in the context of the metropolitan region of Greater Montreal, depicted in Figure 1(a). This Island city is separated from its suburbs by two large rivers; Prairies River and Saint Lawrence River in the north and south of the city, respectively. In this study we focus on trips originating in Montreal with a destination in Laval, the largest suburb of Montreal located north of the city, across the Prairies River. These two islands contain roughly 2.3 million inhabitants, and cover a total surface of 632.3 km^2 (*23*).

Traffic Analysis Zones (TAZ), delineated by Quebec's Ministry of Transport, are used in this study to determine the characteristics of taxi trips' origin and destination points (Figure 1(b)). TAZs are geographical areas, which divide the city into smaller similar areas based on various factors such as population, demography, socioeconomic information, road network, transit access, land use and topography.



(a) Greater Montreal **(b) TAZ of Montreal and Laval**

FIGURE 1 Regional context of the study.

GPS Dataset

We investigated GPS traces collected over a period of one year (2015) by a major Taxi company operating in Montreal. This taxi company constitutes around 25% of Montreal's taxi fleet, and its operation is restricted to trips starting or ending in the central part of the island. Every taxi is equipped by a data logger and GPS data are collected continuously for operational purposes. For our study we extracted a subset of the main dataset, consisting of around 750,000 GPS records collected from 1,746 taxi drivers making a total of 22,394 trips. Drivers are associated with a unique ID so that we can distinguish between trips made by a same car and different drivers. We should mention that personal information on drivers' demographic and socioeconomic characteristics are not available due to privacy issues.

In order to explore factors affecting drivers' route choices, we first need to derive the observed route from the recorded GPS points. Therefore, every record has to be associated with a link in the network, a process which is called map-matching. This step is crucial, since it determines the accuracy of reconstructed trajectories, and accordingly their derived attributes.

Data has been stored in a PostgreSQL database, and the PostGIS spatial extension has been added to support geographical datatypes and queries. To associate GPS records to the road network, a direction-based nearest link point-to-curve map matching algorithm has been adopted, in which every record is matched onto the closest link in the network with respect to its azimuth. To reconstruct the complete trajectory, consecutive GPS points have been connected using a distance-based shortest-path algorithm.

Network Dataset

The road network has been extracted from the OpenStreetMap project in the format of geographical layers ("shapefiles"). It contains more than 156,000 nodes and 89,000 links. The network has been made routable, through a geospatial extension of PostgreSQL named "pgrouting", in order to enable the calculation of shortest paths. It has also been segmented on an intersection-to-intersection basis; therefore, links are defined to be road segments between two consecutive intersections.

METHODS

Process of Deriving Indicators

The first step toward classifying drivers' route choice behaviors is to explore a large range of trips' (choices) characteristics and factors affecting these choices. These factors include temporal and environmental attributes, drivers' attitudes and preferences, network familiarity, personal demographic and socio-economic characteristics, and route related attributes among others. Since personal characteristics are not observable through GPS traces, the main focus of this section is to derive explanatory factors from the GPS dataset to thoroughly describe observed trajectories. Factors are classified into five broad categories, namely *Temporal Indicators*, *Route Specifications*, *Driver Characteristics*, *Land Use*, and *Route Similarities*. A concise description of derived attributes follows.

Temporal Indicators

The timestamp information of the origin point of a trip, containing information regarding the date and time of the record, has been used to derive the following indicators:

- <u>Peak/Off-peak:</u> Trips starting between 6:00 to 9:00, and 16:00 to 19:00 are considered as peak hour trips, while all other trips are regarded as off-peak trips.
- <u>Weekday/Weekends</u>
- <u>Day of the Week</u>
- <u>Day/Night:</u> Day trips consist of trips starting between 6:00 to 21:00 and the rest are considered as night trips.
- <u>Season</u>

Route specifications

These attributes characterize observed routes between given OD pairs, based on their physical specifications and their similarity levels with their respective distance-based shortest paths.

- Length: The total traveled distance.
- Travel time: The timestamp information at origins and destinations have been used to extract the observed travel time.
- % Highway: Specifies the percentage length of trips made on highways.
- # Turns / km: The total number of turns per kilometer.
- # Segments / km: The total number of road segments per kilometer.
- % Shortest path: The percentage of overlap with the distance-based shortest path.

Driver characteristics

The total number of trips per driver and trips made by the same driver between the same TAZ pairs have been inspected and several attributes, pertinent to the regularity of drivers between these TAZ pairs, have been derived:

- Total trips: The total number of trips made by each driver.
- Average number of trip per TAZ pair: The total number of trips is divided by the total number of TAZ pairs, between which a driver has traveled. The lower bound of this attribute is one, indicating only one travel per TAZ for a driver. Higher values signify higher propensity of traveling between the same TAZ pairs.
- TAZ pairs with more than 4 trips: The total number of TAZ pairs between which the driver has made more than four trips, which has been selected based on the distribution of trips between same TAZ pairs, as an indicator of regularity of drivers.
- Maximum same TAZ trips: The maximum number of trips made between a given TAZ pair by the same driver. This factor is also interpreted as an indicator of regular trips between same TAZ pairs.

Land use

Although GPS data allows to locate the exact pick up and drop off position of travellers, we have no information regarding their exact trip purposes. In this work we associated GPS points to their respective TAZ to obtain information regarding the land use of trips' origin and destination points. Land use data was prepared by local government agencies. Four categories of land uses are considered in this study, namely residential, commercial, work/study, and recreational. Accordingly, the following attributes were derived:

- Origin land use
- Destination land use

Route Similarities:

To evaluate the degree of similarity for trips taken by a particular driver between a given pair of TAZ, we have used a measure called path-size (PS). This measure has been proposed by (*24*) to account for similarities between routes in logit based discrete choice models. The following formulation has been adopted in this study:

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \epsilon \varphi_n} \delta_{aj}} \tag{1}$$

where PS_{in} denotes the path-size factor for driver n and route i, L_a and L_i represent the length of link a and route i, Γ_i is the set of road segments in route i, φ_n denotes the observed routes for driver n between the same pair of TAZ, $\sum_{j\epsilon\varphi_n}\delta_{aj}$ indicates the total number of alternatives in φ_n sharing link a (δ_{aj} is the link-path incident binary variable which is 1 if link a is on route i, and 0 otherwise). The upper bound value of this formulation is 1, indicating that observed routes are completely independent and do not share any links. However, smaller values of PS indicate longer overlaps and higher dependencies between trips. The following indicator has been derived:

- PS**:** The average PS factor calculated for all trips made by the same driver, between all TAZ pairs.

Cluster Analysis

Clustering is an unsupervised categorization technique that aims to segregate multivariate data sets into more meaningful clusters, according to their main describing attributes. In transportation studies datasets usually have very high dimensions. Multivariate analysis techniques and dimension reduction techniques, such as Principal Component Analysis (PCA), are adopted to better interpret the data and to improve the quality of cluster analysis. PCA is an unsupervised dimension reduction technique, which preserve most of the initial information within a smaller numbers of mutually uncorrelated factors.

We adopt a two-step procedure in order to classify driver behaviors. First, in order to reduce collinearity among attributes, a PCA for mixed data is performed (*25*). The number of principal components extracted from an analysis depends on the level of correlation between attributes and the amount of variance that can be explained by each principal component. The total number of components is equal to the number of attributes, but only the first few ones are important and are considered in the next step (*26*). Then, a Hierarchical Agglomerative Clustering (HAC) using Ward's criterion is used to classify driver behaviors. In HAC method, clusters are formed hierarchically by merging the two closest clusters at each step. All PCA and clustering analysis are performed in the TANAGRA statistical package (*27*).

To assess the significance level of attributes in the clustering result, the "Test Value" (TV) criterion has been used. Following formulas are used to calculate TV for continuous (t_c), and discrete (t_d) values for each cluster (*28*):

$$t_c = \frac{\mu_g - \mu}{\sqrt{\frac{n-n_g}{n-1} \times \frac{\sigma^2}{n_g})}} \tag{2}$$

$$t_d = \frac{n_{jg} - \frac{n_g \times n_j}{n}}{\sqrt{\frac{n-n_g}{n-1} \times \left(1 - \frac{n_j}{n}\right) \times \frac{n_g \times n_j}{n}}} \tag{3}$$

where μ and μ_g are attributes' means in the cluster and group, respectively; n and n_g denote the size of the cluster and the group, respectively; σ^2 represents the attribute variance in the cluster; and n_{jg} is the number of observations corresponding to the discrete attribute j in cluster g.

RESULTS

Trip Characteristics

The spatial distribution of origin and destination points is used to visualize the dispersion of taxi demand for trips between Montreal and Laval. As depicted in the heat map of Figure 2, taxi trips mostly originate from downtown Montreal, the airport, major commercial centers, and around train stations. Destination regions with high travel densities are more dispersed in Laval, which is probably due to the higher dispersion of commercial centers and overall lower density. Also, considering that Laval is a suburb of Montreal, the higher segregation between its dense residential areas contributes to the dispersion of high density destination points.

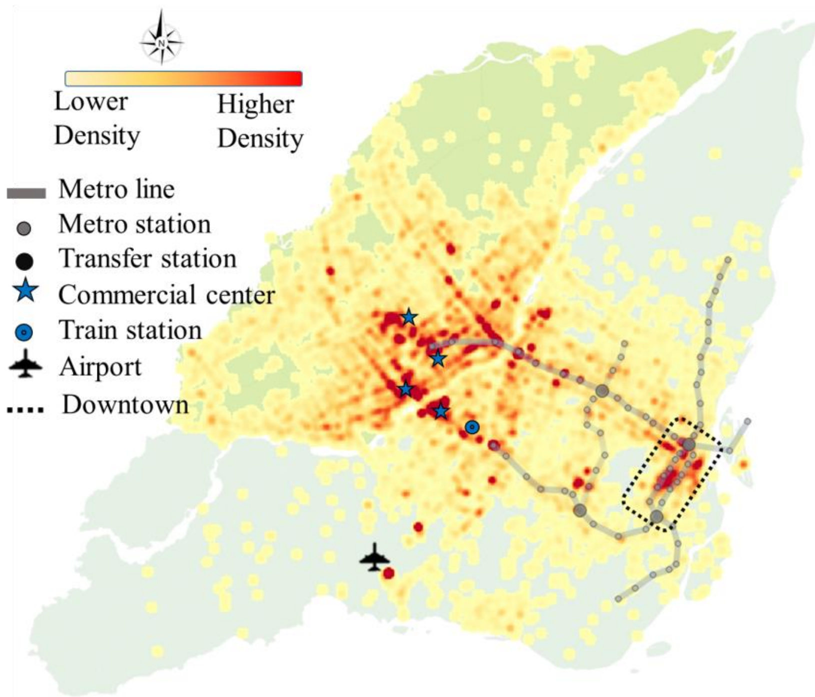
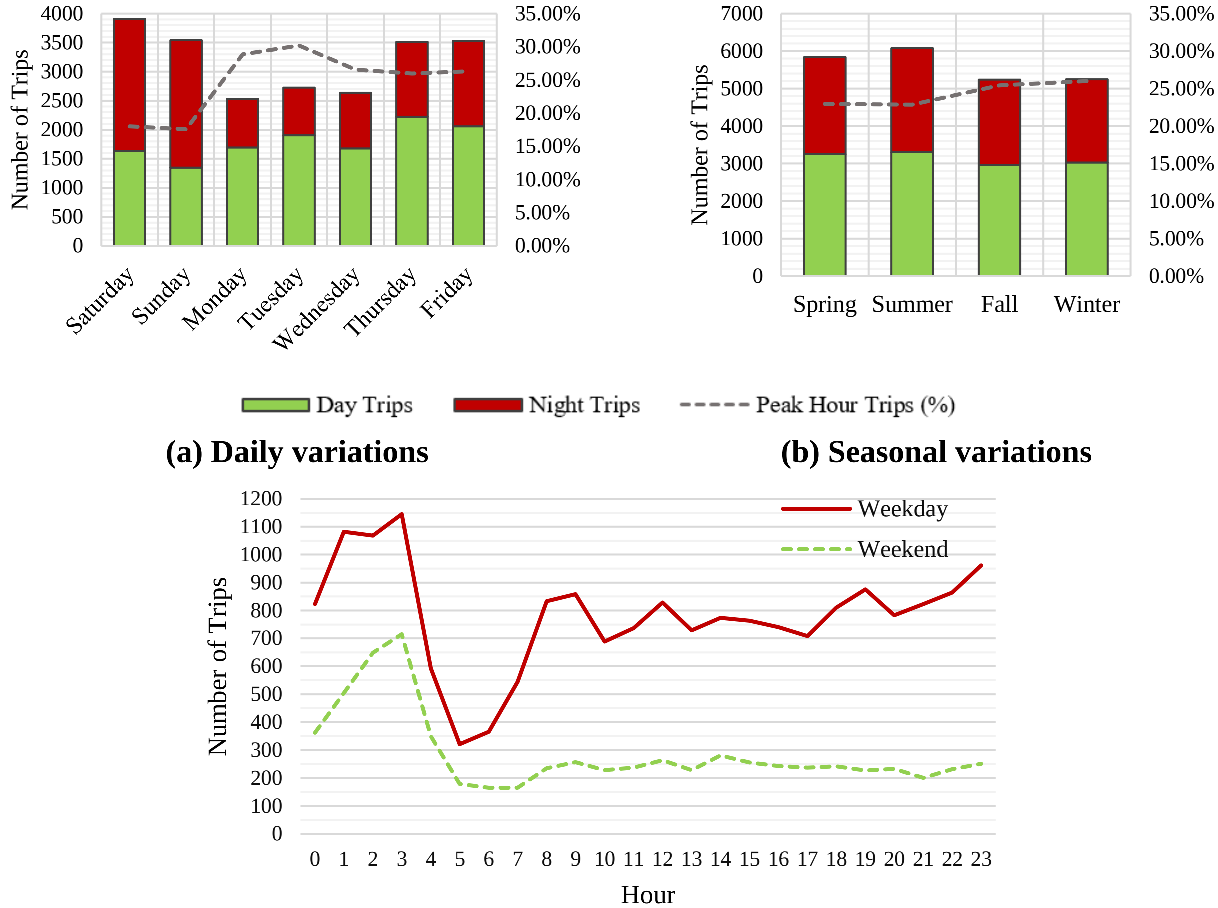


FIGURE 2 Heat map of origin-destination points for trips between Montreal and Laval.

A detailed comparison of trip frequencies across different days, the proportion of day trips versus night trips, and the percentage of trips made in peak hours are presented in Figure 3(a). A quick look reveals that taxi trips are more frequent during weekends, night trips are more common on Fridays and weekends, and peak hour trips are more frequent during weekdays. The same comparison is illustrated for trips made in different seasons (Figure 3(b)). Slightly higher numbers of trips are observed during spring and summer. However, the overall percentage of peak hour trips remains around 24 % throughout the whole year. Overall, around 56 % of trips are day trips and the remaining 44 % are made during nights. A closer look at the hourly distribution reveals that the peak demand occurs around 3 AM, drops significantly early in the morning, augments and stays steady over the day, and starts to increase again by the end of the day, around 9 PM (Figure 3(c)). The same demand pattern holds for both weekdays and weekends, although the overall weekday demand is much higher.



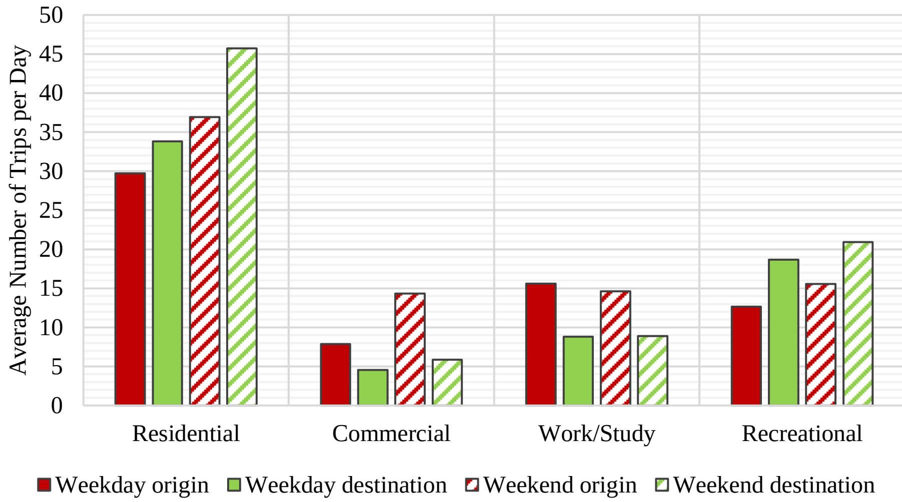
(a) Daily variations **(b) Seasonal variations**

(c) Hourly variations

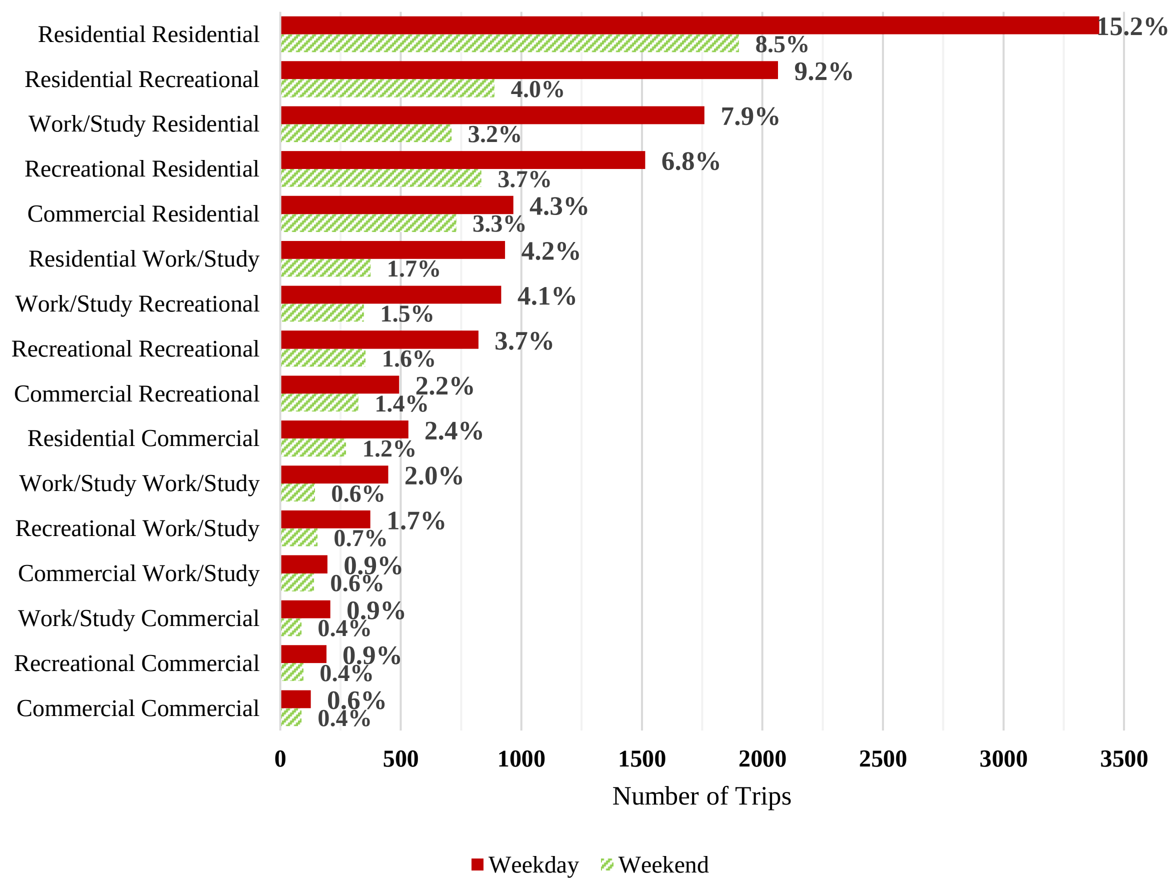
FIGURE 3 Daily, seasonal and hourly variations of taxi trips.

The total traveled distance is around 321,240 km. Around 20 % of taxi trips are shorter than 5 km; another 20 % have a length of 5 to 10 km; 35 % consist of trips from 10 to 20 km long; and the remaining 25 % are longer than 20 km. The mean, median and standard deviation of trip lengths are 14.3, 11.8, and 10.2 km, respectively. Although these values do not show any significant differences between trips made on weekdays compared to weekends, night trips and off-peak trips are slightly longer than day trips and peak hour trips, respectively. Another notable trend is that highway usage is considerably higher for night trips and off-peak trips, which is expected given lower congestion during these periods.

Figure 4(a) illustrates the average number of trips per day, classified based on the land use specification of the TAZ where the passenger was picked-up and dropped-off. Most of the trips originate from and/or have a destination in residential areas. It also highlights the high number of residential-end taxi trips for trips toward Laval. Trips are grouped based on OD land uses and the percentage of trips in each group is shown in Figure 4(b). Number of trips are represented by the length of the bar and labels represent the total share in percentages. A great majority of trips (around 46 %) were residential-based, meaning that trips start from a residential location, while around 20 % start from work/study regions.



(a) Average number of trips per weekday/weekend per land use



(b) Number and percentage of trips based on OD land uses

FIGURE 4 Statistics on trips' land uses.

Driver Characteristics:

The dataset used in this study includes a subset of 1,746 drivers comprising 204 night drivers and 498 day drivers, who only operate during nights and days, respectively. A total of 68 drivers only operate during off peak hours, while 551 drivers work exclusively during peak hours. Around 38 % of peak hour trips take place during morning peak hours and the remaining 62 % occur during evening peak hours. An average of 13 trips and a median of 7 trips per year (between the islands of Montreal and Laval) are recorded for each driver, where the maximum number is 286.

Taking a closer look at trips between same TAZ pairs reveals that certain drivers are more frequent among certain OD pairs, while others have a higher diversity of traveled TAZ pairs. More than 77 % of drivers make a single trip, and only around 5 % of them make four or more trips between a given TAZ pair. Among these more frequent drivers, 52 drivers make more than four trips between more than one pair of TAZ. The maximum number of trips between the same pair of TAZ by the same driver is found to be 40 trips.

Cluster Characteristics

A major challenge in clustering analysis is the selection of a good number of clusters. Although visualization may be an effective way to verify results, it is highly difficult to visualize data with more than three dimensions. In this study, a series of two to eight clusters have been experimented and the optimal number of clusters has been defined by the maximum value of Between-group Sum of Squares (BSS ratio) and GAP-statistic (*29*), as well as the behavioral interpretation that can be associated with the clusters. Furthermore, the dendrogram representing the HAC process, has been inspected to verify the plausibility of results. From this exercise a set of four clusters has been found to provide the best results. These clusters are based on the first two principal components found in the first step (PCA analysis), which explained more than 60 % of the variation between attributes. Clustering has also been performed on more principal components, comprising more than 70 % and 80 % of variations; however, results were not found to be stable, meaningful, and reliable.

Two extra steps were undertaken to measure the results' stability: 1) the dataset was randomly divided in two halves and the clustering process was separately performed on each half for the same parameter settings; 2) observations were randomly permuted in our dataset. Since results were not significantly different in both cases, it was concluded that the four-cluster solution has a high degree of stability and reliability (*30*).

Results showed significant variations of drivers' behavior toward shorter versus longer routes, and routes taken during day versus night. However, the exact hour of the trip, variation of trip days, seasonal variations, and different OD land uses showed no significant impacts on drivers' route choice behaviors in these clusters. Some descriptive statistics of factors, based on which the final clustering has been made, are illustrated in Table 1(a). Correlations between these attributes are presented in Table 1(b). Clusters are presented in Table 1(c) and characterized through their significant attributes. To assess the level of significance, the Test Value criterion is evaluated.

First, two types of drivers are distinguished, representing drivers working during nights and during days. Each of these types are further divided into two separate clusters based on trip lengths. The resulting four clusters are labelled: *Short trips night drivers*, *Long trips night drivers*, *Short trips day drivers*, and *Long trips day drivers*, and represent 20.4 %, 34.9 %, 29.7 %, and 15 % our sample population, respectively.

TABLE 1 Descriptive Analysis, Correlations Matrix, and Clusters' Specifications.

Attribute	Mean (SD)	Median	SD[a]	Min.	Max.
Day	57.3	60.0	36.5	0.0	100.0
Weekday	64.4	66.7	27.9	0.0	100.0
Peak	24.3	20.0	25.5	0.0	100.0
Length (km)	18.5	18.3	7.3	1.2	65.6
Travel time (min)	21.4	20.9	7.3	1.2	62.4
% Highway	54.4	56.7	20.5	0.0	100.0
Path-Size	0.97	1.0	0.1	0.4	1.0
# Links/km	5.69	5.6	1.3	2.5	12.3
# Turns/km	0.64	0.6	0.3	0.0	2.4
% Shortest path	43.9	43.1	18.2	0.0	99.4
# TAZ	12.7	8.0	15.3	1.0	154.0

[a] Standard Deviation

(a) Descriptive analysis of significant factors

	Attribute	1	2	3	4	5	6	7	8	9	10	11
1	Day	-										
2	Weekday	0.31	-									
3	Peak	0.61	0.17	-								
4	Length	-0.11	-0.03	-0.07	-							
5	Travel time	0.08	0.06	0.08	0.69	-						
6	% Highway	-0.08	-0.01	-0.07	0.65	0.19	-					
7	Path-Size	-0.06	-0.09	-0.05	0.20	0.25	0.13	-				
8	# Links/km	0.02	-0.05	0.05	-0.58	-0.14	-0.92	-0.08	-			
9	# Turns/km	0.18	0.07	0.15	-0.35	-0.12	-0.61	-0.14	0.66	-		
10	% Shortest path	0.15	0.08	0.08	-0.39	-0.38	-0.17	-0.12	0.10	-0.05	-	
11	# TAZ	-0.03	0.06	-0.01	-0.35	-0.37	-0.20	-0.59	0.15	0.22	0.08	-

(b) Correlation matrix of significant factors

	Short trip night drivers		Long trip night drivers		Short trip day drivers		Long trip day drivers		All
	TV[a]	Mean (SD[b])	TV	Mean (SD)	TV	Mean (SD)	TV	Mean (SD)	Mean (SD)
Day	-16.7	28.5 (28.2)	-20.0	33.5 (27.7)	22.4	87.5 (16.2)	16.9	92.6 (13.6)	57.3 (36.5)
Weekday	-10.5	50.5 (28.1)	-10.0	55.3 (25.9)	10.3	75.0 (22.9)	12.2	83.7 (21.5)	64.4 (27.9)
Peak	-13.4	8.2 (10.6)	-16.9	10.3 (12.0)	17.5	40.8 (24.4)	15.2	46.4 (24.3)	24.3 (25.5)
Length (km)	-15.1	13.3 (4.6)	18.2	22.8 (5.5)	-17.4	13.8 (5.1)	15.2	24.9 (6.5)	18.6 (7.3)
Travel time	-11.3	17.5 (6.3)	5.4	22.7 (5.3)	-5.9	19.8 (7.2)	13.1	26.8 (8.4)	21.4 (7.3)
% Highway	-17.4	37.5 (15.8)	19.8	67.6 (11.4)	-17.5	41.1 (16.9)	15.7	72.6 (13.6)	54.4 (20.5)
Path-Size	-4.0	0.9 (0.1)	6.6	1.0 (0.5)	-3.8	0.9 (0.1)	6.2	1.0 (0.1)	1.0 (0.1)
# Links/km	17.6	6.7 (1.1)	-17.8	4.9 (0.8)	15.5	6.5 (1.2)	-16.0	4.5 (0.7)	5.7 (1.3)
# Turns/km	7.9	0.8 (0.3)	-15.1	0.5 (0.2)	15.8	0.8 (0.3)	-9.0	0.5 (0.2)	0.7 (0.3)
%SP[c]	4.5	47.8 (18.4)	-9.0	38.6 (16.3)	6.9	48.6 (17.4)	-1.8	42.0 (19.9)	44.0 (18.2)
# TAZ	6.8	17.6 (22.5)	-5.5	10.0 (7.4)	4.6	15.3 (17.3)	-6.2	7.3 (8.5)	12.7 (15.3)
# Obs. (%)	357 (20.4 %)		610 (34.9 %)		517 (29.7 %)		262 (15 %)		1746 (100)

[a] Test Value [b] Standard Deviation [c] Shortest Path

(c) Clusters' specifications

In general, two types of behavior are observed across these clusters. Drivers who most frequently take shorter trips, usually do not use highways and prefer to take local routes with higher number of intersections and turns, which accentuate their familiarity with the road network and their awareness of traffic conditions. These drivers tend to choose similar routes between same OD pairs, which have higher proportions of identical segments with the distance-based shortest path. On the other hand, longer trip drivers showed higher propensity toward using highways to avoid intersections and turns, which is intuitive. These drivers revealed more willingness in taking more diverse routes between the same OD pairs, which are less similar to the distance-based shortest path than routes taken in shorter trips. This might be partly due to the higher number of feasible alternatives between distanced OD pairs compared to closer ones. In order to visualize these clusters, the Z-score (cluster mean minus population mean, divided by population standard deviation), has been calculated (see Figure 5).

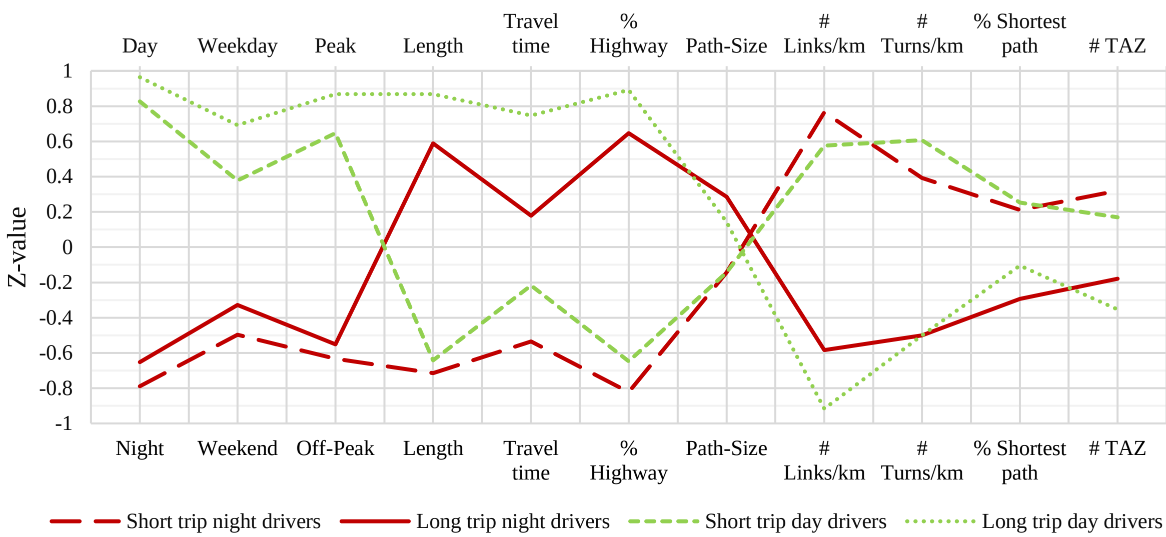


FIGURE 5 Profile plot of the four clusters.

CONCLUSIONS

Understanding heterogeneous route choice behaviors is very important for city and transportation planners. Although previous studies showed that different categories of road users are observable, there is a lack of representative classifications of drivers based on their actual route choices over a long duration of time. New possibilities have opened up due to the availability of very large and rich datasets from mobility services (e.g. taxi companies) that have detailed GPS trajectories of their fleets for a long duration of time.

The main objective of this research is to improve the understanding of drivers' route choice behaviors, by classifying their behaviors based on their observed route choices, using these longitudinal datasets. A GPS dataset comprising more than 22,000 trips, made by 1,746 taxi drivers over a period of one year, for trips originating in Montreal with a destination in Laval, has been studied. It is worth mentioning that the destination choice is not the focus of this research and this specific regional context has been chosen to ensure a wide range of traveled distances.

We first present important statistical properties of taxi trips. Since various degrees of correlation were observed among attributes, a PCA analysis has been performed to improve the efficiency of the clustering algorithm and to reduce the probability of the algorithm getting stuck in a local optima (*26*). Then, a HAC algorithm has been performed to classify drivers' behaviors.

Due to the significant behavioral variations found in trips made during days and nights, and between short trips and long trips, these four clusters were labelled *"Short trips night drivers"*, *"Long trips night drivers"*, *"Short trips day drivers"*, and *"Long trips day drivers"*. Intuitively, drivers prefer highways for longer trips and take local routes with higher number of turns and intersections for shorter trips.

Although it is not possible to encompass all variations of route choice behaviors based on GPS traces alone (due to the lack of some explanatory variables, such as demographics and preferences), this study shed some light on the variation of route choice behaviors and the possibility of classifying them using clustering algorithms.

Findings of this research pave the route for several future directions. A major limitation of this work was the lack of personal information such as demographic and socio-economic variables. Similar longitudinal datasets including these variables can be used to enrich the findings with more personal information. Population synthesis can also be used as a complementary tool to compensate for the lack of these information (*31*).

In this study, we limited our sampled population to taxi drivers. Taxi drivers are considered as a well-informed group of drivers who have acquired higher knowledge of the road network and its travel time variations. From a policy and planning perspective, an improved understanding of taxi drivers' route choice behavior is important, since taxi, as an important mode of transportation in big cities, provides further insights on human mobility patterns and urban structures. A future direction could be the inclusion of regular car drivers and to examine the validity of the four behavioral clusters proposed in this study. Since the experience level of regular car drivers varies widely from taxi drivers, we might expect different driving patterns. Another interesting elaboration of this study could be the incorporation of these clusters into route choice models and to examine the enhancement of their estimation and prediction abilities. Since our findings suggest that discrete route choice behaviors exist, the application of latent class discrete choice models would be appropriate to capture this behavioral heterogeneity. Lastly, the recent rise of ride-hailing services has pushed the cities and transportation planners to develop new laws and policies for urban mobility. This study can be very helpful in giving insights to decision makers and can help them propose new laws and policies which safeguard taxi industry as well as encourage sharing economy.

REFERENCES

[1] Manley, E., J. Addison, and T. Cheng. Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London. *Journal of Transport Geography*, Vol. 43, 2015, pp. 123-139.

[2] Parkany, E., J. Du, L. Aultman-Hall, and R. Gallagher. Modeling stated and revealed route choice: consideration of consistency, diversion, and attitudinal variables. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1985, 2006, pp. 29-39.

[3] Tawfik, A. M., H. A. Rakha, and S. D. Miller. An experimental exploration of route choice: identifying drivers choices and choice patterns, and capturing network evolution.In *Intelligent*

Transportation Systems (ITSC), 2010 13th International IEEE Conference on, IEEE, 2010. pp. 1005-1012.

[4] Ramaekers, K., S. Reumers, G. Wets, and M. Cools. Modelling route choice decisions of car travellers using combined GPS and diary data. *Networks and Spatial Economics,* Vol. 13, No. 3, 2013, pp. 351-372.

[5] Koller-Matschke, I., H. Belzner, and R. Glas. Traffic Information and Individual Driver Behaviour.In *Proceedings of the FISITA 2012 World Automotive Congress*, Springer, 2013. pp. 179-185.

[6] Gan, H., and S. Chen. Why Do Drivers Change Routes? Impact of Graphical Route Information Panels. *Institute of Transportation Engineers. ITE Journal,* Vol. 83, No. 8, 2013, p. 38.

[7] Prato, C. G., S. Bekhor, and C. Pronello. Latent variables and route choice behavior. *Transportation,* Vol. 39, No. 2, 2012, pp. 299-319.

[8] Cools, M., E. Moons, L. Creemers, and G. Wets. Changes in travel behavior in response to weather conditions: do type of weather and trip purpose matter? *Transportation Research Record: Journal of the Transportation Research Board*, No. 2157, 2010, pp. 22-28.

[9] Ben-Akiva, M., A. De Palma, and K. Isam. Dynamic network models and driver information systems. *Transportation Research Part A: General,* Vol. 25, No. 5, 1991, pp. 251-266.

[10] Ulleberg, P. Personality subtypes of young drivers. Relationship to risk-taking preferences, accident involvement, and response to a traffic safety campaign. *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 4, No. 4, 2001, pp. 279-297.

[11] Lucidi, F., A. M. Giannini, R. Sgalla, L. Mallia, A. Devoto, and S. Reichmann. Young novice driver subtypes: relationship to driving violations, errors and lapses. *Accident Analysis & Prevention,* Vol. 42, No. 6, 2010, pp. 1689-1696.

[12] Marengo, D., M. Settanni, and G. Vidotto. Drivers' subtypes in a sample of Italian adolescents: relationship between personality measures and driving behaviors. *Transportation Research Part F: Traffic Psychology and Behaviour,* Vol. 15, No. 5, 2012, pp. 480-490.

[13] Saunier, N., N. Mourji, and B. Agard. Mining microscopic data of vehicle conflicts and collisions to investigate collision factors. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2237, 2011, pp. 41-50.

[14] Higgs, B., and M. Abbas. A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles.In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, IEEE, 2013. pp. 857-862.

[15] Jensen, M. Passion and heart in transport—a sociological analysis on transport behaviour. *Transport Policy,* Vol. 6, No. 1, 1999, pp. 19-33.

[16] Kroesen, M., and S. Handy. The relation between bicycle commuting and non-work cycling: results from a mobility panel. *Transportation,* Vol. 41, No. 3, 2014, pp. 507-527.

[17] Geller, R. *Four types of cyclists*. PortlandOnline, 2009.

[18] Dill, J., and N. McNeil. Four types of cyclists? Examining a topology to better understand bicycling behavior and potential.In *92nd Annual Meeting of the Transportation Research Board, Washington, DC*, 2013.

[19] Reinoso, G., and B. Farooq. URBAN PULSE ANALYSIS USING BIG DATA.

[20] Tawfik, A. M., H. A. Rakha, and S. D. Miller. Driver route choice behavior: Experiences, perceptions, and choices.In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, 2010. pp. 1195-1200.

[21] Tawfik, A. M., J. Szarka, L. House, and H. A. Rakha. Disaggregate route choice models based on driver learning patterns and network experience.In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2011. pp. 445-450.

[22] Liu, L., C. Andris, and C. Ratti. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems,* Vol. 34, No. 6, 2010, pp. 541-548.

[23] Communauté métropolitaine de montréal, C. *An Attractive, Competitive and Sustainable Greater Montreal*, 2012.
[24] Ben-Akiva, M., and M. Bierlaire. Discrete choice methods and their applications to short term travel decisions.In *Handbook of transportation science*, Springer, 1999. pp. 5-33.
[25] Pagès, J. Analyse factorielle de données mixtes. *Revue de statistique appliquée,* Vol. 52, No. 4, 2004, pp. 93-111.
[26] Poucin, G., B. Farooq, and Z. Patterson. Pedestrian Activity Pattern Mining in WiFi-Network Connection Data.In *Transportation Research Board 95th Annual Meeting*, 2016.
[27] Rakotomalala, R. TANAGRA: un logiciel gratuit pour l'enseignement et la recherche.In *EGC*, 2005. pp. 697-702.
[28] Lebart, L., A. PIRON, M. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, 2000.
[29] Everitt, B. S., D. Stahl, M. Leese, and S. Landau. Cluster analysis. 2011.
[30] Mooi, E., and M. Sarstedt. *Cluster analysis*. Springer, 2010.
[31] Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. Simulation based population synthesis. *Transportation Research Part B: Methodological*, Vol. 58, 2013, pp. 243-263.