Public Transport
https://doi.org/10.1007/s12469-022-00305-4

ORIGINAL PAPER

Modulated spatiotemporal clustering of smart card users

Rémi Decouvelaere[1] · **Martin Trépanier**[1] · **Bruno Agard**[1]

Accepted: 11 August 2022

Abstract
Smart card data offers an in-depth understanding of the travel behavior of public transport users. An efficient way to analyze public transport users is to group them into different clusters with similar behaviors. However, this clustering process should take into account space and time because both of these dimensions characterize daily trips. Depending on the outcome, we might wish to give more importance to space or to time, or we might wish to balance the two. In this study, we present a spatiotemporal clustering tool that permits modulation regarding the importance of space versus time. We then test this tool with different values for the space–time balance parameter to evaluate the influence of this parameter on the results. The method has been applied to 769,614 smart card transactions of the *Réseau de transport de la Capitale* (Quebec City, Canada). Results show that the influence of space and time can indeed be controlled, and that the types of clusters obtained vary whether one or both of the dimensions are considered.

Keywords Public transport · Smart card data · Clustering · Spatiotemporal analysis · Travel behavior

1 Introduction

Smart card data has triggered a lot of interest in the past decade, as it has proven to be very useful for different purposes such as long-term planning, network development, and operational management (Pelletier et al. 2011). With the improvement of data processing tools (more computing power and better data mining algorithms),

✉ Martin Trépanier
mtrepanier@polymtl.ca

Rémi Decouvelaere
remi.decouvelaere@polymtl.ca

Bruno Agard
bruno.agard@polymtl.ca

[1] CIRRELT/Polytechnique Montréal, Department of Mathematical and Industrial Engineering, Polytechnique Montréal, C.P. 6079, succ. Centre-Ville, Montréal, Québec H3C 3A7, Canada

Published online: 26 October 2022

many in-depth analyses of this data have been conducted. One particular analysis of public transport smart card data is the study of passengers' travel behaviors, notably by separating users into different groups that have similar behaviors as each other using clustering algorithms, in order to understand the behaviors of each group. Most of the time, only the temporal aspect of passengers' behaviors has been studied (time and number of transactions) during different periods of time (days, weeks, months or years). Yet, other studies have also attempted to take into account the spatial nature of smart card data (place of the transactions). However, the results with spatiotemporal data differ from the results with only temporal data: in studies that only use temporal data, users who live and work in completely different areas can be found in the same cluster, while in spatiotemporal studies, the spatial nature of spatiotemporal data tends to overpower its temporal nature (different groups usually have the same temporal patterns).

Therefore, in this paper, we propose a clustering method where the relative influence of spatial and temporal characteristics can be modulated. This allows this spatial/temporal balance to vary from a temporal model to a strongly spatial model, and it helps to study the variations of the clustering results.

This paper is organized as follows: after a brief literature review on the use of smart card data and clustering methods, we present the spatiotemporal clustering method that we have developed, and finally, we analyze the resulting spatiotemporal behavior of bus users in the city of Québec, Canada, with the different spatiotemporal models that we compare.

2 Literature review

2.1 Use of smart card data in transport analysis

The strength of smart card data is that it provides information about a transportation network not only on a global scale (affluence at each station and on each line) but also on an individual scale. In fact, smart card data can help track the location of a specific user (card) throughout the day, the week, or during a longer period. For this reason, this type of data can truly provide an in-depth analysis of users' behaviors in a transportation network.

Smart card data has been used for many different purposes in the past. (Kurauchi et al. 2017; Pelletier et al. 2011). It can provide information on users' stability and frequency behaviors (Bagchi and White 2005) or about their loyalty (Blythe 2004; Trépanier and Morency 2010). When the information is available, it can provide a demographic profile of the users on each line (Utsunomiya et al. 2006; El Mahrsi et al. 2014; Langlois et al. 2016). It can help study the effect of external factors, such as the weather, on users' behaviors (Arana et al. 2014; Zhou et al. 2017). It can also be used to offer better in-time information on the network to transit users (Ceapa et al. 2012; Yap et al. 2018).

In many cases, smart card users only have to tap their card when they get on a bus, which provides information about boarding, but not when they leave the bus, which means that the destination of their trip is usually not recorded. However, a lot

of work has been done to identify the destinations of these trips, based on transfers and travel habits (Trepanier et al. 2007; Chu and Chapleau 2008). One of the notable difficulties that the authors tackled was to identify the destination of unitary trips (users with only one transaction in a day) (He and Trépanier 2015). In addition to the destination, some authors have also tried to understand the purpose of each trip (Devillaine et al. 2012). The case of multimodal travel (by bus and metro) has been studied in (Seaborn et al. 2009). Smart card data has also confronted other types of data, such as APC (Automated Passengers Counting) and GTFS (General Transit Feed Specification) to improve destination estimation (Giraud et al. 2016), or travel surveys to evaluate the accuracy of such surveys (Kusakabe et al. 2014; Spurr et al. 2014).

2.2 Smart card data clustering

An efficient way to understand the behavior of transit users is to group them into different clusters of similar patterns. This has been done at different levels of time (days, weeks, or months) to identify the temporal patterns of user behaviors on different scales, and to measure the evolution of these behaviors (Asakura et al. 2012; Trépanier 2012). The influence of the fare type on the behaviors of smart card users has also been studied many times (Morency et al. 2007; El Mahrsi et al. 2014; Nishiuchi et al. 2013).

Smart card user clustering studies differ by the type of clustering algorithms and metrics used. K-means (Morency et al. 2007; Trépanier, 2012; Zhao et al. 2017), hierarchical clustering (Langlois et al. 2016; He et al. 2018a, b; 2019; Faroqi et al. 2019) and DBSCAN (Kieu et al. 2014; Ma et al. 2013) are among the most frequently used algorithms. While k-means is faster, hierarchical clustering offers more freedom and a more detailed analysis. Different metrics were tested to evaluate the similarity between different users' temporal behaviors, such as Hamming distance, Euclidean distance, and so forth. Yet, such metrics are not well fitted for time series, as they did not take into account the order (sequence) of the data in vectors. Therefore, different types of time metrics were tested on smart card data to find the most efficient one. The DTW (Dynamic Time Warping) and CCD (Cross-Correlation Distance) were tried on smart card temporal data clustering (He et al. 2018a), and some new metrics were even designed specifically for smart card temporal data (Ghaemi et al. 2017).

In most of the studies, smart card users are clustered based on their temporal patterns only (time of transaction, duration of trips and transfers, etc.). Some authors have also tried to cluster users based on spatial data (position of users) and have compared them with temporal clustering (Zhao et al. 2017). Yet, only a few authors have tried to simultaneously use spatial and temporal data for smart card user clustering. In other fields, such as image processing, traffic management, or seismology, spatiotemporal clustering has been used in many different ways and gave positive results most of the time (Ansari et al. 2019). Li He (He et al. 2018b) used a DTW metric to compare a spatiotemporal series, representing the position of a user throughout the day based on his smart card data. Using this metric, he managed

to create a spatiotemporal clustering of smart card users in Gatineau. In (Faroqi et al. 2019), a temporal metric and a spatial metric are created to compare smart card users in Brisbane, and these users are clustered in three different ways: spatial clustering, then temporal clustering (S-T); temporal clustering, then spatial clustering (T-S), and simultaneous spatial and temporal clustering (ST). The results show that ST clustering gives more robust clusters than T-S or S-T clustering (Faroqi et al. 2019).

Yet, in all of these studies, the relative weight of space and time in spatiotemporal clustering cannot be monitored. The aim of this study is to propose a method to modulate space and time influence in smart card data clustering.

3 Methodology

This section presents the methodology of this study. Please refer to Fig. 1 for the steps explained in the following text.

3.1 Data preparation

The data used for this study is provided by the RTC public transit agency (*Réseau de Transport de la Capitale*). RTC is based in Québec City, Canada, and operates a 616-bus network of 140 bus lines. The dataset contains all of the smart card transactions in this network from June to December 2019, which represent a total number of 20,324,706 transactions. Because users in Québec City only need to validate the card when they get on the bus, each transaction either corresponds to the beginning of a trip or to a transfer. The information associated with each transaction includes the hour and place (bus stop) of the transaction, the type of ticket used, the type of trip (first trip or transfer) and the estimated destination of the bus trip.

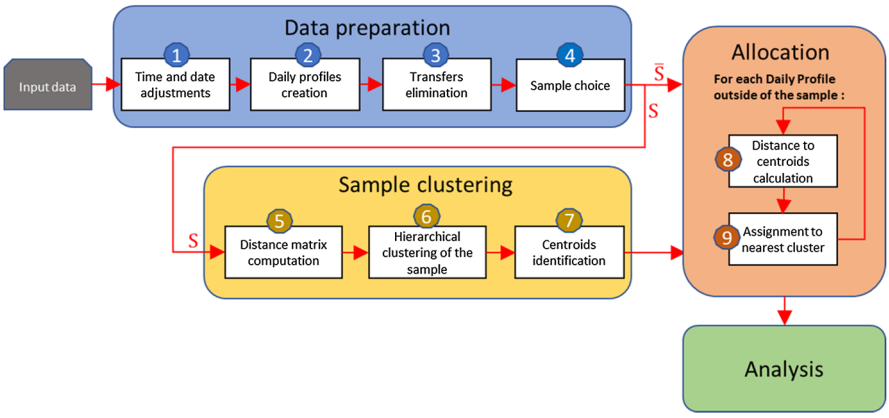

Fig. 1 Overall methodology

3.1.1 Time and date adjustments

A small number of transactions take place between midnight and 4 a.m. (around 0.7% of all transactions on a typical weekday). Almost no transactions (less than 10 per day) take place between 3:30 and 5:00, because there is no service. Therefore, we decide to use this time period to separate temporal patterns from one day to the other. Instead of ending each day at midnight, we start each day at 4:00 and finish it at 3:59 the following day (or 27:59). Then, we assume that for each user, the final transaction before 4 a.m. is always a return home. This way, no round-trip may be cut in two, and the user's home should be correctly identified. We then select only the transactions of the first workweek of December 2019 (from Monday, December 2nd, 2019, 4:00 to Friday, December 6th, 2019, 27:59). The aim is to identify behavior related to work commutes, hence the choice to not use weekends. This represents a total of 769,615 transactions.

3.1.2 Daily profile creation

In Step 2, we create a list of all the transactions for each day and for each card: we call this a daily profile of a card-day. Each daily profile is identified by a key formed by the combination of the user's card ID and the date (for example 1746200_2019-12-04, for card number 1746200 on 2019 December the 4th). In this case, we assume that each card corresponds to a single user and that a user always uses the same card. This is a good assumption for public transport users in the province of Quebec, since many users have a picture on their card, and most cards are associated with monthly or annual passes. Using the 769,615 transactions previously selected, 154,055 Daily profiles are created.

Multimodal travelling may be quite important in some cities: for example, someone might go to work by bus, and go home on foot or by carsharing. In such cases, only the trips by bus can be detected: all other kinds of trips are described as "undetected trips." Some of the Daily profiles (around 32% of the previously selected profiles) only have one transaction, which means they contain undetected trips (assuming that users always depart from home on their first trip and return home on their last trip). These daily profiles have been removed from the database because of missing information (difficulty of knowing whether a trip is to or from home, unknown time spent in each location, etc.) We end up with a total of 105,381 daily profiles describing the daily behavior of 54,753 different card users from Monday, December 2nd, 2019 to Friday, December 6th, 2019. (Daily profiles describing the same card user over different days are not regrouped; therefore, the card id no longer plays an important role).

3.1.3 Transfer elimination

A list of 24 bus stops that match the user's position every hour is created for each Daily profile. Whenever the user makes one transaction, his location is changed to the destination of this trip until the next transaction. The last location of a day has to match the first location (depart from and return to home). For each daily profile A,

we create the following series V^A, that records the location of user A at every hour and whether user A has made a transaction during the hour (1 if yes, 0 if no):

$$\forall i \in [\![T_S, T_E]\!], \begin{cases} s_i^A \ : \ stop\ location\ of\ user\ A\ at\ time\ i \\ n_i^A \ : \ transactions\ of\ user\ A\ during\ hour\ i\ (0\ or\ 1) \\ V_i^A = \left(s_i^A, n_i^A\right) \end{cases} \quad (1)$$

where T_s is the start time of the day and T_E the end of the day. In this study, T_s is set at 4:00 AM and T_E at 3:00 AM the next day (i.e., 27:00).

For example:

On day 1, user A boarded at 7:41 in stop 1362, and came back at 17:03 from stop 1561:

$$\begin{array}{lccccccccccc} & 4{:}00 & 5{:}00 & 6{:}00 & 7{:}00 & \ldots\ldots & 16{:}00 & 17{:}00 & \ldots\ldots & 27{:}00 \\ s_i^A = & [1362, & 1362, & 1362, & \mathbf{1561}, & \ldots\ldots, & 1561, & \mathbf{1362}, & \ldots\ldots, & 1362] \\ n_i^A = & [\ 0, & 0, & 0, & \mathbf{1}, & \ldots\ldots, & 0, & \mathbf{1}, & \ldots\ldots, & 0\ ] \end{array}$$

$$V^A = [(1362,0), \quad \ldots \quad , (\mathbf{1561}, \mathbf{1}), \ldots, (1561,0), (\mathbf{1362}, \mathbf{1}), \ldots, (1362,0)]$$

This spatiotemporal vector characterizes a user profile and is later used to compute the distance between each profile. Only the origin and the final destination of trips are relevant information in our study. Therefore, in each daily profile, transfer transactions are removed (Step 3). Transfers are identified in the following way: if two transactions are made in less than 90 min, the second one is identified as a transfer.

3.2 Sample clustering

Applying our clustering algorithm to all data is time-consuming, so we propose a sampling approach.

3.2.1 Sample choice

In Step 4, a random sample is taken from the 105,381 selected daily profiles. The size of the sample is set to 2000, with regards to previous work (He et al. 2018b) that showed that small sample sizes can be used for such a study. In public transport, sampling size can be very small because the behavior (and the nature of the data vector) are not so different among passengers (people often travel at common times and in common directions); this is validated with a sensitivity analysis that we conducted around this value. In fact, the most time-consuming operation of the clustering algorithm is the computation of the distance matrix, whose complexity is proportional to N^2, where N is the number of points being clustered. This justifies the choice of using a sample to reduce the computation time by reducing N (see, for example, the work by He et al. 2019). With a sample size of 2000, the total

run time is approximately 3 h; with a sample size of 3000 around 6h30; and with a sample size 4000 around 12 h. To keep the computation time under 3 h, we used a sample size of 2000. This sample data forms a dataset S, and the rest of the profiles (103,381) are set aside in a dataset $\overline{S}$, later assigned to the clusters that come from the sample. All of the steps in the sample clustering part are applied to dataset S only.

3.2.2 Distance matrix computation

We calculate a matrix of the distances between each pair of daily profiles (Step 5). The metric used to compare two daily profiles is based on the DTW (Dynamic Time Warping) metric.

3.2.2.1 Traditional dynamic time warping The principle of dynamic time warping is to measure the similarity between two time series by matching the points of the two series in a way that minimizes the distances between each point. Some rules have to be followed:

- the first point of the two series must be matched together, as well as the last point;
- the points must be matched in chronological order (sequence);
- one point can be matched to multiple points if these points are not already matched;
- a window parameter gives the maximum authorized chronological offset between two matched points (for example, if the window is 2, a point of series X at time 8:00 can be matched to points from series Y at time 6:00, 7:00, 8:00, 9:00, or 10:00, but not to other points);
- the way in which the points are matched needs to minimize the sum of the distances between all matched points.

The distance between two points depends on the metric chosen (usually Euclidean distance).

The DTW distance returns the sum of the distances between all matched points. The closer the shape of the two series is, the smaller their DTW distance will be.

For example, if we take the following X and Y time-series, $X=\begin{pmatrix}0\\1\\2\\3\\4\end{pmatrix}$ and $Y=\begin{pmatrix}0\\0\\0\\0\\1\end{pmatrix}$.

With a window of 2, and a Euclidean metric, the optimal DTW solution will be the following (Fig. 2). Point 7:00 of series X is matched to points 7:00, 8:00 and

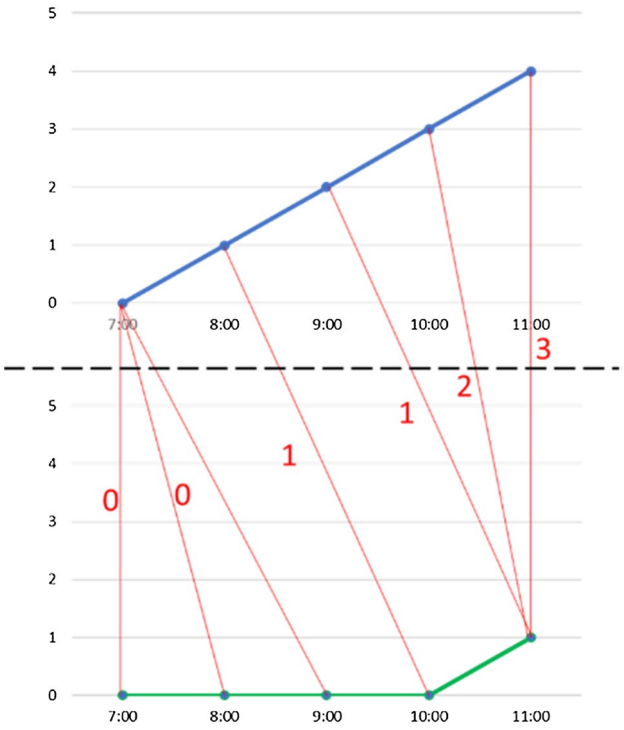

Fig. 2 Basic DTW (X series is the upper part, Y series the lower part)

9:00 of series Y (not to 10:00 because the window parameter is only 2), point 8:00 of series X is matched to point 10:00 of series Y, and points 9:00, 10:00 and 11:00 are matched to series Y's last point.

In this case, the DTW distance between series X and Y is:

$$\begin{aligned} &DTW(X, Y, \text{metric} = euclidean,\ \text{window} = 2) \\ &= 0 + 0 + 0 + 1 + 1 + 2 + 3 = 8 \end{aligned}$$

3.2.2.2 Adaptation of traditional DTW We propose adapting this traditional DTW method to permit a modulation of time and space.

For a daily profile A, we define the following series V^A:

$$\forall i \in [\![T_S, T_E]\!], \begin{cases} s_i^A : \textit{stop location of user A at time } i \\ n_i^A : \textit{transactions of user A during hour } i\ (0 \textit{ or } 1) \\ V_i^A = \left(s_i^A, n_i^A\right) \end{cases} \tag{1}$$

We use the notation: $\begin{cases} V_i^A[1] = s_i^A \\ V_i^A[2] = n_i^A \end{cases}$

We define the same variables for another profile B:

$$\forall j \in [\![T_S, T_E]\!], \begin{cases} s_j^B : \textit{stop location of user B at time j} \\ n_j^B : \textit{transactions of user B during hour j} \ (0 \textit{ or } 1) \\ V_j^B = \left(s_j^B, n_j^B\right) \end{cases} \tag{2}$$

For the example, we take the User A profile presented before and we add User B:

User B boarded between 6:00 and 6:59 in stop 1362, and came back between 16:00 and 16:59 from stop 4111:

	4:00	5:00	6:00	... ...	15:00	16:00	... ...	27:00
$s^B =$	[1362,	1362,	**4111**,	... ...	, 4111,	**1362**,	... ...	,1362]
$n^B =$	[0,	0,	**1**,	... ...	, 0 ,	**1**,	... ...	, 0]

$$V^B = [(1362,0), \ldots, (\mathbf{4111, 1}), \ldots, (4111,0), (\mathbf{1362, 1}), \ldots, (1362,0)]$$

We then define $E\left(s_n, s_m\right)$ as the Euclidean distance between two stops s_n and s_m:

$$\forall n, m \in [\![1\,, N]\!], \begin{cases} E\left(s_n, s_m\right) = \sqrt{\left(x_n - x_m\right)^2 + \left(y_n - y_m\right)^2} \\ \left(x_n, y_n\right) : \textit{spatial coordinates of stop } s_n \\ \left(x_m, y_m\right) : \textit{spatial coordinates of stop } s_m \end{cases} \tag{3}$$

N being the total number of bus stops in Québec City.

We choose the Euclidean distance because it is independent from the nature and the geometry of the public transport and road networks. Based on this distance, we define the following spatial metric:

$$\forall i, j \in [\![T_S\,, T_E]\!], \begin{cases} S\left(V_i^A, V_j^B\right) = \frac{E\left(s_i^A, s_j^B\right)}{E_{max}} \\ E_{max} = \max\limits_{n,m \in [\![1\,, N]\!]} E\left(s_n, s_m\right) = 35162\,m \end{cases} \tag{4}$$

E_{max} *is the largest distance between two bus stops in Québec City.*

And we define the following temporal metric that determines if transactions are made at the same time in the two profiles (equal to 0 if so, and 1 if not):

$$\forall i, j \in [\![T_S\,, T_E]\!], T\left(V_i^A, V_j^B\right) = \left|n_i^A - n_j^B\right| \tag{5}$$

Based on these last two metrics, we define the following spatiotemporal metric:

$$\forall i, j \in [\![T_S\,, T_E]\!], \begin{cases} M\left(V_i^A, V_j^B\right) = \alpha * S\left(V_i^A, V_j^B\right) + (1 - \alpha) * T\left(V_i^A, V_j^B\right) \\ \alpha \in [0, 1] \end{cases} \tag{6}$$

The choice of α balances the comparative weights of the spatial and temporal factors of the equation.

For example, if we want to calculate the spatiotemporal distance $M\left(V_i^A, V_j^B\right)$ between profile A at $i = 5{:}00$ and profile B at $j = 7{:}00$:

- At 5:00, user A is at stop 1362, and at 7:00 user B is at stop 4111. So:

$$S\left(V_i^A, V_j^B\right) = \frac{E(1362, 4111)}{\boldsymbol{E_{max}}} = \frac{10065\,m}{35162\,m} = 0.2862$$

- User A does not make any transactions between 5:00 and 5:59 and user B makes one transaction between 7:00 and 7:59. Therefore:

$$T\left(V_i^A, V_j^B\right) = |0 - 1| = 1$$

If we choose $\alpha = 0.25$:

$$M\left(V_i^A, V_j^B\right) = \alpha * 0.2862 + (1 - \alpha) * 1 = 0.25 * 0.2862 + 0.75 * 1 = 0.82155$$

The final metric D used to compare series V^A and V^B is a DTW distance that uses $M\left(V_i^A, V_j^B\right)$ as a metric, with a window parameter of 2:

$$D\left(V^A, V^B\right) = DTW\left(V^A, V^B, \text{metric} = M\left(V_i^A, V_j^B\right), \text{window} = 2\right) \tag{7}$$

This means that the points of series V^A and V^B will be matched together in a way that minimizes the distance between two matched points, this distance being calculated with metric $M\left(V_i^A, V_j^B\right)$. Every point has to be matched to a point of the other series (following chronological order), and the first points of V^A and V^B have to be matched together, as well as their last points. Two matched points cannot be separated by more than two hours ($|i - j| \leq 2$). Following all of these rules, the DTW algorithm finds a combination of matched points that minimizes the sum of the distances between every pair of matched point. The distance $D\left(V^A, V^B\right)$ returned will be equal to this sum:

$$D\left(V^A, V^B\right) = \sum_{i,j} M\left(V_i^A, V_j^B\right) \tag{8}$$

Every pair (i, j) symbolizing a couple of matched points.
This can be more rigorously defined as:

$$D\left(V^A, V^B\right) = \min_{K \in \mathbb{N}} \sum_{k=1}^{K} M\left(V_{ik}^A, V_{jk}^B\right) / \begin{cases} \forall k_1, k_2 \in [\![1\,,\,K]\!], i_{k_1} \leq i_{k_2} \text{ and } j_{k_1} \leq j_{k_2} (\textit{Chronological order must be respected}) \\ i_1 = j_1 (\textit{The first points of the two series have to match}) \\ i_K = j_K (\textit{The last points of the two series have to match}) \\ [\![T_S\,,\,T_E]\!] \subset \{i_k, k \in [\![1\,,\,n]\!]\} (\textit{all the points in series A have to be matched}) \\ [\![T_S\,,\,T_E]\!] \subset \{j_k, k \in [\![1\,,\,n]\!]\} (\textit{all the points in series B have to be matched}) \\ \forall k \in [\![1\,,\,K]\!], |i_k - j_k| \leq 2 \, (\textit{maximum window between two matched points}) \end{cases} \tag{9}$$

For instance, in the previous example for a window parameter of 2, the optimal combination of matched points will be the following (Fig. 3).

In this case: for $\alpha = 0.25$:

$$D\left(V^A, V^B\right) = 3 * 0 + 10 * \left(\alpha * \frac{E(1561, 4111)}{\mathrm{E}_{\max}} + (1 - \alpha) * 0\right) + 12 * 0 = 10 * 0.25 * \frac{10357\,m}{35162\,m} = 0.7364$$

3.2.3 Hierarchical clustering of the sample

Based on the distance matrix of the sample, it is now possible to classify Daily profiles from the sample into different groups (Step 6). We used HCA (Hierarchical Clustering Algorithm), with the median *ultrametric*, as a clustering algorithm ("Statistics toolbox." API Reference Documentation. The MathWorks.) This algorithm creates recursive groups that embed each other, which means that any group can be repetitively divided into subgroups until groups of only one profile are created.

3.2.4 Choice of the number of clusters

A specific number of clusters must then be chosen (Step 7), by fixing a maximal distance between two profiles of the same group. For each value of **α**, different numbers of clusters have been tested: the optimal number of clusters has been chosen with regards to the evolution of the Silhouette score (Rousseeuw 1987).

3.2.5 Centroids identification

Once the sample is divided into clusters, the centroid of each cluster is then calculated (Step 8)**.** For a given cluster, the centroid is the point of this cluster that minimizes the average distance to all the other points of the cluster. The centroids are

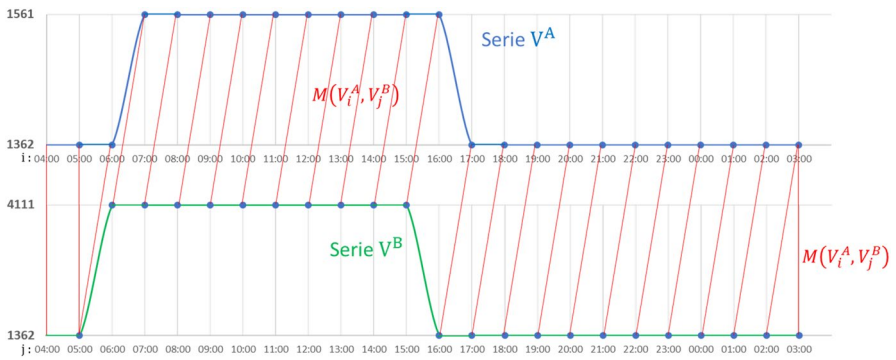

Fig. 3 Proposed DTW algorithm applied to an example

used instead of mean values, because the latter cannot be calculated with the metric used, as this metric does not satisfy the triangle inequality.

3.3 Allocation of remaining profiles

At the end of the sample clustering, every daily profile in the sample is assigned to one cluster. Yet, all the daily profiles from dataset $\overline{S}$ (from step 4) still need to be assigned to a cluster. To find out the closest cluster to a given daily profile, we calculate the distances between this profile and each centroid (Step 9), using the same metric as in Step 5. The cluster with the closest centroid is then chosen, and the profile is then assigned to this cluster (Step 10). At the end of this stage, every daily profile has been assigned to one cluster.

3.4 Analysis

The results are then analyzed. Different types of visualization tools have been used to study the characteristics of each cluster.

4 Results

Three spatiotemporal studies are presented in the following section to highlight the influence of the balance of time and space in the clustering. Three values of alpha are tested on one week of bus trips in Quebec City. The clustering and resulting analyses are presented. Steps 1 to 4 are identical, so S and $\overline{S}$ do not vary. For each alpha, distances from step 5 need to be calculated, and as a consequence the following steps are recomputed, providing different conclusions.

4.1 Alpha = 0.5

We set alpha = 0.5 in Eq. (6) and compute the hierarchical clustering of step 6. We try different numbers of clusters for the hierarchical clustering of the sample and we calculate in each case the average Silhouette score of the clusters. The Silhouette scores for various numbers of clusters is presented in Fig. 4.

At alpha = 0.5, two local maxima are noted for 4 clusters and 7 clusters. However, 4 clusters are not enough to provide a detailed study for the operator, so we chose 7 clusters. This value also seems suitable for alpha = 0 and alpha = 1. We decide to choose the same number of clusters for these values of alpha for comparison purposes, as explained later.

We first study the spatial characteristics of each cluster by plotting, for every bus stop, the cluster with the highest number of departures at that stop (Fig. 5): for each bus stop, we calculated the number of users in each cluster departing from this stop and for each stop we selected the cluster with the highest number. It is clear, according to this map, that the spatial characteristic plays an important role in the classification.

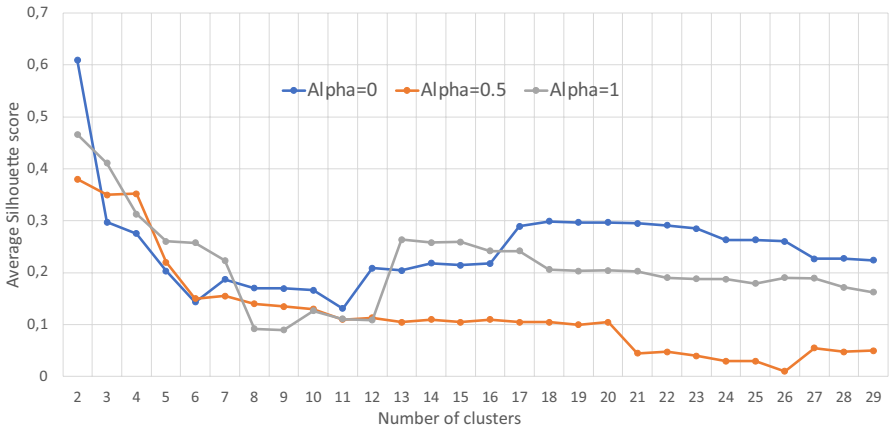

Fig. 4 Silhouette scores for alpha = 0, 0.5 and 1

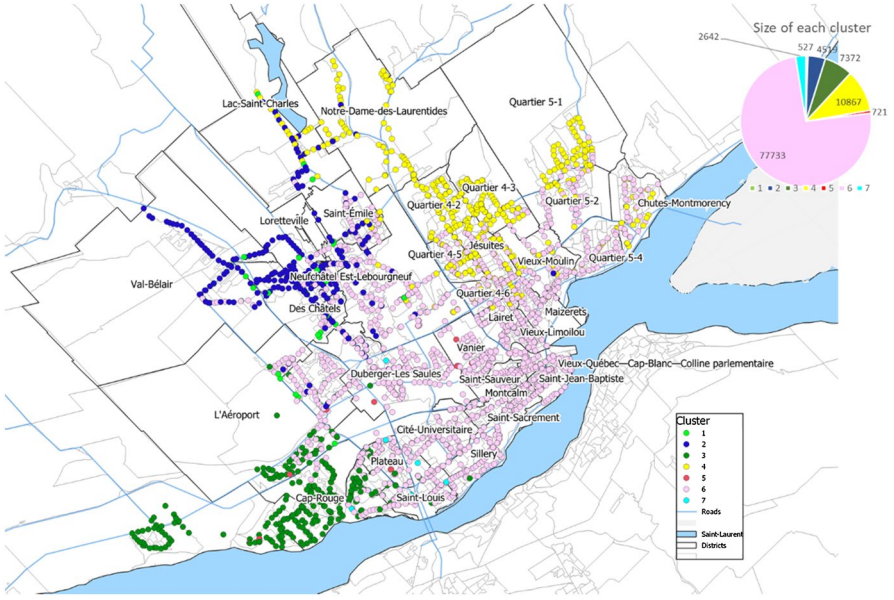

Fig. 5 Most frequent cluster at every departure stop for alpha = 0.5 and 7 clusters

In the four largest clusters (2, 3, 4 and 6) we observe that each has a specific departure zone: the central area of Québec for cluster 6 and the off-center zones for 2, 3 and 4. However, the departure zones of smaller clusters 1, 5, and 6 overlap with the departure zones of the bigger clusters, which suggest that their temporal features differ from those of the bigger clusters.

The next map (Fig. 6) gives the average distance and direction of the first trips of the daily profiles in each cluster: the first trip in a daily profile is the trip between the

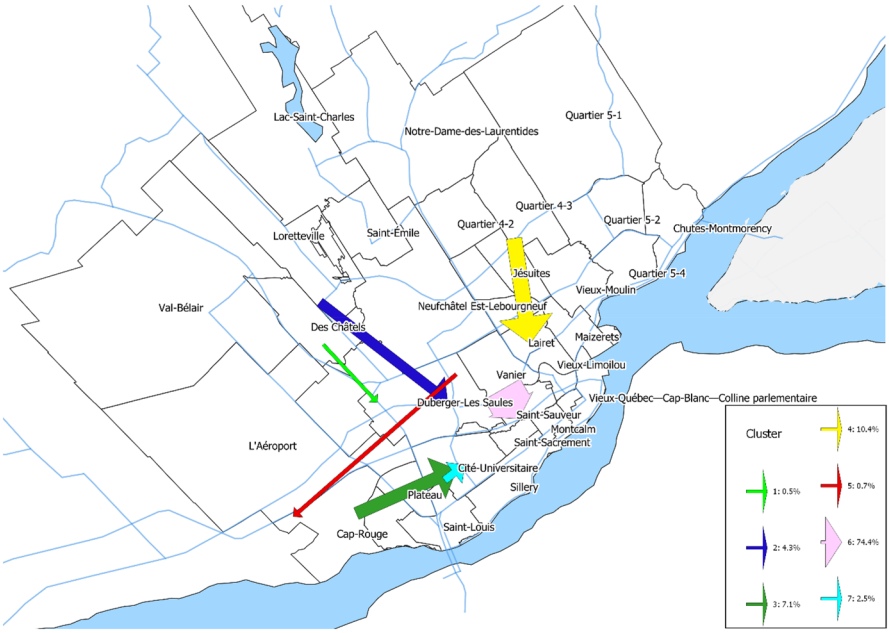

Fig. 6 Average distance and direction of the first trips in every cluster (alpha = 0.5)

first location of the day and the second location. The size of the arrow depends on the size of the cluster: the percentage of the number of daily profiles in each cluster is represented next to each arrow in the legend. This map reveals a unique aspect of cluster 5: unlike other clusters, its users do not travel to one of the two big attractions in Québec City (Cité-Universaire and the downtown area, Saint-Jean-Baptiste), but instead travel to a small attractor in Cap-Rouge.

We then study the temporal patterns in each cluster by plotting the total number of transactions every hour (Fig. 7).

The most common pattern for a RTC user is the one in cluster 6: a departure around 7:00 and a return home around 16:00. Clusters 2, 3, 4 and 5 follow this pattern with a small peak of transactions around lunch time for clusters 2, 3 and 4. As for clusters 1 and 7, they strongly differ from the other clusters on the temporal point of view. Cluster 1 presents one transaction peak between 14:00 and 16:00 and a second peak around 21:00, which suggests people from this group leave home between 14:00 and 16:00 and return home at night (around 21:00). As for cluster 7, its temporal pattern is quite similar to cluster 1, but its users leave home a few hours earlier (between 11:00 and 13:00).

All of this information can be visualized on a 3D path graph (one example at Fig. 8), which represents the average position of each cluster throughout time (each dot on a curve corresponds to one hour, from 4:00 to 27:00).

Apart from cluster 5, all clusters converge towards the center of Québec, but while clusters 4 and 6 are directed towards downtown (Saint-Jean-Baptiste district), clusters 1, 3 and 7 are instead directed towards the university (Cité-Universitaire). To

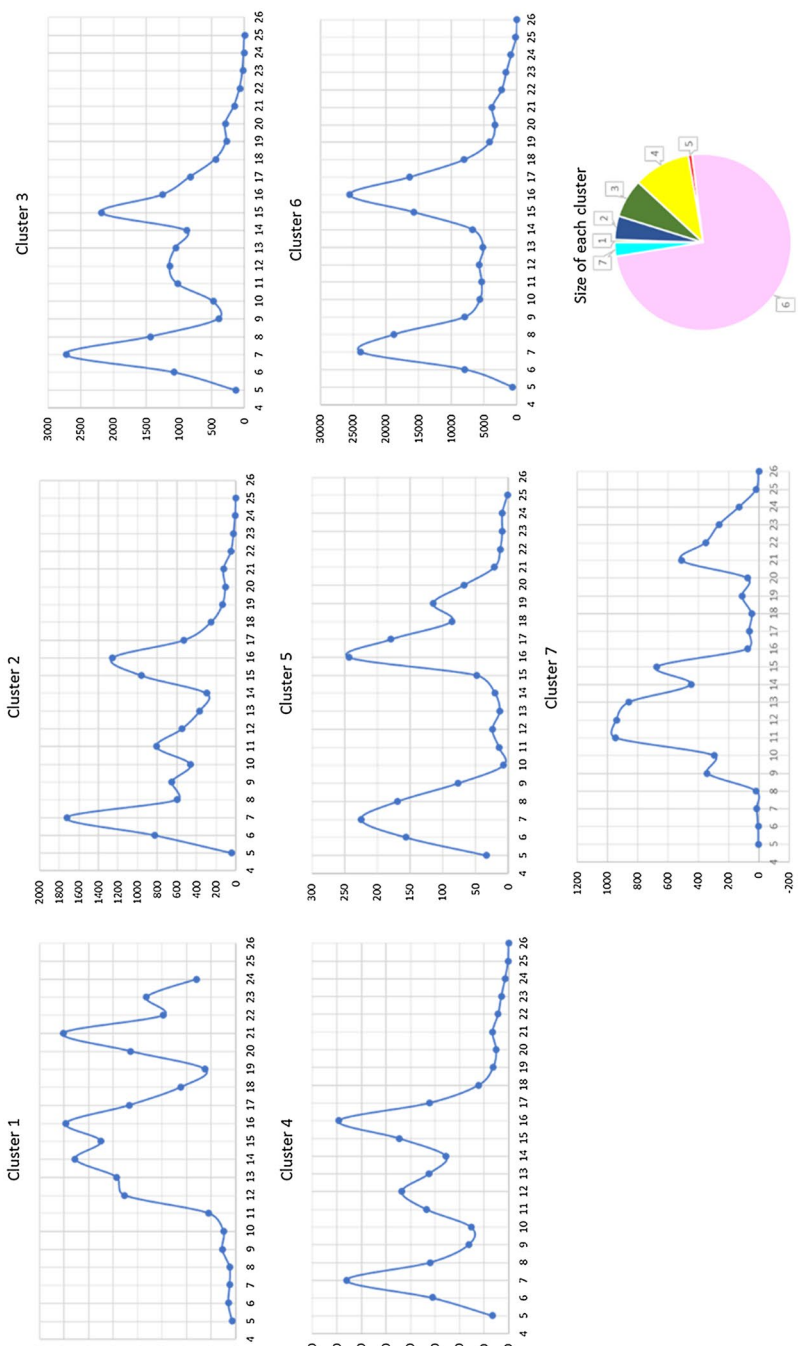

Fig. 7 Number of transactions every hour in each cluster (alpha = 0.5)

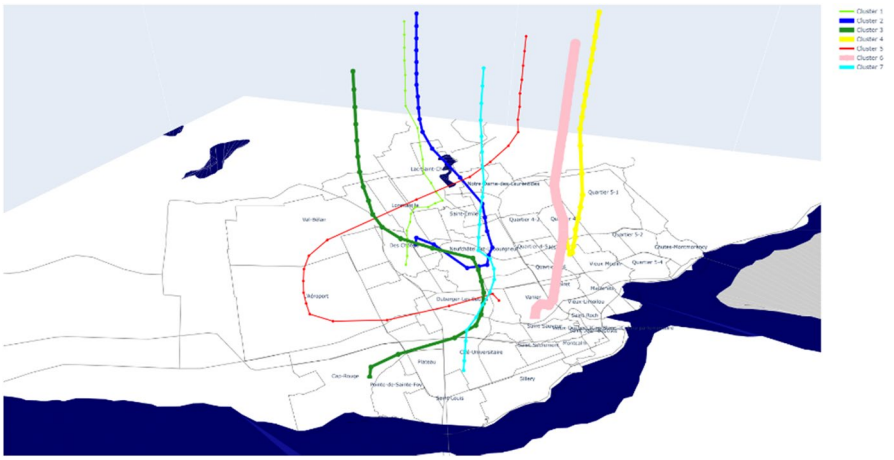

Fig. 8 An example of 3D-path graph

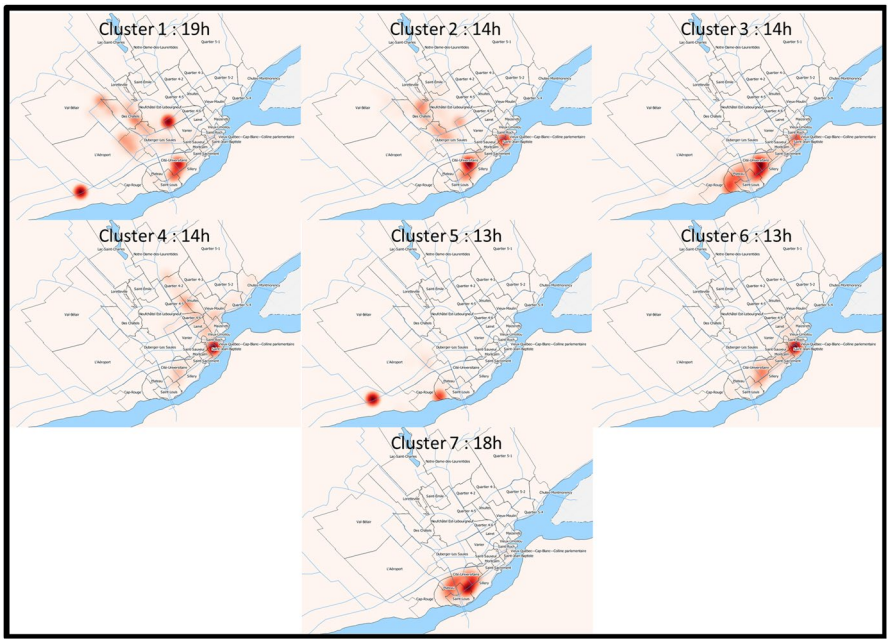

Fig. 9 Heat maps of the users of each cluster at different hours

more precisely identify the destination of the users in each cluster, for each cluster we plotted a heat map of its users at the time between the two main transaction peaks of this cluster (Fig. 9). We noticed two main attractors, Saint-Jean-Baptiste (Quebec City's business district) and Cité-Universitaire (the university). Some clusters (cluster 4 and 6) work almost exclusively in Saint-Jean-Baptiste, while others (cluster 2 and

3) work both in Saint-Jean-Baptiste and in Cité-Universitaire. As for cluster 7, its users mainly go to the university. The smallest clusters (1 and 5) highlight the existence of another small attractor to the outside of Québec City (West of Cap-Rouge): a business district in the small town of Saint-Augustin-de-Desmaures. Although cluster 5 travels almost exclusively to this place, cluster 1 also travels to Cité-Universitaire.

To summarize, with alpha = 0.5, we created 7 clusters with different temporal and spatial features. Cluster 6 represents the most common profile in Québec City: a departure around 7 a.m. to go to work downtown and a return around 4 p.m. Cluster 4 has essentially the same features, except its users live much further from downtown, in the northern and eastern ends of Québec City. Clusters 2 and 3 are also temporally similar to these clusters, since they also depart around 7 a.m. and go home around 4 p.m.; however, they not only work downtown but also at the university (Cité-Universitaire). We can also note in clusters 2, 3 and 4 a small peak of transactions that are made around lunch time, which suggest that some people in this cluster only work in the morning or in the afternoon. As for cluster 5, it also has the same temporal pattern but goes to a completely different place in the western end of Québec City. Finally, clusters 1 and 7 are temporally very different from the other clusters. Cluster 7, composed of a high number of students, depart between 9 a.m. and 1 p.m. and go home late (between 8 p.m. and 1 a.m.): one hypothesis could be that these are students who work all afternoon and evening at the university.

4.2 Alpha = 1

We select the same number of clusters for alpha = 1 as well as for alpha = 0, to be able to compare the balance between space and time in each of the three situations. Then we draw similar maps and graphs for alpha = 0.5. Figures 10, 11, 12 will focus on alpha = 1.

With alpha = 1, we observe that some clusters are similar to the previous clusters of alpha = 0.5:

- Cluster 2 represents a large group of people who live and work near Quebec City's center with typical work hours (a departure around 7:00 and return around 16:00), and can easily be identified with cluster 6 of alpha = 0.5
- The users in clusters 3 and 4 both work downtown and have the same working hours as cluster 2, but they live in remote areas of Quebec City: east, towards Chutes-Montmorency for cluster 3 and north, towards Notre-Dame-des-Laurentides for cluster 4. These two clusters are merged into one single cluster of alpha = 0.5 (cluster 4 of alpha = 0.5), which shows that space plays a more important role in the present case of alpha = 1 than in the previous case of alpha = 0.5.
- Cluster 5 can be identified with cluster 3 of alpha = 0 and represents people who live in the remote western end of Québec City (near Cap-Rouge), and who work or study near the University.
- As for cluster 6, it is then only the "temporal" cluster of alpha = 1, in the sense that its temporal pattern is then only one different from the others. This represents a group of people who live near Des-Châtels and the airport, and who travel all around the day and in the evening (with a small peak at 21:00). It is somewhat similar to cluster 1 of alpha = 0.5.

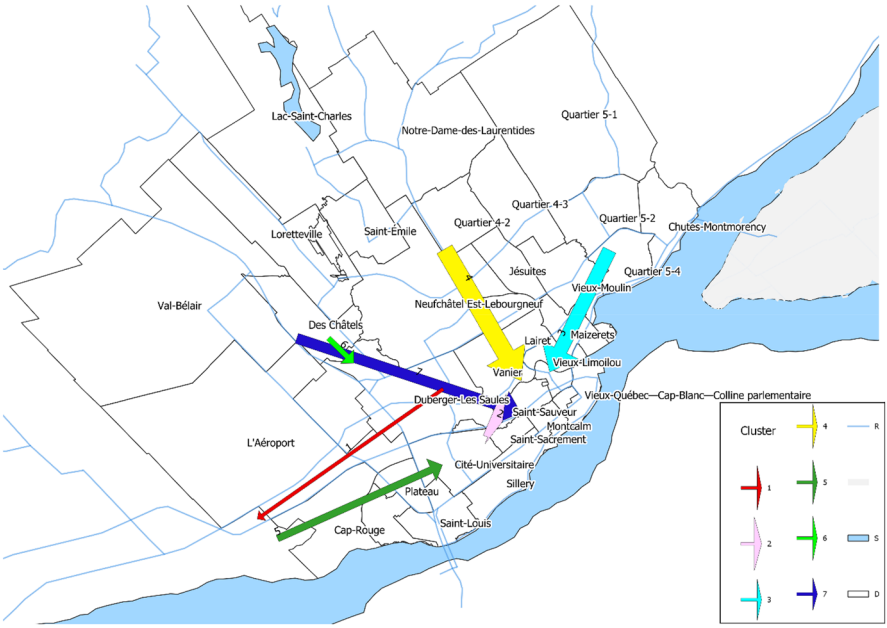

Fig. 10 Average distance and direction of the first trips in every cluster (alpha = 1)

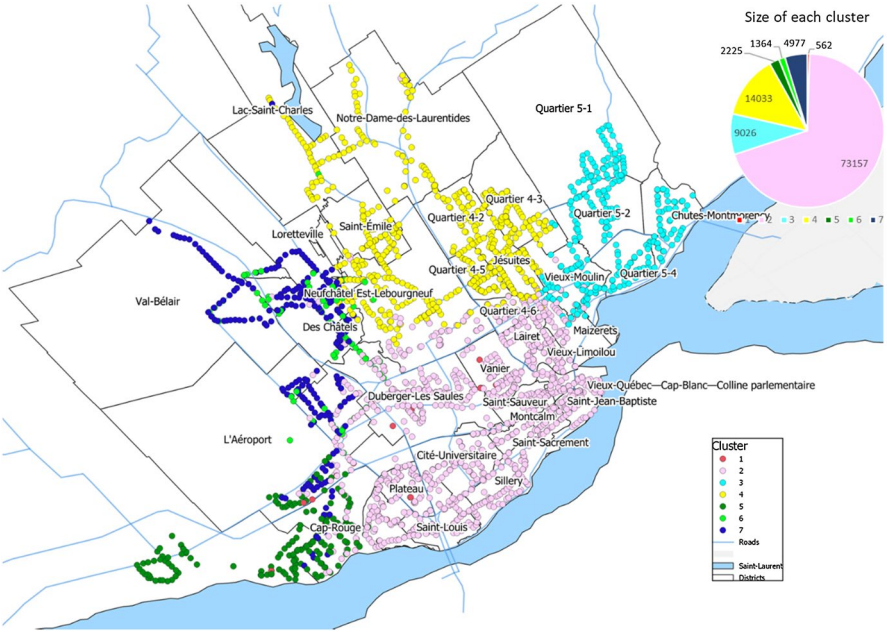

Fig. 11 The most frequent cluster at every departure stop for alpha = 1 and 7 clusters

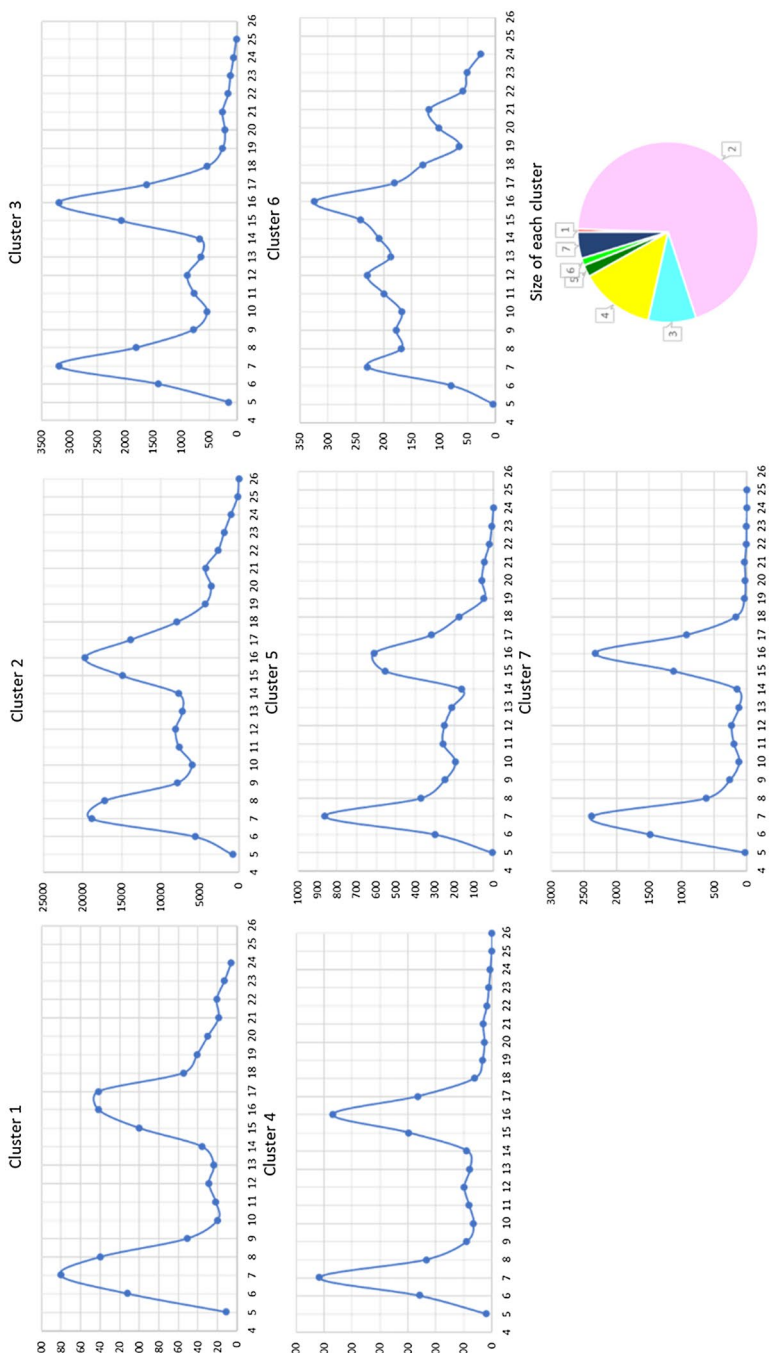

Fig. 12 Number of transactions every hour in each cluster (alpha = 1)

– Finally, cluster 7 represents the people who live in the remote area of Val-Bélair and who work downtown or at the University: it is similar to cluster 2 of alpha = 0.5 with, however, a larger proportion of people who work downtown compared to at the University.
– We can also observe that cluster 7 of alpha = 0.5, which represents a group of students with late working hours, is no longer present for alpha = 1.

To summarize, these results are close to those of alpha = 0.5 with two major exceptions: one "temporal cluster" has been removed (cluster 7 of alpha = 0.5) and one "spatial cluster" has been split into two "spatial clusters" (clusters 3 and 4 of alpha = 1). Overall, the temporal patterns of every cluster differ less for alpha = 1 than for alpha = 0.5 with only one cluster with a slightly different temporal pattern (cluster 6 of alpha = 1). On the other hand, the clusters are more balanced, and the geographical areas of the clusters are more precise. It is clear that space has a greater influence on these results than on the results of alpha = 0.5, yet time still has a small influence.

4.3 Alpha = 0

Now, we set alpha = 0 in Eq. (6) and run all steps as for previous cases. To maintain comparability, we still select a number of clusters equal to 7 for alpha = 0. We then obtain the following map (Fig. 13) and graph (Fig. 14):

With alpha = 0, we observe that the spatial differentiation between the clusters is completely gone. This can be noticed on Fig. 14: it is now impossible to attribute

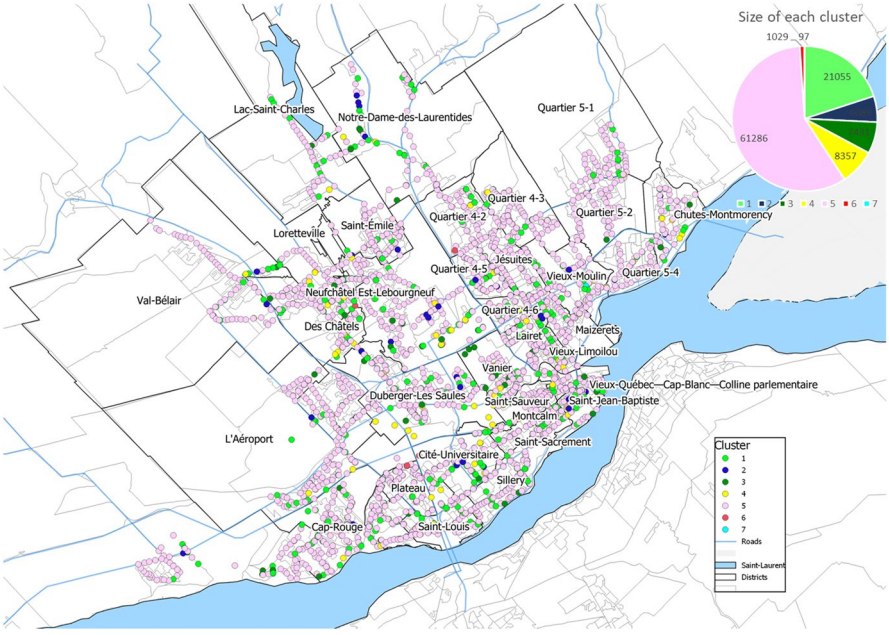

Fig. 13 The most frequent cluster at every departure stop for alpha = 0 and 7 clusters

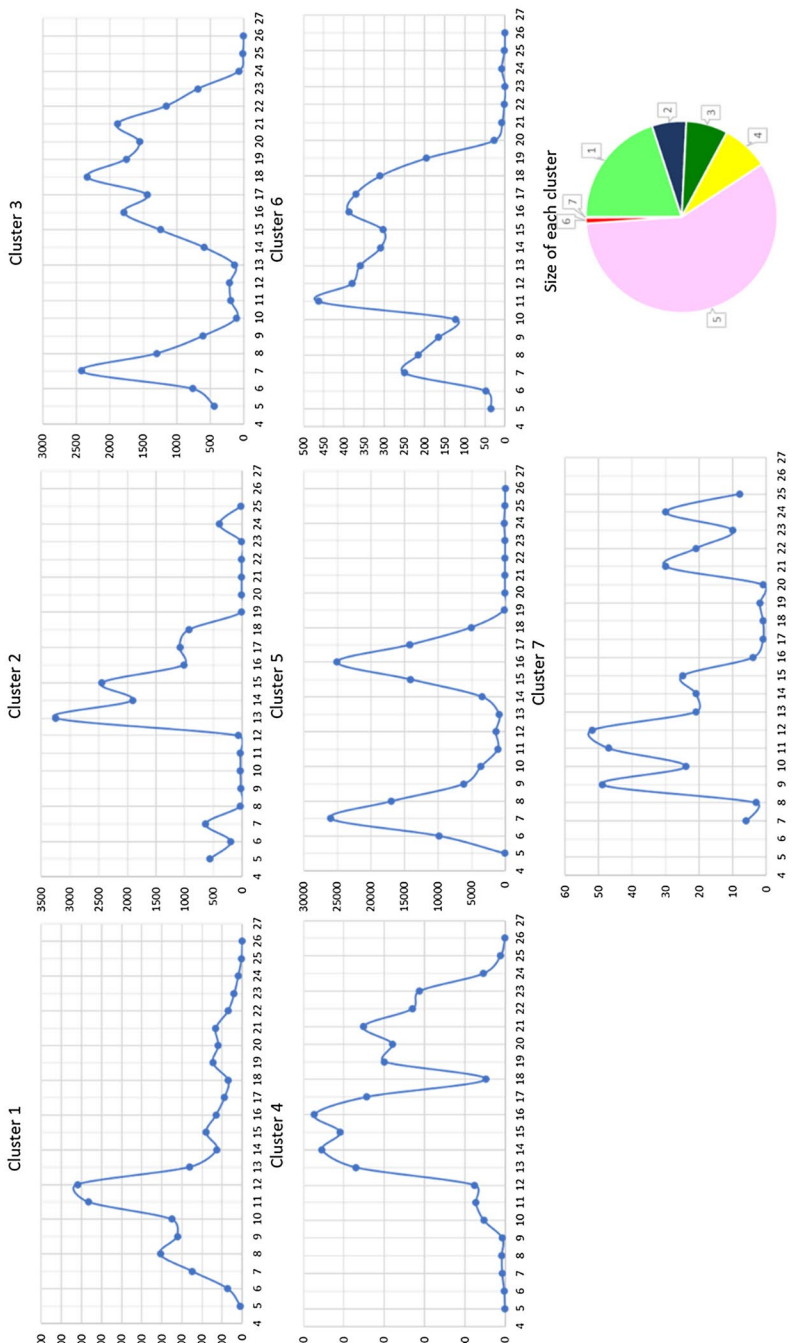

Fig. 14 Number of transactions every hour in each cluster (alpha = 0)

a specific area to any of the seven clusters. However, each cluster now presents a very distinct temporal pattern. The larger cluster, cluster 5, corresponds to the classic working pattern that we also find in alpha = 0.5 and alpha = 1: a departure around 7:00 and a return home around 16:00.

5 Conclusions

In this study, we have proposed a modulated spatiotemporal clustering tool that balances the relative influence of space and time on the spatiotemporal clustering of smart card users. We applied this tool to one week of smart card data in Québec City, with three different values for alpha (the space–time balance parameter). The results show that the choice of alpha plays an important role on the type of clusters that can be obtained: for alpha = 0, the clusters differ only temporally and show several different temporal patterns. For alpha = 1, the clustering is very strong spatially: every cluster has its own living area and work area. As for alpha = 0.5, its clusters present several different temporal and spatial patterns. Some of its clusters can be identified to some of the clusters of alpha = 1 or of alpha = 0, which demonstrates that there is some continuity in the results when the value of alpha is modified. This proves that we have managed to control the influence of space and time in the clustering results. Depending on the application, it might be, in some cases, more appropriate to have a rather time-based (alpha close to 0) or a rather space-based (alpha close to 1) spatiotemporal clustering, while in other cases, we might wish for more balance between space and time (alpha between 0.25 and 0.75). For example, policies regarding the types of fares and their prices at peak hour (or other period) should provide a higher importance to temporal behavior (by selecting alpha = 0.3, for instance), while the design of the bus lines of distance-related fare policies should give more importance to space (by selecting alpha = 0.7, for example). The groups that are created might suggest, depending on the case, different categories of fares for each group (for alpha = 0.3), or a new bus line at a specific period of the day for one particular group (for alpha = 0.7). At this time, there is an important tramway project to be planned in Quebec City, and the results of this study are helpful to identify the main corridors for its implementation.

This study has some limitations. In particular, the number of clusters has been set to 7 for comparability purposes, but the Silhouette score may reveal a different number of clusters for each value of alpha. The dataset processing has also been made with simplified hypotheses, such as the return home location equal to the first stop in the morning. These assumptions could be validated with further analyses. The sample size has been set to 2000 but an advanced sensitivity analysis could be conducted to find a better choice, especially if we change the value of alpha (as for the number of groups, the sample size may be different if we change the balance between temporal, where profiles are more alike, to space, where there is more diversity).

In further studies, it might be of interest to try this method on a larger number of clusters in order to identify more spatiotemporal patterns and to study how the space–time balance varies with a larger number of clusters. It could also be tested with other case studies to see whether the method provides similar results.

Furthermore, the Euclidean metric that was used to calculate the distance between bus stops could be replaced with the travel time between two stops instead for a more precise spatial analysis, but this opportunity would have to be further studied because the travel distance is strongly related to the geometry of the network.

Acknowledgements The authors wish to thank the *Réseau de transport de la Capitale (RTC)* for providing the data. They also thank the Thales group, Cortex Media, Prompt Quebec and the National Science and Engineering Research Council of Canada (NSERC) for providing funding.

Author contributions Creation of spatiotemporal clustering tool that permits modulation of the influence of time and space, in the case of smart card users.

Funding This research was supported by the following organizations: Thales group, Cortex Media, Prompt Quebec and the National Science and Engineering Research Council of Canada (NSERC).

Data availability Confidential.

Code availability Confidential.

References

Ansari MY, Ahmad A, Khan SS, Bhushan G (2019) Spatiotemporal clustering: a review. Artif Intell Rev 53:2381–2423

Arana P, Cabezudo S, Peñalba M (2014) Influence of weather conditions on transit ridership: a statistical study using data from Smartcards. Transp Res Part A Policy Pract 59:1–12

Asakura Y, Iryo T, Nakajima Y, Kusakabe T (2012) Estimation of behavioural change of railway passengers using smart card data. Public Transp 4(1):1–16

Bagchi M, White PR (2005) The potential of public transport smart card data. Transp Policy 12:464–474

Blythe P (2004) Improving public transport ticketing through smart cards. Proc Inst Civil Eng Municipal Eng 157:47–54

Ceapa I, Smith C, Capra L (2012) Avoiding the crowds: understanding tube station congestion patterns from trip data. In: Proceedings of the ACM SIGKDD international workshop on urban computing, ACM, pp 134–141

Chu KA, Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. Transp Res Record J Transp Res Board 2063:63–72

Devillaine F, Munizaga M, Trépanier M (2012) Detection of activities of public transport users by analyzing smart card data. Transp Res Record J Transp Res Board 2276:48–55

El Mahrsi M, Côme E, Baro J, Oukhellou L (2014) Understanding passenger patterns in public transit through smart card and socioeconomic data: a case study in Rennes, France. In: The 3rd International Workshop on Urban Computing (UrbComp 2014)

Faroqi H, Mesbah M, Kim J (2019) Comparing sequential with combined spatiotemporal clustering of passenger trips in the public transit network using smart card data. Math Prob Eng 2019:1–16

Ghaemi MS, Agard B, Trépanier M, Partovi Nia V (2017) A visual segmentation method for temporal smart card data. Transportmetrica A: Transp Sci 13(5):381–404

Giraud A, Légaré F, Trépanier M, Morency C (2016) Data Fusion of APC, Smart Card and GTFS to Visualize Public Transit Use. In: Transportation Research Board 96th Annual Meeting, Washington DC, United States, Jan 8–12

He L, Trépanier M (2015) Estimating the destination of unlinked trips in transit smart card fare data. Transp Res Record J Transp Res Board 2535:97–104

He L, Agard B, Trépanier M (2018a) A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. Transportmetrica A: Transp Sci 16:56–75

He L, Agard B, Trépanier M (2018b) Space-time classification of public transit smart card users' activity locations from smart card data. In: Conference on Advanced Systems in Public Transport and TransitData 2018, paper 62
He L, Trépanier M, Agard B, Munizaga M, Bustos B (2019) Comparing transit user behaviour of two cities using smart card data. In: Annual Meeting of the Transportation Research Board, Washington, DC. No. 19-05564
Kieu LM, Bhaskar A, Chung E (2014) Transit passenger segmentation using travel regularity mined from Smart Card transactions data. In: Transportation Research Board 93rd Annual Meeting, Jan 12–16, Washington, DC
Kurauchi F, Schmöcker JD (eds) (2017) Public transport planning with smart card data. CRC Press
Kusakabe T, Asakura Y (2014) Behavioural data mining of transit smart card data: a data fusion approach. Transp Res Part C: Emerg Technol 46:179–191
Langlois GG, Koutsopoulos HN, Zhao J (2016) Inferring patterns in the multi-week activity sequences of public transport users. Transp Res Part C: Emerg Technol 64:1–16
Ma X, Wu YJ, Wang Y, Chen F, Liu J (2013) Mining smart card data for transit riders' travel patterns. Transp Res Part C: Emerg Technol 36:1–12
Morency C, Trepanier M, Agard B (2007) Measuring transit use variability with smart-card data. Transp Policy 14(3):193–203
Nishiuchi H, King J, Todoroki T (2013) Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. Int J Intell Transp Syst Res 11(1):1–10
Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C: Emerg Technol 19(4):557–568
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
Seaborn C, Wilson NH, Attanucci J (2009) Analyzing multimodal public transport journeys in London with smart card fare payment data. Transp Res Record J Transp Res Board 2121(1):55–62
"Statistics toolbox." API Reference Documentation. The MathWorks. http://www.mathworks.com/access/helpdesk/help/toolbox/stats/. Accessed 3 Jan 2021
Spurr T, Chapleau R, Piché D (2014) Discovery and partial correction of travel survey bias using subway smart card transactions. Transp Res Record J Transp Res Board 2405(1):56–67
Trépanier M (2012) Use of smart card data to plan urban public transport. RTS-Recherche Transp Securite 28(2):139
Trépanier M, Morency C (2010). Assessing transit loyalty with smart card data. In: Paper presented at the 12th World Conference on Transport Research, Lisbon, Portugal
Trépanier M, Tranchant N, Chapleau R (2007) Individual trip destination estimation in a transit smart card automated fare collection system. J Intell Transp Syst 11(1):1–14
Utsunomiya M, Attanucci J, Wilson N (2006) Potential uses of transit smart card registration and transaction data to improve transit planning. Transp Res Record J Transp Res Board 1971:119–126
Yap M, Cats O, van Arem B (2018) Crowding valuation in urban tram and bus transportation based on smart card data. Transportmetrica A: Transp Sci 16:23–42
Zhao J, Qu Q, Zhang F, Xu C, Liu S (2017) Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE Trans Intell Transp Syst 18(11):3135–3146
Zhou M, Wang D, Li Q, Yue Y, Tu W, Cao R (2017) Impacts of weather on public transport ridership: results from mining data from different sources. Transp Res Part C: Emerg Technol 75:17–29