POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Characterising Annual Behaviour of Carsharing Users in Montreal
Auteurs: Authors:	Hanieh Baradaran Kashani, & Martin Trépanier
Date:	2019
Type:	Communication de conférence / Conference or Workshop Item
Référence: Citation:	Kashani, H. B., & Trépanier, M. (mai 2019). Characterising Annual Behaviour of Carsharing Users in Montreal [Communication écrite]. World Conference on Transport Research (WCTR 2019), Mumbai, India. Publié dans Transportation Research Procedia, 48. https://doi.org/10.1016/j.trpro.2020.08.175

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10641/
Version:	Version officielle de l'éditeur / Published version Révisé par les pairs / Refereed
Conditions d'utilisation: Terms of Use:	Creative Commons Attribution-Utilisation non commerciale-Pas d'oeuvre dérivée 4.0 International / Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND)

Document publié chez l'éditeur officiel
Document issued by the official publisher

Nom de la conférence: Conference Name:	World Conference on Transport Research (WCTR 2019)
Date et lieu: Date and Location:	2019-05-26 - 2019-05-31, Mumbai, India
Maison d'édition: Publisher:	Science Direct
URL officiel: Official URL:	https://doi.org/10.1016/j.trpro.2020.08.175
Mention légale: Legal notice:	© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 48 (2020) 1435–1449

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Characterising Annual Behaviour of Carsharing Users in Montreal

Hanieh Baradaran Kashani[a], Martin Trépanier[a] *

[a]*Polytechnique Montréal and Interuniversitary Research Centre on Entreprise Networks, Logistics and Transportation (CIRRELT)*
P.O. box 6079, station Centre-Ville, Montréal, Québec, H3C 3A7, Canada

Abstract

Carsharing is one of the emerging modes of transportation in the recent years, as a pointer to the fact that owning a private car is not all we need for the sake of travelling. Using k-means clustering and Principal Component Analysis, the purpose of this paper is to study the behaviour of the Communauto regular-service carsharing users in Montreal over a year (2014) and find the usage patterns in each cluster. Also, by having the Communauto customer features available, the characteristics of the customers in each cluster will be defined, which eventually ends in customer profiles. The k-means clustering results show that the Communauto regular-service carsharing users are divided into nine different clusters. Some of the clusters patterns are similar during the weeks and some are similar during a year, but none of them are similar during week and year. So, each cluster has a unique usage pattern over the weeks and the year. Furthermore, the resulting clusters are ordered from highly intensive users to the occasional ones based on the frequency of carsharing usage over one year.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the World Conference on Transport Research – WCTR 2019

Keywords: Carsharing; User Behaviour, Clustering, Principal Component Analysis (PCA)

1. Introduction

Carsharing (Autopartage in French) sits within the emerging class of 'mobility services' that draw on modern technology to enable access to car-based mobility without the consumer owning the physical asset (a car) (Le Vine, Zolfaghari et al. 2014). As described by the carsharing association in USA, the mission of carsharing is to reduce car ownership, provide easy access to automobiles for publics, reduce vehicle distant travelled and so on. Easy access to carsharing services has a wide sense with many aspects such as affordability, little or no paper work, 24/7 access.

* Corresponding author. Tel.: +1-514-343-7240; fax: +1-514-340-4173.
E-mail address: mtrepanier@polymtl.ca

2352-1465 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the World Conference on Transport Research – WCTR 2019
10.1016/j.trpro.2020.08.175

An important advantage of carsharing could be the fact that the users enjoy this mode of transportation like a private one, without having to take the responsibilities of owning a car. This might be one of the main reasons that people are more and more attracted to this mode of transportation. Consequently, it is more important to study about carsharing users' behaviours, to better understand their needs as well as carsharing role in the urban transportation system. This way the carsharing enterprises would find the best ways of improving their services to their customers.

Communauto, a carsharing company based in Montreal, Quebec, Canada, also operates in three other cities in Quebec: Quebec City, Gatineau and Sherbrooke. This company, which is founded in Quebec City in 1994, by Benoît Robert, offers two types of services: 1) Regular service, which offers station-based vehicles that should be reserved up to a month in advance and be brought back to the same station it is picked up. 2) Free-floating which offers « Automobile » vehicles, that needs no reservation and can be released anywhere in the service area. This carsharing operator wants a better knowledge of its clientele, especially their behaviour, to be able to offer more adapted fare packages.

By means of K-means clustering, the aim of this paper is to explore the Communauto regular service reservation dataset of the year 2014, to uncover the usage patterns of the customers. Because of the nature of the data and the k-means assumptions, Principal Component Analysis or PCA is applied on the data before k-means. Therefore, the k-means clustering is implemented on the principal components. Since k-means is an unsupervised learning which attempts to cluster the observations, the original data can adopt the resulting clusters from PCA, so that the results would be interpretable.

In this paper, we will first review some of the related studies on carsharing systems, the customers' usage behaviours as well as the studies which relate PCA to k-means clustering. Next, we will focus on the methodology of this study by introducing the datasets and the adopted statistical methods, the results of this study will come afterwards and the paper concludes with section 5 which is devoted to the discussion part of the study.

2. Related Works

In this section, some studies on carsharing systems are reviewed in three parts. The first part, reviews the studies on carsharing system in general, while the second part provides a brief look on more specific studies about carsharing users' behaviours using statistical methods such as k-means clustering. The third part reviews the studies on k-means clustering methods.

2.1. Carsharing in General

Even until the late 20s, no one could imagine sharing a vehicle could be happening among all the fellow citizens. Nowadays, this has been turned to a popular mode of transportation. People now support such a mode of transportation that takes away the concerns of car ownership and at the same time provides them with the comfort of driving a private car that is also environmentally friendly. According to the carsharing missions, various studies have been conducted that can show to what extent these missions have been achieved so far.

As the car ownership reduction is one of the missions of the carsharing companies, a very recent study in London, UK, (Le Vine and Polak 2017) established the early stage impact of free-floating carsharing on private car ownership. The results of this study showed, 37% of the users revealed that free-floating carsharing impacted their ownership of private vehicles. This study also exposed that the frequent service users, had a higher level of education and income than the average of the population.

A study in Canada (Klincevicius, Morency et al. 2014) assessed the reduction of car ownership in an area in Montreal served by Communauto regular-service carsharing. Using the historical data from the population and the users' behaviours, the authors examined the relation between the people car ownership and exposure to carsharing. The results by a linear regression model showed that car ownership had a reverse correlation with the number of carsharing vehicles in a 500-m radius of the local households.

Carsharing aims to reduce the greenhouse gas by decreasing the number of vehicles in use in the cities. A survey in Montreal, Canada, examined the contribution of Communauto carsharing, in reducing greenhouse gas emission. They examined the quantity of CO2, as the main source of GHS, emitted by Communauto vehicles. Considering the usage habits of carsharing users, their results confirmed that each carsharing vehicle replaces ten to fourteen private cars (Trépanier et al. 2013). Therefore, each carsharing user produces 1160 kg of CO2 less than when they were not subscribed to this service. Also, a case study (Sioui, Morency et al. 2013) based on two comparative surveys showed that carsharing members did not reach the level of car use of typical residents owing one or more cars. This is another indication that the carsharing users' contribution in greenhouse gas is less than other people owning a car.

2.2. Vehicle-Sharing Users' Behaviour Studies

Getting to know the customers' behaviours and requirements is essential as it helps the companies to better improve their services and adopt their strategies to the customers' needs. A study in California, US, (Shaheen and Cohen 2008) based on 33 carsharing international surveys, claimed that cost saving, convenient locations and the guaranteed parking are the main motivations of using the carsharing services worldwide.

The customers' behaviour and their frequency of usage have been attractive to the researchers and the carsharing companies. A study identified typical patterns of carsharing use, by k-means clustering (Morency, Trépanier et al. 2007). Eight clusters were found, in each of which the users had some favourite weekdays of using the carsharing system. They also classified the carsharing users from different aspects. For instance, based on frequency of use, the users were classified to the occasional and frequent users, and from the aspect of trip length to the short-distance users and long-run users. De Luca and di Pace (2015) proposed a stated-preference approach to study users' behaviour and found that access time to parking slot is an important attribute. Some evidences have also been brought by Koop and al. (2015) on the differences between free-floating carsharing users and non-car-sharers. Hwang and Griffiths (2016) have studied the factors that attracts younger generations to carsharing schemes.

Like carsharing, the usage behaviour of bikesharing users have also been studied worldwide. Research in Lyon, France, (Vogel, Hamon et al. 2014) was carried out based on a large-scale behavioural dataset of bicycle sharing users. Exploiting cluster analysis, they produced user typology based on annual weekly, monthly and daily patterns. They found nine clusters of users with a unique profile for each of them, using the characteristics of the customers.

Another study on Montreal's Bixi bikesharing members (Morency, Trepanier et al. 2017), indicated that people living near carsharing systems have a higher possibility of being a recurrent user. They also showed that the weather conditions influence the users' behaviours in using the bikesharing system. Some other studies are conducted to compare the usage behaviours in carsharing and bikesharing systems. The results of a study in Montreal, Canada (Wielinski, Trépanier et al. 2017), confirmed that the bikesharing users are mostly men and younger and have higher income, whereas most of the carsharing users have more children and fewer cars. Using a multinomial logit model, they found that the carsharing-only users have the lowest income among the others and tend to use public transport systems more than the others, while the bikesharing-only users have the highest income and tend to use private cars more. They also identified two-system users whose income and transportation usage behaviour are in-between these two groups.

2.3. K-Means Clustering and PCA

Cluster or segmentation analysis is a kind of exploratory analysis that seeks to find some structure in the data. It divides the data points into the clusters in which they are similar, but dissimilar with the other data points in the other clusters. Because of its nature of exploratory, it has a wide application in different fields of studies.

PCA or Principal Component Analysis, sometimes referred by factor analysis, is a statistical procedure that shrinks the high-dimensional data to a lower dimension by maintaining most of the information of the data. In many applications, the data analysts prefer to lower the dimension of data by using a shrinkage method like PCA, especially for k-means clustering. The main reasons to do so will be discussed more in detail in the methodology section.

One of the issues that is sometimes discussed about PCA is that, applying other statistical techniques on the principal components instead of the original variables, cause the loss of interpretability of the results. However, this is not always true and depends on the methods that we desire to apply on the components. Research in Atlanta, USA, (Liang, Balcan et al. 2013) introduced a distributed PCA algorithm, and theoretically proved that any good approximation solution on the projected data by distributed PCA for k-means clustering, would also be a good approximation on the original data.

Some studies in this field are devoted to the connections of PCA and k-means clustering. A study in California, (Ding and He 2004) indicated that PCA as an unsupervised dimension reduction is very close to k-means clustering as an unsupervised learning. They showed that principal components are relaxed (without constraint) solutions of the cluster indicator vector in k-means clustering. They also proved that the cluster centroid subspace is spanned by the first k-1 principal component directions.

Since one of the problems in k-means is that a change in the initial values of the centroids changes the results of the clustering, some articles attempt to solve this initialisation problem. For instance, a study in Boston, (Su and Dy 2004) focused on this and suggested that by utilising the eigenvectors of the covariance matrix in Principal Component Analysis as the k-means' centroid initialisation, the clustering results produce smaller Sum of Squared Errors (SSE), also they showed that k-means converges faster by this approach. However, at the end they confirmed that, because of some limitations, this initialisation approach might sometimes fail.

For the problem of initialisation, a study whose results are very popular now, (Arthur and Vassilvitskii 2007) introduced a method of initialisation named "k-means++". In this study, the authors propose a method by which the initial centroids are first chosen randomly, but to choose the next centroids, the data points are weighted according to their squared distance from the closest previously chosen centroid. Finally, they empirically show that k-means++ initialisation, is very often faster and gives more accurate results than the random initialisation. Similarly, in this study we utilize k-means++ initialisation for k-means.

3. Methodology

In this section, the aim is to describe the adopted methods for this study. As discussed, the objective is to cluster the customers, based on the frequency of their usage (number of reservations), and in each cluster, find the usage patterns. Figure 1 gives a summary of the methods we went through to achieve this objective.

First, the relevant variables in Communauto reservation datasets are selected and the data is cleaned. Then, the vector of attributes is produced using the pivot tables counting the number of reservations daily, weekly and monthly. Afterwards, the k-means assumptions and PCA prerequisites are verified. Accordingly, at the next step the data would be pre-processed. When the data is appropriately pre-processed, PCA projects the data into a smaller subspace. Next, K-means clusters the new projected data and consequently, the usage patterns in each cluster are defined. Finally, using the produced clusters and the customers' dataset, the customer profiles are defined.

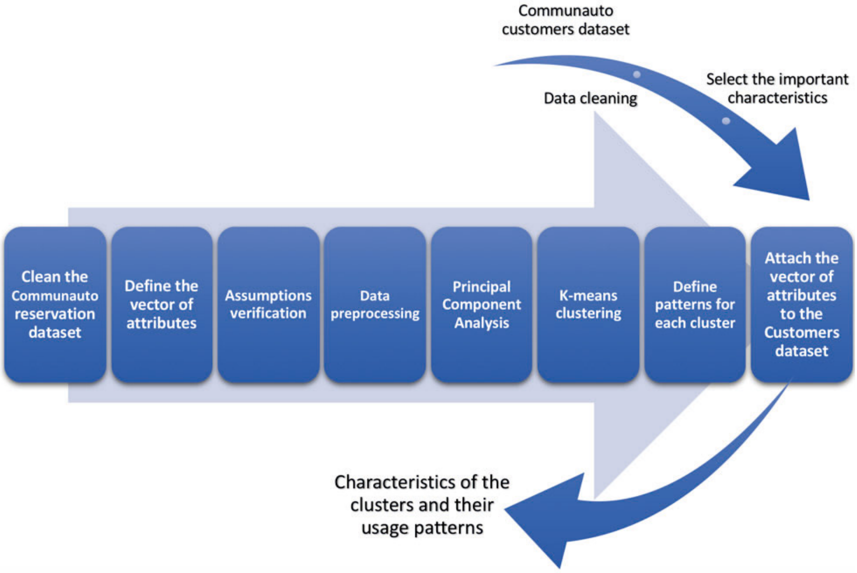
Communauto customers dataset
Data cleaning
Select the important characteristics
Clean the Communauto reservation dataset
Define the vector of attributes
Assumptions verification
Data preprocessing
Principal Component Analysis
K-means clustering
Define patterns for each cluster
Attach the vector of attributes to the Customers dataset
Characteristics of the clusters and their usage patterns

Fig. 1. Diagram of the study methodology.

3.1. Communauto Dataset

According to the two types of Communauto carsharing services, there are also two types of datasets of the carsharing transactions, regular (station-based) and Auto-mobile (free-floating), each of which contains its features related to the type of service. The available data in this study are the transactions happened only in the year 2014, and we chose to work only on the regular-service carsharing as for the other, Auto-mobile, the approach would be the same. Because these datasets contain several columns that are not relevant to our objective, Table 3.1, describes the relevant variables.

Table 1. Description of the relevant variables in the regular-service reservation dataset.

Variables	Description	Range
CustomerID	The identity number for the customers, which is unique for each customer	[4, 82636]
datDateDebutReservation	The date and time that the customer reserves the vehicle	2014-01-01 00:00:00 to 2014-12-31 23:45:00
intDebutKilometrage	The mileage of the car, at the time of the reservation	[0, 234747]
intFinKilometrage	The mileage of the car, after the reservation is ended	[0, 234837]

In this dataset, each row belongs to each reservation, so we might encounter a customer ID repeating several times for several reservations during the year. The reservation date helps to find out about the time of the carsharing usages. Which months have the most frequent usage? Are the customers using the carsharing system regularly or irregularly? This variable can disclose the usage pattern of the customers by answering such questions. The car's mileage uncovers the reservations without usage, in which we are not interested and must be removed from the data.

Also, certain customers' specifications like gender, age, language, etc. were available for some of the customers. These specifications helped us at the end, to better describe each cluster and the usage patterns.

3.2. Vector of Attributes

For our objective to be met, the number of reservations, per weekdays, weeks and months for each customer must be counted. To do so, the year (2014), month (1, …,12), weekday (Mon, …, Sun) and the number of the week in the year (1, …,53) are extracted from one variable, "reservation date". Next, a pivot table for each of them is made to count the number of reservations per customer. Therefore, in the new data frame the number of rows equals the number of customers in the regular service reservation dataset. In this section, the vector of attributes i.e. the variables of the new dataset are described:

- $X_1, \ldots, X_{12}$: Average monthly use, i.e. the average number of trips per month.
- $X_{13}, \ldots, X_{19}$: Average daily use, i.e. the average number of trips per days of the week, Monday to Sunday.
- X_{20}: Averaged weekly use, i.e. the average number of trips per weeks of the year 2014, calculated over all the weeks during which user travels at least once. Divided by seven to make it consistent with the previous attributes.
- X_{21}: Normalised total trips, i.e. Total number of trips made over the year 2014, summed over all weeks, normalised dividing by 1.5 times its interquartile range of the distribution for all users. The interquartile range is used as a robust measure of scale. That is, it is an alternative to the standard deviation and it is less affected by extremes than the standard deviation.

This way, out of only one variable: "Reservation date", 21 variables are created. Then they are attached to build the new dataset containing these 21 variables as the columns and the customer ID's as the indexes. Now that the data is ready, they must be verified if they need to be pre-processed before any analysis. For this reason, the assumptions of k-means clustering must be verified.

3.3. Assumptions Verification

Assumptions play an important role in the statistical methods. Without verifying the assumptions, any result from the analysis would not be reliable.

K-means clustering assumes that the distribution of the data is spherical. This means when looking at the scatter plot we should not observe any ellipse suggesting some correlation between two variables. In other words, sphericity means the variables are uncorrelated (covariance = 0), and they all have an equal variance of one, which means the covariance matrix is equal to the identity matrix. Bartlett Sphericity test, also suggests the same approach for verifying the sphericity:

- H0: The correlation matrix of the data is equal to the identity matrix
- H1: The correlation matrix of the data is different from the identity matrix

If we reject the null hypothesis (P-Value<0.05), then the correlation matrix is different from the identity matrix and the data is not spherical.

The above-mentioned test is highly sensitive to the number of samples, n, i.e. if n is very large, it rejects the null hypothesis even if the correlations are very close to zero, but not zero. This test also assumes that the multivariate distribution of the data is normal. (Sarmento and Costa 2017)

Despite these restrictions, we considered this test to verify the sphericity of the data, we also kept observing the scatter plot of the variables at every step of the pre-processing the data.

On the other hand, Principal Component Analysis or PCA which will be described in the section 3.5, is sensitive to the noises. Besides, it needs the variables to follow the same measurements. The data pre-processing section will go through the pre-processing steps to make the data ready for the analysis.

3.4. Data Preprocessing

According to the previous section, the data should be verified if the assumptions are already met. To verify this, figure 2 illustrates how the pairwise scatter plot of our data looks like. Also, the diagonal line plot shows how each variable is distributed. According to this figure there are big outliers in the data.

The outliers cause many deficiencies in the data. They are one of the reasons that the data distribution is not spherical and it is extremely skewed. Moreover, the big outliers distract k-means clustering from finding the appropriate centroids.

Thus, the first step in pre-processing the data would be the best to remove particularly the big outliers to study them later. So, we need to keep them apart and verify the assumptions again in the rest of the data.

Figure 2 reveals that the distribution of the variables is strongly right-skewed. The scatter plot of the last two variables, i.e. X20: Averaged weekly use and X21: Normalised total trips, are not visible, and this is due to the very large outliers that exist in these two variables.

For PCA, it is very important to ensure that the variables follow the same measurements. Hence, the following operations were applied to the dataset:

- Outlier removal using Mahalanobis distance. An outlier, or noise, is an observation that is distant from other observations (Maddala, G. S. 1992). It may be due to variability in the measurement or an experimental error (Grubbs 1969). The outliers might also exist because some of the observations show different behaviours. In our data, beside the fact that some of the customers shown to have totally different usage patterns, the major cause of the outliers is that, the measuring of the last two variables is different from the other 19 variables. The other variables are the averages that vary between 0 and 1, but the last two are positive variables ranging in an interval of (0, ∞). Using an estimate of the location of each observation, Mahalanobis distance, locates the data points, that are significantly distant from the rest of the data. (Franklin, Thomas et al. 2000) Here we present a simple definition of Mahalanobis distance. Let x ⃗= {x_1, x_2, …, x_n)^T, represent a set of data points with the mean μ ⃗=(μ_1,μ_2,…,μ_n)^T, and the covariance S. The Mahalanobis distance would be as follows: (De Maesschalck, Jouan-Rimbaud et al. 2000)

$$D_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1} (\vec{x}-\vec{\mu})} \tag{1}$$

- Log transformation. Log transformation replaces all the data points by $z_i = f(x_i)$, where f is a logarithm function usually with the base of 2, 10 or e=2.718. By maintaining the order of the data points, log transformation has the biggest effect on the largest values, i.e. it reduces the distances between the large numbers more. The skewness of our data is already shown in figure 2. Therefore, at the same time of resolving the skewness of the data, log transformation, smooths the noises and makes the distribution of the data more spherical. This helps the k-means assumptions and the PCA requirements to be met.

- Standardisation. Standardisation is a process that transforms the mean and the standard deviation of the data to zero and one, respectively. This is always done by subtracting the mean from the data and dividing them by the standard deviation of the data:

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{2}$$

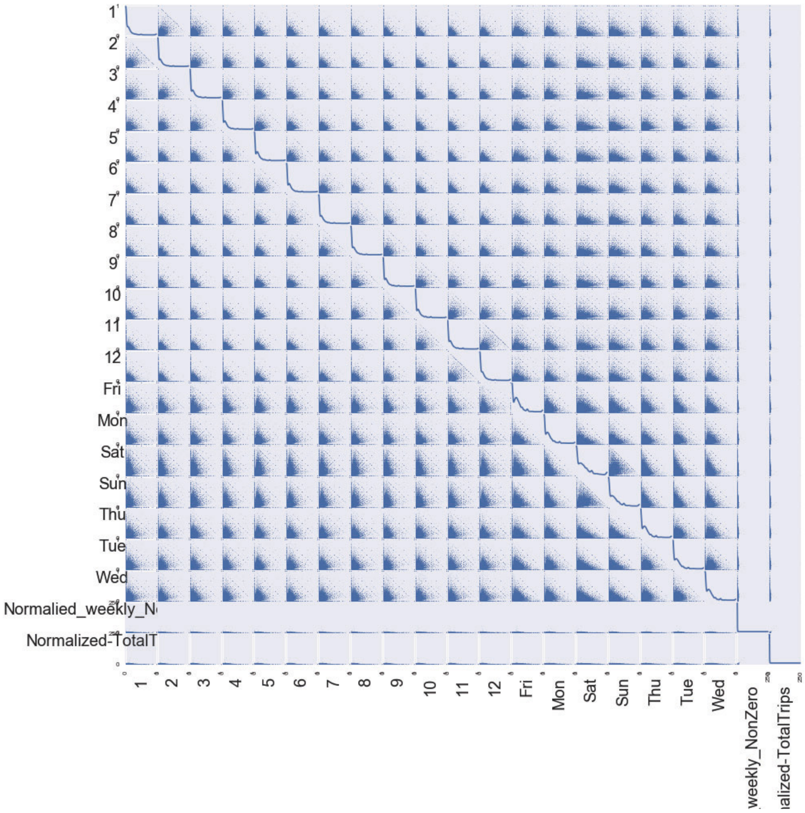

Fig.2. The pairwise scatter plots for the vector of variables created from Communauto reservation dataset (2014).

3.5. Principal Component Analysis (PCA)

Principal Component analysis or PCA is an orthogonal linear transformation of the data which projects the variables to a lower-dimensional subspace by which the first new variable (or the first principal component) holds the highest variance of the data, the second variable holds the second-highest variance and so on. (Jolliffe 2002) In other words the new coordinate system is a combination of all the original variables in a way that it can capture the maximum variance it can from the data. Principal component analysis like many other statistical procedures has both its pros and cons. However, there are strong reasons to apply PCA on the data, before k-means clustering:

- As the dimensionality increases, the accuracy of k-means decreases. This is called the “curse of the dimensionality”. PCA, projects all the variables of a dataset to a lower-dimensional subspace.
- At the same time of reducing the dimension, PCA is building new variables which are not redundant or correlated.
- K-means is sensitive to the noises and outliers, PCA helps improve the clustering accuracy by smoothing the noises.

To implement the principal component analysis, the eigenvalues and eigenvectors should be calculated and to choose the top- K subspace, the value of K must be selected, depending on the variance of the data that we choose to keep.

Recall that PCA tries to minimise the average squared projection error:

$$\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)} - x_{approx}^{(i)}\right|^2 \tag{3}$$

Which means it tries to minimise the squared distance between x and its projection onto that lower-dimensional surface. The total variation in the data is given by:

$$\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)}\right|^2 \tag{4}$$

This second formulation defines how far the training examples are from the vector, i.e. from being all zeros. To choose a K, a common rule of thumb is to calculate the ratio between (3) and (4) to be less than a certain value, c:

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)}-x_{approx}^{(i)}\right|^2}{\frac{1}{m}\sum_{i=1}^{m}\left|x^{(i)}\right|^2} < c \tag{5}$$

This ratio defines the amount that the data varies. For example, if the c value is 0.01, we retain 99% of the variance of the data. Depending on the data, the analyst might choose another percentage for the variance of the data to be retained. Therefore, the value of K is chosen in a way that the ratio (5) is satisfied.

A more efficient way to do this is to calculate the singular value decomposition (SVD(Σ)=USV*) of the covariance matrix. Considering the elements of the decomposed diagonal matrix S, the following equation can be calculated:

$$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} < c \tag{6}$$

Where S_{ii} is the i'th diagonal element of the matrix S. Equation (6) ensures that (1- c) % of the variance of the data is retained. We used the latter equation to decide about the number of principal components, which will be discussed in results.

3.6. K-Means

K-means clustering is a method for finding clusters and cluster centres in a set of unlabelled data. One chooses the desired number of cluster centres, say K, and the K-means procedure iteratively moves the centres to minimise the total within cluster variance. Given an initial set of centres, the K- means algorithm alternates the two steps:

- for each centre, we identify the subset of training points (its cluster) that is closer to it than any other centre;
- the means of each feature for the data points in each cluster are computed, and this mean vector becomes the new centre for that cluster.

These two steps are iterated until convergence.

The K-means algorithm is one of the most popular iterative descent clustering methods. It is intended for situations in which all variables are of the quantitative type, and squared Euclidean distance:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p}\left(x_{ij} - x_{i'j}\right)^2 = \|x_i - x_{i'}\|^2 \tag{7}$$

is chosen as the dissimilarity function (Friedman, Hastie et al. 2001). K-means method uses K prototypes, the centroids of clusters, to characterise the data. They are determined by minimising the sum of squared errors:

$$J_k = \sum_{k=1}^{K}\sum_{i\in C_k}(x_i - m_k)^2 \tag{8}$$

where $(x_1, \ldots, x_n) = X$ is the data matrix and $m_k = \sum_{i \in C_k} x_i / n_k$ is the centroid of the cluster C_k and n_k is the number of points in C_k. (Ding and He 2004).

One of the drawbacks of k-means clustering is that, it requires a priori specification of the number of clusters, k. There are several methods to find the proper number of clusters. This will be discussed in the Results section.

4. Results

The described methods were applied on the Communauto datasets, and the results are described here. As discussed, after building the vector of attributes, we pre-processed the data according to the data and the assumptions of the methods stated before.

4.1. Application of PCA

Figure 3 shows the accumulated explained variance by the principal components. As already calculated, if we retain 72% of the data's variance, the number of components would be K=10, which empirically proves to be small enough to treat the "curse of dimensionality" for k-means clustering. On the other hand, preserving 72% of the whole variance of the data is reasonable for the percentage of information we desire to maintain.

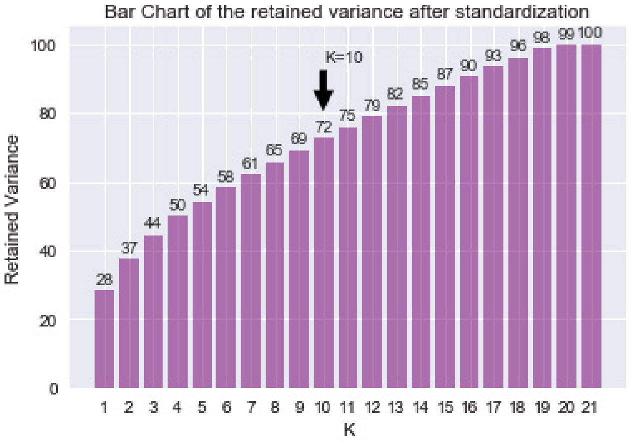

Fig.3. The accumulated retained variance for each number of principal components.

Here we look back to the assumptions. We would like to observe if the distributions of the variables are spherical now. Figure 4 shows that the pairwise scatter plots of the variables is spherical. The histograms of the variables are now closer to the normal distribution and less skewed. Recall that because of sphericity, the principal components must have a correlation of zero between them and variance of one.

Now that the assumptions of k-means are met, the k-means clustering can be applied to the principal components.

4.2. K-means Clustering

Selecting the number of clusters prior to k-means clustering is essential and Silhouette score (Rousseeuw 1987) is introduced as one of the methods of k selection in k-means. Figure 5, shows the silhouette score for each number of clusters, which is an average of all the silhouette scores of the observations. Since the closer to 1 the silhouette score is, the better the clustering is supposed to be, the candidates for the number of clusters according to Figure 5, could be K=5, 7, 9. At these three points, this score drops afterwards. Especially at the points 5 and 9. Although the silhouette score is higher for 2 or 3 clusters, they do not seem appropriate numbers of clusters, because they tend to have higher errors.

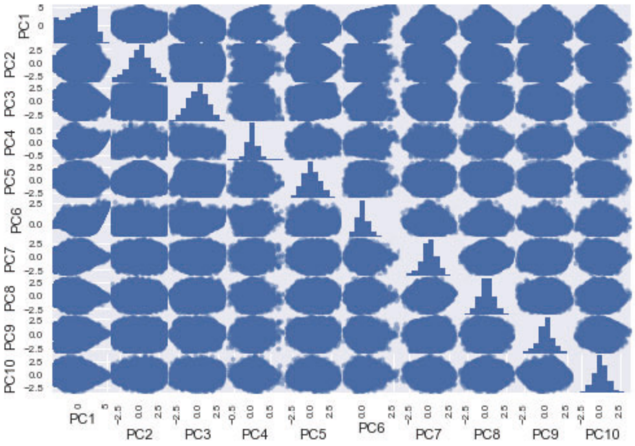

Fig.4. The scatter matrix of the principal components.

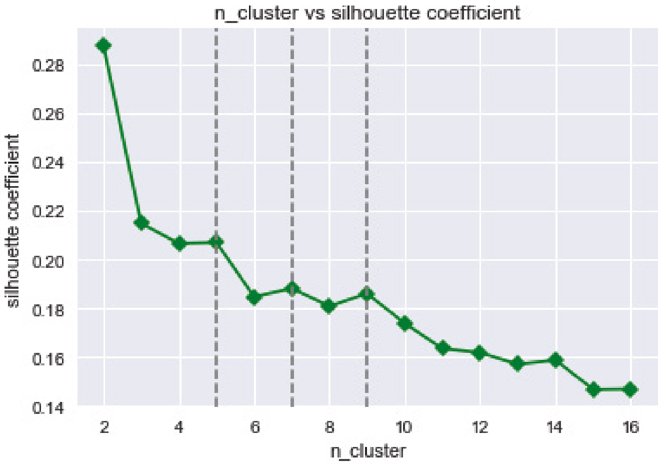

Fig.5. Silhouette score for each number of clusters.

To finally resolve the number of clusters, among the range of possible K’s from cross-validation results (K= 7,8,9,10,11,12), we pick the ones with the highest silhouette scores which are K=7 and K=9. We also examine K=12 because of the lowest MSE. Figure 6 illustrates customer monthly usage patterns in each cluster for different number of clusters in k-means, i.e. for K=5,7,9,12.

This figure shows that clustering with 5 and 7 clusters, build three distinct look-alike clusters and the other clusters seem to follow close patterns. To be exact, clusters “0, 2, 3” in K=5, are the same as clusters “1, 2, 4”, in K=7. Whereas, clustering with 9 clusters, reveals one more distinct cluster, cluster 4 – yellow, and the other five clusters follow very close patterns. Clustering with 12 clusters, does not add anything to the previous ones. In contrary, it shows less distinct clusters.

Because there is a trade-off between the mean square error and the silhouette score, we must be careful about the number of clusters, to keep the balance. As a result, K=9 can prove to be the best, because it produces less error than K=5,7 and it has the same silhouette score as K=7. Besides, the clusters in clustering with 9 clusters, are more distinct than in clustering with 12 clusters, and hold a higher silhouette score.

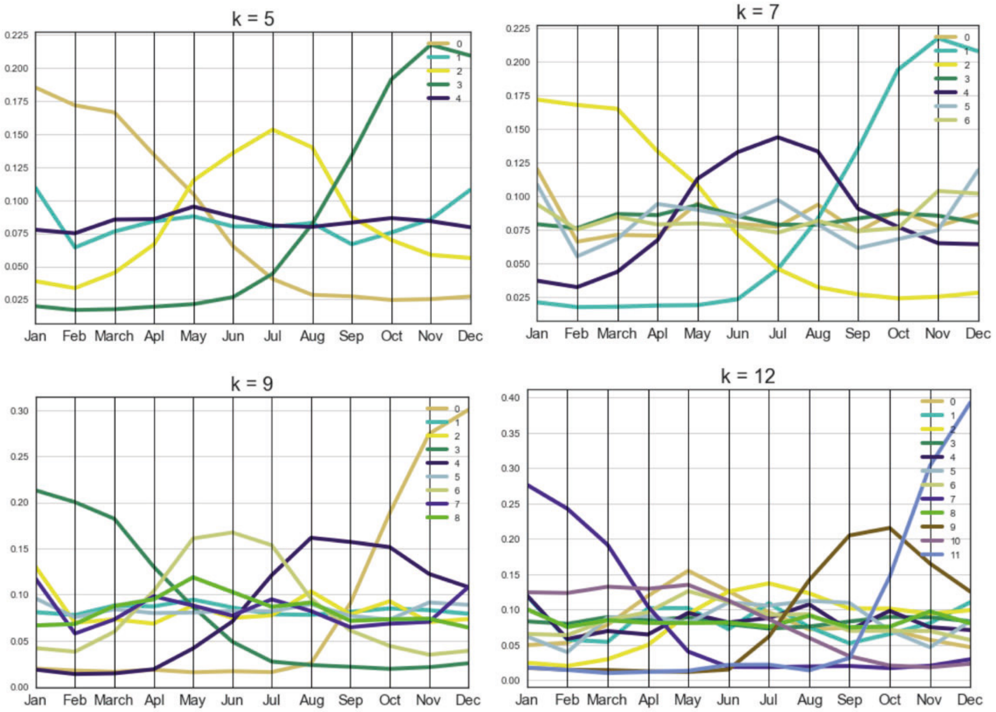

Fig.6. Customers monthly usage patterns for different numbers of clusters.

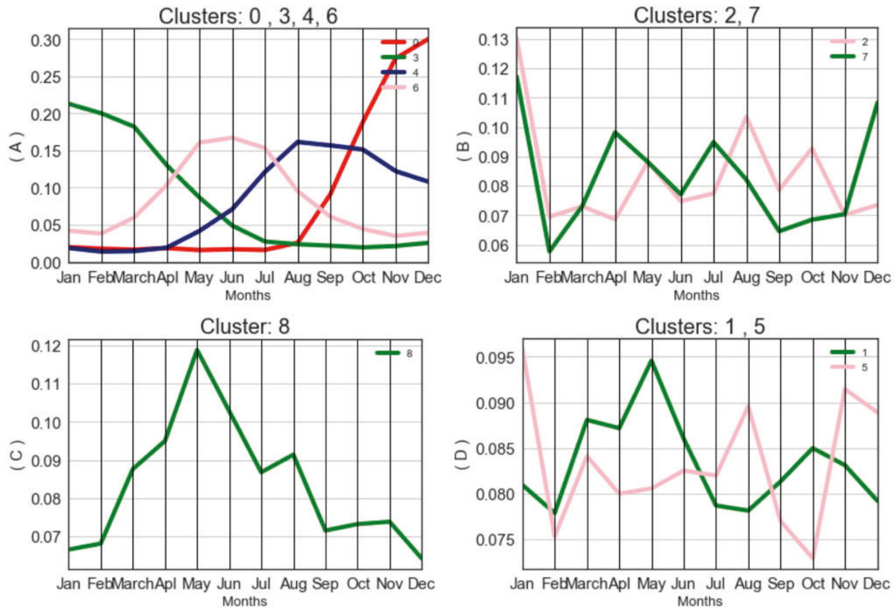

Fig.7. Customer monthly usage patterns for k-means clustering with nine clusters.

4.3. K-Means With 9 Clusters

Now that we are confident about the number of clusters, we study more in deep the behaviours of the customers. As already illustrated in figure 6, the customers' usage variation in different clusters is different. Figure 7 helps to observe better these variations in months. In this illustration, figures from A to D, are separated for clusters from higher variations to the lower variations respectively.

The usage patterns in (A), are extremely clear and there is no sudden jump. The users in (A) choose some consecutive months to use carsharing systems, each cluster shows a different season, and the average usage goes down to near zero in the other months. The usage patterns in (C), are also showing some months as the most preferable ones, which are happening in spring and summer. The users in (B) and (D) happen to show up in any season. However, they also have some preferable months.

Figure 7 displays such variations in weekdays. The five illustrations in this figure are separated according to the range of variations and the similarity of patterns. In this figure, clusters in A, are mainly Saturday and Sunday users, while in B they are almost only-Friday users. The users who are less interested in weekends are grouped in C, whereas users in D, are mostly Saturday users, however they also tend to show up on Fridays and Sundays to some degree. We must note that the variations from the illustrations A to D, decrease. So, although the users in D (clusters 1 and 4) seem to be the weekend users, their usage frequency over the week does not change as much as it does for the clusters in A (clusters 5 and 8).

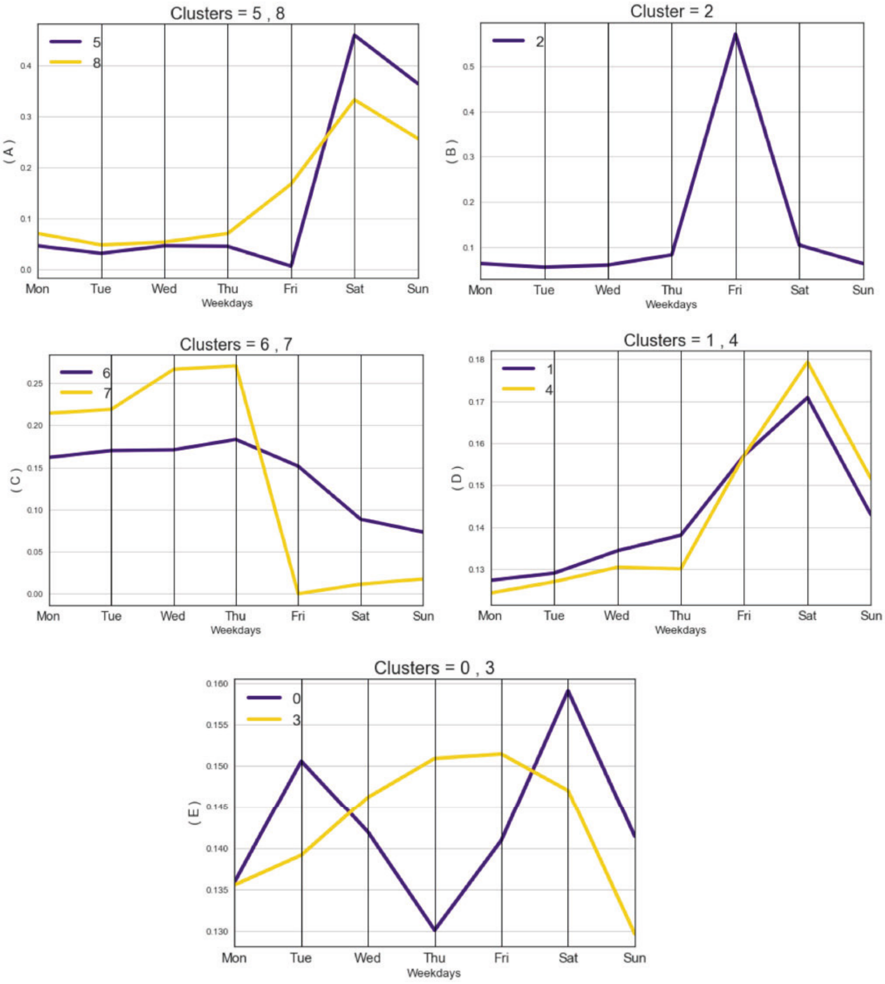

Fig 8. Customer weekday usage patterns for k-means clustering with nine clusters.

4.4. Customer's Analysis

As explained in the methodology section, customers' dataset is attached to the vector of attributes, to better describe each cluster according to the customers' characteristics. Out of 28,464 customers only 19,309 of them could be found in the customers' dataset and the personal specifications of the others were not available. Table 2 describes the characteristics of each cluster, according to the intensity and patterns of usage, as well as the available customers' features.

Table 2. Cluster characteristics based on the Communauto regular-service customers' dataset and the 9-means clustering patterns.

Clusters	*User Class*	*Number of users*	*Avg. no. Trips/user*	*Gender Ratio (W/M)*	*Median Age*	*Average years of membership*	*Language Ratio (Fr/En)*
1	Extreme - Regular	6528 (34%)	49.53	1.07	41	5.19	3.76
4	Intensive – summer, fall	1208 (6%)	15.99	0.98	40	4.88	4.25
3	Intensive – winter, Spring	2170 (11%)	15.19	1.17	38	4.25	3.84
8	Very Frequent – weekends	2254 (12%)	12.85	1.12	39	4.52	5.85
6	Very Frequent – spring, summer	1588 (8%)	10.88	0.83	43	4.94	4.84
0	Frequent – Fall	790 (4%)	7.65	1.06	42	4.73	5.75
2	Occasional – Fridays	1340 (7%)	2.87	1.04	40	4.29	4.88
5	Occasional– weekends	1938 (10%)	2.50	1.21	41	4.12	5.61
7	Occasional – Mon to Thu	1493 (8%)	1.95	0.93	42	4.73	5.07
Total		19309 (100%)	13.27	1.05	41.00	4.63	4.87

The clusters are ordered with respect to the usage intensity in each cluster. The "average number of trips per user" as an indicator displays this order. Also, the intensity of usage per user is named by "Extreme" users to the "Occasional" users. Regular in the user class column means that the users in this cluster are using the carsharing system almost all the months and all the days of the week. Whereas an indicated month or day means that the users are showing up in some specific days or months more than the other times. For instance, the customers' usage behaviour in cluster 4, is very intensive in summer and fall, but they don't show a very specific pattern for the weekdays. In contrary, the users in cluster 2 who are occasional ones, tend to show up on Fridays much more than the other days. However, they might be using the regular carsharing system in any month of the year.

The gender ratio indicates that in all the clusters, except for 4, 6 and 7, women are the dominant customers of Communauto carsharing regular service. The median age of all the users is around forty-one and the average year of membership in all the clusters is more than four years. Plus, the French speakers are very dominant in all the clusters which is quite expected in Quebec, Canada.

5. Conclusion

The objective of this study was to find the usage patterns of Communauto carsharing regular-service customers. Using k-means clustering, nine unique user profiles were found. These profiles were ordered from the most frequent users to the most occasional ones. Each cluster or user profile is identified with the most favourite season or days of the week for using the service. The low value of Silhouette score was not a major issue, as the main purpose of this study was to exploit the k-means clustering via PCA and discuss the methodology. The clusters found gives the carsharing operator a better knowledge of its clientele and a lever to manage the rates accordingly to their behaviour.

5.1. Contributions

K-means is one of the most popular clustering methods, because of its speed and simplicity. However, it has some assumptions and limitations on the data. It assumes that the distribution of the data to be clustered is spherical and consequently the variables are uncorrelated and have a variance of one. This assumption was not met on our raw data. On the contrary, the data contained big outliers and the distribution of the variables was strongly right-skewed. The cloud of the data had no sphericity, but it was more diagonal, meaning that there were correlations among the variables. This issue needed to be resolved before clustering.

Therefore, the big outliers were kept apart from the data using Mahalanobis distance and were analysed separately. This was a very important task to be done before everything, since the very big outliers could distract any other task on the data. But still the data distribution was strongly right-skewed, so, log transformation helped on this issue and made the distribution of the variables closer to normal and even the cloud of the data closer to spherical but not for all the variables.

Since some of the variables were correlated, Principal Component Analysis was chosen to be applied on the data, to have uncorrelated variables. At the same time, this reduced the noise and the number of variables to which k-means clustering was sensitive. However, PCA is also sensitive to the data measurements and had to be performed on the standardised data. So, the data was transformed in three steps: first log transformation, then standardisation and afterwards PCA. Subsequently, k-means clustering was performed on the transformed data.

Several works have been done on the vehicle sharing datasets, but none talked about the k-means clustering issues. In this study, we examined and found that k-means clustering on the PCA transformed data had smaller mean-squared error than the k-means clustering on the original data.

Some of the similar works that preferred not using PCA transformation for k-means, addressed the interpretability of the results of k-means on PCA transformed data, as an issue. Whereas interpretability should not be an issue when thinking of k-means clustering as an unsupervised learning.

To discuss this thought recall that, k-means attempts to cluster the observations of the unlabelled data, and PCA transforms the data according to the variables (columns). Consequently, the observations (rows) remain the same in the new transformed data. Thus, the original data would adopt the resulting cluster labels, and they could be simply interpreted according to the original variables.

5.2. Future Works

The resulting customer profiles were built based on some limited features of the customers. The customers' distances from the car stations or the number of cars they own, are some of the absent features which could lead to more constructive user profiles. More available customer features would be an effective way to improve the customer profiles in the similar future studies.

Statistical methods always have difficulties dealing with most types of data. There are always assumptions to be met before analysing. However, there are robust methods that are less sensitive and can handle the data conditions. For instance, in "Robust and sparse k-means clustering" (Xu, Han et al. 2016), a k-means approach has been proposed that can treat the outliers. There are also some other works that propose alternative methods to handle the data that is not spherical. Like in "What to do when K-means clustering fails: a simple yet principled alternative algorithm" (Raykov, Boukouvalas et al. 2016). Since data transformations can alter the accuracy of the results, one of the advantages of utilising the robust methods is that the original data would be clustered without any transformation. As a future work on the similar data, the analyst could consider the robust methods to improve the results accuracy.

Acknowledgements

The authors wish to acknowledge the support of Communauto and the Natural Resources and Sciences Research Council of Canada (NSERC) for providing the financial support.

References

Arthur, D. and S. Vassilvitskii (2007). k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics.

De Amorim, R. C. and C. Hennig (2015). "Recovering the number of clusters in data sets with noise features using feature rescaling factors." Information Sciences 324: 126-145.

De Luca, S. and Di Pace R. (2015). Modelling users' behaviour in inter-urban carsharing program: A stated preference approach, Transportation Research Part A: Policy and Practice, 71:59-76.

De Maesschalck, R., D. Jouan-Rimbaud and D. L. Massart (2000). "The mahalanobis distance." Chemometrics and intelligent laboratory systems 50(1): 1-18.

Ding, C. and X. He (2004). K-means clustering via principal component analysis. Proceedings of the twenty-first international conference on Machine learning, ACM.

Fachinger, J., 2006. Behavior of HTR Fuel Elements in Aquatic Phases of Repository Host Rock Formations. Nuclear Engineering & Design 236.3, 54.

Fachinger, J., den Exter, M., Grambow, B., Holgerson, S., Landesmann, C., Titov, M., Podruhzina, T., 2004. Behavior of spent HTR fuel elements in aquatic phases of repository host rock formations, 2nd International Topical Meeting on High Temperature Reactor Technology. Beijing, China, paper #B08.

Franklin, S., S. Thomas and M. Brodeur (2000). Robust multivariate outlier detection using Mahalanobis' distance and modified Stahel-Donoho estimators. Proceedings of the Second International Conference on Establishment Surveys, American Statistical Association Buffalo, NY.

Friedman, J., T. Hastie and R. Tibshirani (2001). The elements of statistical learning, Springer series in statistics New York.

Hwang, J. and Griffiths M.A. (2016). Share more, drive less: Millennials value perception and behavioral intent in using collaborative consumption services, Journal of Consumer Marketing, 34(2), 132-146.

Jolliffe, I. T. (2002). "Springer series in statistics." Principal component analysis 29.

Klincevicius, M., C. Morency and M. Trépanier (2014). "Assessing impact of carsharing on household car ownership in Montreal, Quebec, Canada." Transportation Research Record: Journal of the Transportation Research Board(2416): 48-55.

Kopp, J., Gerike, R. and Axhausen, K.W. (2015). Do sharing people behave differently? An empirical evaluation of the distinctive mobility patterns of free-floating car-sharing members, Transportation 42: 449

Le Vine, S. and J. Polak (2017). "The impact of free-floating carsharing on car ownership: Early-stage findings from London."

Le Vine, S., A. Zolfaghari and J. Polak (2014). "Carsharing: evolution, challenges and opportunities." Scientific advisory group report 22.

Liang, Y., M.-F. Balcan and V. Kanchanapally (2013). Distributed PCA and k-means clustering. The Big Learning Workshop at NIPS.

Morency, C., M. Trepanier, A. Frappier and J.-S. Bourdeau (2017). Longitudinal Analysis of Bikesharing Usage in Montreal, Canada.

Morency, C., M. Trépanier, B. Agard, B. Martin and J. Quashie (2007). Car sharing system: what transaction datasets reveal on users' behaviors. Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE, IEEE.

Raykov, Y. P., A. Boukouvalas, F. Baig and M. A. Little (2016). "What to do when K-means clustering fails: a simple yet principled alternative algorithm." PloS one 11(9): e0162259.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20: 53-65.

Sarmento, R. and V. Costa (2017). Comparative Approaches to Using R and Python for Statistical Data Analysis, IGI Global.

Shaheen, S. A. and A. P. Cohen (2008). "Worldwide carsharing growth: An international comparison." Transportation Research Record Journal of the Transportation Research Board 1992 (458718).

Sioui, L., C. Morency and M. Trépanier (2013). "How carsharing affects the travel behavior of households: a case study of montréal, Canada." International Journal of Sustainable Transportation 7(1): 52-69.

Su, T. and J. Dy (2004). A deterministic method for initializing k-means clustering. Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, IEEE.

Trépanier, M., Morency, C., Nouri P., Braham A. (2013), Impacts of carsharing on urban mobility: environmental and behaviourial evidences, 13th World Conference on Transport Research, Rio de Janeiro, Brésil, 15-18 juillet

Vogel, M., R. Hamon, G. Lozenguez, L. Merchez, P. Abry, J. Barnier, P. Borgnat, P. Flandrin, I. Mallon and C. Robardet (2014). "From bicycle sharing system movements to users: a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset." Journal of Transport Geography 41: 280-291.

Wielinski, G., M. Trépanier and C. Morency (2017). Carsharing vs Bikesharing: Comparing Mobility Behaviors.

Xu, J., J. Han, K. Xiong and F. Nie (2016). Robust and Sparse Fuzzy K-Means Clustering. IJCAI.