POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Méthodes de fouilles spatio-temporelles des données de cartes à puce en transport urbain

LI HE

Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiae Doctor*

Génie industriel

Juin 2019

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée:

Méthodes spatio-temporelles de fouilles des données de cartes à puce en transport urbain

présentée par **Li HE**

en vue de l'obtention du diplôme de *Philosophiae Doctor*

a été dûment acceptée par le jury d'examen constitué de :

Jean-Marc FRAYRET, président
Martin TRÉPANIER, membre et directeur de recherche
Bruno AGARD, membre et codirecteur de recherche
Geneviève BOISJOLY, membre
Angelo FURNO, membre externe

DÉDICACE

Porter haut levé le grand drapeau du petit-garçonisme basé sur le chant de Faye Wong et la danse de Sam Mikulak avec les caractéristiques de Kumamoto, et lutter sans relâche pour devenir un petit-garçon chinois comme riz comme cool.

REMERCIEMENTS

J’aimerais dédier cette thèse à mes parents et mes amis qui m’ont tous soutenu pendant toute la durée de mon doctorat. Merci à tous, du fond du cœur.

Je tiens à exprimer mes profonds remerciements à mes directeurs de recherche, le professeur Martin Trépanier, et le professeur Bruno Agard du département de mathématiques et de génie industriel de Polytechnique Montréal, pour m’avoir encouragé, assisté et subventionné durant ce projet de recherche au sein de mon doctorat à Polytechnique Montréal. Vous étiez toujours très gentils à m’aider lorsque que je rencontrais les difficultés. Grâce à vous, je me rends enfin à la fin de la Route Doctorat.

Je tiens aussi à exprimer mes remerciements à mes directeurs de recherche lors de mon séjour au Chili, la professeure Marcela Munizaga du département de génie civil, et le professeur Benjamin Bustos du département d'informatique de Universidad de Chile, pour m’aider à intégrer dans la vie et la recherche au Chili.

Mes remerciements vont également à la Société de Transport de l’Outaouais, TranSantiago qui ont fourni les données de cartes à puces que j'ai utilisées, ainsi que le Conseil de Recherches en Sciences Naturelles et en Génie du Canada, les Fonds de Recherche du Québec sur la Nature et les Technologies, le groupe de Thalès, le Fonds de Cortex, Instituto Milenio Fundamentos de los Datos pour m’avoir subventionné.

Ensuite, j’aimerais remercier l’ensemble du CIRRELT, du département de transport de Polytechnique Montréal, du département de génie industriel de Universidad de Chile pour toute l’aide et le support qu’ils m’ont apportés.

En fin, j’aimerais remercier Faye Wong (une chanteuse chinoise) et Sam Mikulak (un gymnaste artistique américain), même s’ils ne me connaissent pas, mais j’espère qu’ils me connaîtront un jour. Leurs performances m’ont soutenu à survivre dans les moments les plus difficiles, et leur esprit m’encourage toujours à être une meilleure personne dans l’univers.

RÉSUMÉ

Les données des cartes à puce du système de transport en commun sont utiles pour comprendre le comportement des usagers du réseau du transport en commun. De nombreuses recherches pertinentes ont déjà été menées concernant : (1) l'utilisation de données de cartes à puce, (2) les techniques de fouille de données et (3) l'utilisation de la fouille de données avec des données de cartes à puce. Dans ces recherches, la classification des comportements des usagers est basée sur des déplacements pour lesquels les classifications temporelles et spatiales sont considérées comme des processus séparés. Nos partenaires de recherche ont exprimé le souhait de pouvoir examiner les comportements des usagers en considérant simultanément les dimensions spatiales et temporelles. Dans cette thèse, nous développons des méthodes, basées sur les comportements quotidiens des usagers, prenant en compte à la fois les comportements spatiaux et temporels. La méthodologie développée pour classifier les comportements des utilisateurs de cartes à puce s'appuie sur la méthode de distance corrélation croisée (*cross correlation distance*, ou CCD), sur la déformation temporelle dynamique (*dynamic time warping* ou DTW), sur la classification hiérarchique et sur l'échantillonnage. De plus, une méthode basée sur la densité est aussi abordée.

Cette thèse est contribuée de quatre articles plus d'autre résultats présentés dans un chapitre distinct: (1) Afin de commencer la classification temporelle, une comparaison entre CCD et DTW est faite en vue de choisir la meilleure métrique et développer une méthode de classification des séries temporelles en utilisant la classification hiérarchique, et CCD a été prouvé meilleur dans ce cas-ci. Avec cette méthode proposée, un morceau des comportements temporels peut être classifié. (2) Afin de réaliser la classification temporelle pour les données massives, une méthode d'échantillonnage permettant de traiter les grands volumes de données provenant des systèmes de cartes à puce de transport en commun ainsi qu'un indicateur de calibration de cette méthode sont proposés. Cette méthode d'échantillonnage nous permet de classifier tous les comportements temporels d'usagers dans un réseau de transports en commun, et cet indicateur nous permet de choisir les meilleurs paramètres dans l'algorithme. (3) Afin de regrouper les comportements spatiaux et spatio-temporels d'usagers en transport en commun, des méthodes de classification spatiale et spatio-temporelle de comportements des usagers en ajustant l'algorithme de DTW sont développées, et des méthodes de visualisation des résultats en appliquant un graphique spatio-temporel en 3 dimensions sont aussi développées, en vue de montrer l'efficacité de l'algorithme. La visualisation des résultats nous montre l'effectivité de ces deux méthodes. (4) Afin de tester si la

méthode de classification développée dans une ville s'applique dans une autre ville, nous développons une méthode de reconnaissance et de comparaison des comportements de deux villes entre le Canada et le Chile. Les résultats montent qu'environ 66% de comportements temporelles peuvent être reconnu donné un profile de transaction d'un jour, et l'exactitude de reconnaissance est environ 70%. (5) Afin d'analyser les résultats de les classifications spatiale et spatio-temporelle plus profonde, des analyses sont faits incluant la proportion de métro, le moyen et la déviation de trajectoire espace-temps etc, et ces analyse nous permet d'identifier les différences de demande entre les groupes obtenus. (6) En outre, des méthodes de classification de zones géographiques basées sur la densité pour la mesure du changement de comportements des usagers sont développés.

Afin de tester ces méthodes, des données massives provenant des systèmes de perception automatique de la Société de Transport l'Outaouais (STO) de Gatineau et de TranSantiago de Santiago (Chili) sont utilisées. Concernant l'implémentation, les méthodes proposées sont programmées en Python. Les résultats des méthodes, non seulement permettent de regrouper les profils des usagers du transport en commun en quelques groupes et de mieux connaître les caractéristiques de chacun, mais aussi de développer une série de méthodes de visualisation, avec lesquelles les données peuvent être traitées automatiquement pour que des graphiques soient générés. Grâce à ces graphiques, les autorités de transport en commun peuvent traduire les données recueillies automatiquement pour illustrer la demande de transport. Par conséquent, des chercheurs espèrent ces contributions aideront les autorités pour planifier les transports en commun afin de mieux répondre aux demandes des citoyens.

ABSTRACT

Transit smart card data is useful for understanding the behavior of transit users. Numerous relevant research has been conducted on: (1) the use of smart card data, (2) data mining techniques and (3) the use of data mining with smart card data. In this research, the classification of user behavior is based on travel in which temporal and spatial classifications are considered as separate processes. We develop methods, based on the daily behaviors of users, taking into account both spatial and temporal behaviors. The methodology developed to classify the behavior of smart card users is based on the cross correlation distance (CCD) method, dynamic time warping (DTW), hierarchical classification and sampling method. In addition, the density-based method is also affected.

This thesis is presented with four articles plus other results in a separate chapter: (1) In order to start the temporal classification, a comparison between CCD and DTW is made in order to choose the best metric and develop a method of classification of time series using hierarchical classification. CCD has been proved better in this case. A piece of temporal behaviors can be classified with this proposed method. (2) In order to achieve temporal classification for Big Data, a sampling method for processing large volumes of data from transit smart card systems and a calibration indicator for this method are proposed. This sampling method allows us to classify all the users' temporal behaviors in a public transport network, and this indicator allows us to choose the best parameters in the algorithm. (3) In order to classify the spatial and spatio-temporal behavior of users in public transport, methods of spatial and spatio-temporal classification of user behaviors by adjusting the DTW algorithm is developed, and a method of visualization of the results by applying a 3-dimensional spatio-temporal graph is also developed, to show the efficiency of the algorithm. The visualization of the results shows us the effectiveness of these two methods. (4) In order to test whether the classification method developed in one city applies in another city, we develop a method to recognize and compare the behavior of two cities between Canada and Chile. The results show that about 66% of temporal behaviors can be recognized given one-day transaction profiles of two cities, and the recognition accuracy is about 70%. (5) For a deeper view of the spatio-temporal classifications results, analyzes are made including the proportion of metro utilisation, the mean and the deviation of space-time trajectory etc, and these analyses allow us to identify the differences of demands between the clusters obtained. (6) In addition, density-based geographic classification methods for measuring the change of user behavior are developed.

To test these methods, massive data from the Automated Collection System of the la Société de Transport l'Outaouais (STO) and the TranSantiago of Santiago de Chile are used. Regarding the implementation, the proposed methods are programmed in python. The result of these methods not only allows the profiles of transit users to be grouped in a few groups and better understand the characteristics of each, but also creates a series of visualization approaches with which data can be directly transferred to the graphs. With these graphs, transit authorities can translate automatically collected data into traveler demand. As a result, researchers hope that these contributions help the authorities to plan public transit by better meeting the demands of citizens.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

BIRCH	Balanced iterative reducing and clustering using hierarchies
CASPT	Conference on Advanced Systems in Public Transport
CCD	Cross correlation distance
CLIQUE	Clustering in quest
CLP	Chilean peso
CURE	Clustering using representative
DBSCAN	Density-based spatial clustering of applications with noise
DTW	Dynamic time warping (distance)
FRQNT	Fonds de Recherche du Québec – Nature et technologies
GPS	Global Positioning System
GTFS	The General Transit Feed Specification
IMFD	Millennium Institute for Foundational Research on Data
NSERC	Natural Science and Engineering Research Council of Canada
O-D	Origine-Destination
OPTICS	Ordering points to identify the clustering structure)
PAM	Partitionnement autour de medoids
STING	Statistical information grid-based method
STO	Société de Transport l'Outaouais

CHAPITRE 1 INTRODUCTON

La caractérisation des comportements des usagers à partir de grandes séries de données à composantes temporelles et spatiales est précieuse pour de nombreuses applications dans le domaine du transport (Devillaine et al., 2012; Tranchant et al., 2005). En fait, nous retrouvons de nombreux travaux de recherche où de larges séries de données de transport sont analysées pour extraire des comportements qui sont utilisés dans de nombreux domaines d'application (Faroqi et al., 2019; Briand et al., 2017).

Des données sont générées à chaque fois qu'un voyageur passe sa carte à puce sur une machine de perception installée dans un bus, par exemple. Ces données contiennent des informations sur l'heure et l'endroit d'embarquement (parfois aussi pour le débarquement), et d'autres informations telles que le numéro et la direction du bus utilisé, le type de la carte, etc. Nous utiliserons de telles données pour conduire notre recherche. L'utilisation de ces données de cartes à puce, obtenues à partir d'un système de perception automatique, permet une identification des comportements de déplacement de chaque utilisateur. L'efficacité et l'utilité des données de cartes à puce pour analyser le comportement de déplacement des usagers du transport ont été rapportées dans Pelletier et al., 2011 et dans plusieurs travaux subséquents.

Du côté de la fouille des données, des méthodes de classification ont été développées (Rokach et al., 2005). Le profil d'utilisation d'un usager peut être représenté par un vecteur. Pendant le processus de la classification, il est nécessaire d'avoir une métrique de distance en vue de mesurer la (dis)similarité entre deux de ces vecteurs. Il existe de nombreuses métriques utilisées dans les méthodes pour mesurer la dissimilarité de deux séries temporelles (Deza et al., 2009). Pourtant, il existe peu de méthodes en vue de classifier les séries temporelles (présentant une séquence de transactions dans une période de temps donnée). C'est un défi de choisir les meilleures méthodes pour ce genre de données, incluant la « meilleure » méthode de classification et la « meilleure » métrique de distance pour s'adapter à la nature des phénomènes observés en transport; et plus particulièrement dans cette étude, les comportements des usagers de cartes à puce.

Des méthodes visant à analyser le comportement des voyageurs basées sur chaque transaction des cartes à puce ont été proposées (Agard et al., 2006; Ghaemi et al., 2015). Ces recherches avaient pour but de classifier les comportements en termes de l'heure de transaction, du jour de semaine de transaction, du type de carte, etc. Ces recherches avaient aussi pour but de développer les

méthodes pour mesurer la dissimilarité de comportements des usagers de cartes à puce. Cependant, les recherches actuelles traitent les comportements d'une manière discrète, au lieu des séries de comportements. Bien que ces recherches proposent des méthodes sur une base de la transaction individuelle, peu d'études proposent une méthode pour évaluer les comportements sur la base de l'usage individuel du réseau de transport collectif pendant une journée. L'analyse des comportements d'un jour, basée sur l'usager individuel, permettra d'enrichir la compréhension du phénomène de manière plus fine que le traitement des données agrégées.

Dans cette recherche de doctorat, nous allons développer une approche, concernant les usagers des réseaux de transport collectif urbains, pour classifier individuellement les comportements spatio-temporels à partir de l'analyse des données de cartes à puce, concernant les usagers des réseaux de transport collectif urbain. Pour ce faire, nous allons d'abord présenter la revue de littérature, puis présenter des données, analyser les types d'algorithmes disponibles, transposer les caractéristiques du transport en commun dans les algorithmes proposés dans d'autres domaines, et enfin développer les algorithmes et les outils pour caractériser les comportements spatiaux et temporels des usagers de transport collectif. Les données utilisées pour cette recherche sont celles du système de paiement par cartes à puce de la Société de Transport d'Outaouais (STO) et de TranSantiago de Santiago, au Chili. La base de données de carte à puce a plusieurs avantages par rapport aux autres sources des données. Par rapport aux données de GTFS, les données de carte à puce nous permettent de mesurer les comportements des usagers individuels de transport en commun. Par rapport aux données de GPS, les données de carte à puce contiennent moins d'entrées et sont ciblées sur les utilisateurs, et cela rend cette base de données plus facile à manipuler.

Au niveau de la classification temporelle, une technique de fouille des données sera développée, basée sur diverses métriques de distance de séries temporelles, et une méthode de classification. Ensuite, une comparaison des méthodes permettra de trouver la meilleure métrique (couplée à la meilleure méthode) pour ce cas d'étude. Enfin, une méthode sera développée pour classifier les usagers de transport en commun ayant une distribution similaire d'heure de départ (de transactions par cartes à puce).

Au niveau de la classification spatiale, une comparaison des métriques de distance des séries temporelles permettra de développer un algorithme de classification spatiale à partir de l'algorithme purement temporel. Enfin, une méthode sera développée pour classifier les usagers

de transport en commun ayant une distribution similaire d'itinéraires de transactions par cartes à puce (en termes de séquence de points d'embarquement et débarquement). En outre, tenant compte des grandes quantités de données à traiter, nous proposerons une méthode d'échantillonnage et de réaffectation pour calculer les regroupements, et nous utiliserons de grandes puissances de calcul pour montrer l'efficacité de cette méthode.

Les premiers trois chapitres se concentrent sur l'introduction, la revue de littérature et la démarche de travail (voir Figure 1-1 (a)). La thèse sera articulée autour de cinq grandes contributions, chacune étant relative à un article scientifique (voir Figure 1-1 (b)). Dans le mémoire de maîtrise, nous avons réalisé un algorithme d'estimation des destinations à partir de données de cartes à puce où seules les localisations des montées étaient connues. Il nous permet d'estimer la plupart des destinations et cela a permis d'enrichir les données utilisées dans cette thèse (essentiel pour la classification spatiale et spatio-temporelle). La première contribution dans cette thèse a pour but d'analyser les métriques de distances entre séries chronologiques dans le cadre de l'analyse des données de transactions par cartes à puce. Des comparaisons pédagogiques, ainsi que sur un sous-ensemble de données réelles, sont faites pour obtenir la meilleure méthode de classification temporelle pour le cas d'étude. Dans la deuxième contribution, nous appliquons la méthode sur des données complètes et proposons une méthode d'échantillonnage en vue de traiter la grande quantité de données. Dans la troisième contribution, une méthode en vue de classifier les comportements spatiaux sera développée, et le résultat appliqué aux données réelles sera aussi présenté. À cette étape, les méthodes de distance de corrélation croisée (Cross Correlation Distance, CCD) et la déformation temporelle dynamique (Dynamic Time Warping, DTW) sont utilisées pour la classification temporelle et spatiale respectivement. Les résultats obtenus sont les groupes d'usagers de cartes à puce. Dans la quatrième contribution, les données de transactions de deux villes sont mélangées, l'idée principale est de développer une méthode classification temporelle en vue de reconnaître le comportement d'une ville. Avec l'algorithme proposé, les comportements des usagers de Gatineau et Santiago sont testés et comparés. Il existe d'autres contributions dans la thèse. Nous développons les méthodes de visualisation en vue d'analyser la différence de demande entre les groupes, nous développons aussi l'algorithme de la classification sur la base de densité en vue de mieux connaître le changement de comportement entre les groupes obtenus.

L'hypothèse de base est que dans chaque groupe, les comportements temporels ou/et spatiaux des usagers sont similaires, et les usagers faisant partie de deux groupes différents ont des

comportements différents. Ces résultats nous permettront de mieux « comprendre » (ou du moins identifier) les comportements des usagers de cartes à puce, ce qui permettra, nous le souhaitons, à l'autorité de transport en commun de mieux planifier et opérer son réseau de transport en commun, incluant l'horaire, le réseau, la capacité des véhicules, etc., en vue de répondre à la demande (qui s'exprime par les comportements).

Cette thèse est organisée de la manière suivante. La revue de littérature (Chapitre 2) se concentre sur les données de cartes à puce, l'état de l'art des méthodes de fouille de données, les métriques (pour les méthodes de segmentation de données), et l'application des méthodes de fouille des données dans la recherche de comportement des usagers de cartes à puce de transport en commun. La démarche de l'ensemble du travail de recherche et organisation générale du document indiquant la cohérence des articles par rapport aux objectifs de la recherche seront présentées au Chapitre 3. Les quatre contributions seront présentées dans le Chapitre 4 à Chapitre 7 sous la forme de quatre articles. Pour les contributions supplémentaires, ses méthodologies et résultats seront introduits au Chapitre 8. Une discussion générale sur l'ensemble de contributions fera au Chapitre 9. Finalement, le Chapitre 10 est une conclusion concernant la contribution, la limitation et la perspective.

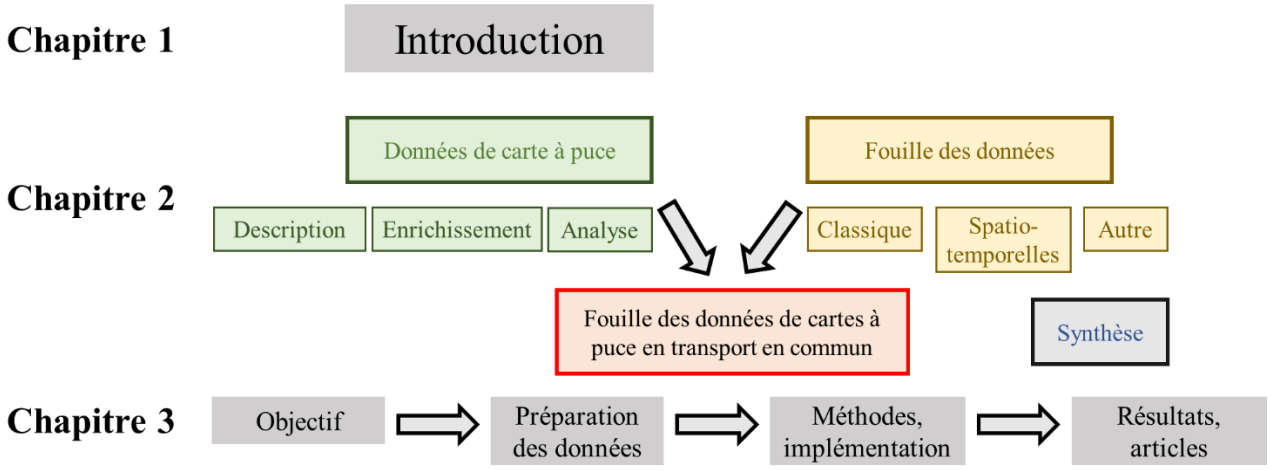


Figure 1-1 (a) Introduction, revue de littérature et démarche de travail

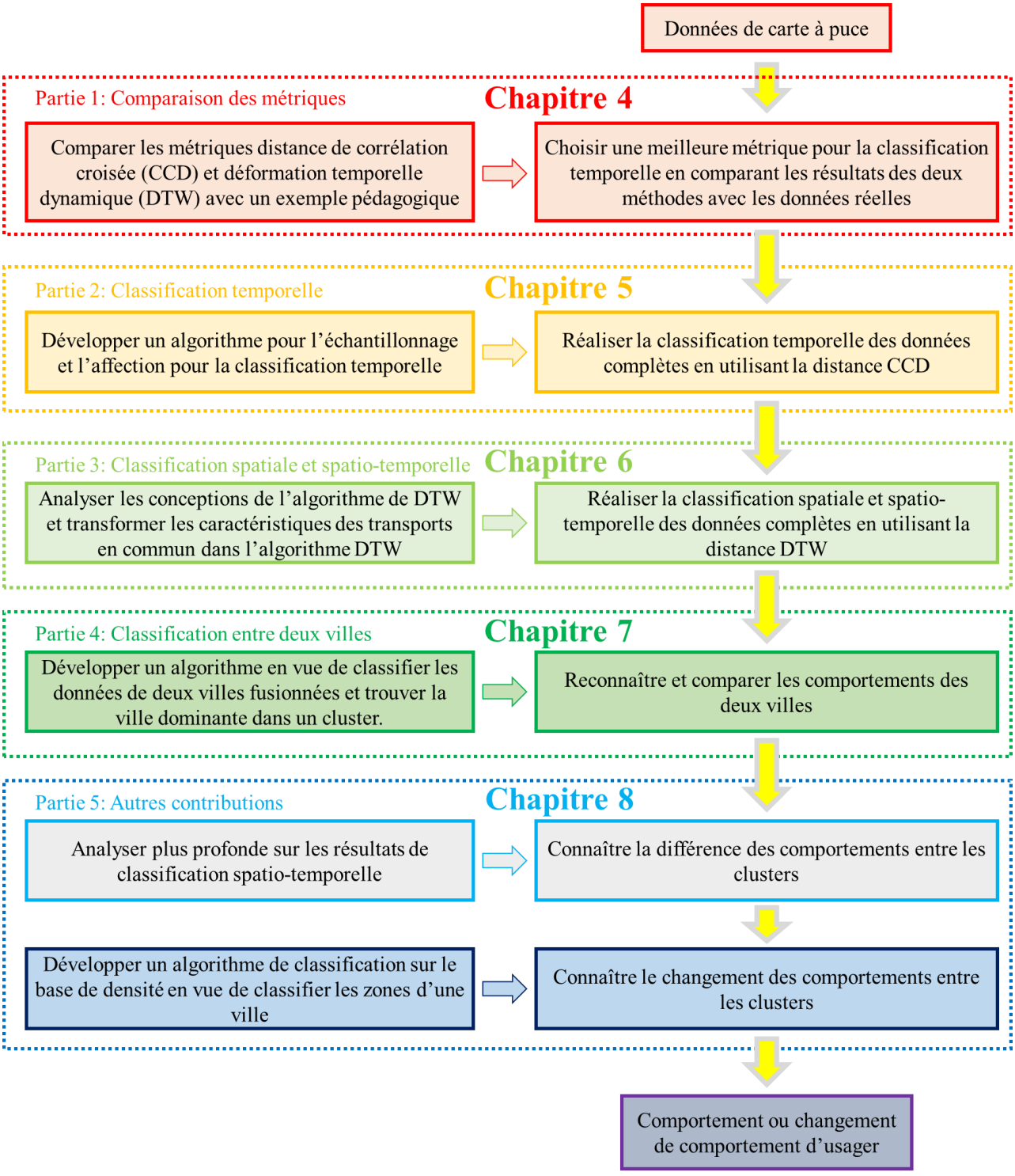


Figure 1-1 (b) Contributions

Figure 1-1 : Contributions de la thèse

CHAPITRE 2 REVUE DE LA LITTÉRATURE

La revue de littérature se compose de trois parties. D'abord, les données de cartes à puce et l'utilisation générale de ces données sont présentées. Ensuite, les méthodes de fouille de données, incluant les méthodes de classification et les métriques de distances, sont expliquées. Enfin, en combinant le côté cartes à puce et le côté fouille des données, les méthodes de fouille des données de cartes à puce sont développées, en vue de mieux connaître les comportements des usagers de cartes à puce. Une revue de ces applications est présentée à la fin de cette partie.

2.1 Données de cartes à puce en transport en commun

Cette partie a pour but de répondre à deux questions : (1) Quelles sont les données de cartes à puce disponibles ? et (2) Comment utiliser les données de cartes à puce pour comprendre les comportements des usagers ?

Pour répondre à la première question, la génération des données de cartes à puce et des extraits de tableaux faisant partie de ces données seront présentés. Pour répondre à la deuxième question, un modèle-objet traitant des données et des méthodes d'enrichissement des données seront présentés. En outre, des méthodes développées pendant ces années permettent de mieux comprendre les comportements des usagers de cartes à puce ainsi que les prévoir.

2.1.1 Description des données de cartes à puce

2.1.1.1 La génération des données

La Figure 2-1 montre le fonctionnement typique d'un système de transactions à base de cartes à puce. Les flèches représentent les flux de données. La base de données des transactions de cartes à puce contient les données sur les usagers, les données sur les validations et un serveur de données. Les usagers peuvent charger les cartes à puce auprès des points de vente émetteurs ou des points de vente rechargeurs. En même temps, les informations des usagers ainsi que des titres vont être transférées à la base de données. Une fois montés dans le bus, les usagers valident la carte à puce au sein du système. À ce moment-là, la base de données va valider l'état de cette carte à puce. Concernant la validation du métro, si un usager passe la carte sur le lecteur d'un tourniquet, l'état de cette carte à puce sera validé. Notons que le type de titre est aussi enregistré (régulier,

occasionnel, mensuel, tarif réduit, etc.) dans la base de données. À la fin de la journée, les autobus équipés de lecteurs et de GPS échangeront éventuellement les données portant sur les tracés, les voyages et les validations (He, 2014; M. Trépanier et al., 2004).

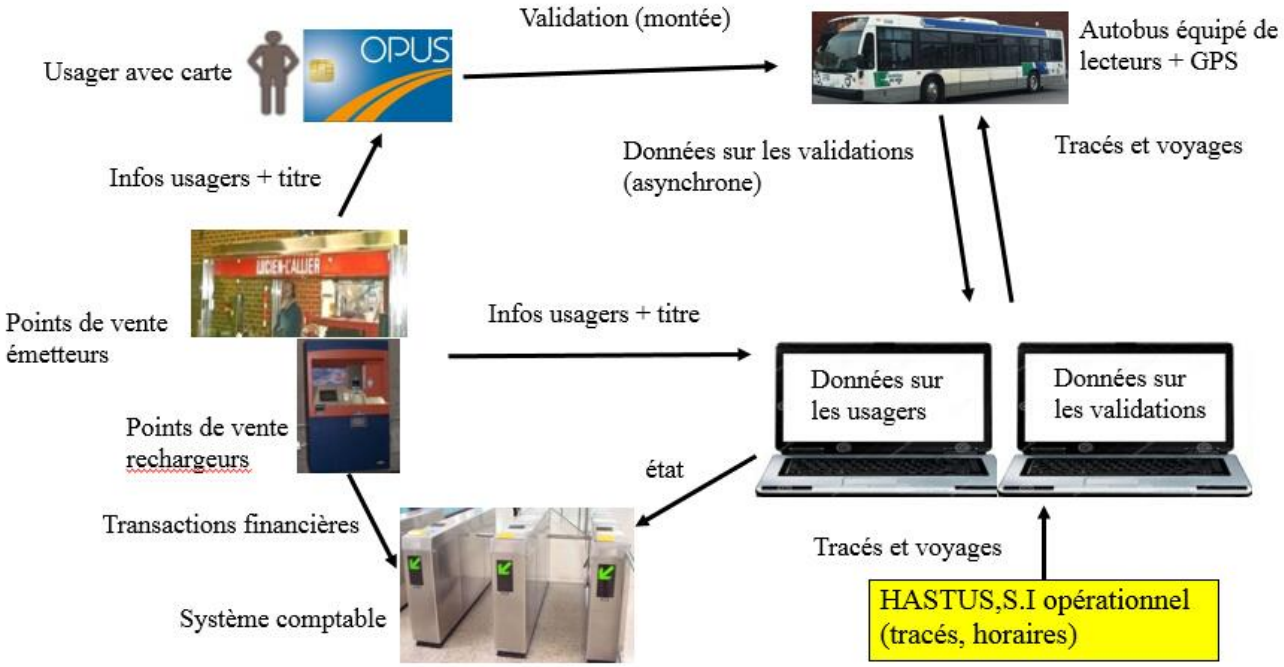


Figure 2-1: Diagramme fonctionnel du paiement par cartes à puce de STO (M. Trépanier et al., 2004)

2.1.1.2 L'information des données

En ce qui concerne la genèse des données de cartes à puce, ces données ont été catégorisées de la façon suivante (Bagchi et al., 2004):

- Information spatiale : emplacement d'embarquement, emplacement de débarquement dans certains cas, durée et itinéraire du trajet (dans le cas où les emplacements d'embarquement et de débarquement sont tous connus) ;
- Information temporelle : date et heure d'embarquement, date et heure de débarquement (dans le cas où les emplacements d'embarquement et de débarquement sont tous connus) ;
- Information structurale : mode de transport et service emprunté (ligne, sens) ;
- Information personnelle sur l'usager : nom, sexe, âge, adresse, possession automobile ;

- Information d'achat : type de billet, prix du billet, emplacement d'achat (l'emplacement d'achat et d'embarquement est différent dans certains cas).

Au total les données de cartes à puce contiennent de grandes quantités d'informations, certaines seront inutilisées dans cette recherche. Plus de détails sur les champs que nous utilisons dans cette recherche seront présentés à la Section 3.1.2.1.

2.1.1.3 Modèle-objet traitant des données

Pour caractériser les données, une approche orientée-objet en transport a été proposée (M. Trépanier et al., 2001) (Figure 2-2). Cette approche vise à observer des données de transport, notamment en définissant et quantifiant les concepts inhérents. Ce modèle est en relation avec les processus de gestion.

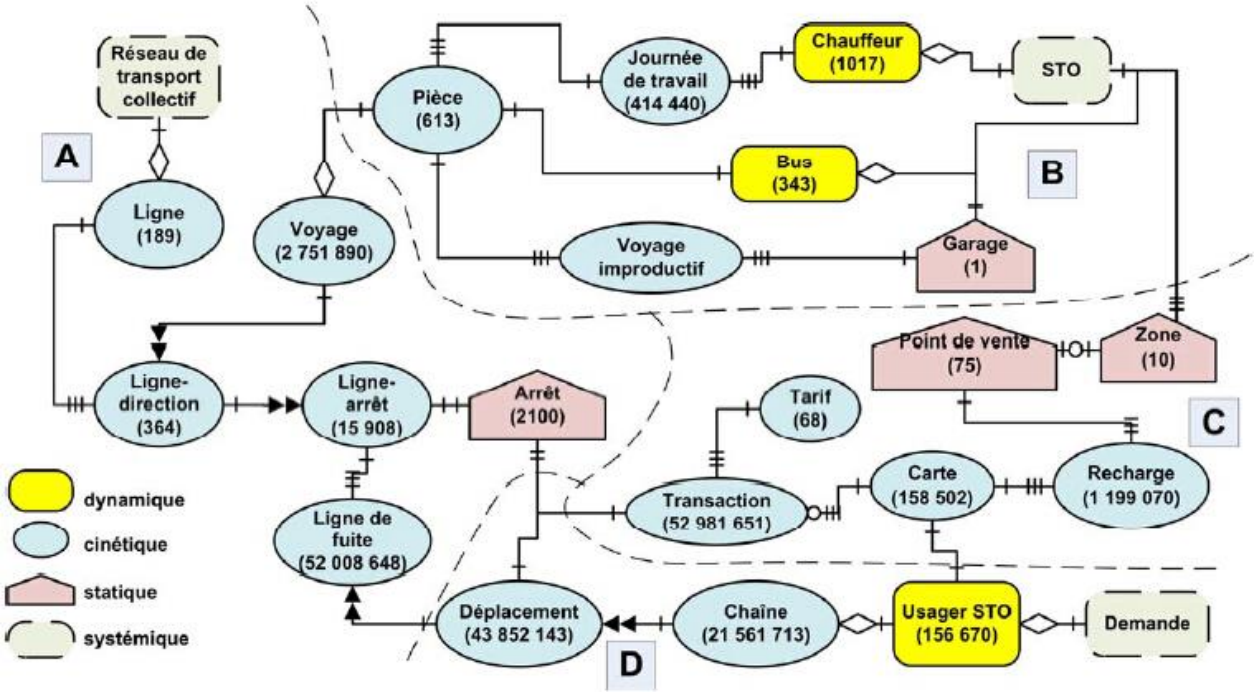


Figure 2-2: Modèle-objet correspondant pour une période limitée (M. Trépanier et al., 2001)

Ce modèle-objet en transport comprend quatre types d'objets :

- Dynamiques (jaune), représentant ce qui se déplace, tels que les véhicules et les usagers de l'agence de transport ;

- Cinétiques (bleu), représentant des descripteurs du mouvement, tels que les lignes, les lignes-arrêts, les recharges et les transactions ;
- Statiques (rouge), représentant les supports immobiles faisant partie du système de transport, tel que les arrêts et les points de vente ;
- Systémiques (vert), étant les fédérateurs des autres objets : Réseau de transport collectif, STO et demande.

Selon leurs fonctions, ces objets ont été regroupés pour concevoir quatre groupes (M. Trépanier et al., 2001). Les objets du réseau de transport collectif (A) représentent les éléments « visibles » du réseau de transport. Ils contiennent les lignes, les arrêts et les objets qui en sont dérivés au sein de ce système de transport en commun.

- Les objets opérationnels (B) définissent la mécanique de fonctionnement du service. Basés sur les éléments « bus » et « chauffeur », les plans de travail (par exemple : trajet et horaire de chaque chauffeur) pendant une journée sont produits. Parmi eux, une partie des voyages sont improductifs (en transit en laissant le véhicule vide); dans ce cas, les bus peuvent retourner au garage, ce qui constitue tous les éléments des objets (B).
- Les objets administratifs (C) regroupent tous les éléments portant sur la gestion financière.
- Les objets liés à la demande (D) regroupent les éléments associés à la planification des transports.

Ce modèle nous aide à comprendre tous les concepts de base pertinents sur les données de carte à puce. Il nous permet de chercher un mot clé pour lier un tableau à l'autre dans la base de données de carte à puce. Ensuite, ce modèle-objet est aussi utile lors que nous fusionnons les différentes bases de données concernant le transport en commun. Par exemple, le mot clé sera utilisé pour lier les bases de données de GTFS et de cartes à puce, ensuite, cette fusion nous permet de calculer les indicateurs tels que passagers/(véhicules*heures), passager/(véhicules*kilomètres), etc.

2.1.2 Enrichissement des données

Les données de cartes à puce de certains systèmes de transport en commun ne contiennent pas la destination des déplacements, car les usagers n'ont pas à valider à la sortie (*tap-in* seulement).

Plusieurs travaux ont été réalisés pour développer des algorithmes d'estimation des destinations, ainsi pour que les valider.

2.1.2.1 Estimation des destinations

Les données que nous utilisons ne contiennent pas les informations de débarquements. Afin d'enrichir ces données, une série de méthodes d'estimation des destinations ont été proposées dans la littérature : l'estimation traditionnelle, l'estimation supplémentaire et la calibration de la méthode. Ces méthodes sont importantes pour la classification spatiale et spatio-temporelles de la thèse, parce que les localisations dans l'algorithme de classification spatiale et spatio-temporelle sont basées sur les résultats de l'algorithme d'estimation des destinations.

2.1.2.1.1 Estimation traditionnelle

Un modèle est proposé afin d'estimer les lieux de débarquement en supposant que l'individu va rembarquer au prochain trajet à l'arrêt le plus près du lieu où il a débarqué (Trépanier et al., 2007), cette recherche d'estimation est basée sur la distance entre les arrêts des séquences des transactions. La Figure 2-3 présente la possibilité d'estimer l'emplacement de débarquement en quatre phases:

- Séquences de déplacement. Nous supposons que les usagers descendent à l'endroit où la distance entre l'emplacement d'embarquement de la prochaine transaction et l'emplacement de débarquement du déplacement actuel est minimale. En même temps, cette distance minimum devrait être inférieure à un seuil de tolérance.
- Retour à domicile. Nous supposons que les usagers retournent à l'origine du premier déplacement.
- Déplacement du prochain jour. S'il est impossible de trouver une solution après les deux phases précédentes, il faut tenir compte du prochain déplacement du jour suivant. Nous vérifions s'il existe un lien entre l'emplacement d'embarquement de déplacement du prochain jour et les emplacements de débarquement possibles du déplacement actuel.
- Déplacements unitaires. Si les trois phases précédentes ne fonctionnent pas, nous pouvons chercher une transaction similaire dans l'historique de la carte, c'est-à-dire une transaction avec le même emplacement d'embarquement et presque la même heure. L'emplacement de débarquement sera déterminé en fonction de la destination utilisée dans l'historique.

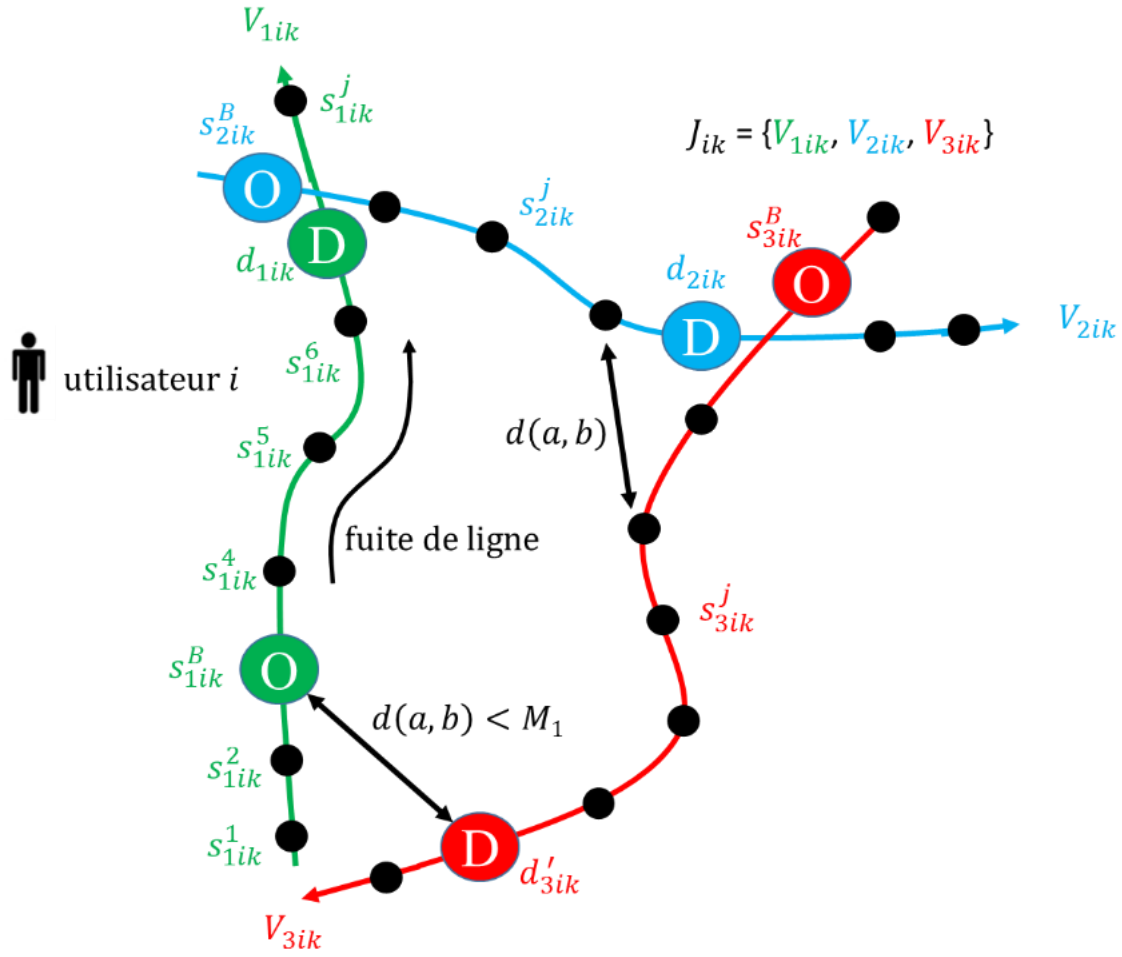


Figure 2-3: Modélisation des phases « séquence de déplacement » et « retour à domicile » (Trépanier et al., 2007)

2.1.2.1.2 Estimation supplémentaire

Afin de trouver les destinations de déplacement unitaire, un algorithme utilisant l'historique des déplacements de l'usager est développé (He et Trépanier, 2015). Au niveau de la spatialité ainsi que la temporalité, nous pouvons utiliser l'estimation par noyau pour évaluer le niveau d'impact de la transaction historique sur la transaction actuelle. L'estimation par noyau est une méthode non-paramétrique d'estimation de la densité de probabilité d'une variable aléatoire. Elle se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support (Wasserman, 2005).

Puisque l'estimation par noyau peut transférer la probabilité des variables discrètes en variables continues, il est possible de superposer la distribution normale de probabilité de débarquement à

chaque emplacement de débarquement potentiel. De cette façon, nous pouvons obtenir une courbe de probabilité pour tout le tracé de la ligne où se trouvent tous les emplacements de débarquement potentiels.

En ce qui concerne la spatialité, puisque la distance est la seule mesure de cet espace à une dimension, la distance à partir de l'emplacement d'embarquement aux emplacements de débarquement potentiels est prise comme l'axe horizontal des abscisses. En ce qui concerne la temporalité, puisque le temps est la seule mesure de cet espace à une dimension, l'heure de débarquement est également prise comme l'axe horizontal des abscisses.

Lors de l'estimation, nous multiplions les probabilités spatiale et temporelle pour prédire la probabilité de se rendre à cet emplacement de débarquement potentiel à cette heure. La Figure 2-4 présente ce processus de l'estimation au niveau de l'estimation temporelle (He, 2014).

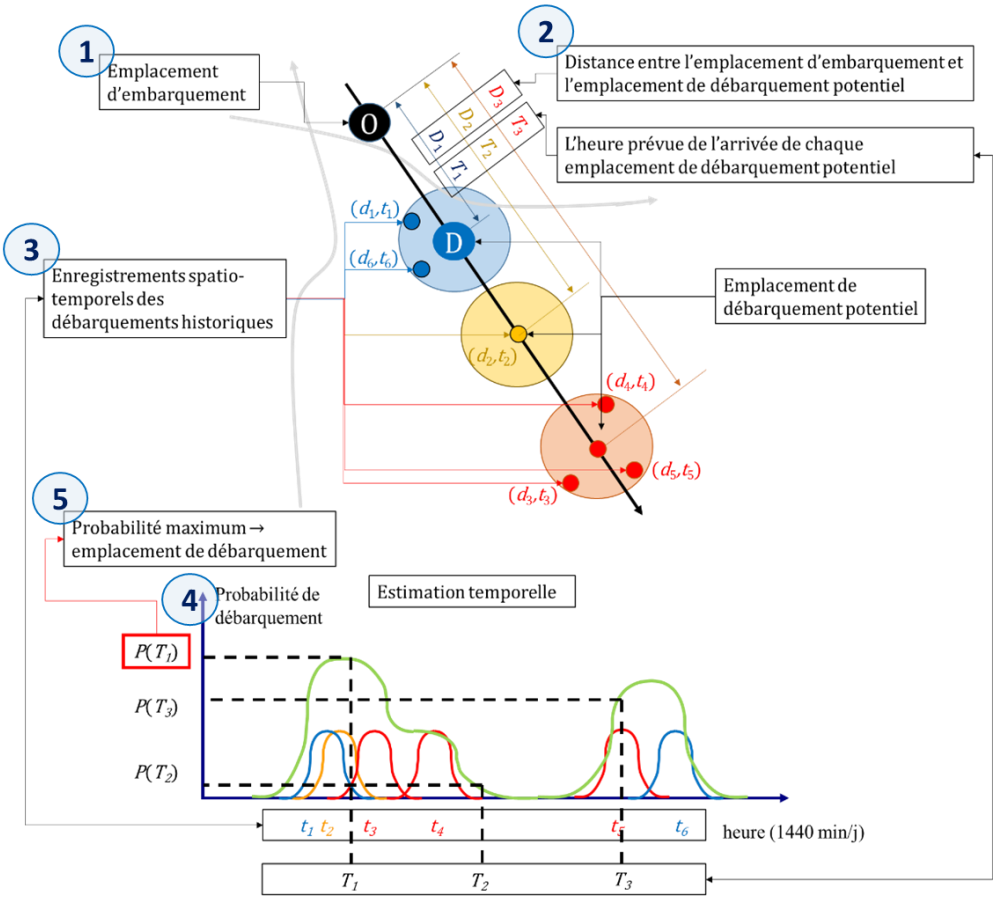


Figure 2-4 : Estimation par noyau pour le traitement temporel du déplacement unitaire (He & Trépanier, 2015)

2.1.2.1.3 Exactitude des méthodes

La Figure 2-5 présente la distribution des résultats pour l'ensemble des données d'octobre 2009 de la Société de transport de l'Outaouais. Ceux-ci sont codés:

- Code 11 : Séquence de déplacement
- Code 12 : Retour à domicile
- Code 13 : Déplacement du prochain jour
- Code 21 : Déplacement unitaire avec plusieurs emplacements de débarquement potentiels
- Code 22 : Déplacement unitaire avec emplacement de débarquement potentiel unique
- Code 30 : Pas de résolution encore

L'algorithme précédent résout 80,64% des déplacements et il reste 19,36% de déplacements sans aucune approche pour les résoudre. Les résultats de calcul démontrent que la contribution de ce projet, surtout pour le déplacement unitaire, résout 10,9% de déplacements additionnels. Somme toute, 56,30% des déplacements unitaires sont résolus.

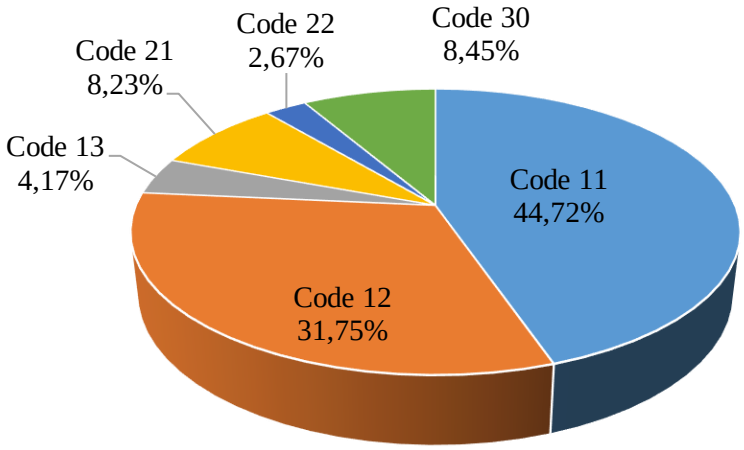


Figure 2-5: Distribution des types d'estimation obtenue par l'algorithme des destinations

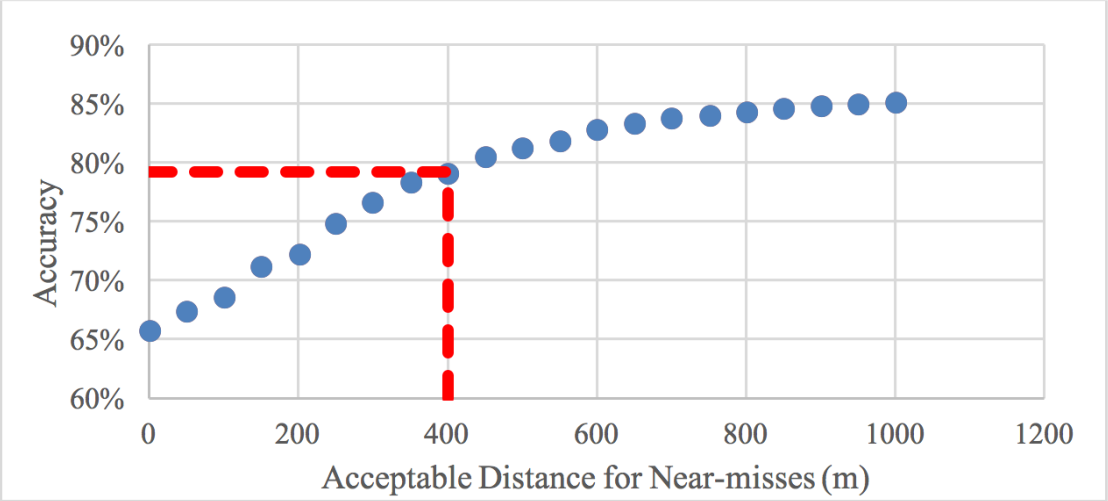


Figure 2-6: Calibration pour la distance acceptable

Afin de calibrer et valider un algorithme de prévision de destination de voyage développé pour des données canadiennes, nous utilisons des données tap-in/tap-out de carte à puce de Brisbane, Australie. La Figure 2-6 présente la relation entre le seuil de distance de correspondance et l'exactitude d'estimation de destinations. Les résultats montrent que l'algorithme a une précision de 79% à la distance acceptable de 400 mètres (He et al., 2015).

2.1.2.2 Détection des correspondances

Certaines recherches se concentrent sur la réalité que certains passagers n'utilisent pas la carte à puce. Il est donc intéressant de comparer la différence entre les comportements des usagers de cartes à puce et des autres passagers. Sur la base des travaux précédents, des méthodes sont proposées pour estimer l'heure d'arrivée des opérations de bus au niveau de l'arrêt en utilisant des contraintes temporelles et pour identifier les déplacements liés en utilisant des concepts spatio-temporels. Ces enrichissements conduisent à la reconstruction d'itinéraires individuels, à l'analyse de l'activité de transfert et à la synthèse des profils de charge du véhicule. Ce dernier fournit aux planificateurs une progression spatio-temporelle détaillée de chaque exécution, des arrêts d'origine et de destination pour chaque transaction individuelle, et l'activité d'embarquement et de descente à chaque arrêt. L'étude s'appuie sur plus de 37 000 transactions d'embarquement de carte à puce d'un jour de semaine moyen d'une agence de transport en commun de taille moyenne. Les résultats suggèrent que les voyages liés représentent un peu plus de 10% du nombre total de transactions

dans le réseau et le système de carte à puce surestime la proportion de voyages liés de près de 40% (Chu et al., 2008).

2.1.2.3 Inférences du motif de déplacement

Les données obtenues (enrichies) sur les transactions par carte à puce peuvent être exploitées pour obtenir des informations utiles sur le comportement des passagers en transport en commun, à savoir le but du déplacement ou de l'activité. Les données de collecte automatique des tarifs (par exemple, cartes à puce) peuvent être utilisées pour inférer le but du voyage et révéler des modèles de déplacement dans une zone urbaine. Une étude de cas démontre le processus d'inférence des fins de déplacement en fonction des données de cartes à puce de Metro Transit dans la Région métropolitaine de Minneapolis / St. Paul (Lee et al., 2014).

2.1.3 Analyse du comportement des usagers et le réseau

Beaucoup de recherches ont été introduites en vue d'analyser les usagers et le réseau. D'abord, des méthodes sont développées pour la caractérisation des usagers et du réseau. Notamment, des méthodes de fouilles des données sont implémentées en vue de classifier les comportements des usagers. Ensuite, les facteurs externes sont introduits à voir comment ils influencent le réseau. À la fin, quelques méthodes sur la prédiction en utilisant des données de cartes à puce seront présentées.

2.1.3.1 Caractérisation des usagers

Cette section a pour but de mieux connaître les comportements (les pattern) des usagers. C'est important car c'est avec la connaissance des comportements des usagers que nous développer les méthodes pour optimiser les réseaux et améliorer le service du transport en commun.

En analysant la base de données, il est possible d'obtenir les informations concernant des clientèles. Cette analyse contient beaucoup de choses : l'objectif de déplacement, l'analyse de l'origine et de la destination (la spatialité), la définition des heures de pointe (la temporalité), la fidélité des usagers, etc. (He, 2014) :

En tenant compte des types de tarifs, il est possible de calculer la proportion des usagers selon leur appartenance aux groupes (Morency et al., 2007). Le type de tarif est associé à la carte (régulier,

étudiant, express, etc.). L'analyse de types de tarifs permet de caractériser les comportements de différents groupes afin de fournir un meilleur service pour chaque groupe d'usagers.

En tenant compte des heures d'embarquement et de débarquement, il est possible d'estimer le motif du déplacement. Par exemple, les règles suivantes nous aident à distinguer les objets du déplacement (Devillaine et al., 2012).

Les critères suivants sont utilisés pour déduire le motif de déplacement 'travail'.

- Type de carte est adulte.
- Un des cas suivants:
    - Durée de l'activité est supérieure à 5 h et
    - Durée de l'activité est comprise entre 2 et 5 h, et de la zone d'arrêt de bus n'est pas l'hôpital ou de loisirs.
- Déplacement avant de l'activité n'était pas le dernier de la journée.

Les critères suivants sont utilisés pour déduire le motif de déplacement 'étude'.

- Type de carte est étudiant ou mineur.
- Un des cas suivants:
    - Durée de l'activité est supérieure à 5 h et
    - Durée de l'activité est comprise entre 2 et 5 h, et de la zone d'arrêt de bus est éducatif.
- Déplacement avant de l'activité n'est pas le dernier de la journée.

De la même manière, les critères de l'objectif de déplacement tels que domicile et autres sont précisés.

Le résultat de l'estimation du motif du déplacement permet de caractériser la temporalité des transactions de chacun des motifs d'activité. La distribution de l'heure de début du déplacement (l'heure d'embarquement) de lundi à jeudi à Gatineau (la STO) pour deux années différentes peut être présentée. La méthode permet d'analyser les heures de pointe des transactions pour chaque motif. Elle permet aussi d'évaluer la proportion des types de déplacement par heure. En comparant les courbes des deux années, il est possible d'identifier l'évolution des comportements, dans ce cas-ci l'étalement des heures pour motif 'étude' (Devillaine et al., 2012).

Une meilleure connaissance de l'achalandage aux heures de pointe du transport en commun facilite l'optimisation du service par rapport à la demande. Il est nécessaire de trouver un moyen pour mesurer et déterminer l'achalandage basé sur les données de cartes à puce. Une méthode a été créée et a été mise en œuvre à la ville Shenzhen, Chine (Shi et al., 2014). Dans cette recherche, les auteurs combinent le nombre de transactions d'une heure du matin et d'une heure de l'après-midi et calculent la somme des transactions de ces deux heures. Ensuite, ils comparent les combinaisons où elles comprennent le plus de transactions. Pour toutes les paires d'heures de transactions, ils trouvent les heures les plus fréquentées et obtiennent les heures de pointe de matin et de l'après-midi. Selon cette méthode, ils identifient que l'heure de pointe du matin est entre 7 heures et 9 heures, et l'heure de pointe de l'après-midi est entre 18 heures et 19 heures. Connaître l'heure des pointes du matin et du soir permet de réagir aux demandes des usagers. Par exemple, offrir les transports en commun de grande capacité pendant ces périodes, planifier les voies spécifiques pour les véhicules collectifs, etc.

Outre la temporalité, il est aussi possible de trouver les informations utiles par rapport à la spatialité. Il est relativement facile de calculer le nombre de transactions à chacun des arrêts. Il est ainsi possible de trouver les arrêts les plus utilisés à chacune des heures. En plus, l'existence de cette base de données nous permet de valider l'exactitude des résultats des autres sondages tels que l'enquête de déplacement (Spurr et al., 2014).

Concernant la spatialité, les données de carte à puce permettent non seulement de connaître l'origine des transactions, mais aussi les destinations si le système l'enregistre. Les origines et les destinations des transactions d'une ville peuvent être montrées. Ces distributions aident à savoir les zones principales de génération des déplacements du jour de la semaine ainsi que le week-end (Shi et al., 2014). Cela facilite la planification des systèmes de navettes pour transporter les gens qui se rendent au travail et retournent à domicile. De plus, la connaissance de la différence de ces O-D (Origine-Destination) entre les jours de la semaine et le week-end aide à redistribuer les trajets ou les horaires du transport en commun durant le week-end.

Il est intéressant d'analyser la fidélité des titulaires d'une carte à puce. Une vérification de la possession longitudinale de chacun des types de cartes permet de connaître la fidélité des clientèles en leur attribuant un taux de survie. Nous remarquons que les titres avec paiement bancaire (PB) ont un taux de rétention plus élevé que les autres. Les usagers des lignes régulières sont moins

fidèles que ceux des lignes express. Pour les étudiants collégiaux et universitaires, nous constatons des chutes régulières de rétention, d'après (Trépanier et al., 2010).

Notamment, les méthodes de fouilles des données ont été intégrées dans la caractérisation des usagers. La classification du comportement des usagers de cartes à puce est maintenant une application importante des données. Pourtant, tenant compte qu'il s'agit des méthodes de classification, cette partie sera présentée après la présentation des méthodes de fouilles des données, soit dans la partie 2.3.

En conclusion, deux types de caractéristiques sont introduits dans cette section : l'une des deux portes sur les caractérisations des usagers eux-mêmes, ce qui permet de connaître leurs motifs de déplacements, leur fidélité, etc. L'autre porte sur les caractérisations des spatio-temporalité des usagers. Les recherches proposent des méthodes en vue de connaître les comportements spatio-temporels des usagers en transports en commun. Pourtant, peu de recherches dans cette section permet de classifier les comportements en un nombre limité de groupes.

2.1.3.2 Caractérisation du réseau

En analysant les mêmes bases de données, il est possible d'obtenir des informations concernant le service offert par l'agence de transport (He, 2014; Tranchant, 2005; Trépanier et al., 2009) :

Premièrement, un diagramme espace-temps peut être établi pour chaque voyage. Une fois reconstitué l'ensemble des arrêts des voyages associés à leur heure de passage, un algorithme de détermination des heures de passage aux différents arrêts peut être appliqué. Nous pouvons alors en déduire un diagramme espace-temps de chacun des voyages déclarés. Il est alors possible d'en déduire la vitesse commerciale approximative du voyage.

Deuxièmement, un profil de charge peut être établi pour représenter le taux de l'utilisation d'un voyage du véhicule. La procédure précédente permet de savoir quel véhicule a été emprunté, et la base de données comprend directement les informations spatio-temporelles d'embarquement. Il est alors possible de déterminer la charge maximale de chaque voyage, le nombre de montées du voyage et de calculer le nombre de passagers-kilomètres.

Troisièmement, en combinant les deux figures précédentes, un graphique sur la densité de l'achalandage spatio-temporel a été obtenu. Ce diagramme permet d'évaluer où se trouvent temporellement et spatialement les pointes de charge de la journée.

Quatrièmement, une matrice O-D peut être obtenue grâce à la destination des usagers. En effet, la réalisation de la matrice O-D nous indique quelles paires origines-destinations sont les plus utilisées lors des déplacements des usagers.

Cinquièmement, pour conclure sur les perspectives d'analyse étudiées jusqu'à maintenant, un outil d'aide à la planification a été réalisé à partir du logiciel Excel, de programmation VB et de la base de données traitée dans MS Access. Cet outil expose principalement toutes les analyses présentées précédemment de manière dynamique.

Une autre recherche illustre l'utilisation de données de cartes à puce pour estimer diverses mesures de performance de transport en commun. Combinées à des processus d'évaluation établis, ces mesures peuvent aider les opérateurs à surveiller leurs réseaux plus en détail. La performance de l'approvisionnement du réseau (véhicules-kilomètres, véhicules-heures, vitesse commerciale, etc.) et les statistiques sur le service passagers (passagers-kilomètres, passagers-heures, durée moyenne de trajet, etc.) peuvent être calculées à partir de ces ensembles de données pour tout niveau de résolution spatiale ou temporelle, y compris les niveaux de route et d'arrêt de bus (Trépanier et al., 2009).

2.1.3.3 Facteurs externes qui influencent l'utilisation du réseau

D'autres études visent à illustrer que les comportements des passagers dans un réseau de transport peuvent être liés à différents paramètres externes (Briand et al., 2017).

Une recherche (Arana et al., 2014) analyse l'influence des conditions météorologiques sur le nombre de déplacements en transport en commun pour les loisirs, les magasins et les affaires personnelles à Gipuzkoa, en Espagne. Les résultats multiples de la régression linéaire ont montré que le vent et la pluie pouvaient entraîner une diminution du nombre de voyages, alors que l'augmentation de la température provoquait une augmentation du nombre de voyages, en accord avec les résultats d'études antérieures basées sur des enquêtes. De plus, les voyageurs réguliers et occasionnels se sont révélés partager ce modèle de comportement.

Au Québec, ce type d'étude pose deux défis méthodologiques: premièrement, il faut disposer de données complètes et longitudinales sur l'achalandage; deuxièmement, il doit être possible d'isoler l'effet météorologique des autres causes de variation. Dans le cadre de ce projet, le chercheur tente d'aborder ces deux problèmes à l'aide de données de cartes à puce continues provenant des réseaux

de transport en commun de Gatineau et de Montréal (Québec). Les résultats montrent que les conditions météorologiques influencent surtout les voyageurs effectuant des déplacements sans contrainte (comme pour les aînés de Gatineau) et entraînent un transfert modal du réseau d'autobus au métro de Montréal (Trépanier et al., 2012).

(Chu, 2015) propose une stratégie d'échantillonnage pour extraire des observations longitudinales à partir d'une base de données de validation de cartes à puce. La cohérence interne et la comparabilité avec la population sont évaluées. Il est révélé que les pratiques opérationnelles, plutôt que la durée de vie théorique de la carte, sont le facteur déterminant de la durée d'observation, de la taille de l'échantillon et de la présence de biais spatial et temporel. En utilisant les indicateurs de la mobilité et de la diversité de la localisation, les observations longitudinales sont analysées individuellement et globalement pour comprendre les comportements de déplacement au cours de la journée, de la saison et de l'année à l'année. Puisque l'activité des passagers évolue avec le temps, cette recherche démontre l'impact des facteurs externes sur la durée d'observation.

Les facteurs externes influencent toujours et parfois soudainement le réseau ou/et le comportement des usagers, par conséquent, le développement de méthodes dynamiques est important. (Tao et al., 2014) propose une méthodologie multi-étape afin de voir les modèles spatio-temporels du comportement de déplacement des passagers. En s'appuyant sur le réseau de bus de Brisbane, en Australie, l'ensemble de données de cartes à puce a été traité en combinaison avec la spécification générale pour les flux relatifs aux transports en commun (GTFS) pour reconstituer les trajectoires de déplacement des passagers de bus au niveau de la granularité spatiale. La méthodologie proposée a dévoilé visuellement la dynamique du comportement spatial temporel des passagers.

2.1.3.4 Prédiction des comportements en utilisant des données de cartes à puce

L'utilisation de données de cartes à puce a pour but de générer des informations sur la prévision des habitudes de déplacement (Briand et al., 2017):

- En utilisant les données de cartes à puce, (Ceapa et al., 2012) concentrent leur travail sur la congestion du trafic et établissent qu'il existe une certaine régularité spatiale et temporelle dans la congestion qui facilite la prévision. Ces résultats démontrent que les informations sur les niveaux de congestion peuvent être incorporés dans les systèmes avancés

d'information des voyageurs, afin de fournir aux voyageurs des plans de voyage plus personnalisés.

- (Lathia et al., 2010) étudient les données de voyage individuels sur les métros de Londres (offre de services personnalisés) pour estimer les déplacements personnels et développer une méthode de prévision des heures de voyage personnalisées pour les passagers qui classent les stations en fonction des modèles de mobilité futurs.
- L'approche proposée par (Foell et al., 2014) est appliquée à un réseau de bus et utilisé pour prédire les déplacements.

La section 2.1 a présenté des informations sur des données de cartes à puce, incluant de nombreuses méthodes développées pour l'utilisation de ces données. Cependant, ces méthodes sur des données de transports touchent peu les méthodes de fouille des données. Par conséquent, la section 2.2 dévoile des méthodes de fouilles des données et ses fonctionnalités.

2.2 Méthodes de fouille des données

Suite à l'introduction sur les données de cartes à puce, les techniques de fouilles de données liées à ce projet se structurent en 2 catégories : (1) les méthodes de classification et (2) les métriques de similarité.

2.2.1 Méthodes de classification

Les méthodes de classification ont pour but de regrouper un ensemble d'observations en clusters (groupes). Une bonne méthode de segmentation va produire des clusters de haute qualité avec une forte similarité intraclasse et une faible similarité d'interclasse. Plusieurs approches principales de segmentation existent, les principales sont présentées ci-après (Subbiah, 2011).

2.2.1.1 Algorithmes de partitionnement

Pour les algorithmes de partitionnement, la structure classificatoire recherchée est la partition. L'objectif est de trouver, parmi l'ensemble fini de toutes les partitions possibles (*k* classes), une partition qui optimise un critère défini a priori (Chevalier et al., 2013). Elles touchent essentiellement deux méthodes heuristiques.

- K-means: Chaque groupe est représenté par la valeur moyenne du cluster. K-means est une méthode qui vise à partitionner n observations en k clusters dans lesquelles chaque observation appartient au cluster avec la moyenne la plus proche (Srimani et al., 2013). C'est l'une des méthodes de classification les plus connues et utilisées.
- K-medoids ou PAM (Partitionnement autour de medoids): Chaque cluster est représenté par l'un des objets du cluster (Park et al., 2009). K-means tente de minimiser l'erreur carrée totale, tandis que k-medoids minimise la somme des différences entre les points étiquetés à affecter dans un cluster et un point est désigné comme le centre de ce cluster (Mirkes, 2011). K-medoids est une technique classique de partitionnement de clustering.

En outre, il existe des méthodes dérivées telles que et k-modes (Huang & Ng, 2003), une approche non paramétrique permettant de dériver des grappes à partir de données catégorielles à l'aide d'une nouvelle procédure de regroupement, et k-medians (Bradley et al., 1997), une variante de k-means où, au lieu de calculer la moyenne de chaque cluster pour déterminer son centroïde, on calcule plutôt la médiane.

2.2.1.2 Algorithmes hiérarchiques

Dans la fouille des données et les statistiques, l'algorithme hiérarchique est une méthode d'analyse de cluster qui cherche à construire une hiérarchie de clusters. En général, il y a deux types de stratégies pour le regroupement hiérarchique général (Rokach et al., 2005).

- Agglomérative: il s'agit d'une approche "ascendante". Chaque observation commence dans son propre cluster, et des paires de clusters sont fusionnées comme on se déplace vers le haut de la hiérarchie.
- Divisive: c'est une approche "descendante". Toutes les observations commencent dans un cluster, et la division est effectuée de manière récursive que l'on se déplace vers le bas de la hiérarchie.

En outre, il existe des méthodes dérivées telles que CURE (clustering using representative), BIRCH (balanced iterative reducing and clustering using hierarchies), etc.

- CURE utilise un algorithme de regroupement hiérarchique qui adopte un terrain intermédiaire entre les centroïdes et tous les points extrêmes, afin d'éviter les problèmes de taille des clusters ou de forme non uniformes (Guha et al., 1998).
- BIRCH. L'avantage de cette méthode est sa capacité à regrouper progressivement et dynamiquement les points de données entrants de métriques multidimensionnelles (Zhang et al., 1996).

En conclusion, l'algorithme hiérarchique est une méthode d'analyse de clusters qui cherche à construire une hiérarchie de clusters.

2.2.1.3 Algorithmes basés sur la densité

Ces algorithmes sont basés sur des fonctions de connectivité et de densité.

- DBSCAN (density-based spatial clustering of applications with noise) est une méthode de classification basée sur la densité spatiale des points (Ester et al., 1996) : étant donné un ensemble de points dans un espace, il regroupe des points étroitement groupés (points avec de nombreux voisins proches), marquant comme points aberrants se trouvant seuls dans des régions à faible densité dont les zones les plus proches (les voisins sont trop loin).
- OPTICS (ordering points to identify the clustering structure) est une méthode dont la base est similaire à celle de DBSCAN, mais elle s'attaque à l'une de ses principales faiblesses: le problème de la détection de clusters significatifs dans des données de densité variable. Pour ce faire, les points de la base de données sont ordonnés (linéairement) de sorte que les points spatialement les plus proches deviennent voisins dans l'ordre (Ankerst et al., 2008).

Dans le clustering basé sur la densité, les clusters sont définis comme des zones de densité plus élevée que le reste de l'ensemble de données (Kriegel et al., 2011).

2.2.1.4 Autres Algorithmes

D'autres méthodes de segmentation existent, telles que celles basées sur une structure de granularité de niveau multiple (STING, CLIQUE) et celles basées sur des modèles proposés pour chacun des clusters, l'idée étant de trouver le meilleur ajustement de ces modèles.

Les méthodes basées sur une structure de granularité à niveaux multiples (Liao et al., 2004) :

- STING (statistical information grid-based method): L'idée est de capturer les informations statistiques associées aux cellules spatiales de manière à pouvoir répondre à des problèmes de clustering sans recourir aux objets individuels (Wang et al., 1997).
- CLIQUE (clustering in quest): Cette méthode identifie les unités de densité dans les sous-espaces de l'espace de données dimensionnel élevé, et utilise ces sous-espaces pour fournir un clustering plus efficace (Agrawal et al., 1998).

Concernant la méthode de classification basée sur les modèles, un modèle est proposé pour chacun des clusters, et l'idée est de trouver le meilleur ajustement de ce modèle l'un pour l'autre (Yeung et al., 2001). Les différentes méthodes de classification permettent de choisir le meilleur ensemble de modèle pour résoudre le problème spécifique.

2.2.1.5 Synthèse de classification

Chaque méthode de segmentation présente son avantage et son inconvénient. K-means est une méthode populaire dans la segmentation d'exploration de données. Cependant, la distance calculée par k-means est généralement une distance euclidienne. Cette méthode utilise rarement d'autres types de distances, et elle a besoin d'un nombre prédéfini de clusters k. Les algorithmes hiérarchiques peuvent fonctionner sans déterminer le nombre de clusters prédéfinis, et ils peuvent s'utiliser avec la plupart des types de distance. L'inconvénient est qu'il ne convient pas d'utiliser des algorithmes hiérarchiques avec de gros ensembles de données en raison du temps de calcul. Pour les méthodes de classification basée sur la densité, il ne faut pas spécifier le nombre de clusters. Cependant, la qualité de la méthode basée sur la densité dépend de la mesure de distance utilisée dans l'algorithme, et elle ne convient pas aux grandes données établies avec des différences importantes dans les densités. En outre, pour la méthode basée sur une structure de granularité, il faut construire des grilles spécifiques pour chaque cas, etc.

En conclusion, il existe trois facteurs principaux qui influent sur l'efficacité d'une méthode de classification:

- Si le nombre de clusters est déterminé automatiquement ou non;
- Si la méthode peut être appliquée à de grands ensembles des données ou non (temps de calcul);

- Si l'algorithme est limité par le choix de métriques de distance. Par exemple, k-means est sur la base de la distance Euclidienne, et certaines métriques ne peuvent pas s'appliquer dans l'algorithme k-means. Dans la prochaine section, les métriques de similarité seront introduites.

2.2.2 Métriques de similarité

Un élément essentiel dans la majorité des méthodes de classification est l'utilisation d'une métrique qui vise à évaluer la distance entre 2 objets. Différentes métriques existent en fonction du contexte d'utilisation pour mesurer la (dis) similarité entre deux instances (vecteurs liés aux observations). Quatre types de métriques seront présentés ici. D'abord, la distance euclidienne et la distance Manhattan sont présentées. Ce sont les métriques classiques et elles sont populaires dans le domaine de fouille des données. Par contre, elles ne conviennent pas nécessairement dans le cas des séries temporelles. Deux métriques sont donc présentées en vue de résoudre ce problème. Ces autres métriques sont la distance corrélation croisée et la distance déformation temporelle dynamique.

2.2.2.1 Distance euclidienne

La distance euclidienne mesure la distance entre deux points dans l'espace euclidien (Deza et al., 2009). Soit x_i et y_j chaque être un vecteur *P*-dimensionnel. La distance euclidienne est calculée comme:

$$d_E = \sqrt{\sum_{k=1}^{P} (x_{ik} - y_{jk})^2} \qquad (1)$$

La distance euclidienne est largement utilisée dans de nombreux domaines d'application. En particulier, une application a été développée pour le système de transport en commun de Gatineau (Canada) qui contient de très grands ensembles de données, où environ 600 000 entrées sont collectées chaque mois. La distance euclidienne a été utilisée pour analyser ces données avec des résultats intéressants (Agard et al., 2006), comme présentés dans la section 2.3.1.

2.2.2.2 Distance Manhattan

La distance Manhattan, appelée aussi taxi-distance, est la distance entre deux points parcourus par un taxi lorsqu'il se déplace dans une ville où les rues sont agencées selon un réseau ou quadrillage. Un taxi-chemin est le trajet fait par un taxi lorsqu'il se déplace d'un nœud du réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau (Black, 2006).

Soit x_i et y_j chaque être un vecteur P-dimensionnel, la distance de Manhattan est définie par (Mori et al., 2016):

$$d_E = \sum_{k=1}^{P} |x_{ik} - y_{jk}| \tag{2}$$

Selon les formules (1) et (2), pour les deux distances, le résultat de la distance ne changera pas si l'ordre de *k* est modifié, par exemple, lorsque les valeurs de k_1 et k_2 sont échangées, la distance sera la même. Cependant, une série temporelle inclut une relation entre les observations du temps *t* elles-mêmes. De cette façon, les séries temporelles sont différentes en comparant des autres vecteurs. Pour une série temporelle, si les valeurs de k_1 et k_2 sont échangées, le résultat de la distance doit être modifié. Par conséquent, la distance euclidienne et la distance Manhattan ne sont pas adaptées aux séries temporelles (He et al., 2017).

2.2.2.3 Distance corrélation croisée (CCD)

La distance de corrélation croisée est basée sur la corrélation entre deux séries chronologiques. Nous mesurons la similarité entre deux séries chronologiques en décalant une des séries dans le temps afin de trouver une corrélation croisée maximale avec l'autre série chronologique. La corrélation croisée entre deux séries chronologiques au décalage k est calculée comme suit (Mori et al., 2016):

$$CC_k(X,Y) = \frac{\sum_{i=0}^{N-1-k}(x_i - \bar{x})(y_{i+k} - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_{i+k} - \bar{y})^2}} \tag{3}$$

Où $\bar{x}$ and $\bar{y}$ sont les valeurs moyennes de la série. Sur cette base, la mesure de distance est définie par:

$$CCD(X,Y) = \sqrt{\frac{(1 - CC_0(X,Y)^2)}{\sum_{k=1}^{max} CC_k(X,Y)^2}} \tag{4}$$

Dans le logiciel R (https://www.r-project.org), la mesure de la distance peut être calculée en utilisant une fonction. Cette fonction retourne la distance entre deux séries chronologiques en spécifiant deux vecteurs numériques (*x* et *y*) et un décalage maximum.

Le décalage maximum est un paramètre qui représente le délai maximal accepté pour comparer une série temporelle à une autre. La Figure 2-7 illustre le paramètre "décalage maximal" pour la corrélation croisée. Supposons deux séries temporelles identiques, telle que la première série temporelle soit décalée vers la droite par rapport à la deuxième pour qu'elles soient identiques. Le décalage peut être expliqué comme le nombre d'unités nécessaires pour décaler, de sorte qu'une série temporelle sera alignée sur une autre. La configuration du décalage maximal vise à limiter l'unité de décalage. Si elle n'est pas définie, ce paramètre continuera jusqu'à ce que la première valeur d'une série temporelle corresponde à la dernière valeur d'une autre série temporelle. Ensuite, la distance calculée peut omettre de nombreux points de temps, et la distance calculée peut ne pas refléter la dissemblance réelle des deux séries temporelles. Dans la Figure 2-7, si le décalage maximal est de 1, la 1ère série temporelle peut être décalée de 1 unité pour aligner la 2ème série temporelle. De cette façon, la 2ème série temporelle peut être acceptée par la 1ère série temporelle, de sorte que la 1ère et la 2ème série temporelle seront dans le même groupe, et la 3ème série temporelle sera dans un autre groupe.

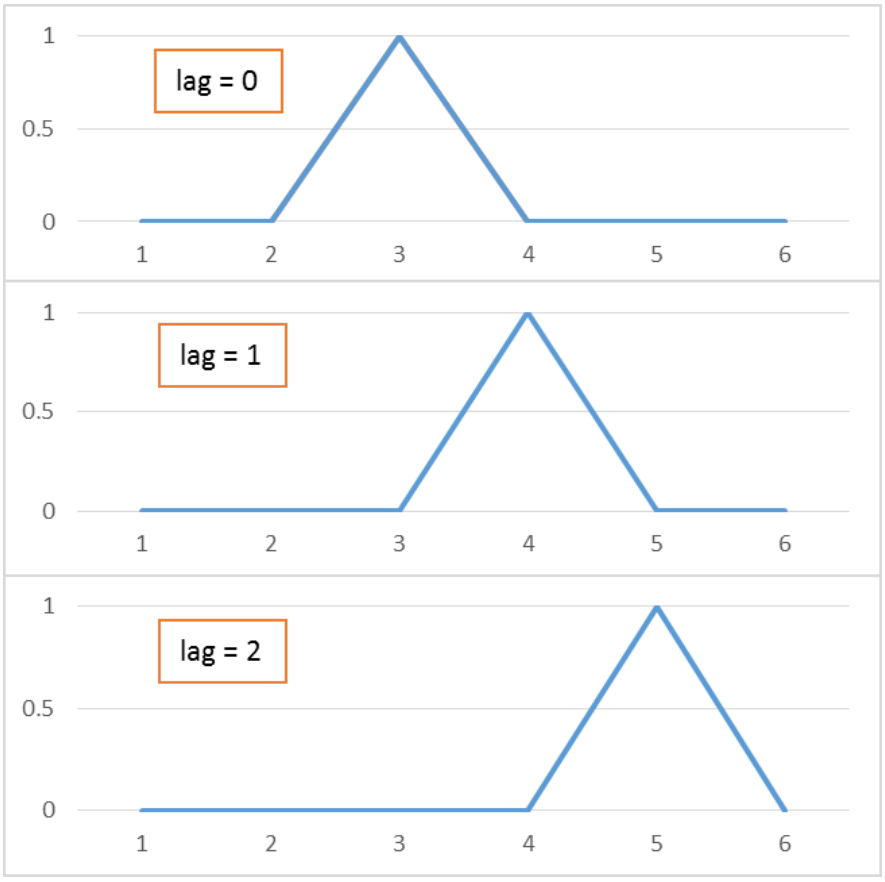


Figure 2-7: Calibration pour la méthode de la distance de la série temporelle: décalage maximal pour la corrélation croisée

2.2.2.4 Distance déformation temporelle dynamique (DTW)

La déformation temporelle dynamique (DTW) (Kruskall, 1983) a longtemps été utilisé pour trouver l'alignement optimal de deux signaux. L'algorithme DTW calcule la distance entre chaque paire possible de points sur deux séries temporelles en termes de valeurs de caractéristiques associées. Il utilise ces distances pour calculer une matrice de distance cumulée et trouve le chemin le moins coûteux à travers cette matrice. Ce chemin représente la déformation idéale - la synchronisation des deux signaux, qui permet de minimiser la distance entre leurs points synchronisés (Ten Holt et al., 2007).

Pour la déformation temporelle dynamique, le paramètre fenêtre maximale représente le nombre maximal de fois qu'une série temporelle peut être déformée. La Figure 2-7 illustre le paramètre "fenêtre maximale" pour la déformation temporelle dynamique. Lors de l'utilisation de la corrélation croisée, les séries temporelles sont décalées pour être comparées, tandis que pour la

déformation du temps dynamique, les valeurs des éléments des séries temporelles sont déformées (changées), de sorte que deux séries temporelles de différentes tailles peuvent être comparées. Par exemple, si la valeur du 4ème point du temps de la première série temporelle est déformée de 0 à 1, les première et deuxième séries temporelles seront identiques. La fenêtre des paramètres peut être expliquée comme le décalage maximal autorisé causé par la déformation.

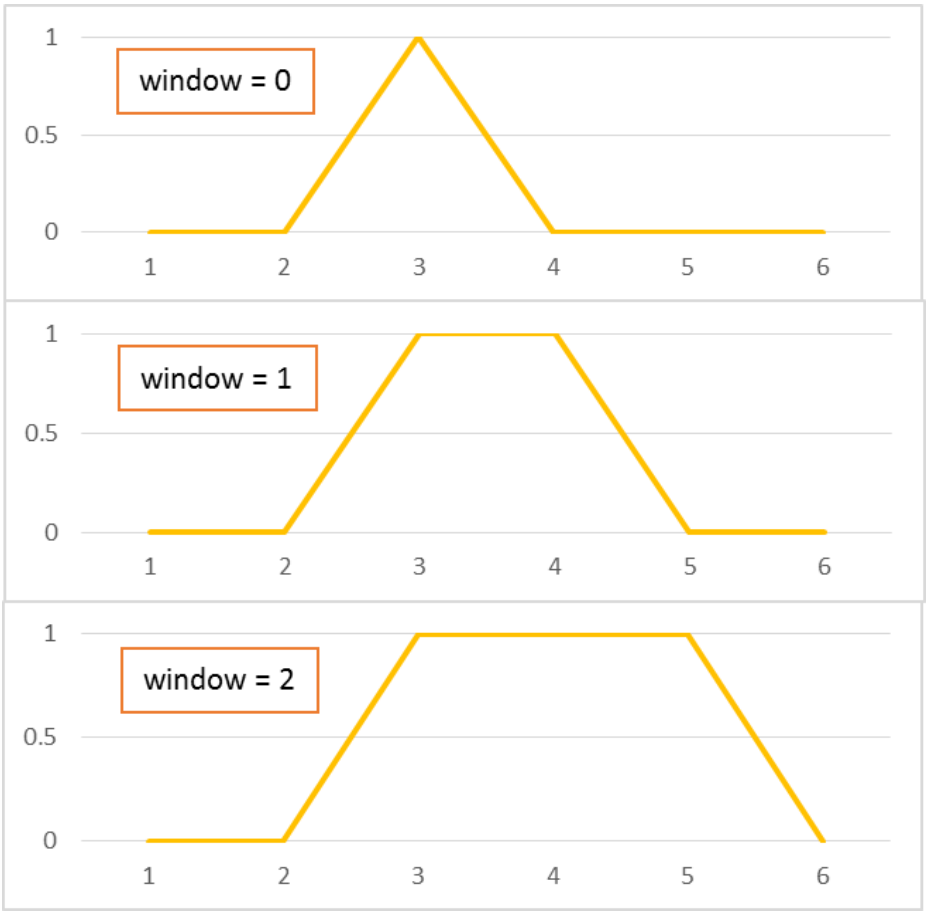


Figure 2-8: Calibration pour la méthode de la distance de la série temporelle: fenêtre maximale pour la déformation temporelle dynamique

Dans la Figure 2-8, si la fenêtre est 1, la 1ère série temporelle peut être déformée 1 unité afin que la 1ère série temporelle et la 2ème soient identiques, de cette façon, la 2ème série temporelle peut être acceptée par la première série temporelle, de sorte que la 1ère et la 2ème série temporelle seront dans le même groupe, et la troisième série temporelle sera dans un autre groupe.

En fait, la déformation temporelle dynamique est une technique populaire pour comparer les séries chronologiques, fournissant à la fois une mesure de distance qui est insensible à la compression et

des étirements locaux et la déformation qui se déforme de manière optimale une des deux séries d'entrée sur l'autre (Giorgino, 2009). La méthode pour calculer la déformation dynamique la distance de temps est la suivante (Berndt et al., 1994). Soit :

$$S = s_1, s_2, \dots s_i, \dots, s_n \quad (5)$$

$$T = t_1, t_2, \dots t_j, \dots, t_m \quad (6)$$

Les séquences *S* et *T* peuvent être disposés pour former un plan ou d'une grille n par m, où chaque point de la grille, *(i, j)*, corresponds à un alignement entre les éléments s_i et t_j. Un chemin de déformation, *W*, aligne les éléments de *S* et *T*, de telle sorte que la "distance" entre eux soit minimisée.

$$W = w_1, w_2, \dots w_k, \dots, w_P \quad (7)$$

C'est-à-dire que *W* est de la séquence de points de grille, où chaque w_k correspond à un point $(i,j)_k$.

Afin de formuler ce problème de programmation dynamique, nous devons avoir une mesure de distance entre deux éléments. De nombreuses mesures de distance sont possibles. Deux candidats pour une fonction de distance *d*, sont la grandeur de la différence (comme montré dans la fonction (8)) ou le carré de la différence (comme montré dans la fonction (9)).

$$d(i,j) = |s_i - t_j| \quad (8)$$

$$d(i,j) = (s_i - t_j)^2 \quad (9)$$

Une fois une mesure de distance sélectionnée, nous pouvons formellement définir le problème de déformation temporelle dynamique comme une minimisation sur les chemins de déformation potentiels basés sur la distance cumulative pour chaque chemin, où d est une mesure de distance entre deux éléments de séries temporelles.

$$\mathrm{DTW}(S,T) = \min w \left[\sum_{k=1}^{P} d(w_k) \right] \quad (10)$$

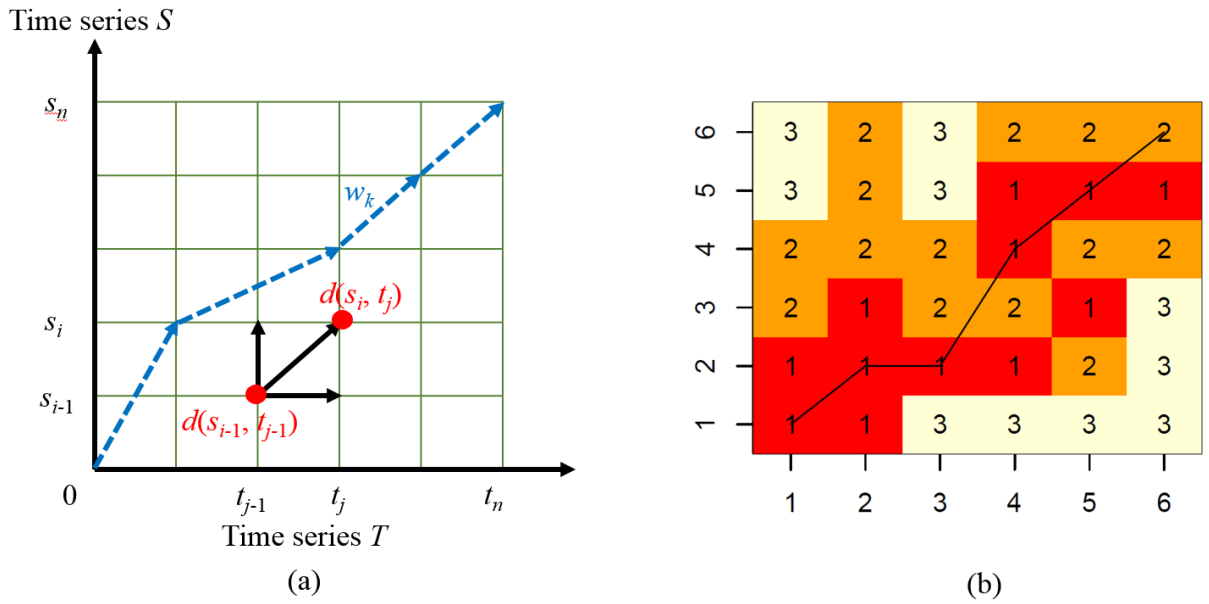


Figure 2-9: La fonction de la déformation temporelle dynamique

La Figure 2-9 montre comment fonctionne la déformation temporelle dynamique. Dans la Figure 2-9(a), pour obtenir une distance cumulative minimale, la série temporelle peut être déformée au prochain moment. Par exemple, le point de grille (S_i-1, T_j-1) peut être déformé à (S_i, T_j-1), (S_i-1, T_j), (S_i, T_j) pour calculer chaque distance. Et une séquence de points de grille Wk peut être un chemin de (S_0, T_0) à (S_m, T_n). Sur chaque point de la grille, la distance entre deux points de temps (moments) d (S_i, T_j) doit être calculée, comme le montre la Figure 2-9(b). Ensuite, tous les chemins possibles du point de grille (1, 1) à (6, 6) sont calculés, pour trouver le chemin avec une distance cumulative minimale. Dans cette grille de la Figure 2-9(b), la distance de la déformation temporelle dynamique est de 7.

2.3 Fouille des données de cartes à puce en transport en commun

Dans cette partie, nous présenterons d'abord les méthodes de base de classification des comportements d'usagers de cartes à puce. Ensuite, nous présenterons les recherches plus récentes en termes de fouilles spatio-temporelles des données de cartes à puce.

2.3.1 Fouilles classiques

Cette recherche présente une méthode générale de « planification des transports par la fouille de données » pour l'analyse du comportement des usagers. Des expériences ont été effectuées sur des

données provenant d'une autorité de transit canadienne. Ces expériences démontrent qu'une combinaison d'outils de planification et de fouille de données permet de produire des indicateurs de comportements de déplacement, principalement en ce qui concerne la régularité et les modèles quotidiens, à partir de données issues du système opérationnel et de gestion. Les résultats montrent que les utilisateurs des transports en commun de cette étude peuvent être divisés rapidement dans quatre groupes de comportement majeurs, quel que soit le type de ticket qu'ils utilisent (Agard et al., 2006). L'analyse de la classification montre le comportement général des usagers.

Le traitement de l'ensemble de données produit 4 clusters de semaines d'utilisateurs avec des modèles similaires. Deux des clusters ont des comportements de déplacement facilement interprétables.

À la Figure 2-10, le groupe (45,6% des semaines-usager) se rapporte clairement aux personnes ayant des déplacements réguliers à destination et en provenance d'activités contraignantes telles que le travail, car elles se déplacent principalement pendant les heures de pointe. En moyenne, 79,4% de ces usagers voyagent pendant l'heure de pointe du matin et 71,0% pendant l'heure de pointe de l'après-midi en semaine. La proportion d'usagers voyageant pendant les autres périodes est assez faible, 6,4% pendant la journée et 2,6% pendant la soirée.

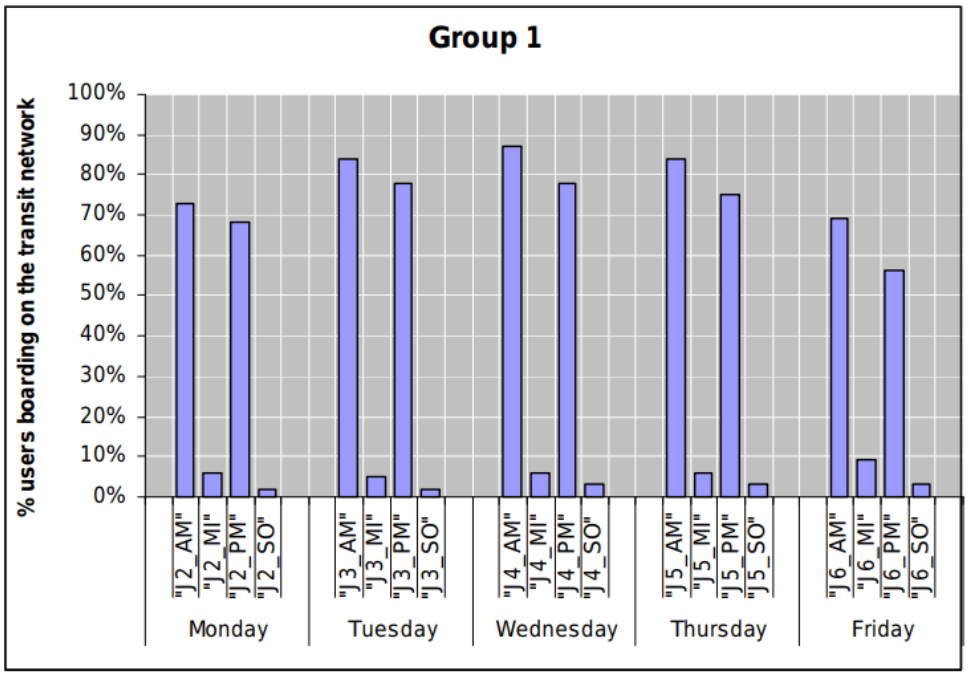


Figure 2-10: Comportement général des usagers – Groupe 1 (Agard et al., 2006)

Les tableaux suivants résument la composition des quatre grappes créées par la méthode de fouille de données. D'une part, le Tableau 2-1 présente la répartition des semaines d'usagers dans les quatre groupes en fonction du type de prise de carte par le voyageur. Par exemple, il montre que près de 80% des semaines de déplacement des détenteurs de cartes anciennes sont classés dans le groupe 4. Ce n'est pas une surprise, car ce groupe a un faible niveau de mobilité. Il confirme également que les déplacements symétriques observés pour les usager-semaines du groupe 1 s'expliquent par des activités des adultes, avec près de 60% des titulaires de cartes adultes appartenant à ce groupe.

Tableau 2-1: Répartition des usager-semaines dans les quatre clusters selon le type de carte (Agard et al., 2006)

Card type	Gr1	Gr2	Gr3	Gr4	TOT
Adult	**58,8%**	13,9%	9,2%	18,1%	100%
Student	21,0%	17,7%	26,4%	34,8%	100%
Elderly	6,2%	6,4%	7,9%	**79,5%**	100%

Une meilleure compréhension de ces deux groupes est obtenue par l'étude de la variabilité de l'appartenance des utilisateurs au cours des 12 semaines.

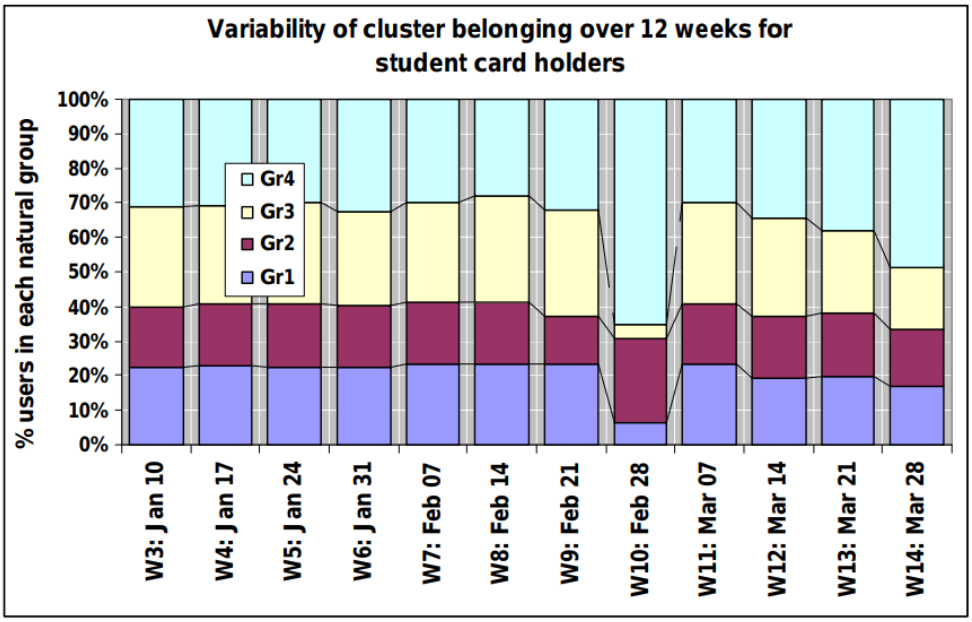


Figure 2-11: Variabilité du groupe appartenant à plus de 12 semaines des titulaires de carte d'étudiante (Agard et al., 2006)

Une étude de l'appartenance à ces groupes au cours des 12 semaines, présentée à la Figure 2-11, montre clairement une semaine de voyage irrégulière. La distribution atypique observée à la semaine 10 est due à la pause scolaire où les élèves adoptent temporairement d'autres modèles d'activité.

2.3.2 Fouilles spatio-temporelles

Les données des cartes à puce se composent généralement de deux types d'informations : spatiale et temporelle. Les données spatiales comprennent des coordonnées de l'arrêt de bus ou des stations, par exemple la latitude et la longitude qui peuvent être les données GPS ou les valeurs de localisation. Les données temporelles comprennent l'heure de début de chaque déplacement. Nous calculons cette information comme un vecteur 0-1, où le début du déplacement est identifié par 1. Selon ces informations, l'analyse du mode d'utilisation du transport public basé sur les données de la carte à puce peut être divisée en trois catégories: (1) motif spatial, (2) motif temporel et (3) motif temporel spatial (Ghaemi et al., 2015).

2.3.2.1 Fouille temporelle

Cette section discute surtout la classification des comportements des usagers en tant que des séries temporelles. Une technique innovante est introduite pour regrouper et caractériser les usagers des transports publics à partir des données temporelles (Agard et al., 2013). Ensuite, une nouvelle technique de calcul de distance est proposée par les auteurs pour appliquer la méthode de clustering k-means (Ghaemi et al., 2016).

Supposons qu'un vecteur 0 - 1 des données temporelles soit donné dans l'entrée comme il est indiqué dans le Tableau 2-2. L’idée principale de cette méthode est de projeter les données temporelles dans un demi-cercle comme indiqué dans la Figure 2-12. Ensuite, la dissimilarité de deux vecteurs est mesurée par la distance de deux localisations de points de vecteur.

Tableau 2-2: Séquence des données temporelles pour le calcul de la distance (Ghaemi et al., 2016).

User	1	2	3	4	5	6	7	...	24
X_1	1	0	0	0	0	0	0	...	0
X_2	0	1	0	0	0	0	0	...	0
X_3	0	0	1	0	0	0	0	...	0
X_4	0	0	0	1	0	0	0	...	0
X_5	0	0	0	0	1	0	0	...	0
X_6	0	0	0	0	0	1	0	...	0
X_7	0	0	0	0	0	0	1	...	0
X_8	1	1	0	0	0	0	0	...	0
X_9	1	0	1	0	0	0	0	...	0
X_{10}	0	1	1	0	0	0	0	...	0
X_{11}	1	0	0	1	0	0	0	...	0
X_{12}	0	0	0	0	1	1	0	...	0
X_{13}	0	0	0	0	0	1	1	...	0

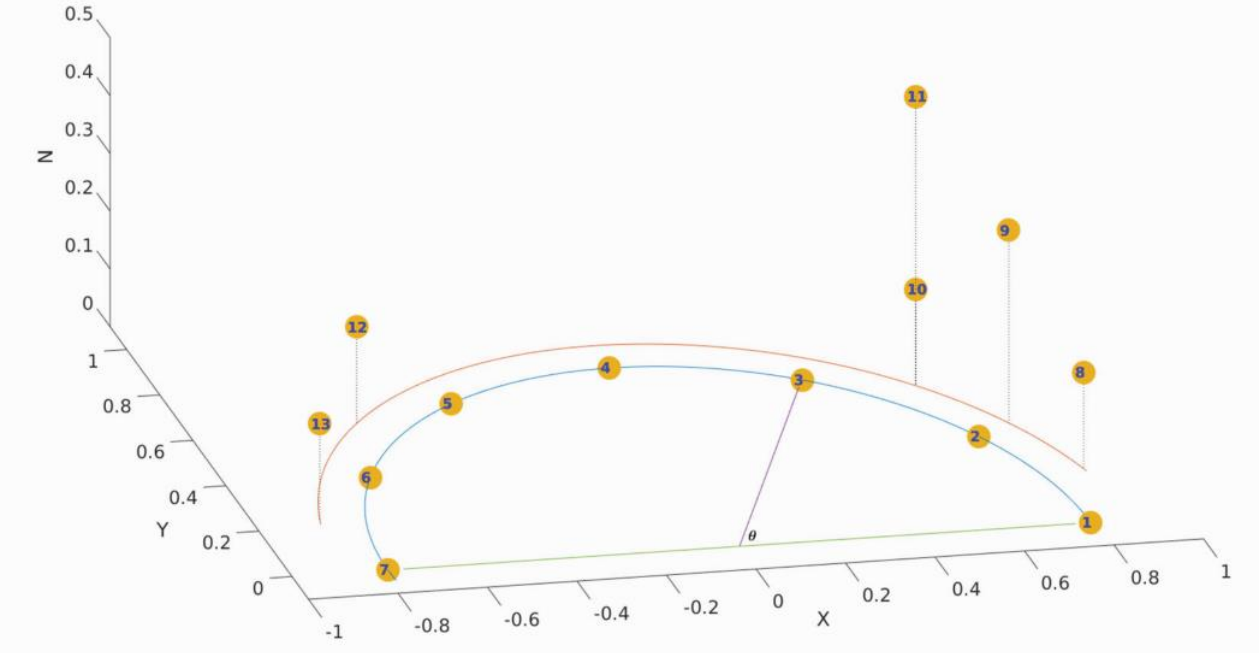


Figure 2-12: Mapping des données temporelles dans les coordonnées sphériques (Ghaemi et al., 2016)

2.3.2.2 Fouille spatiale

Concernant la fouille spatiale, l'idée principale est d'apparier les arrêts de bus pour la mesure de la dissimilarité des usagers. La Figure 2-13 montre trois lignes et des flèches de différentes couleurs, où les séries temporelles rouge et orange sont comparés à l'usager bleu. Les points représentent les valeurs à chaque moment donné, et les flèches représentent la séquence des séries temporelles.

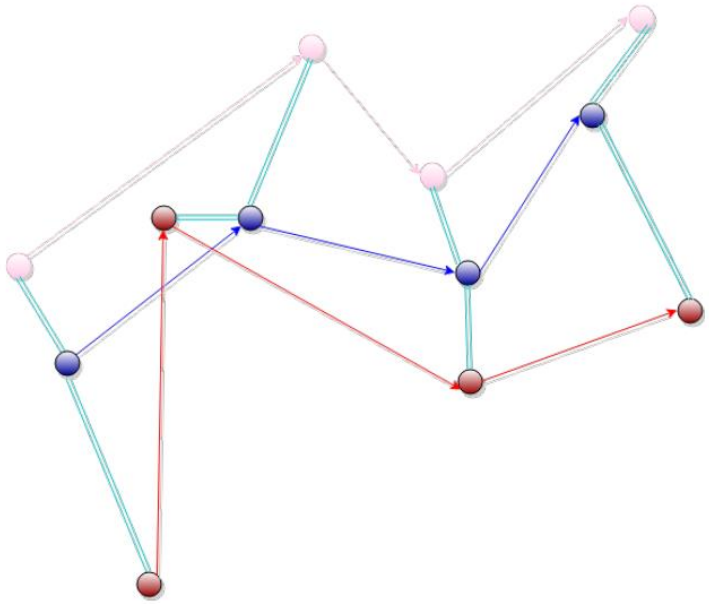


Figure 2-13: Match des arrêts de bus pour la mesure de la dissimilarité des usagers (Ghaemi et al., 2015)

La somme des différences entre toutes les paires d'arrêts de bus entre les cercles bleu et rouge (lignes vertes) identifie la dissimilarité de l'usager bleu et l'usager rouge. De manière analogue, nous pouvons calculer la dissimilarité de l'usager bleu et de l'usager orange. Les auteurs présentes 10 patterns de match :

(1) Deux usagers avec le même point de départ et point de terminus.

(2) Deux usagers prennent les mêmes bus dans les directions opposées.

(3) Deux usagers ayant le même motif directionnel (les arrêts sont différents).

(4) Deux usagers ayant le même motif directionnel.

(5) Deux usagers ayant le même motif symétrique.

(6) Deux usagers ayant le même motif, sauf pour certains arrêts tels que leur deuxième match.

(7) Deux usagers partiellement similaires, mais avec le nombre différent des arrêts de bus.

(8) La même distance parcourue avec les différents arrêts de bus.

(9) Deux usagers prennent les mêmes bus, mais avec la séquence différente.

(10) Le même motif des deux usagers vivant dans les différents endroits.

2.3.3 Autres travaux de fouilles

En plus des travaux précédents, plusieurs recherches se concentrent sur les autres objets de classifications et diverses méthodes de classifications, afin d'obtenir les résultats ci-après :

- K-means s'applique en vue de classifier les arrêts basés sur la dissimilarité de l'heure de transaction, afin d'identifier les arrêts populaires (Morency et al., 2006). K-means s'applique également en vue de classifier les passagers basé sur la dissimilarité de l'heure de transaction, afin d'identifier les passagers réguliers (Morency et al., 2007).
- Réseau neuronal peut également s'appliquer afin d'identifier les passagers réguliers, dans ce cas-ci, nous mesurons la similarité par l'heure et la localisation de transaction (Ma et al., 2013).
- DBSCAN (Density-based spatial clustering of applications with noise) peut également s'appliquer afin d'identifier les passagers réguliers, dans ce cas-ci, nous mesurons la similarité par le dernier arrêt puis premier arrêt d'embarquement du jour puis embarquement (Kieu et al., 2014).
- La fouille de données peut viser à identifier l'objectif de déplacement, en utilisant la méthode Naïve Bayes (Kusakabe et al., 2014). La fouille de données peut également viser à identifier le centre urbain, en utilisant la méthode hiérarchique agglomératif (Cats et al., 2015).
- Les travaux récents de Aghabozorgi et al. (2015) peuvent être envisagés pour approfondir des solutions classification pour un ensemble assez grand de métriques de distance. De même, les nombreuses recherches sur les diverses sources à grande échelle de données de

mobilité individuelle (GPS, données de téléphone mobile, données Bluetooth) sont présentées (Naboulsi et al., 2017, Ketabi et al., 2019).

- Une solution basée sur DBSCAN et une métrique spatio-temporelle sont proposées pour identifier des groupes de trajectoires de véhicules similaires du point de vue spatio-temporel à partir de données GPS (Ketabi et al., 2019). En ce qui concerne les trajectoires de la téléphonie mobile, Kang et al. adoptent une variante de la distance de *Longest Common Subsequence* (LCSS) pour trouver des groupes de trajets (Kang et al., 2009), et Yuan et al. proposent d'utiliser une définition modifiée de *Edit Distance* dans une fonction de coût spatio-temporel (Yuan et al., 2014).

En conclusion, les fouilles des données sont largement utilisées pour la classification des comportements des usagers de cartes à puce. Les fouilles classiques se concentrent notamment sur la méthode k-means. Les usagers sont regroupés dépendamment de leurs heures de transactions. Concernant les fouilles spatio-temporelles, les auteurs commencent à considérer les comportements des usagers pendant une journée comme des séries temporelles. Pourtant, au lieu de traiter les séries temporelles directement, ils projettent les comportements à un espace 3D pour utiliser les méthodes de fouilles classiques. Ces méthodes fonctionnent, cependant, le processus de projection de 3D peut causer la perte d'informations sur les comportements des usagers. Les autres méthodes ne peuvent pas résoudre ce problème quand même, même si beaucoup d'eux ont été développées en vue de classifier une série de types de comportements des usagers en transport en commun.

2.4 Synthèse de la revue de la littérature

Dans la section de la revue de la littérature, l'introduction des données de cartes à puce est premièrement présentée. Avec ces données, des méthodes de prétraitement sont développées pour l'enrichissement des données. Ensuite, afin de mieux connaître les comportements des usagers et les caractéristiques des réseaux de transport collectif, une variété de méthodes sont illustrées. Beaucoup de ces méthodes se basent sur les statistiques et dans cette partie peu de recherches a utilisé la méthode de la fouille des données. Cependant, classifier les comportements des usagers pour les regrouper dans un nombre limité de clusters, c'est un sujet important, parce que cela nous permet de fournir les services de transport en commun différemment pour répondre leurs besoins.

Dans la section 2.2, les principales méthodes de classification et métriques de similarité appliquée à l'analyse des données de cartes à puce sont présentées. Avec les métriques de similarité, une matrice dans laquelle la distance (dissimilarité) de n'importe quelles deux observations peuvent être générées. Les méthodes de classification peuvent regrouper les observations en quelques clusters en utilisant cette matrice. Le problématique principal de la thèse est la classification des séries temporelles, par conséquent, les métriques traitant les séries temporelles (CCD, DTW) sont soulignées. Nous avons aussi expliqué la raison pour laquelle les méthodes de classification ne fonctionnent pas avec l'objectif de cette thèse. (Les méthodes de classification existantes sont sur la base de la distance euclidienne, qui ne fonctionne pas lors que nous traitons les séries temporelles). Il existe des centaines de méthodes de fouilles des données, mais la chose la plus important pour un(e) chercheur(se), est de trouver une méthode la plus pertinente dans son domaine et sa problématique, car il n'existe pas encore de méthode parfaite pour tous les cas d'études.

En combinant les sujets des section 2.1 et 2.2, la section 2.3 présente une variété de méthodes pour la classification des comportements des usagers en transport en commun en utilisant les méthodes de fouilles des données. Le problème reste toujours lors de la classification des séries temporelles, même s'il existe une méthode qui projette les comportements à un espace 3D pour utiliser les méthodes de fouilles classiques, qui risque de perdre des informations des comportements des usagers. Dans nos cas d'études, les problèmes principaux seront de trouver la meilleure méthode de classification ainsi que la meilleure métrique de similarité pour la classification temporelle, la classification spatiale, et la classification spatio-temporelle.

CHAPITRE 3 DÉMARCHE DE L'ENSEMBLE DU TRAVAIL DE RECHERCHE ET ORGANISATION GÉNÉRALE DE LA THÈSE

Dans ce chapitre, nous présentons la démarche de l'ensemble du travail, incluant la définition des objectifs (Section 3.1.1), la préparation des données (Section 3.1.2), et cinq contributions (de Section 3.2.1 à Section 3.2.5). Pour chaque contribution, nous allons présenter le design des méthodes, l'implémentation, et l'analyse des résultats. Nous allons indiquer la cohérence des articles par rapport aux objectifs.

3.1 Démarche de l'ensemble du travail de recherche

Afin de présenter la démarche de l'ensemble du travail de recherche, nous allons introduire les parties suivantes respectivement : la définition des objectif, la préparation des données, le design des méthodes, l'implémentation et l'analyse des résultats.

Parmi elles, la définition des objectifs et la préparation des données seront présentées dans cette section (Section 3.1). Cependant, le design des méthodes, l'implémentation, et l'analyse des résultats seront fusionnés à la prochaine section (Section 3.2).

3.1.1 Définition des objectifs

3.1.1.1 Défi général

Le défi général de cette thèse est de proposer des méthodes pour mieux analyser les données de cartes à puce afin d'améliorer les services de transport en commun. Dans cette recherche, en analysant les données de carte à puce, nous pouvons mieux comprendre les comportements des usagers. Dans la thèse, les aspects comportementaux concernent plutôt l'heure de transaction et la localisation où un usager demeure pendant une certaine période, même si parfois nous pourrions considérer les autres objets (tels que les motifs de déplacement) comme aspects comportementaux. Concernant la temporalité, les autorités de transport pourront répondre à la demande en optimisant les horaires des véhicules. Concernant la spatialité, les autorités de transport pourront répondre à la demande en optimisant la géométrie des réseaux des transports en commun.

3.1.1.2 Défi méthodologique

Le défi méthodologie consiste à développer des méthodes de classification qui tiennent compte du caractère spécifique des déplacements sur un réseau de transport collectif en tenant compte de l'espace et du temps. Pour montrer le problème méthodologique de classification spatio-temporelle, un exemple est présenté.

La Figure 3-1 illustre le problème de classification des séries d'heures de transactions du jour. Dans cette figure, l'axe X représente les heures d'une journée, soit entre 4:00 du matin et 1:30 du matin (le lendemain). Il y a 3 usagers de carte à puce dans cet exemple :

- Le 1er usager part à 6:30 du matin pour se rendre à l'école, il fait une transaction avec sa carte à puce. Ensuite, à 16 :00, l'après-midi, il rentre chez soi avec une autre transaction de carte à puce.
- Le 2ème usager part un peu plus tard, à 7:00 du matin pour se rendre au travail. Puis, il rentre chez soi un peu plus tard que le 1er usager, à 18:00, l'après-midi.
- Le 3ème usager fait les premières deux transactions à même heure que le 1er usager. Pourtant, avant de rentrer chez lui, il se rend au supermarché, et il donc fait une troisième transaction. L'heure de dernière transaction de 3ème usager et 2ème usager est le même, soit à 18:00, l'après-midi.

Actuellement, le problème est de savoir comment classifier ces trois comportements du jour. Est-ce que le comportement de 1er usager est plus proche du 2ème usager ou du 3ème usager? C'est-à-dire que : si le 1er usager et le 2ème usager doivent être dans le même groupe, et le 3ème usager doit-il être dans un autre groupe?

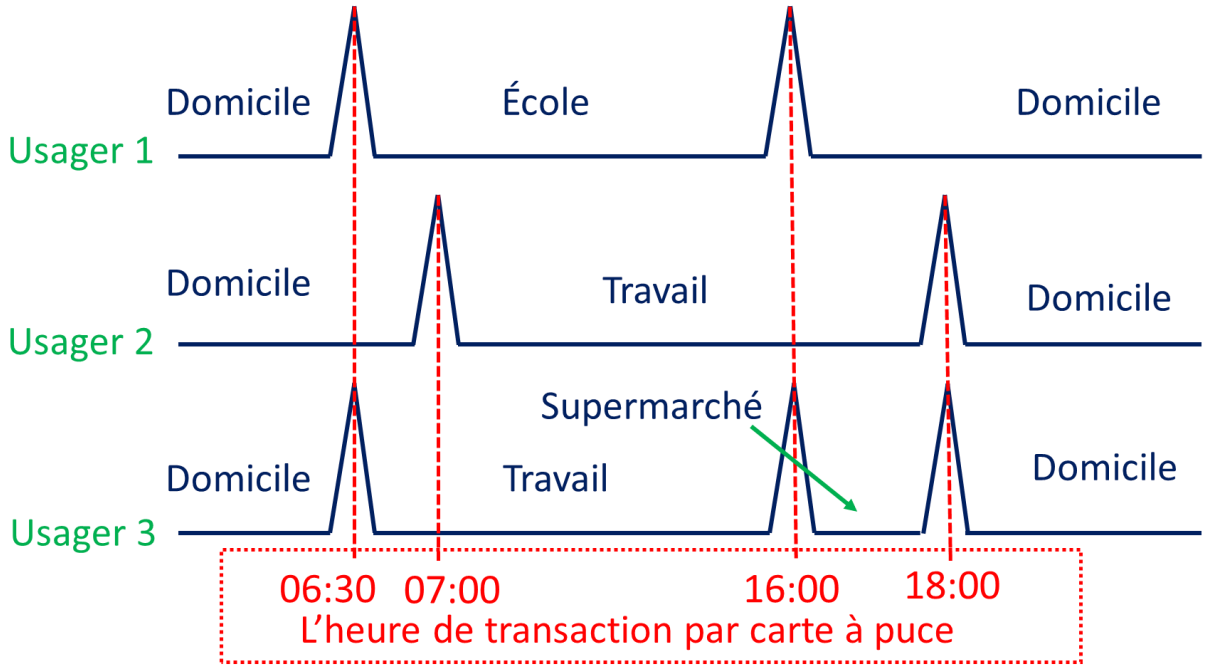


Figure 3-1: Problème de classification de série d'heures de transactions du jour

En termes de spatialité, la Figure 3-2 présente le problème de classification de séries de localisations des transactions du jour. Les comportements de trois usagers sont les mêmes que la Figure 3-1. Par contre, pour la classification spatiale, ce qui est important est la distance spatiale entre deux usagers à un moment donné. Par exemple, à 08:00 du matin, nous utilisons la distance euclidienne des coordonnées de localisation en vue de mesurer la dissimilarité spatiale de ces deux usagers à ce moment-là. Nous devons comparer les dissimilarités spatiales pendant ce jour, afin de classifier le comportement spatial de ce jour.

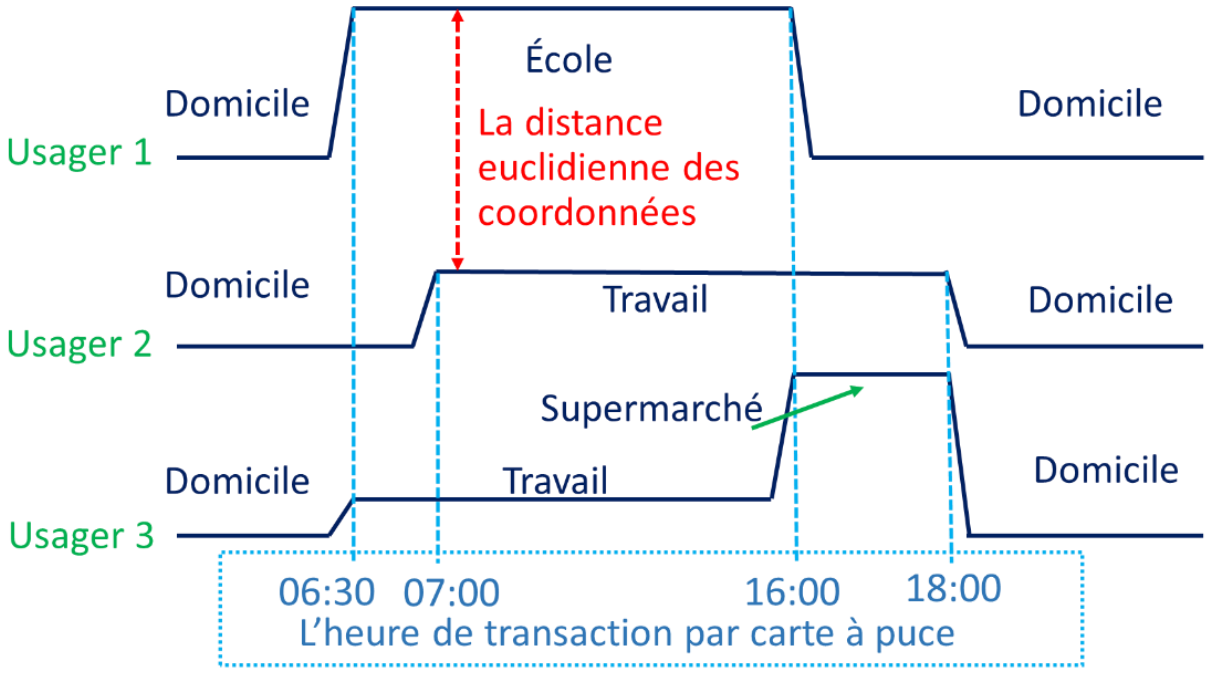


Figure 3-2: Problème de classification de série de la localisation des transactions du jour

Dans cette thèse, nous réutilisons des éléments de la méthodologie actuelle pour expliquer le comportement de l'usager en utilisant les données de la carte à puce. Cependant, nous nous basons sur chacune des transactions de l'usagers de la carte à puce, et non seulement sur l'allure de leur série temporelle. En plus de l'analyse pour chaque transaction distincte par les usagers, une question d'intérêt est d'agréger toutes les transactions pour un usager, puis de classer le comportement quotidien de cet usager. En fait, la méthode de classification actuelle n'est pas adaptée pour résoudre ce problème car elle n'est pas conçue pour mesurer la similarité entre les séries temporelles. L'introduction de la technique de classification de séries temporelles aidera à développer une méthode pour ce problème.

3.1.2 Préparation des données

Dans cette partie-ci, nous allons d'abord présenter les données de notre partenaire la Société de Transport de l'Outaouais (STO). Une méthode similaire a été utilisée pour Santiago. Ensuite, un résumé des erreurs dans ces données est montré. À partir des dernières deux parties, une méthode d'estimation des destinations est introduite.

3.1.2.1 Données testées

Dans le projet de doctorat, les données testées font partie de la base de données de cartes à puce de la STO (sauf le Chapitre 8). Un extrait de tableau des données de la STO sera donc montré et expliqué. Les enregistrements de la base de données de cartes à puce contiennent quatre tables : « transaction », « ligne-arrêt », « arrêt » et « ligne ».

3.1.2.1.1 Tableau « transaction »

Le Tableau 3-1 est un extrait des enregistrements de la table transaction. L'explication des attributs est comme suit:

Tableau 3-1: Enregistrements de la table « transaction »

Identificateur	NumCarte	DateComp	HeureComp	NUM_LI	NUM_SENS	NUM_ARRÊT
65510379	173355871	2014-09-01	0651	25	0	4370
65510388	71446948	2014-09-01	0954	84	0	3806
65510405	342504612	2014-09-01	1704	12	0	3864
65510644	69394235	2014-09-02	0522	98	0	4600
65510658	205283461	2014-09-02	1534	68	0	4746
65510673	80273467	2014-09-03	0655	65	0	3001
65510751	203657861	2014-09-03	1049	288	0	4060
65510763	182995042	2014-09-03	1151	288	0	4058
65510778	213291653	2014-09-03	1655	696	0	4706

- Identificateur : La clé primaire de cette table, qui représente un enregistrement unique pour une transaction.
- NumCarte : Le numéro de la carte, représentant chaque usager qui se déplace de façon unique. Chaque carte contient une photo de l'usager, ce qui assure l'unicité d'individu. Ceci est un avantage de la base de données de la STO pour analyser l'activité de déplacement de chaque individu. Cependant, pour des raisons de confidentialités, nous n'avons pas accès à ces données nominatives, seulement le numéro de carte.
- DateComp : La date de la transaction sous forme de « année-mois-jour ». La période s'étend du 2014-09-01 au 2014-09-30.
- NUM_LI : Le numéro de la ligne. Dans le réseau de transport en commun de la STO, il a 138 lignes de bus à traiter.

- NUM_SENS : Le numéro du sens. Pour une ligne régulière, elle a deux sens. Les numéros « 0 » et « 1 » représentent les différents sens d'une ligne. Pour une ligne en boucle, il n'a qu'un numéro « 0 » ou « 1 » représentant le sens unique.
- NUM_ARRET : Le numéro de l'arrêt de la transaction, représentant l'emplacement d'embarquement du déplacement. Il existe, au moment des traitements effectués, un total de 2007 arrêts au sein du réseau examiné.

Le tableau « transaction » enregistre donc principalement des informations individuelles sur les transactions des cartes.

3.1.2.1.2 Tableau ligne-arrêt

Le Tableau 3-2 est un extrait des enregistrements de la table "ligne-arrêt". L'explication des attributs est comme suit:

- NUM_SSLI : Le numéro du sens de cette ligne de bus, comme l'attribut « NUM_SENS » dans la table « transaction ».
- NUM_ORDRE : Le numéro de séquence de l'arrêt dans un sens de la ligne. Le numéro « 0 » représente le point de départ de ce sens de cette ligne, et le numéro le plus grand de ce sens de cette ligne représente le terminus.
- Distance : La distance entre le point de départ d'un sens de cette ligne et l'arrêt donné du même sens de la même ligne.

Le tableau « ligne-arrêt » enregistre donc principalement des informations de l'espace à une dimension en intégrant les informations utiles des lignes et des arrêts.

Tableau 3-2: Enregistrements de tableau « ligne-arrêt »

NUM_LI	NUM_SSLI	NUM_ARRÊT	NUM_ORDRE	Distance
1	0	5532	0	0
1	0	7048	1	1063
1	0	7046	2	2090
1	0	7044	3	3082
1	0	5520	4	3321
1	0	7042	5	3974
1	0	7040	6	5345
1	0	7038	7	6020
1	0	7036	8	6669
1	0	7034	9	7435
1	0	5516	10	7728
1	0	…	11	…

3.1.2.1.3 Tableau « arrêt »

Le Tableau 3-3 est un extrait des enregistrements de la table arrêt. L'explication des attributs est comme suit :

- NUM_ARRET : le même sens que dans la table « transaction », étant la clé primaire de cette table.
- COORD_X et COORD_Y : La latitude et la longitude. Ce système utilise la projection universelle de Mercator (Universal Transverse Mercator, UTM) pour décrire les coordonnées des arrêts. Le "18" au début de la coordonnée X indique qu'il s'agit de la zone 18, il faut donc enlever ce nombre avant de représenter les données.
- LIBEL_ARRET : Le nom de l'arrêt. Cet attribut vise à décrire chaque arrêt en mots.

Notons que dans cette table, certains enregistrements sont factices. Par exemple, avec les coordonnées « 0 », les arrêts « 1 » et « 2 » n'existent pas en réel. Il faudra donc les supprimer lors de l'implantation.

Le tableau « arrêt » enregistre donc principalement des informations de l'espace à deux dimensions concernant les coordonnées des arrêts.

Tableau 3-3: Enregistrements de la table « arrêt »

NUM_ARRET	COORD_X	COORD_Y	LIBEL_ARRET
1	0	0	DUMMY 1
2	0	0	DUMMY 2
1000	18433093	5028194	FRONT/CORMIER
1001	18433003	5028579	FRONT/DE LA TERRASSE-EARDLEY
1002	18433113	5028181	FRONT/CORMIER
1003	18432982	5028570	FRONT/DE LA TERRASSE-EARDLEY
1004	18433091	5028050	FRONT/PEARSON
1005	18433089	5028433	FRONT/DE LA TERRASSE-EARDLEY

3.1.2.1.4 Tableau « ligne »

Le Tableau 3-4 est un extrait des enregistrements de la table transaction. L'explication des attributs est comme suit :

- « Num_li » et « num_ssli » : Les mêmes que « NUM_LI » et « NUM_SSLI » dans la table de « ligne-arrêt ». Notons que ces deux attributs construisent une clé primaire de la table « ligne », ils déterminent ensemble l'attribut « direction ».
- Direction : une valeur pour distinguer le sens dans une ligne.

Le tableau « ligne » enregistre donc principalement des informations sur les lignes sans mentionner les arrêts. Par rapport à la table « ligne-arrêt », elle nous fournit le sens réel, au lieu d'un numéro « 0 » ou « 1 » dans la table « ligne-arrêt ».

Tableau 3-4: Enregistrements de la table « ligne »

num_li	num_ssli	direction
1	0	Sud
1	1	Nord
3	0	Sud
3	1	Nord
5	0	Sud
5	1	Nord
6	0	Sud
6	1	Nord
11	0	Sud
11	1	Nord

3.1.2.1.5 Modèle relationnel

La Figure 17 présente la relation entre les tables. La table « transaction » et celle de « ligne-arrêt » sont liées par les trois attributs : « NUM_LI », « NUM_SENS » (soit « NUM_SSLI ») et « NUM_ARRET ». La table « ligne-arrêt » et celle de « ligne » sont liées par les attributs « NUM_LI » et « NUM_SSLI ». La table « ligne-arrêt » et celle de « arrêt » sont liées par l'attribut « NUM_ARRET ».

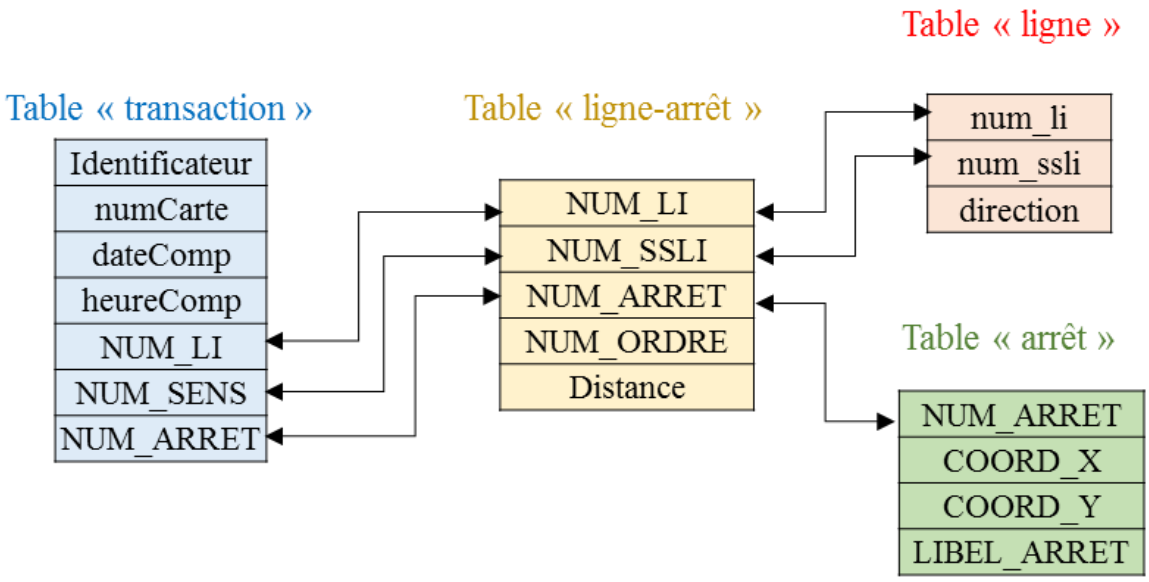


Figure 3-3: Modèle Relationnel de la base de données (He, 2014)

En fonction des relations entre les tables, il est possible d'aborder la stratégie de l'implantation pour utiliser toutes les données utiles dans cette base de données. D'abord, à l'aide de la table « transaction », il est possible de savoir qui se déplace, quand et où il fait la transaction, ainsi que quelle ligne de bus et quel sens cet usager utilise pour se déplacer. Ensuite, en utilisant les attributs « NUM_LI », « NUM_SENS » et « NUM_ARRET », ce déplacement peut être lié à la table « ligne-arrêt ». Nous pouvons donc savoir les informations de tous les arrêts dans le sens de cette ligne. Finalement, en liant à la table « arrêt » par l'attribut « NUM_ARRET », il est possible d'obtenir les coordonnées de tous les emplacements de débarquement potentiels. Notons que l'attribut « direction » dans la table « ligne » n'est pas nécessaire lors de l'implantation, parce que le numéro de l'attribut « sens » peut représenter la direction d'une ligne. La table « ligne » peut donc être ignorée au niveau de l'implantation (He, 2014).

3.1.2.2 Prétraitement des données

Avant l'utilisation des données, il faut d'abord vérifier la validité des données. La validation individuelle vise à détecter les erreurs en vérifiant l'enregistrement des données une par une. On peut par exemple trouver les erreurs suivantes (Trépanier et al., 2004).

- Des lignes non existantes (géoréférencées en coordonnées nulles).
- La définition du réseau est incomplète par rapport à notre base de données.
- L'arrêt n'est pas référencé dans la base de données.
- Des arrêts-ligne des données de montées n'appartiennent pas à la ligne considérée.
- Certaines montées sont effectuées au dernier arrêt du tracé de la ligne considérée.
- L'heure de départ du voyage (indice caractérisant le numéro de voyage) est nulle.

3.1.2.3 Algorithme d'estimation des destinations

L'algorithme pour les déplacements unitaires et les méthodes de calibration est présenté dans la partie 2.1.2.1. Basé sur ces méthodologies (la méthode de prétraitement des données, celle d'estimation des destinations traditionnelle et celle pour les déplacements unitaires), ces divers types d'estimation des destinations sont obtenus (Giraud, 2016):

- Code 11 : Séquence de déplacement
- Code 12 : Retour à domicile
- Code 13 : Déplacement du prochain jour
- Code 21 : Déplacement unitaire avec plusieurs emplacements de débarquement potentiels.
- Code 22 : Déplacement unitaire avec emplacement de débarquement unique.
- Code 30 : Pas de destination trouvée
- Code 31 : Pas de destination trouvée – impossible de créer l'historique des transactions
- Code 400 : Id arrêt vide
- Code 404 : Id arrêt inconnu
- Code 41 : Ligne inexistante

- Code 412 : Impossible de retrouver la ligne la plus proche
- Code 42 : Arrêt et ligne incompatible
- Code 43 : Arrêt embarquement soit le terminus de la ligne

À la fin de cette étape, le prétraitement des données est terminé. Nous pourrons ensuite utiliser les données pour la classification des comportements des usagers.

3.1.3 Design des méthodes, implémentation et analyse des résultats

Toutes ces démarches seront fusionnées dans les cinq sections de contributions à la prochaine section (section 3.2).

3.2 Organisation générale du document

Dans la section 3.2, une discussion sur les quatre articles sera présentée. Pour chaque contribution, nous allons présenter l'objectif et la problématique, la méthodologie, le cas d'étude et le résultat. En outre, les autres contributions pendant le doctorat seront aussi illustrées.

3.2.1 Première contribution : comparaison des méthodes CCD et DTW pour la classification des séries temporelles provenant de données de cartes à puce

Cette contribution a pour but de comparer les métriques distance de corrélation croisée (CCD) et déformation temporelle dynamique (DTW). Un exemple pédagogique sera introduit pour démontrer leur application sur des séries temporelles en vue de bien connaître comment fonctionnent les deux méthodes. Une comparaison sera faite en appliquant différents paramètres de chaque méthode. Cette calibration nous aidera à mieux comprendre les avantages et les inconvénients de ces méthodes. Une portion des données de cartes à puce faisant partie de la STO a été testée. À la fin de l'implémentation, nous avons réalisé la classification des données et analyser les résultats. Le test nous permet également de choisir la meilleure métrique pour la classification temporelle en comparant les résultats des deux méthodes.

L'article résultant de cette contribution est publié dans la revue *Transportmetrica A : Transport Science* (He et al., 2018a). L'article est présenté au Chapitre 4.

Cet article a pour but de commencer la partie ‘classification temporelle’ dans l’objectif général de la thèse ‘classification spatio-temporelle’. La comparaison des méthodes de classification et des métriques nous permet de réaliser la partie ‘classification temporelle’ avec un morceaux des données.

3.2.2 Deuxième contribution : Méthode d’échantillonnage pour la classification temporelle de grandes quantités de données provenant de cartes à puce

Cette contribution a pour but de développer un algorithme pour l’échantillonnage et l’affection des séries de comportements horaires des usagers. Avec de grands ensembles de données, la matrice de distances par paires peut être extrêmement longue à calculer, en particulier lorsque le calcul de la distance entre deux éléments demande du temps (ce qui est le cas avec la corrélation croisée). Dans la situation actuelle, nous devons utiliser une méthode d'échantillonnage. Pendant l’implémentation, nous avons réalisé le calcul pour les données complètes avec une grille de calcul (CIRRELT-Calcul Québec) afin de déterminer les performances de la méthode d’échantillonnage (comparer le complet avec différents paramètres d’échantillons).

Pour cette contribution, nous avons amorcé un article qui a été publié dans les proceedings du congrès annuel du *Tranportation Research Board* en janvier 2017 (He et al., 2017).

Nous avons continué ces travaux, notamment avec l’ajout des analyses de l’effectivité de l’échantillonnage, et avons proposé un nouvel article, présentement en cours de révision pour la revue *Transportmetrica A : Transport Science*. L’article est présenté en Chapitre 5.

Cet article a pour but de finir la partie ‘classification temporelle’ dans l’objectif général de la thèse ‘classification spatio-temporelle’. Basé sur la méthode proposée dans la section 3.2.1, la méthode d’échantillonnage nous permet de finir la partie ‘classification temporelle’ avec les données massives.

3.2.3 Troisième contribution : Classification spatio-temporelle des données provenant des cartes à puce

Cette contribution a pour but de développer une méthode de classification spatio-temporelle. Pendant le processus de design de la méthode, nous avons bien connu l’effet des paramètres dans la méthode DTW, analysé et remplacé les métriques de base de l’algorithme de DTW, en vue

d'intégrer les caractéristiques des déplacements de transports en commun dans l'algorithme DTW. Ensuite, une méthode de classification des comportements en utilisant l'échantillonnage et la méthode DTW a été développée. Enfin, nous avons développé un algorithme de classification spatio-temporelle pour les profils quotidiens d'utilisation du réseau, et utilisé les données réelles pour tester ces algorithmes.

Pour cette contribution, nous avons amorcé un article qui a été présenté à *Conference on Advanced Systems in Public Transport* (CASPT) (He et al., 2018b).

Nous avons continué les travaux et avons proposé un nouvel article dans une revue avec comité de lecture. L'article résultant de cette contribution est présentement en cours de révision pour la revue *Public Transport*. L'article est présenté en Chapitre 6.

Cet article contribue pour la partie 'classification spatiale' et enfin 'classification spatio-temporelle' dans l'objectif général de la thèse 'classification spatio-temporelle'. Basé sur la méthode proposée dans la section 3.2.1 (les méthodes de classification et les métriques) et la section 3.2.2 (la méthode d'échantillonnage), la métrique de déformation temporelle dynamique est introduite pour la partie 'classification spatiale' et 'classification spatio-temporelle', avec les données massives.

3.2.4 Quatrième contribution : Reconnaissance et comparaison des comportements de différentes villes

Cette contribution a pour but de développer une méthode pour reconnaître et comparer des comportements dans des villes différentes. L'objectif de cette recherche est de tester si cet algorithme développé avec les données de Gatineau peut être transféré à une autre ville, et basé sur cela, comparer les comportements similaires et différents entre les deux villes en vue de transposer l'expérience de planification des transports en commun d'une ville à l'autre.

Cette recherche présente une méthode, basée sur des métriques de séries chronologiques, un algorithme de classification hiérarchique et une méthode d'échantillonnage, permettant de comparer le comportement des usagers de cartes à puce en transport en commun de Gatineau (ville nord-américaine) et de Santiago (ville sud-américaine). Le résultat montre que 66,24% des comportements quotidiens des utilisateurs peuvent être reconnus différemment dans les deux villes.

L'analyse des résultats montre que le comportement des utilisateurs de Gatineau est plus concentré le matin et rentre plus tôt à la maison que celui de Santiago.

Pour cette contribution, nous avons présenté un article à *Transportation Research Board* (He et al., 2019) et nous sommes en train d'essayer de le publier dans une revue.

Nous avons continué les travaux et avons proposé un nouvel article dans une revue avec comité de lecture. L'article résultant de cette contribution est présentement en cours de révision pour la revue *Journal of Transport Geography*. L'article est présenté au Chapitre 7.

Cet article est une extension de la partie 'classification temporelle' dans l'objectif général de la thèse 'classification spatio-temporelle'. Il vérifie si la méthode développée avec des données de carte à puce de Gatineau fonctionne pour une autre ville. Basée sur la méthode proposée dans la section 3.2.1 (les méthodes de classification et les métriques) et la section 3.2.2 (la méthode d'échantillonnage), une méthode de classification temporelle est développée en fusionnant les données des deux villes. Cela nous permet non seulement comparer les différences de comportements entre deux villes, mais aussi de développer une méthode en vue de reconnaître les comportements entre deux villes.

Il y a des applications potentielles de cette méthode proposée. D'abord, pour les clusters où les deux villes partagent les comportements similaires, les méthodes développées dans une des villes (tant académique que pratique) peuvent être transférées à une autre ville en vue de résoudre un problématique similaire de cette ville. Ensuite, pour les clusters où les comportements d'une ville dominent, une vérification de facteurs externes peuvent se faire, pour trouver la raison pourquoi ces comportements existent dans une ville, mais pas dans l'autre. Dans le futur, lorsque le facteur important d'une ville change, nous pourrons prévoir le changement des comportements des usagers de cette ville.

3.2.5 Autres contributions

En plus, nous proposons des contributions supplémentaires : des résultats de recherche que nous avons obtenus, mais nous rédigeons les articles pertinents dans le futur. Ils contiennent principalement des analyses profondes sur la méthode proposée et une méthodologie de classification basé sur la densité.

3.2.5.1 Analyse de résultats de classification spatio-temporelle

Avec la classification spatio-temporelle, une visualisation d'espace-temps permet de voir les comportements quotidiens de chaque usager de carte à puce de transport collectif, cependant, il faudrait faire une analyse profonde sur les comportements de différents groupes. La différence entre des groupes et la raison des différences seront illustrées.

Les proportions d'utilisation du métro sont différentes, notamment parce que son tarif est plus élevé que le bus à Santiago. Il est donc intéressant de voir qui (quel groupe) utilise plus le métro, et inférer pourquoi c'est le cas (voir la section 8.1.1).

La coupe transversale de trajectoire espace-temps permet de voir les localisations des usagers à un moment donné. Ensuite, la densité ces localisations à certaines heures données permet de définir les zones de domicile et de travail. Des cartes de chaleur démontrent la concentration de localisations de travail et la déconcentration des lieux de domicile à Santiago du Chili (voir la section 8.1.2).

Dans la section suivante, quelques exemples de comportements des usagers de cartes à puce sont présentés dans un graphique 3D (les chemins d'espace-temps) (voir la section 8.1.3).

Une visualisation est proposée, sur la moyenne des trajectoires espace-temps du groupe. Avec ces trajectoires espace-temps, il est facile de voir la différence de comportements entre les habitants des zones riches de Santiago et les autres (voir la section 8.1.4).

La visualisation de la déviation des trajectoire espace-temps est aussi réalisée. En comparant les trajectoires espace-temps réelles et la déviation de trajectoire du groupe, nous proposons une méthode de planification de transport en commun (voir la section 8.1.5).

Ces contributions sont les extensions de la partie 'classification temporelle' dans l'objectif général de la thèse 'classification spatio-temporelle'. Elles nous permettent de mieux comprendre les demandes de chaque groupe d'usagers en utilisant une analyse plus profonde.

3.2.5.2 Classification des zones basé sur la densité

Les méthodes indiquées jusqu'à maintenant sont toutes basées sur l'algorithme de partitionnement. Cependant, l'algorithme basé sur la densité est aussi utile, surtout l'analyse de zones. À la section 8.2, trois méthodes de classification des zones seront présentées.

La coupe transversale permet non seulement de définir les localisations de domicile et de travail, mais aussi de réaliser le zonage des arrondissements. Les arrondissements représentant le comportement des usagers de transport en commun de Santiago sont indentifiés(voir la section 8.2.1).

Cette classification a pour but de choisir l'heure de première transaction dominant dans une zone donnée. Avec la carte de chaleur, nous pouvons voir la grande diversité de temps de départ à différentes zones, ainsi que la demande d'optimisation de l'horaire du métro Santiago (voir la section 8.2.2).

Lorsque le service est modifié, tel que l'implémentation d'un système de bus rapide, il est intéressant de voir si les comportements des usagers changent. La comparaison d'une carte de chaleur avant et après l'implémentation permet de voir s'il existe un changement significatif. La section 8.2.3 présente un exemple appliqué à l'implémentation du Rapibus de la STO à Gatineau.

Ces contributions sont un supplément de l'objectif général de la thèse 'classification spatio-temporelle'. Contrairement aux autres sections qui sont sur la base de l'algorithme hiérarchique, la classification de ces contributions sont basées sur la densité. Elles permettent également de mieux comprendre les demandes de chaque groupe d'usagers.

CHAPITRE 4 ARTICLE 1: A CLASSIFICATION OF PUBLIC TRANSIT USERS WITH SMART CARD DATA BASED ON TIME SERIES DISTANCE METRICS AND A HIERARCHICAL CLUSTERING METHOD [1]

4.1 Abstract

Classification of smart card users' daily behavior is important in the field of public transit demand analysis. It allows an understanding of people's sequence of activities within a period of time. However, the classical metrics such as Euclidean distance is not appropriated when dealing with time series classification. To solve this problem, in this article, a method is presented for the classification of public transit smart card users' daily transaction represented in time series. The chosen approach uses cross correlation distance (CCD), hierarchical clustering, and subgroups by metric parameter, to understand the user' temporal patterns. The clustering results are compared with dynamic time warping distance (DTW, a common method to measure time series distance). After a small pedagogical example to explain the DTW and CCD concepts, a program is developed in R to validate the method on a real dataset of smart card data transactions. The dataset concerns the use of the public transit system in the city of Gatineau on September 2013. The results demonstrate that CCD performs better than DTW to classify the time series, and that the classification method permits to identify different public transit users' daily behaviors. The result will help transit authority to offer a better service for smart card users from diverse groups.

Keywords: Public transportation data, smart card users' behavior, time series classification, cross correlation, dynamic time warping

[1] He, L., Agard, B., & Trépanier, M. (2018). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. *Transportmetrica A: Transport Science*, 1-20.

4.2 Introduction

The extraction of customer behavior in public transit systems is of great interest in the scientific communities (Jou et al., 2007). Having a better understanding of travelers' behavioral patterns is helpful in assessing the demand for transportation services (Joh et al., 2006). We now take advantage of automated payment based smart card technology that generates and stores gigabits for data about day-to-day activities of users. The time series technique is widely used in customer behavior forecasting (Chen et al., 2009). A time series represents a collection of values obtained from sequential measurements over time (Esling and Agon, 2012). The segmentation of transit users permits to synthetize users' activity in a limited number of groups of typical behaviors. Knowledge from those groups may then be used to improve the service.

Data from automatic fare systems of public transit can be analyzed in many directions. For example, the data helps to evaluate and predict public transit users' demand (Kurauchi and Schmöcker, 2017). Some analyses have been done to cluster smart card users' temporal behaviors using data mining (Agard et al., 2006) and (Nishiuchi et al., 2013) designed a method to mine the pattern of card users' daily frequency. Many authors have suggested definitions (Das and Pandit, 2015), metrics, tools (Bordagaray et al., 2014), models (Li et al., 2015), algorithms (Chang et al., 2010) and methods (Del Castillo and Benitez, 2013) to help for a better understanding of the mobility of users during different time periods.

In data mining, most classification methods are based on distance metrics between observations. The traditional methods to measure the distance of samples includes Euclidean distance (Berkhin, 2006) or Manhattan distance (Bakar et al., 2006). Some other derivative method such as Minkowski distance (Jain et al., 1999), is a generalization of both the Euclidean and Manhattan distance (Lhermitte et al., 2011). However, these methods do not adhere to a conception of time process, which prevents the methods to analyze smart card users' behavior series. Some other distance measure methods can deal with near-time points, such as cross correlation distance (Liao, 2005) or dynamic time warping distance (Berndt and Clifford, 1994). All these distance-calculating methods have been implemented in R (Buchta, 2015). The selection of a pertinent metric is critical for the clustering methods. Various performant clustering methods are actually available, such as partition methods and hierarchical clustering methods (Langfelder et al., 2008).

Classification of transit users in subgroups is of great interest for transportation planners. Schedules and transit networks can be better fit to travelers' needs if we are able to identify them. This can also serve to define fare strategies and offer market-oriented services. This could be an advantage for both transit users and authorities. The objective of this paper is to propose a transit user classification approach to classify time series from the smart card user's profile, by combining time series metrics and data mining methods. Therefore, section 2 provides a state of the art on the application of data mining on public transit smart card data, and then classification methods are outlined as well as distance metrics between time series. In section 3, we proposed a method to use cross correlation and dynamic time warping distances combined with hierarchical clustering, to extract users' temporal behavior. Then, in section 4, the performance of the developed algorithms is compared. Finally, in section 5, a real database is tested to see how it performs.

4.3 State of the art

4.3.1 Use of Data Mining in Public Transit Smart Card Data

Data collected from an automatic collection system (in this case, smart card data) can be used to understand characteristics of public transit card users (Pelletier et al., 2011). Many researches have been done to explore the potential information from smart card data.

4.3.1.1 Data Completion and Preparation

Due to the characteristic of smart card data, some preparation must be made before the analysis of user's behavior. An algorithm has been developed based to estimate the alighting location given a smart card user's boarding location (Trépanier et al., 2007). This algorithm has been improved using kernel density estimation to estimate the unlinked trips (He and Trépanier, 2015). Furthermore, this algorithm has been calibrated to obtain a more accurate estimation of destinations (He et al., 2015). Besides, some researches focus on transfer detection (Chu and Chapleau, 2008) trip purpose inferences (Lee and Hickman, 2013), etc.

4.3.1.2 Classification for Transit Smart Card User's Behavior

An issue of great interest to transport operators involves partitioning network passengers into groups based on their transportation network activity. Clustering approaches are used such as the Hierarchical Ascendant Classification (HAC) or k-means algorithm (Agard et al., 2006). Based on

temporal analysis (Ghaemi et al., 2016) and spatial analysis (Ghaemi et al., 2015), the public transit card user's temporal patterns and spatial patterns are discovered and analyzed. Data mining even helps to predict user demand (Nuzzolo et al., 2016). Moreover, by using data mining, especially the classification technique, a methodology has been developed to analyze the quality of transit service level (de Oña et al., 2015). Langlois et al. (2016) perform analyses of multi-week activity patterns using clustering, and these authors propose a representation of longitudinal activity sequences using temporal and spatial activity (area of validation). The groups created by the clustering are associated with distinct sequence structures, thus allowing for better knowledge of several weeks of passenger activity (Briand et al., 2017).

4.3.1.3 Limitation of the Current Methods

The papers present a pertinent methodology of classification to explain user behavior using smart card data. However, the research is based on each individual smart card user's transactions instead of daily behavior time series. For example, when clustering using k-means, the algorithm considers only the value of vector elements, not the position of these elements in the vector. The interest of transportation planners is to consider the time of the day in the boarding sequence. In fact, the current classification methods are not suitable to solve this problem, because they are not designed to measure the dissimilarities between time series. The introduction of the time series classification technique helps develop a method to analyze a smart card user's daily profile, so that public transit authorities can offer better service that will satisfy passengers' daily requirements.

4.3.2 Classification

In this document, classification and clustering will be used as synonyms. A cluster is a collection of data objects so that one object is similar to one another within the same cluster and dissimilar with the objects in other clusters. A classification method is grouping a set of data objects into clusters. A good clustering method will produce high quality clusters with high intra-class similarity and high inter-class dissimilarity in level of mathematic (the real class may should be adjusted to each situation). The quality of a clustering result depends on both the similarity measure used by the method and its implementation (Subbiah, 2011). A class means that in which all the elements share similar characteristics, and between which elements share distinct characteristics. There are several major approaches to classification.

4.3.2.1 Partitioning Algorithms

Partitioning algorithms construct various partitions and then evaluate them by certain criteria. The major idea is to find a partition of *k* clusters that optimizes the chosen partitioning criterion given a *k* number of partitions. The two main heuristic methods are k-means and k-medoids (Subbiah, 2011). In the first one, each cluster is represented by the center of the cluster, while in the second one, each cluster is represented by one of the objects in the cluster. Partitioning algorithms have their advantages and limits when treating a time series. For example, the k-means can deal with large datasets, but it uses traditional distance metrics between vectors.

4.3.2.2 Hierarchical Algorithms

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis that seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types (Rokach et al., 2005): agglomerative and divisive. For the first one, each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. This is a "bottom-up" approach. For the second, all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. It is a "top-down" approach. Compared to partitioning algorithms, hierarchical clustering is available for a variety of distances but it cannot deal easily with large datasets, due to high computational costs.

4.3.2.3 Other algorithms

There are other methods that were used in the case of transit data:

- DBSCAN is used to classify regular passengers based on the dissimilarity of last alighting stop then first boarding stop of the day then boarding time (Kieu et al., 2007).
- Neural networks can also be used to classify regular passengers, and this is based on boarding stops, then boarding time (Ma et al., 2013).
- Naïve Bayesian network is used to estimate trip purpose based on Time and location of boarding and alighting stops (Kusakabe and Asakura, 2014).
- Continuous hidden Markov model subgroup all the users into eight groups based on their start time and duration of activity, land use around stops (Han and Sohn, 2013).

4.3.3 Distance Between Time Series

A time series is a set of observations x_t, each one being recorded at a specific time t. A discrete-time series (the type to which this article is primarily devoted) is one in which the set T_0 of times at which observations are made is a discrete set (Brockwell et al., 2002). Comparing it to other vectors, a time series contains a relationship among the time t itself. For example, for a time series $x_1, x_2, x_3, \ldots, x_n$, with the corresponding specific time $t_1, t_2, t_3, \ldots, t_n$, we know that t_1 is closer to t_2 than to t_n regarding time, regardless of the value of x_1, x_2 and x_n.

Various distance metrics exist to measure the (dis)similarity between two vectors (He et al., 2017). In this part, four types of distance are presented: Euclidean distance, Manhattan distance, Cross Correlation Distance (CCD) and Dynamic Time Warping distance (DTW). The first two distances are basic ones traditionally used in the classification methods presented earlier. Even if those metrics are not dedicated to be used to measure the distance between time series, it is still a common practice in transportation research (Agard et al., 2006; Morency et al., 2007; Ghaemi et al., 2015; Yuan et al., 2014; Kang et al., 2009). On the opposite, CCD and DTW are designed to compare the (dis)similarity of time series but are not actually incorporated in classification methods.

4.3.3.1 Euclidean and Manhattan Distances

Euclidean distance is the straight-line distance between two points in Euclidean space (Deza, 2009). Let x_i and v_j each be a P-dimensional vector. The Euclidean distance is computed as (Liao, 2005):

$$d_E = \sqrt{\sum_{k=1}^{P}(x_{ik} - v_{jk})^2} \qquad (1)$$

Manhattan distance is computed between the two numeric series using the following formula (Mori et al., 2016):

$$d_M = \sum_{k=1}^{P}\left|x_{ik} - v_{jk}\right| \qquad (2)$$

According to functions (1) and (2), for both distances, the result of distance would not be changed if the order of k is changed; for example, if the positions of k_1 and k_2 are exchanged, the distance remains the same. However, a time series contains relationship among the time t itself; this is a characteristic that makes time series different from other vectors. For a time series, if the values of k_1 and k_2 are exchanged, the distance result should change. Therefore, the Euclidean distance and

Manhattan distance are not suitable for time series. Besides, some efforts (Ghaemi et al., 2016) have been done to classify smart card users' daily transaction time.

4.3.3.2 Cross Correlation Distance (CCD)

This distance is based on the cross correlation between two time series (Mori et al., 2016). The similarity of two time series is measured by shifting one time series to find a maximum cross-correlation with another time series. The CCD between two time series at lag *k* is calculated as:

$$CC_k(X,Y) = \frac{\sum_{i=0}^{N-1-k}(x_i-\bar{x})(y_{i+k}-\bar{y})}{\sqrt{(x_i-\bar{x})^2}\sqrt{(y_{i+k}-\bar{y})^2}} \qquad (3)$$

Where $\bar{x}$ and $\bar{y}$ are the mean values of the series. Based on this, the distance measure is defined as:

$$CCD(X,Y) = \sqrt{\frac{(1-CC_0(X,Y))}{\sum_{k=1}^{max} CC_k(X,Y)}} \qquad (4)^1$$

In R, the distance measure can be calculated by using a function. This function will return the distance between two time series by specifying two numeric vectors (*x* and *y*) and maximum lag.

4.3.3.3 Dynamic Time Warping (DTW)

DTW is a popular technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches and warping, which optimally deform one of the two input series on the other (Giorgino, 2009). The method to calculate the DTW is as follows (Berndt et al., 1994):

$$S = s_1, s_2, \dots s_i, \dots, s_n \qquad (5)$$

$$T = t_1, t_2, \dots t_j, \dots, t_m \qquad (6)$$

The sequences *S* and *T* can be arranged to form a n-by-m plane or grid, where each grid point (*i*,*j*) corresponds to an alignment between elements s_i and t_j. A warping path, *W*, maps or aligns the elements of *S* and *T*, such that the "distance" between them is minimized.

$$W = w_1, w_2, \dots w_k, \dots, w_P \qquad (7)$$

---

[1] Should be $CCD(X,Y) = \sqrt{\frac{(1-CC_0(X,Y)^2)}{\sum_{k=1}^{max} CC_k(X,Y)^2}}$

That is, W is a sequence of grid points, where each w_k corresponds to a point $(i,j)_k$.

To formulate a dynamic programming problem, a distance measure between two elements is indispensable. Two possible distance measures are usually used for a distance function d. They are the magnitude of the difference (8) or the square of the difference (9),

$$d(i,j) = |s_i - t_j| \quad (8)$$

$$d(i,j) = (s_i - t_j)^2 \quad (9)$$

Once a distance measure is selected, the DTW problem can be defined as minimization over potential warping paths based on the cumulative distance for each path, where d is a distance measure between two time-series elements.

$$DTW(S,T) = \min w \left[\sum_{k=1}^{p} d(w_k)\right] \quad (10)$$

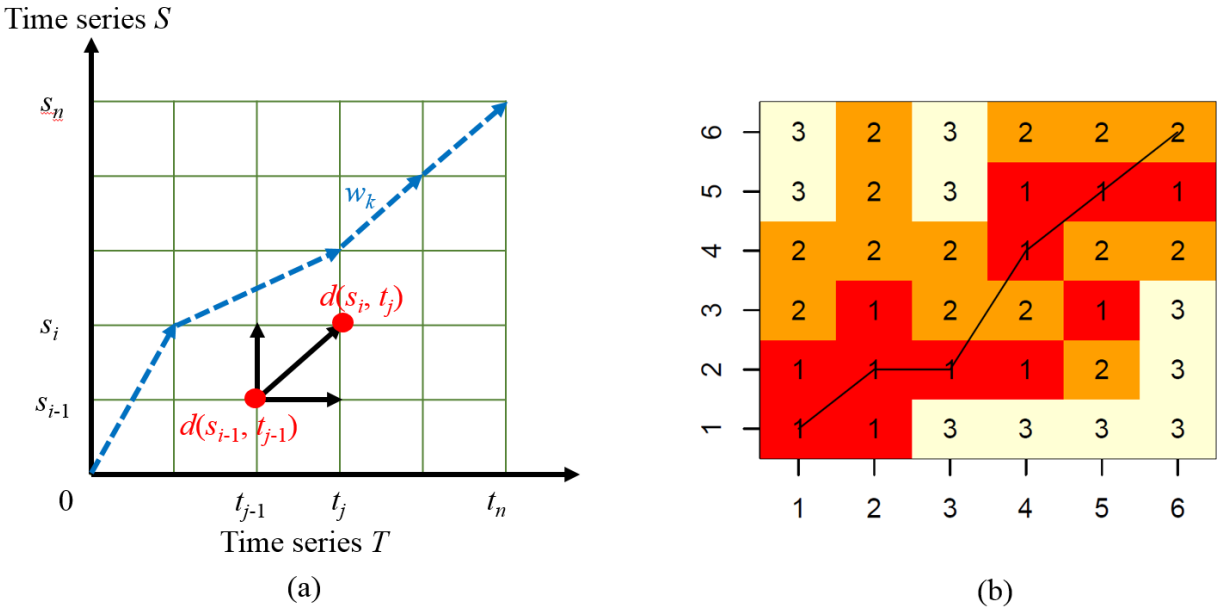


Figure 4-1: Dynamic time warping example (Giorgino, 2009)

Figure 4-1 illustrates DTW method. In Figure 4-1(a), to obtain a minimum cumulative distance, the time series can be warped to the next time point (moment). For example, grid point (s_{i-1}, t_{j-1}) can be warped to (s_i, t_{j-1}), (s_{i-1}, t_j), (s_i, t_j) to compute each distance. A sequence of grid points w_k can be a path from (s_0, t_0) to (s_m, t_n). On every grid point, the distance between two time points (moments) d (s_i, t_j) should be computed, as shown in the Figure 4-1(b). Then, all possible paths

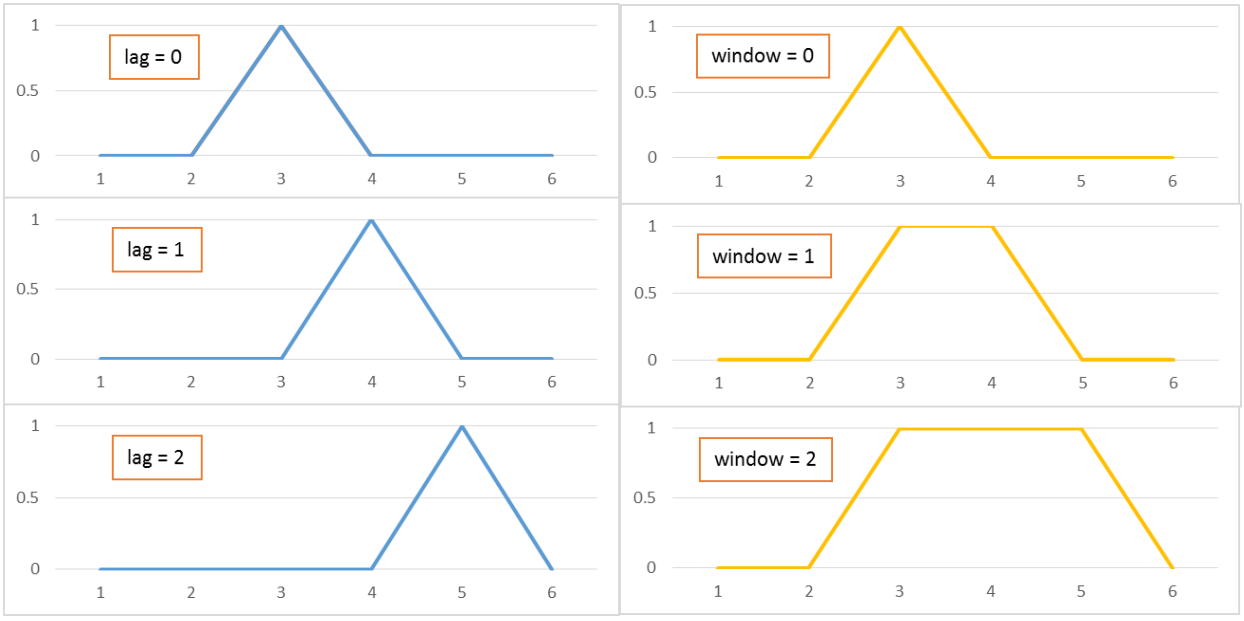
from grid point (1, 1) to (6, 6) are calculated, to find the path with minimum cumulative distance. In this grid in Figure 4-1(b), the distance of DTW is 7.

4.3.4 CCD and DTW Parameters

4.3.4.1 CCD Parameters

(1) Correlation coefficient. The Pearson correlation coefficient measures the strength of the linear association between two variables (Sedgwick and Philip, 2012). A correlation coefficient close to +1 or -1 represents a strong correlation.

(2) Max lag. This argument represents the maximum delay accepted to compare one time series to another. Figure 4-2(a) illustrates parameter "lag" for CCD. The first time series can be shifted to the right one unit, so that the first and second time series will better correspond. The lag can be explained as the number of units needed to shift, so that one time series will be aligned to another. In the Figure 4-2(a), if the max lag is 1, then the 1st time series can be shifted 1 unit to align 2nd time series. In this way, the 2nd time series can be accepted by the 1st time series, so that the 1st and 2nd time series will be in the same group, and the 3rd time series will be in another group.



(a) (b)

Figure 4-2: Time series metrics parameter - (a) lag for CCD (b) window for DTW

4.3.4.2 DTW Parameter

The parameter window represents the maximum time that a time series can be warped. Figure 4-2(b) illustrates parameter "window" for DTW. The values of elements in time series are warped, so that two time series can be compared. For example, if value of the 4th time point of the first time series is warped from 0 to 1, then the 1st and 2nd time series will be the same. The parameter window can be explained as the maximum delay allowed caused by warping. In the Figure 4-2(b), if the window is 1, then the 1st time series can be warped 1 unit so that the 1st time series and the 2nd will be the same. In this way, the 2nd time series can be accepted by the 1st time series, so that the 1st and 2nd time series will be in the same group, and the 3rd time series will be in another group.

To avoid that the first point of a time series compares to the last point of another time series, the parameters (max lag, window) have been set. This ensures to compute as more as possible points of two time series, and ensure to measure the dissimilarity of two users' behaviors as possible.

4.4 Proposed methodology for the classification of time series

4.4.1 Algorithm Design

The following three steps are proposed for the time series classification, as presented in Figure 4-3.

Step 1 The input data is time series on transit smart card activities' data. On the first part, pairwise distances are computed with CCD. The output is a distance matrix between any two time series. On the other part, DTW is used to perform this step.

Step 2 Hierarchical clustering is computed to cluster the time series using the distance matrices of CCD and DTW. The results of hierarchical clustering are presented in a dendrogram, from which the clusters are selected. The output is the clusters, including all the observation points. At the end of this step, we obtain the result of series classification by CCD on one side and by DTW on the other side. For CCD, a finer result will be obtained with step 3.

Step 3 Each cluster from the CCD side is split using the CCD parameters: correlation coefficient and maximum lag.

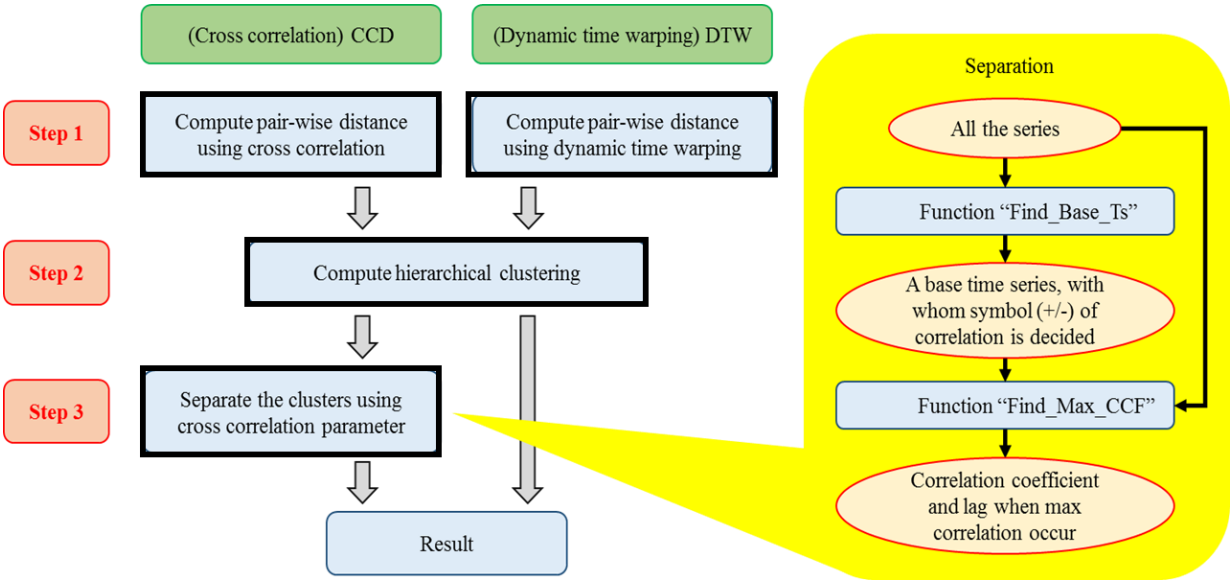


Figure 4-3: Proposed algorithm for the time series classification

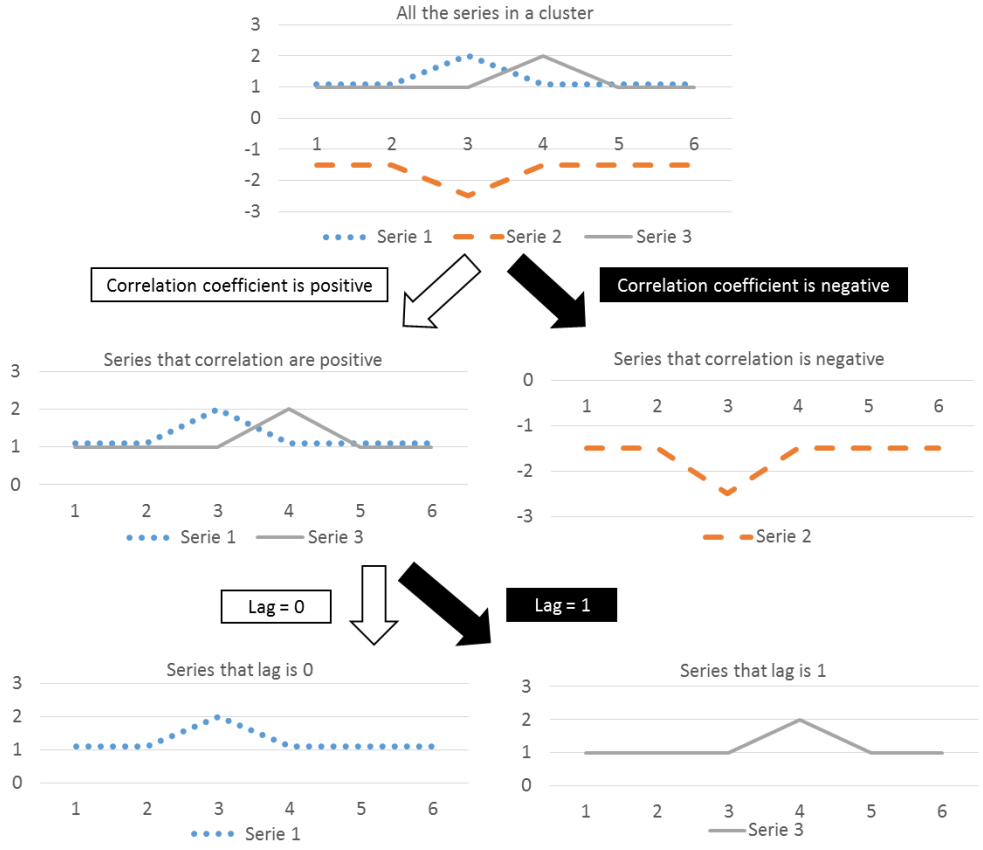


Figure 4-4: Separating by parameters in a cluster

In conclusion, both ways (DTW and CCD) were used to make clusters of time series from a mathematical point of view. Comparison of results and performances will permit to define the most efficient method from a practical point of view, in the classification of public transit smart card data temporal profiles.

The time series are first separated by the correlation coefficient. In this example, if series 1 and 3 are positive, then series 2 is negative (it's based on the correlation coefficient of a time series comparing to the others). Therefore, the pattern of series 2 is opposite to that of series 1 and 3. Then a separation is done by maximum lag (*k* in the function (4)) in the middle left figure of Figure 4-4. The bottom figures of Figure 4-4 show the results. If a lag of series whose peak occurs in the time point 3 is 0, then lag of series whose peak occurs in the time point 4 is 1. That means, the series 1 has been shifted by 1 time unit (lag) to align series 3, then the lag of the series 3 compared to the series 1 is 1. In this way, the relation between series 1, 2 and 3 in the cluster is obtained by CCD and the hierarchical method.

4.4.2 Implementation

The implementation contains 3 main steps as shown in the Figure 4-3:

Step 1 First, in some cases, a pre-treatment is needed to deal with the original database. For example, series whose values are all 0 or "NA" are removed. However, giving all of the values a scale is not necessary; if we do so, some values treated are not original. Then, the CCD is computed to calculate the dissimilarity of any two time series.

Step 2 Compute hierarchical clustering method. At the end of this step, the clusters in which the correlation coefficient and lag are not separated are obtained.

Step 3 Separate the clusters using the CCD parameter (correlation coefficient and lag). The most important part is in the right rounded corner rectangle:

Step 3.1 Firstly, a function "Find_Base_Ts" is applied to all the series in a certain cluster. It will return a base time series whose correlation is positive and lag is 0.

Step 3.2 Based on this time series, another function "Find_Max_CCF" is applied. This function will return correlation coefficients and lags relating to the base time series of all the other time series in a cluster. With the correlation coefficient and lag, a

minimum CCD between the base time series and a given series in the cluster can also be obtained.

Finally, three values are obtained for a time series: (1) Cluster, in which this time series has the best correlation with the other time series in the same cluster. (2) Correlation, the symbol of the base series. (3) Lag, the best one with which a minimum CCD can be obtained.

4.5 Comparison between CCD and DTW for classifying transit smart card data

4.5.1 A pedagogical example

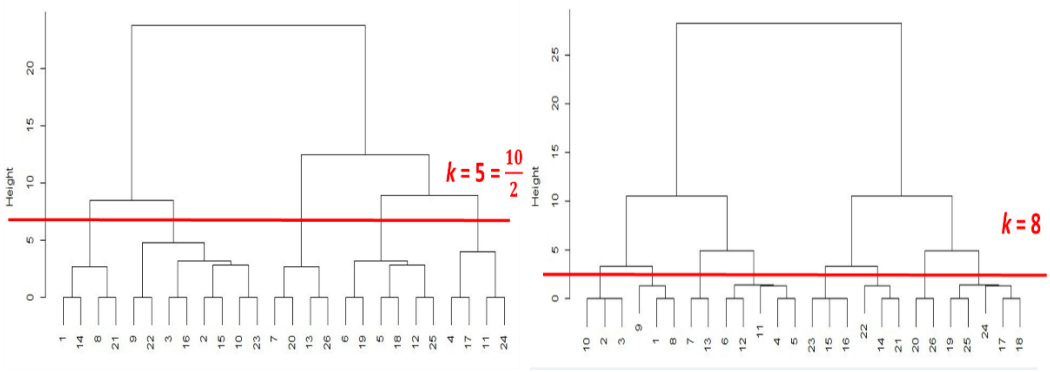
In this section, we apply the methods on public transit smart card data coming from an automated fare collection system. We need them to classify the transit users accordingly to their travel behavior. In the following section, an application to the real public transit smart card data in the city of Gatineau, in Canada, permits to show the efficiency of classification of smart card users' daily transaction time series. A sample, which contains 26 time series (user's daily behavior) of 7 time periods, is exemplified. By similarity with the real dataset, the values in these time series are 0 or 1, as shown in

Table 4-1. "1" at a time period (TP_i) means that a transaction has been registered during this time period while "0" means no transaction.

In this table, for example, the value of the first time series is T_1 = (1, 0, 0, 0, 0, 0, 0), means that a transaction of the smart card number 1 happened in the 1st time period and nothing in the remaining periods. In this table, smart cards series 14-26 have symmetric behavior to the series 1-13. "symmetric" here means that if a value of a certain time point is 0 in a series, then the value in the same time period in the other series is 1.

Figure 4-5(a) is the dendrogram of hierarchical clustering resulting from CCD, in which lag is 2. All the series are separated into 5 clusters as shown in Table 1 (column "lag =2"). Figure 4-5(b) is the dendrogram of hierarchical clustering resulting from DTW distance, in which the parameter window is 2. All the series are cut into 6 clusters, as shown in

Table 4-1(column "window =2").



(a) (b)

Figure 4-5: Hierarchical clustering dendrogram (a) with CCD (max lag = 2) (b) with DTW (window = 2)

In

Table 4-1, for CCD, the lag varies from 1 to 2, and for DTW, the window varies from 1 to 2. The test of calibration helps to understand the sensitivity of each parameter and metric. Each result consists of a first number, a plus or minus sign, and second number. The first number means the cluster affectation, the sign means whether the correlation coefficient is positive or negative, and the second number is the lag (of this time series compared to the lag 0 in the same group).

Table 4-1: A pedagogical example. Left half: Sample. 0 - 1 sample data (26 smart cards' data for 7 time periods TPi). Right half: Sample result. The No. of groups of CCD (calibrated by "lag") and DTW (calibrated by "window")

Sample								Sample result			
Smart Card	TP 1	TP 2	TP 3	TP 4	TP 5	TP 6	TP 7	Cross correlation		Dynamic time warping	
								max lag = 1	max lag = 2	window = 1	window = 2
1	1	0	0	0	0	0	0	1+0	1+0	1	1
2	0	1	0	0	0	0	0	2+0	2+0	2	2
3	0	0	1	0	0	0	0	2+1	2+1	2	2
4	0	0	0	1	0	0	0	3+0	3+0	3	2
5	0	0	0	0	1	0	0	5+0	5+0	3	3
6	0	0	0	0	0	1	0	5+1	5+1	3	3
7	0	0	0	0	0	0	1	4+1	4+1	4	4
8	1	1	0	0	0	0	0	1+1	1+1	1	1
9	1	0	1	0	0	0	0	2+1	1+2	1	1
10	0	1	1	0	0	0	0	2+0	2+0	2	2
11	1	0	0	0	1	0	0	3+1	3+1	3	1
12	0	0	0	0	1	1	0	5+0	5+0	3	3
13	0	0	0	0	0	1	1	4+0	4+0	4	4
14	0	1	1	1	1	1	1	1-0	1-0	5	5
15	1	0	1	1	1	1	1	2-0	2-0	6	6
16	1	1	0	1	1	1	1	2-1	2-1	6	6
17	1	1	1	0	1	1	1	3-0	3-0	7	6
18	1	1	1	1	0	1	1	5-0	5-0	7	7
19	1	1	1	1	1	0	1	5-1	5-1	7	7
20	1	1	1	1	1	1	0	4-1	4-1	8	8
21	0	0	1	1	1	1	1	1-1	1-1	5	5
22	0	1	0	1	1	1	1	2-1	1-2	5	5
23	1	0	0	1	1	1	1	2-0	2-0	6	6
24	0	1	1	1	0	1	1	3-1	3-1	7	5
25	1	1	1	1	0	0	1	5-0	5-0	7	7
26	1	1	1	1	1	0	0	4-0	4-0	8	8

For example, for 22nd smart card in the column “lag =2” of

Table 4-1, the result of CCD with lag of 2 is the cluster 1, with a negative correlation coefficient. Thus, it is presented as “1 (cluster) – (negative) 2 (lag)”. For the 22nd smart card user’s daily transaction series, four results can be obtained by using different metrics and different calibration.

(1) By using CCD, if the max lag is configured as “1”, the 22nd smart card user’s daily transaction series will be grouped in cluster 2. Compared to the other smart card user’s daily transaction series in the same group, this behavior has a negative correlation coefficient, and a lag of 1 time period.

(2) By using CCD, if the max lag is configured as “2”, the 22nd smart card user’s daily transaction series will be grouped in cluster 1. Compared to the other smart card user’s daily transaction series in the same group, this behavior has a negative correlation coefficient, and a lag of 2 time period.

(3) By using DTW, if the window is configured as “1”, the 22nd smart card user’s daily transaction series will be grouped in cluster 5.

(4) By using DTW, if the window is configured as “2”, the 22nd smart card user’s daily transaction series will also be grouped in cluster 5.

To explain the lag parameter in depth, the 2nd and 3rd user’s daily profile can be compared. They are in the same group and the lag of 2nd time series is 0, while the 3rd one is 1. It is to say the 2nd time period of 2nd time series can be shifted by 1 time unit, to align the 3rd time period of the 3rd time series. Then the 3rd time series can be accepted to be in the same group of the 2nd group.

From this pedagogical sample containing 26 smart card users’ behavior time series, four types of results have been obtained by using CCD, DTW and each parameter.

4.5.2 Comparison Between Cross Correlation and Time Warping

The classification result is impacted by selected metrics and parameters, it is of interest to discover the difference with the influence of each. Based on

Table 4-1, given a method (CCD or DTW), the size of a given cluster can be known, and the intersection size of two methods or parameters can also be known. For example, for the smart card data series in the cluster 1+ of CCD, there are two smart card data series that correspond in cluster 1 of DTW (with window =1). In

Table 4-1, these two smart card data series are series 1 and 8. Therefore, in Table 4-2(a), the horizontal axis is the size of each group of DTW. The parameter window is 1. The vertical axis is the size of each group by CCD. The parameter lag is also 1. Based on the same logic, Table 4-2(b)(c)(d) are built, for the comparison of other metrics and parameter values.

(1) Table 4-2(a) shows the comparison between CCD (max lag = 1) and DTW (window = 1). Almost the same result can be obtained by using these two methods. For group 3 and group 7 of DTW, these groups can be divided into two groups when using CCD. Groups 1 and 5 of DTW have minor differences with CCD. Moreover, the group sizes that are given by DTW contain large numbers; CCD divides these groups into smaller numbers, so that group sizes are more uniform.

(2) Table 4-2(b) shows the comparison between CCD (max lag = 1 and 2). The results are almost the same. Moreover, the augmentation of max lag can lead to a more comparable size. This means that even though augmentation of the max lag will more significantly shift the time series, the best correlation coefficient should be matched when the max lag is equal to 1. Therefore, the max lag of 2 not only maintains satisfactory results that do not need to significantly shift time series, but it also makes the group size similar. In conclusion, the result of a bigger lag has a minor change compared to the result of the smaller lag.

(3) Table 4-2(c) shows the comparison between DTW (window = 1 and 2).Unlike the comparison between CCD, almost all the groups have been changed if the parameter window has been changed. This means that DTW is more sensitive when changing parameters. Even though a substantial change in this case can make a 50% change in the group (4 groups out of 8 have been changed in Table 4-2(c)).

(4) Based on (1)(2)(3), the result by CCD (max lag = 2) and DTW (window = 1) are better. It is of interest to compare the results by these two conditions, as shown in Table 4-2(d). It shows that if the parameter is carefully chosen for each of the methods, the result through two methods will be nearly the same. However, CCD is easier to calibrate because it’s less sensitive to parameter values.

Table 4-2: Comparison between metrics and parameters

(a) Comparison of CCD (max lag = 1) and DTW (window =1)

(b) Comparison of CCD (max lag = 1 and max lag = 2)

(c) Comparison of DTW (window = 1 and window = 2)

(d) Comparison between CCD (max lag = 2) and DTW (window = 1)

* 1+ represents the group number 1 of CCD with positive correlation coefficient.

(a)

Size of group		Group No. by time warping (window =1)								
		1	2	3	4	5	6	7	8	Total
Group No. by cross correlation (max lag = 1)	1+	2								3
	2+	1	3							3
	3+			2						2
	4+				2					2
	1-					2				3
	2-					1	3			3
	3-							2		2
	4-								2	2
	5+			3						3
	5-							3		3
	Total	3	3	5	2	3	3	5	2	26

(b)

size of group		Group No. by cross correlation (max lag = 2)										
		1+	2+	3+	4+	5+	1-	2-	3-	4-	5-	Total
Group No. by cross correlation (max lag = 1)	1+	2										2
	2+		3									3
	3+			2								2
	4+				2							2
	5+				1	3						4
	1-						2					2
	2-							3				3
	3-								2			2
	4-									2		2
	5-									1	3	4
	Total	2	3	2	3	3	2	3	2	3	3	26

(c)

Size of group		Group No. by time warping (window = 1)								
		1	2	3	4	5	6	7	8	Total
Group No. by time warping (window = 2)	1	3		1						4
	2		3	1						4
	3			3						3
	4				2					2
	5					3		1		4
	6						3	1		4
	7							3		3
	8								2	2
	Total	3	3	5	2	3	3	5	2	26

(d)

Size of group		Group No. by time warping (window = 1)								
		1	2	3	4	5	6	7	8	Total
Group No. by cross correlation (max lag = 2)	1 +	3								3
	2 +		3							3
	3 +			2						2
	4 +				2					2
	1 -					3				3
	2 -						3			3
	3 -							2		2
	4 -								2	2
	5 +			3						3
	5 -							3		3
	Total	3	3	5	2	3	3	5	2	26

4.5.3 Comparison result

After comparing the application of the CCD and DTW in 0 – 1 sample data, the CCD is determined to be better in the classification of smart card data because of the following reasons:

(1) CCD is easier to calibrate. As presented in Table 4-2(b), when using CCD, the choice of parameters (lag) has a minor impact on the result, and almost the same result can be obtained when using 1 or 2 as the lag. However, as presented Table 4-2 (c), the choice of parameters (window) of DTW has a larger impact on the result than CCD.

(2) The result of CCD contains information on the correlation coefficient and lag. As presented from Table 4-2(a) to Table 4-2(d), besides the parameter "lag", for each result of CCD, there is another factor "correlation coefficient" (positive or negative). That means the number of groups can be adjusted depending on our need. For example, group 1+ and 1- can be combined into group 1 if fewer groups are needed. However, the DTW has only one choice in the group number.

(3) Group size is more similar and better defined when using CCD. Table 4-2(a) illustrates the group size of each method. For CCD, all the group sizes are 2 or 3, comparing to the size of 2 to 5 for DTW. The grouping of CCD is more even than DTW in the case of temporal transaction profiles.

4.6 Application to real public transit smart card data

In this section, real smart card data are used and classified with the method proposed in section 4.4. The results are presented and analyzed to show how this method performs.

4.6.1 Presentation of the Case Study

The dataset has been provided by the *Société de Transport de l'Outaouais* (STO), a transit authority serving 280,000 inhabitants in Gatineau, Quebec. The STO authority is a Canadian leader in user transit using smart cards fare collection (Morency et al., 2007). Table 4-3 shows an excerpt of the raw smart card dataset; it contains a variable of a user's trip information. Every line of Table 4-3 is obtained automatically once a transaction is done by a smart card user. Apart from the card identification (which has been made anonymous), there is the ticket type (fare categories such as junior, regular, senior, etc.), the date and the time of the transaction, the line (route) number and the direction. All transactions are made on a bus network; the location of the transaction is also

available (He et al., 2017). In this experiment, 100 000 transactions of 3095 card holders have been tested.

Table 4-3: Excerpts of the raw smart card dataset (He et al., 2017,)

Card id	Ticket type	Date	Time	Line	Direction	Weekday	Stop id
1150629967111800	140	2013-09-03	65232	44	Sud	2	1140
1273590714804090	110	2014-09-02	71909	224	Sud	1	2801
1273590714804090	110	2014-09-02	154607	224	Nord	1	2610

The objective is to make clusters of smart card data with similar daily behaviors. In each group the boarding time of day of a user should be similar to that of another user in the same group: this could mean not exactly at the same period, but "around" the same period.

4.6.2 Results

First, the data from Table 4-3 are transformed into a 0 – 1 table (see Table 4-4in which every line is a user's daily profile ("card id_date" combination), and every column is a time period, for example, the second column "05_30" means the period from 05:30 to 05:59. In the table, "1" represents that a transaction happened in this time period. For example, for the user whose card id is 1150312817303160, in 2013-09-03, he had a transaction in the time period 05:30-05:59.

Table 4-4: Example dataset of users-day (0-1 table)

Combination	05_30	06_00	06_10	06_20	06_30	06_40	…
1150296033731200_2013-09-04	0	0	0	1	0	0	…
1150312817303160_2013-09-03	1	0	0	0	0	0	…
1150320729466490_2013-09-03	0	0	0	0	0	0	…

With Table 4-4, the distance of every two combinations of lines is calculated by using the CCD and DTW. For the CCD, the parameter "lag" is 2. For the DTW, the parameter "window" is 2. Then, the CCD and DTW are computed and a distance matrix of any two combinations is obtained for each method. With that distance matrix, the combination ("user-date") is computed using hierarchical clustering.

By observing the dendrogram, we cut the dendrogram depending on the circumstances, to obtain the subgroups with a more even size. Then, 11 groups are cut for CCD, and 6 groups for DTW.

Finally, the sum of transactions for each cluster is calculated; this result is shown in Figure 4-6 and Figure 4-7.

Figure 4-7 shows the classification results by using DTW. By comparing Figure 4-6 and Figure 4-7, the DTW is not effective in our case. Firstly, the size of cluster 5 is so large that cluster 5 contains most of the transaction profiles, which can lead to an uneven size between all of the clusters. Secondly, comparing cluster 5 and cluster 6 in Figure 4-7, even though the size is different, the "peak hours" of these two clusters are almost the same. This means the users who have different behaviors cannot be separated by using DTW. Two criteria are used to judge which distance metric is better: Exclusivity and Homogeneity.

(1) Exclusivity:

This makes it possible to distinguish the different behaviors. It means that if one group occurs in one period, the other group will not occur in the same time period. In Figure 4-6, for CCD, the first smart card user transaction in Group 6 is between 05:30 and 06:20. In this period, the sum of transactions of group 6 always exceeds 4000. On the contrary, the sum of any other group is less than 600. The situation of the other groups is similar. In Figure 4-7, for the DTW, the group 5 and the group 6 appear in the period between 05:00 and 07:30, whose duration is 2.5 hours. This means that the DTW cannot well separate the smart card users' behaviors. Therefore, with respect to the exclusivity criterion, it is preferable to classify user behaviors using CCD. Exclusivity is the priority. If two metrics produce the same result, homogeneity is considered later.

(2) Homogeneity:

When classifying, we try to get as homogeneous groups as possible. It means that the group size (the amount of user profiles of the smart card in the group) should be roughly uniform. In Figure 4-6, for CCD, the maximum sum of all groups is 2500 to 8000, and there is no group whose size represents more than 50% of the number of profiles of smart card users. However, in Figure 4-7, for DTW, group 5 represents more than 50% of all profiles of smart card users, which means that it is an unequal classification. It is not necessarily that users are distributed evenly among all groups. However, CCD gives more uniform size groups compared to DTW, it's more interesting in this case, even though it's not absolutely necessary all the time.

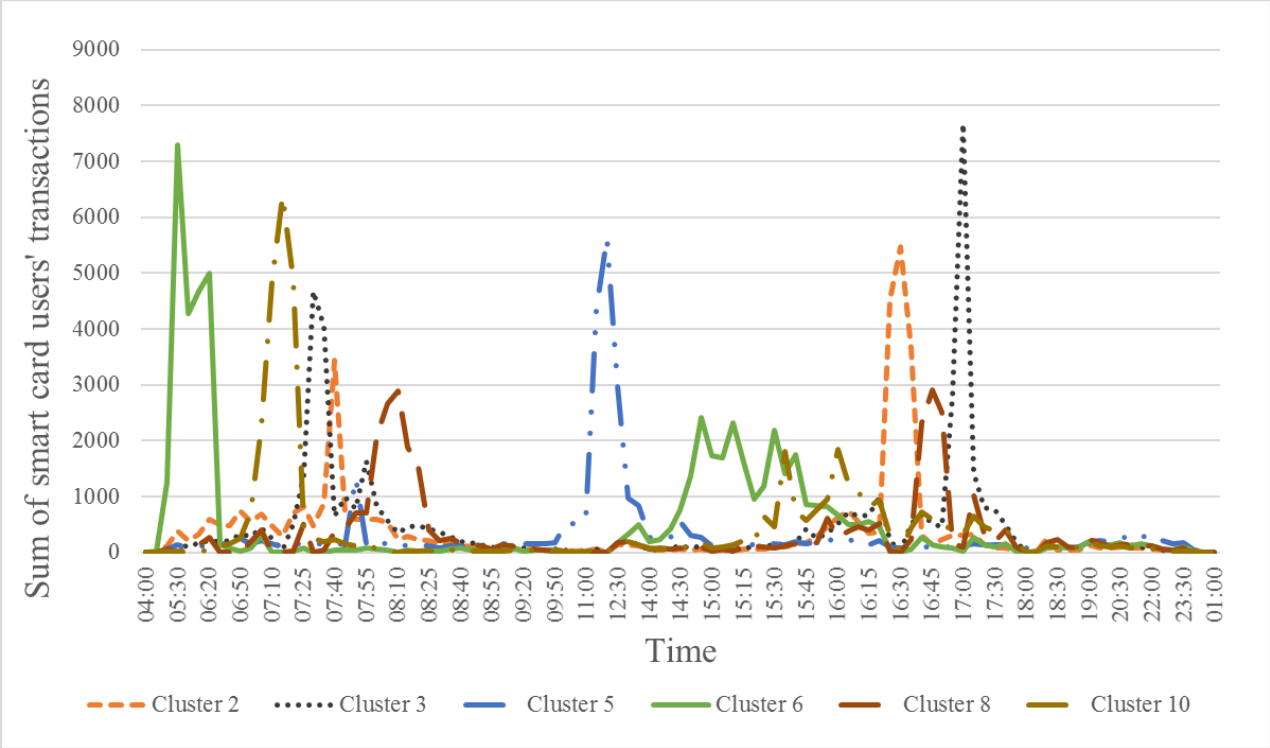


Figure 4-6: Sum of transaction time of each group (CCD)

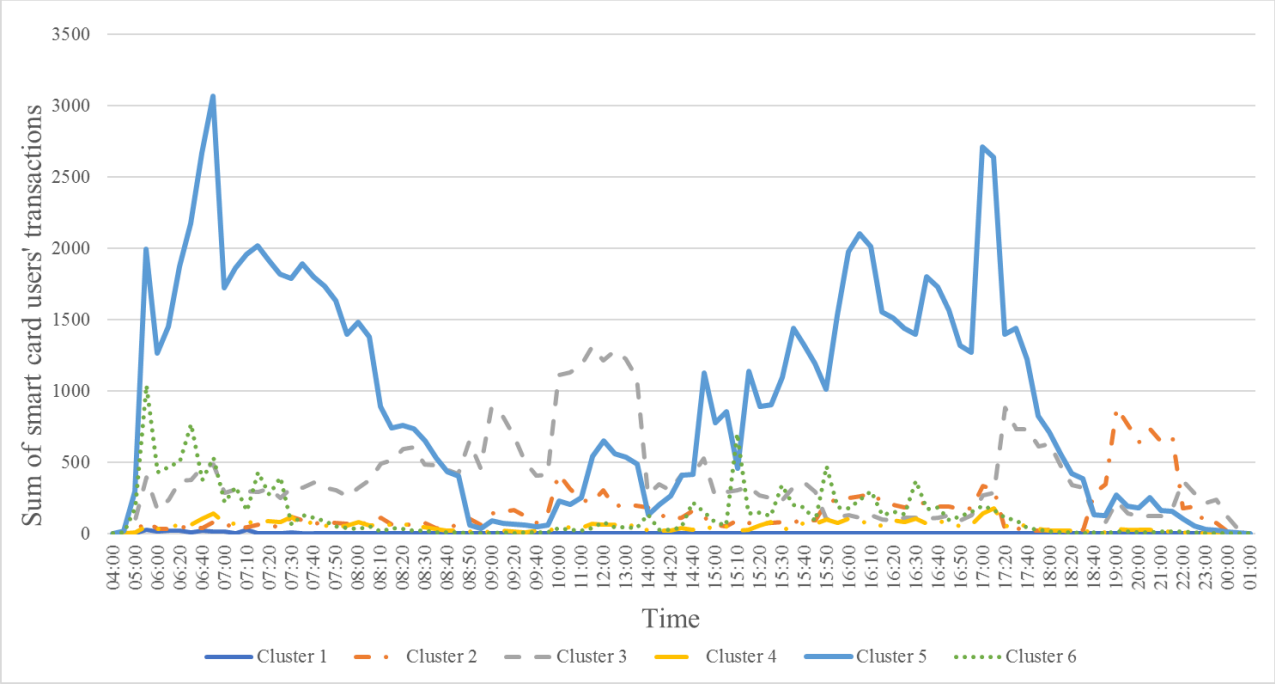


Figure 4-7: Sum of transaction time of each group (DTW)

Finally, it is of great interest to discuss how the proposed method and classification results can be used for improving transit features. In general, the result will help transit authority to offer a better service for smart card users from diverse groups. Firstly, based on the behavior of different group, the transit authority is able to optimize schedules to satisfy the demand of groups respectively but save vehicle turns. For example, based on the Figure 4-6, for a period 04:00 - 07:25, more vehicle turns may be scheduled during 04:00 - 06:20 and 06:50 - 07:25, to respond the demand of group 6 and 10 respectively. Drivers could take a break during 06:20 – 06:50. Furthermore, if this optimization result in a less total vehicle turns, it will help to save energy and reduce exhaust of greenhouse gases. Secondly, even though activity episodes are not encoded, we know the boarding time of every group, that means, if a user's daily profile contains only two transactions, we can infer users' home and work or study place. In fact, in Figure 4-6, the groups 2, 3, 8 contain and only contain two peaks, which means that most of the users in these groups contain only two transactions in that day. Thirdly, taking advantage of off-peak periods, vehicles can be allocated to relieve burdens of other bus line. For example, a vehicle who serve the group 8 may finish the service at 09:00, then it can be allocated to serve group 5 at 11:00. In conclusion, the main idea is to look for different characters among groups, then serve them differently.

4.7 Conclusion

An analysis of public transit smart card users' daily profiles needs a method that permits to classify time series. Because of the limitations of traditional distance metrics, a method has been designed with combining time series metric and hierarchical clustering. The results show that cross correlation is better adapted for the classification of public transit smart card data temporal profile. The test using the data from a middle-sized public transit association shows a clear separation of card users' daily transaction profiles. This may permit a better information about each subgroup of users and then a better pacification of the transit system.

With regards to the limitations and perspectives. The first is the calculation time when trying different parameters (lag for cross correlation and windows for DTW). To solve this problem, a new algorithm could be developed in order to avoid certain calculations in which the calculation of the CCD between certain vectors is canceled out by assuming that the distance of these two vectors is too large. The second limit is the choice of metrics when dealing with transportation issues. In this case, the CCD is suitable because the delay of a smart card user's transaction time is

like the parameter “lag” in the CCD. However, when dealing with other time series in transportation problems, other distances that may be applied For example, the Fourier transformation distance based on the analysis of fluctuations could be tested to explain the fluctuations in transactions, etc. Overall, the objective is still to find the best metric for a specific transportation issue.

4.8 ACKNOWLEDGMENTS

The authors wish to acknowledge the support of the *Société de transport de l’Outaouais (STO*) for providing data, the Thales group and the National Science and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

CHAPITRE 5 ARTICLE 2: SAMPLING METHOD APPLIED TO THE CLUSTERING OF TEMPORAL PATTERNS OF PUBLIC TRANSIT SMART CARD USERS [1]

5.1 Abstract

The study of temporal patterns has been applied to represent the various behaviours of transport users. Smart card data is useful for characterizing travel behaviours. The behaviours identified can be analysed and thereby transport services can be improved. For large datasets of transactions, the traditional method is to segment the data into several groups and use one unique pattern to represent each group. However, classifying very large datasets is still challenging. Here we propose a method to classify the temporal patterns of all the users of a public transit system. We recommend a clustering method that combines a sampling method and cross-correlation distances. This method was applied to classify the temporal patterns of public transport users from Gatineau, Canada. An indicator was developed to validate the efficiency of the proposed method. Compared to current methods, the proposed method is faster and better able to deal with very large datasets.

Keywords: Public transit big data, Smart card user behaviour, Time series classification, Sampling method

5.2 Introduction

Extracting temporal behaviours from large time series datasets has proven to be valuable for many applications in the domain of transport. In fact, large sets of transport data are analysed in order to determine the temporal behaviours that are used in numerous application domains. Analysing temporal behaviours allows public transit patterns to be understood (Liu & Cheng, 2018), user

[1] Soumis à *Transportmetrica A: Transport Science* le 6 avril 2019. Auteurs : Li He, Martin Trépanier, Bruno Agard.

behaviour to be predicted (Yang et al., 2018) and crowding levels in public transit vehicles to be measured (Yap et al., 2018) so that transit authorities can intervene and improve levels of service.

Using data mining techniques on transport data helps us to better understand user behaviour (Mohamed et al., 2017). Data mining techniques make it possible to measure the travel time of different user groups and different fare types (Ma et al., 2013). It helps estimate the origin-destination demand by integrating pattern matching methods (Chen et al., 2015), and travel trajectory can be analysed in order to understand the behaviour of different traveler groups (Zheng, 2015). Data mining techniques help deduce/predict the purpose of trips (Kusakabe and Asakura, 2014) and they can also uncover traffic bottlenecks in the urban network through spatiotemporal analysis (Lee et al., 2011). Moreover, data mining can also help analyse the impacts of weather on public transport ridership (Zhou et al., 2017).

Data mining can be used with diverse sources of data. For example, when used with bike sharing data, the spatiotemporal behaviour of bike rentals can be understood (Bordagaray et al., 2016). A similar method can also be implemented in carshare systems (Morency et al., 2007). Data mining can also be applied to GPS data to understand freight trip chaining behaviour (Ma et al., 2016). Finally, the data from smart cards is relevant to the various uses of data mining and it have long been proven to be effective and useful for analysing the mobility behaviour of transit riders (Pelletier et al., 2011).

The temporal pattern analysis of public transit users is challenging when dealing with large datasets. When using large datasets, many authors (Agard et al., 2006; Ghaemi et al., 2015; Ghaemi et al., 2017) propose classifying data into groups (user behaviours, for example) in order to deal with a limited, but still representative, set of generic behaviours that can be analysed. The recent work from Aghabozorgi et al., 2015 investigates novel clustering solutions to deal with fairly larger set of distance metrics for time series mining. Similarly, some work has been done on mining diverse large-scale sources of individual mobility data (GPS, mobile phone data, Bluetooth data) (Naboulsi et al., 2017, Ketabi et al., 2019). This practice makes it possible to simplify the data in a useful way and reveal the hidden patterns in a huge volume of individual data. The level of classification (number of groups) gives analysts the control to either get more or less detail on certain information, depending on the accuracy needed. However, the classification of large sets of temporal data (time series) is still a challenge and needs further development.

In this paper, we propose a new method for classifying temporal patterns, using cross-correlation distances and sampled hierarchical clustering. The method is used to analyse a large dataset of smart card transactions from a public transport service company and to classify the behavioural patterns of all the smart card users in the public transit system.

The paper is organized as follows. The following section presents the literature review. Section 3 describes the proposed classification method by first explaining the classical method, which combines cross-correlation distance and hierarchical clustering. Then, the framework for method we developed is described by combining the classical method, a sampling method and an assignment process. In section 4, the proposed framework is applied to a real case study. Furthermore, an indicator is designed to check which sample sizes are large enough. Finally, section 5 concludes the paper and introduces future research perspectives.

5.3 Literature review

This section describes works relevant to the proposed method. Section 5.3.1 focuses on traditional classification methods. Section 5.3.2 points out the notion of distance. Two distance metrics, in the context of the analysis of smart card data, are highlighted. Section 5.3.3 outlines classification methods and distances in smart card data research.

5.3.1 Traditional Classification Methods

Data classification aims to solve the following problem: Given a set of training points with the associated training labels, determine the class label for a test instance that is also not labeled (Aggarwal, 2014). In other words, these methods aim to group a set of observations into clusters. There are many clustering approaches. The principle of intra-class distance (dissimilarity) is to maximize the similarities between objects in the same class. Inter-class distances aim to minimize the similarity between objects of different classes (Lakshmi & Raghunandhan, 2011). Some classification methods are presented as follows:

5.3.1.1 Partitioning Algorithms

Partitioning methods try to find the best partitions (k) from a given number of objects (n) (Ng & Han, 1994). The following introduces two popular partitioning algorithms:

K-means: K-means clustering (Lee & Hickman, 2014; MacQueen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting the initial cluster centers of the k groups and then iteratively refines them as follows: (1) Each instance (di) is assigned to its closest cluster center. (2) Each cluster center (Cj) is then updated as the mean of the instances (di) that were assigned to it (Wagstaff et al., 2001). This is one of the classification methods that best addresses the well-known problems in data clustering.

K-medoids: Each cluster is represented by one of the objects in that cluster (Park & Jun, 2009). The K-means algorithm is sensitive to outliers, such as inputs with extremely large values that may substantially distort the distribution of data. To solve this issue the K-medoids method uses a medoid instead of using the mean value of objects in a cluster as the reference point. A medoid is the most centrally located object in a cluster (Velmurugan & Santhanam, 2010).

5.3.1.2 Hierarchical Algorithms

Hierarchical algorithms create a hierarchical decomposition of the set of data (or objects) using some hierarchical criteria (Karypis et al., 1999).

Hierarchical clustering organizes objects into a dendrogram, which branches out into the desired clusters. In a dendrogram, pairs of vertices that are more closely related have common ancestors that are located lower in the tree. Those of more distantly related pairs are located higher in the tree (Clauset et al., 2008). The process of cluster detection is referred to as tree cutting, branch cutting, or branch pruning (Langfelder et al., 2007).

Hierarchical algorithms fall into two types. One is based on repeatedly merging two smaller clusters into a larger one (agglomerative, or bottom-up). The other is based on splitting a larger cluster into smaller ones (divisive, or top-down) (Ding & He, 2002). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters (Rokach et al., 2005).

There are two basic advantages for hierarchical algorithms. First, the number of clusters does not need to be specified a priori. Second, they are independent of the initial conditions (not necessary to initial cluster centers as partitioning algorithms) (Frigui and Krishnapuram, 1999). However, the

disadvantage is that data-points may fail to separate overlapping clusters due to a lack of information about the global/overall shape or size of the clusters (Jain et al., 1999).

Furthermore, there are some derivative methods such as CURE (clustering using representatives) and BIRCH (balanced iterative reducing and clustering using hierarchies), among others. CURE employs a hierarchical clustering algorithm that adopts a middle ground between the centroid and all point extremes, in order to avoid problems relating to clusters that are not uniformly shaped or sized (Guha et al., 1998). For BIRCH, the advantage is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points (Zhang et al., 1996). In conclusion, hierarchical algorithms are a method of cluster analysis which seeks to build a hierarchy of clusters.

5.3.1.3 Density-Based

Density-based methods are based on connectivity and density functions (DBSCAN (density-based spatial clustering of applications with noise), OPTICS (ordering points to identify the clustering structure)) (Ester et al., 1996). In density-based clustering, clusters are defined as areas of higher density than the remainder of the dataset (Kriegel et al., 2011).

5.3.1.4 Other Methods

Other classification methods exist such as grid-based and model-based methods.

The grid-based method is based on a multiple-level granularity structure (Liao et al., 2004). Examples include STING (statistical information grid-based method) and CLIQUE (clustering in quest). For STING, the idea is to capture statistical information associated with spatial cells in a way in which whole classes of queries and clustering problems can be answered without recourse to the individual objects (Wang et al., 1997). CLIQUE identifies the dense units in the subspaces of high-dimensional data space, and uses these subspaces to provide more efficient clustering (Agrawal et al., 1998).

In the model-based classification method, a model is hypothesized for each of the clusters and the idea is to find the best way these models fit together (Yeung et al., 2001). The various classification methods help to choose the best one to solve a specific issue.

Each classification method has its advantages and disadvantages. K-means is a popular method in data mining classification that is applied to various fields (Cui et al., 2017). However, k-means

usually uses Euclidean distance calculations. This method rarely uses other types of distances, and it needs a predefined number of clusters (k). Hierarchical algorithms can be used without a predefined number of clusters, and it can be used with most of the distance types. The disadvantage of hierarchical algorithms is that it is not suitable for large datasets because of its extended/long computational time.

Concerning classification of large database, dimensionality reduction techniques for machine learning tasks on large datasets have been developed (Ding et al., 2015). Some work has been done to increase hierarchical clustering computation efficiency. For example, Gilpin et al., 2013 propose an approach based on identifying objects with similar distances and reducing the number of distances computed during hierarchical clustering via hashing and hierarchical organization of the data. Some other efficient clustering techniques have been also developed. For instance, Fast DBSCAN via distance computation pruning (Chen et al., 2018) appears to be a very efficient solution for high-dimensional distance computation on very large dataset.

In conclusion, there are four main factors to consider when choosing a classification method: whether the number of clusters is determined automatically, whether it can be applied to the entire dataset, the computational time and how the results of the algorithm are impacted by the choice of distance metric. Beside, the usefulness of the results for the planning of public transport would be another factor after obtaining the classification result.

5.3.2 Distance Calculation Methods

Various distance metrics exist to measure the (dis)similarity between two instances (vectors related to observations). In this section, two types of distance are compared: Euclidean distance and cross correlation distance. Euclidean distance is largely used in various application domains. Cross-correlation distance is better adapted for time series but is much more time consuming, which limits its use for real and larger datasets.

5.3.2.1 Euclidean Distance

The Euclidean distance is the straight-line distance between two points in Euclidean space (Deza, 2009). Let x_i and y_j each be a P-dimensional vector. The Euclidean distance is computed as (Liao, 2009):

$$d_E = \sqrt{\sum_{k=1}^{P}(x_{ik} - y_{jk})^2} \qquad (1)$$

The Euclidean distance is widely used in many application domains. In the case of the transit system in Gatineau (Canada), the very large dataset collects about 600,000 entries each month. Data mining techniques have been used to analyse this data with valuable results (Agard et al., 2006).

The Euclidian distance, however, is not well adapted to time series analyses. According to the function (1), the result of the distance would not change if the order of k is changed. For example, when the values of $k = 1$ (x_{i1}) and $k = 2$ (x_{i2}) are exchanged, the distance will be the same. However, a time series represents the relationship between time (t) observations themselves, which is a characteristic that distinguishes it from other vectors. For a time series that represents the moment a transport system is used during a day, if the values of $k = 1$ and $k = 2$ are exchanged, the results of the distances should also change (He et al., 2018).

5.3.2.2 Cross-Correlation Distance

Cross-correlation distance is based on the correlation between two time series. The similarity between two time series is measured by shifting one time series to find a maximum cross-correlation with another time series. The cross-correlation between two time series at lag k is calculated as (Mori et al., 2016):

$$CC_k(X,Y) = \frac{\sum_{i=0}^{N-1-k}(x_i - \bar{x})(y_{i+k} - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_{i+k} - \bar{y})^2}} \qquad (2)$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the series. Based on this, the distance measure is defined as:

$$CCD(X,Y) = \sqrt{\frac{(1 - CC_0(X,Y)^2)}{\sum_{k=1}^{max} CC_k(X,Y)^2}} \qquad (3)$$

In the R software package, the distance measure can be calculated by using a function. This function will return the distance between two time series by specifying two numeric vectors (x and y) and maximum lag.

5.3.3 Classification Methods and Distances in Smart Card Data Research

Through the years, several authors have proposed the use of data mining techniques to analyse smart card transaction data (Diab & El-Geneidy, 2013), such as by analysing the characteristics of smart card users (Sun et al., 2016), and the behaviour changes of travelers (Asakura et al., 2012). A previous study showed the variability of the travel behaviour of riders in a bus network using the k-means technique (Morency et al., 2007). In most cases, k-means is used when a mean value for each cluster needs to be computed (Vicente & Reis, 2016), making it difficult to consider smart card data as a time series. More recent works used DBSCAN (Density-based spatial clustering of applications with noise) to assess the mobility behaviour of smart card users (Kieu et al., 2015; Ma et al., 2013). Other research used dynamic time warping (DTW) to unify the time references of smart card transactions (Li & Chen, 2016). However, DTW is not well adapted to large sets of data, due to its computational complexity. Another study used a mix of techniques to analyse data from the city of Rennes, France (El Mahrsi et al., 2014). However, there are still challenges associated with characterizing behaviours based on temporal distribution, because the methods used to calculate the distances between observations are not well suited to the kinds of analyses that are required by transit authorities. To address these challenges, efforts have been made to incorporate spatial data in order to measure the spatial-temporal data dissimilarity (Ghaemi et al., 2015). Furthermore, a classification-related technique has also been used to analyse the quality of transit service (de Oña & de Oña, 2015), to gauge passenger type (Legara & Monterola, 2018) and to analyse transit network performance (Parzani et al., 2017).

In this study, we preferred using cross-correlation distance to Euclidean distance, for the following reasons. First, the result of the distance would not change if the order of time points is changed when using Euclidean distance. Second, transaction times can be represented by a point with the value “1” at the time of the transaction, while the values are all “0” for other times. An earlier or later shift of a transaction time is represented by the shift of value “1”. This shift corresponds to the parameter “lag” seen in cross-correlation distance. Therefore, cross-correlation distance is preferred, as we have shown in previous work (He et al. 2018). Our study also prefers the use of hierarchical clustering to k-means because hierarchical clustering us to compute the classification algorithm by using a distance matrix of time series metrics, and this also permits to avoid using Euclidean distance between points and cluster centers (Jain, A. K. et al., 2010).

5.4 Proposed methodology

This section introduces the method for clustering the temporal behaviour of public transport users with the help of smart card transaction data. The method consists of three elements: the distance metric (cross-correlation distance), the hierarchical clustering algorithm and the sampling method. Figure 5-1 presents the structure of the proposed method. Please note that steps 6, 7 and 8 are added here to validate the results of the method and therefore are not necessary when the method is implemented.

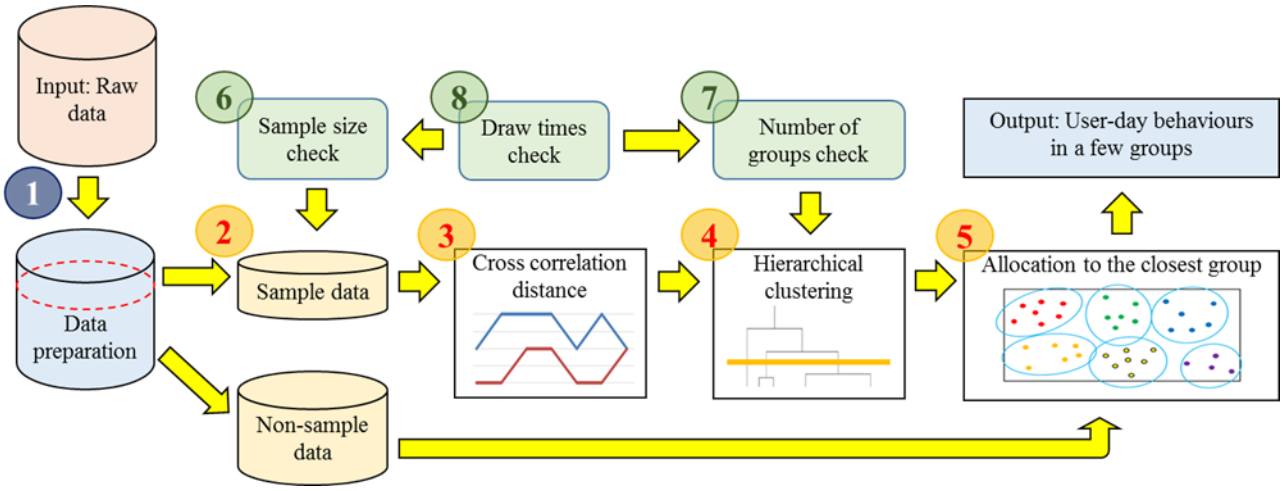


Figure 5-1: Overall proposed method

The unit of analysis is the temporal daily departure time profile of public transport users. For each user, and for each day, we created a vector of binary values stating whether or not the system was used. These “user-day” vectors were used as raw data. The following explains each of the steps in Figure 5-1.

1. The smart card transaction data is subjected to a series of pre-treatments in order to create the daily departure time profiles of users.
2. Some user profiles are selected for the sample dataset, while the other user profiles will be in the non-sample dataset.
3. Cross-correlation distance (CCD) is applied to the sample data to measure the dissimilarity between any two daily profiles.

4. A hierarchical algorithm is applied to regroup all the sample data profiles according to their similarities. At the end of this step, a dendrogram is used to arrange/separate the user profiles from the sample dataset into a number of groups.
5. Non-sample data observations are assigned to their nearest group. The data from each user-day is used to compute the distance to the closest cluster. The dissimilarity is also measured by cross-correlation distance. Finally, a group consisting of the daily profiles of smart card users is obtained as the output in Figure 5-1. The main classification processes is from step 2 to step 5.
6. For this paper, we experimented with several sample sizes to conduct a sensitivity analysis.
7. Because a dendrogram is used, the number of groups is also a factor that we may control. We also changed the numbers of groups to test their effects on the results.
8. The contents of the samples were randomly selected. We also tested multiple numbers of draws in order to check the reliability of the results obtained by the sampling method. Note that the sequence of the test is from step 6, to step 7, finally to step 8.

Figure 5-2 presents an overview of the sampling method mentioned in steps 4 and 5.

a. Start with all observations (seen here as points).
b. Sample random points (red points).
c. Apply cross-correlation distance and hierarchical clustering algorithms to these sample points. Clusters are made with this sample.
d. Calculate the distance between each remaining point and all the other points of the sample group. Then, use these distances to allocate the remaining points to the nearest group. In the end, all the points are grouped.

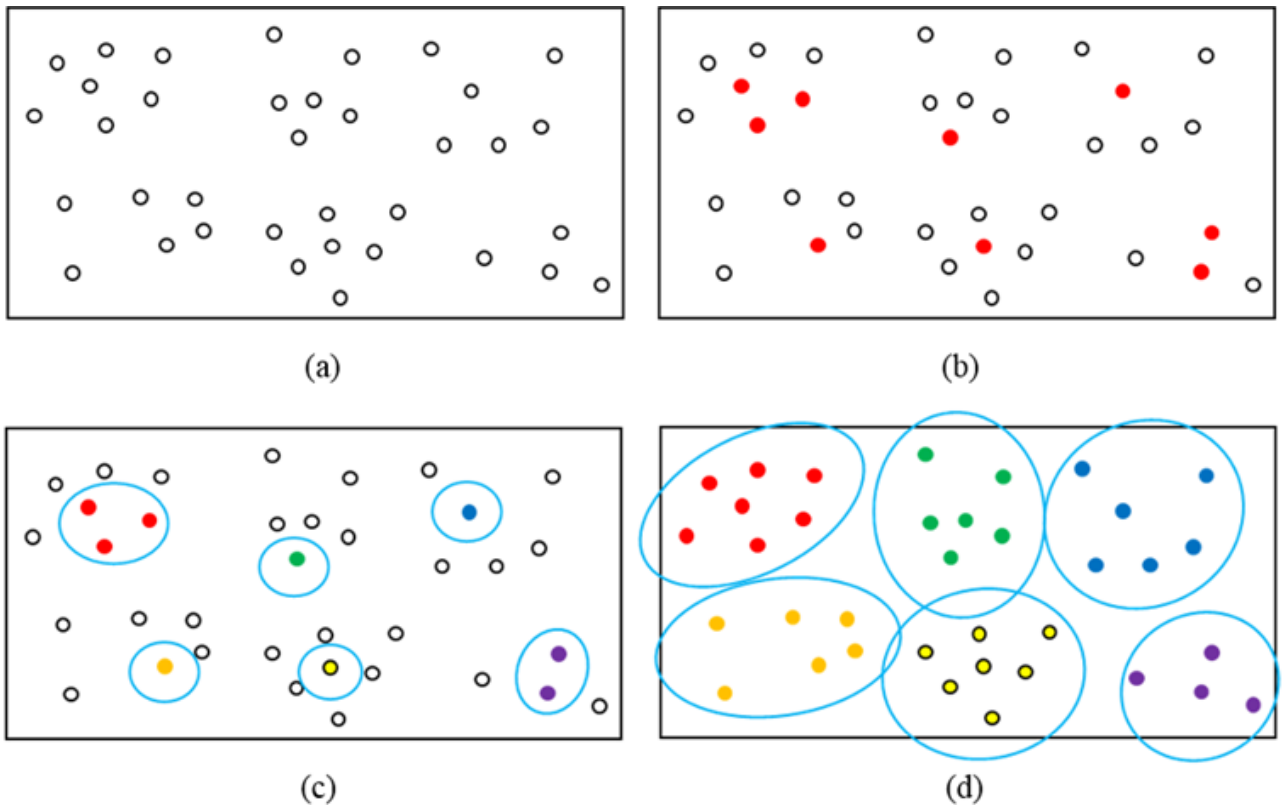


Figure 5-2: Sampling and allocation process

5.5 Case study

For the case study, the dataset was provided by the Société de Transport de l'Outatouais (STO), a transit authority serving the 280,000 inhabitants/residents of Gatineau, Quebec. The STO operates a bus-only service with a smart card automated fare collection system (Agard et al., 2006). This system has been in use since 2001, and a large proportion (over 80%) of STO users have a smart card. The following sections describe how the method was implemented.

5.5.1 Data preparation

This is the first step in Figure 5-1. The raw smart card transaction dataset is available in a flat file and contains the following information: card identification (anonymized), fare category, date and time of the transaction, location of the transaction (bus stop), route number and direction. Only the data from weekdays were selected in order to better characterize the travel behaviour of regular workers and students. The dataset contains 1,707,192 transactions that were registered in September and November 2013 from 26,320 cards.

5.5.1.1 Timeframes setting

In order to create temporal profiles based on transaction times, we organized the days that were studied into periods of time. For example, 09: 21 and 09:28 belong to the 09:20 - 09:29 time period, and 09:32 belongs to the 09:30 – 09:39 time period. The periods are of different lengths depending the time of day. From 04:00 to 01:30 (of the next day), there are a total of 35 time periods in each day. Table 5-1 shows the timeframes (time periods) for one day.

Table 5-1: Timeframes for the daily distribution of transactions

Time	Type of period	Interval
00:00 - 05:59	Off-peak hours	30
06:00 - 06:59	Regular hours	10
07:00 - 08:59	Peak hours	5
09:00 - 09:59	Regular hours	10
10:00 - 13:59	Off-peak hours	30
14:00 - 14:59	Regular hours	10
15:00 - 15:59	Peak hours	5
16:00 - 17:59	Regular hours	10
18:00 - 23:59	Off-peak hours	30

5.5.1.2 Vectors of Observations

In this step, we created a table that represents the temporal profile of each card (the transaction times for each card). The STO uses photo-ID cards, so each card represents one user. Each profile is thus called a "user-day". Unique IDs were created by concatenating the card number with the date. For example, in Table 5-2, the numerical identification displayed on the first line (1150296033731200_2013-09-04) represents the card user's profile followed by the date (04 September 2013). The number 1 that is shown in this row under column 06_20 indicates that a transaction occurred between 06:20 and 06:29 that morning. A total of 337,745 user-day profiles were created. When represented in its entirety, the table contains 337,745 × 35 elements (0 or 1), which is too large to be segmented by hierarchical clustering in a suitable amount of time on regular computers. For this reason, we suggest using a sampling method.

Table 5-2: Timeframes for the daily distribution of transactions

Combination	05_30	06_00	06_10	06_20	06_30	06_40	...
1150296033731200_2013-09-04	0	0	0	1	0	0	…
1150312817303160_2013-09-03	1	0	0	0	0	0	…
1150320729466490_2013-09-03	0	0	0	0	0	0	…

5.5.2 Application of the Method

The following three sections present the proposed methods applied to STO data.

5.5.2.1 Sampling

This is step 2 in Figure 5-1. In the sampling process where user-days are randomly selected, each user was only present once in the sample. For example, if "user A date A" was chosen for the sample set, then "user A date B" was not added. We tested a range of sample sizes. More details about sample size will be discussed in Sections 5.5.3 and Section 5.6.

We avoid more pre-selections of the data. The algorithm we develop ensures that this user-day is representative. (1) If the preselected user-day represents the regular behavior of this user, this would be the ideal case. (2) If no and if this user-day is part of the other regular behavior group, the classification result would be reasonable. (3) If not, this user-day will be abnormal in the dendrogram, and we will remove this user-day from the sample and replace with another, so that it does not impact the result of the classification.

5.5.2.2 Sample Clustering

These are steps 3 and 4 in Figure 5-1. In this process, cross-correlation distances and hierarchical clustering algorithms were applied to the sample data. We then analysed the dendrogram obtained by the hierarchical clustering algorithm, and chose the number of groups to create the sample groups. A total of 25 groups was first tested. The branches cut by red line in Figure 5-3 present the results. There were still many groups that did not have enough points. Therefore, the number of groups was gradually reduced and the dendrogram was updated. In the end, 11 groups were selected (more detail about selecting the number of groups will be discussed in Part 5.5.3 and Part 5.6). Note that in the general method, fewer groups can be chosen. However, because this is a sampling

method, we chose to conserve a higher number of clusters for the sample and to link the remaining observations to these clusters.

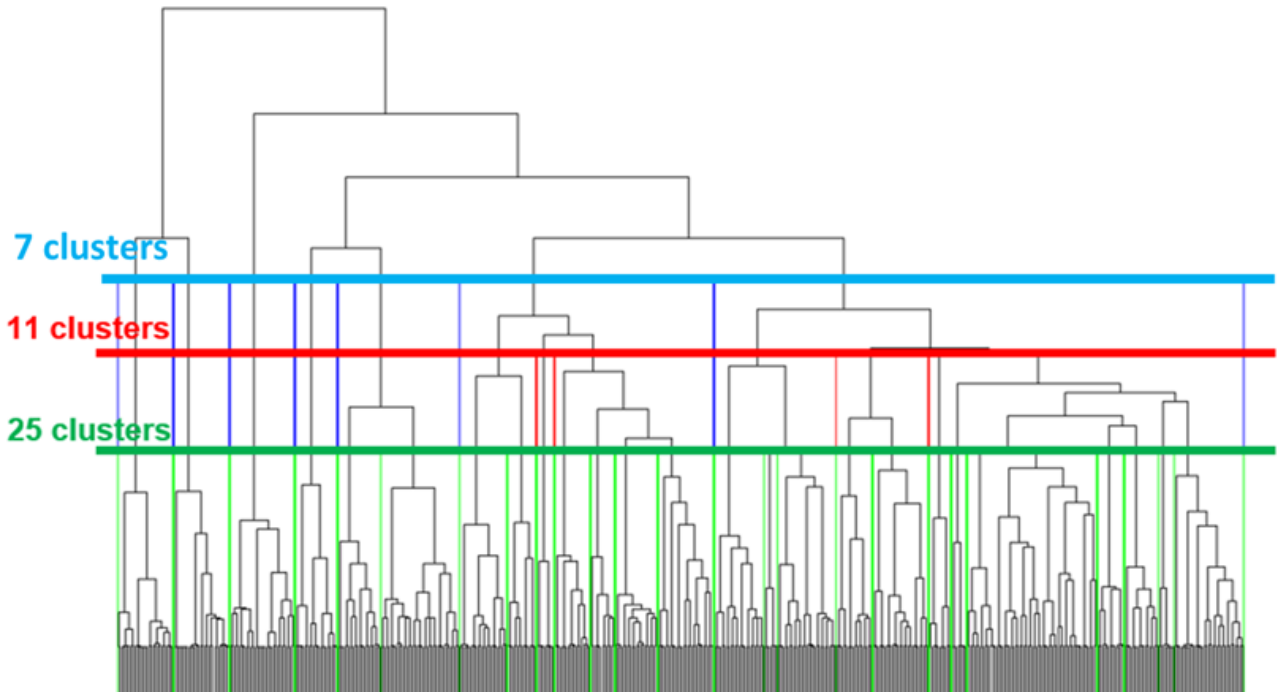


Figure 5-3: Dendrogram of sample data

5.5.2.3 Allocating the Remaining Observations

This is step 5 in Figure 5-1. In this process, the rest of the user-days were allocated to their nearest cluster. The allocation method is presented in section 5.4, Figure 5-2. The "nearest cluster" means the cluster with the shortest average distance from a remaining point to all points in that cluster. After this step, all the daily transaction time profiles of the smart card users were added to clusters.

5.5.3 Evaluating Sampling Performance

In this section, we proposed a methodology to measure the effectiveness of different sample sizes and different numbers of groups. This approach proposes using the variance of inter-group distances (dissimilarity) and the variance of intra-group distances. The variance is based on a series of tests made using the same parameters. For example, if we choose 10 groups and a sample size of 1 000, we will execute the method a number of times (we later call it the number of draws, because each execution of the method will lead to a different draw in the sample). Ideally, the variance of inter-group distances (dissimilarity) and intra-group distances would decrease as the

sample size increases. The objective is to determine a sample size with a small enough variance for inter-group and intra-group distances, so that the sample size does not influence the classification method results.

The following are the steps proposed to evaluate the sampling performance. Let us define S as the sample size, N as the number of groups and D as the number of draws:

1. Take a sample of S user-day and classify them into N groups.
2. Calculate the inter-group and intra-group distances with the method presented above.
3. Repeat steps 1 and 2, D number of times.
4. Based on the tests using D, calculate the variance of inter-group and intra-group distances average of D times. Next, calculate the combined inter-group and intra-group distance variance.

In the following section, we present the results determined using this approach with different values of S, N and D, to test the sampling method's performance in our case study.

5.6 Results

In this section, we first present the results of the tests and then adjust the parameters to obtain the desired clustering of the temporal profiles.

5.6.1 Variance analysis by inter-group distance, intra-group distance, and their combined variances

Figure 5-4 (a) illustrates the combined inter-group and intra-group distance variances for sample sizes from 200 to 2 000. The results show that for the groups where N=5, 11 and 20, the value of the indicator stabilizes around S=700, 900 and 1 500, respectively. For example, for N = 5, the variation of the indicator between S = 600 and 700 is about 10^{-5}. However, for S = 700 to 2 000, the difference between the maximum and the minimum value of the indicator does not exceed 5×10^{-4}. In this case, for N = 5, a sample size of 700 is sufficient to apply the clustering method. These tests were conducted using 20 random draws (D=20).

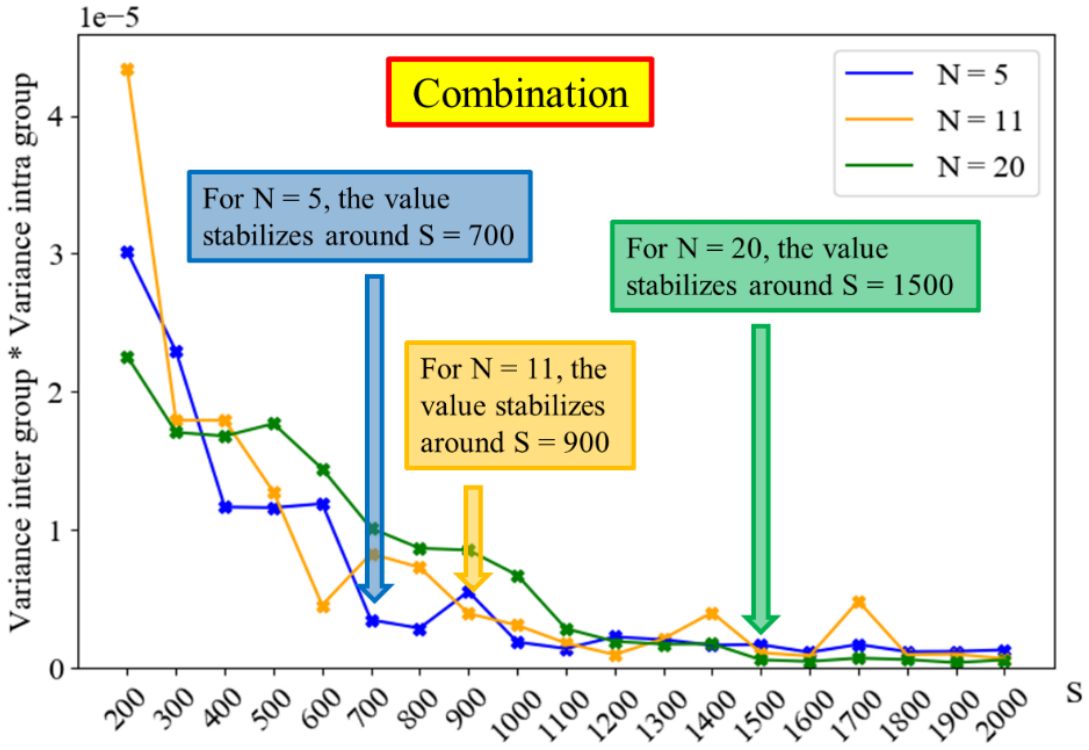


Figure 5-4 (a)

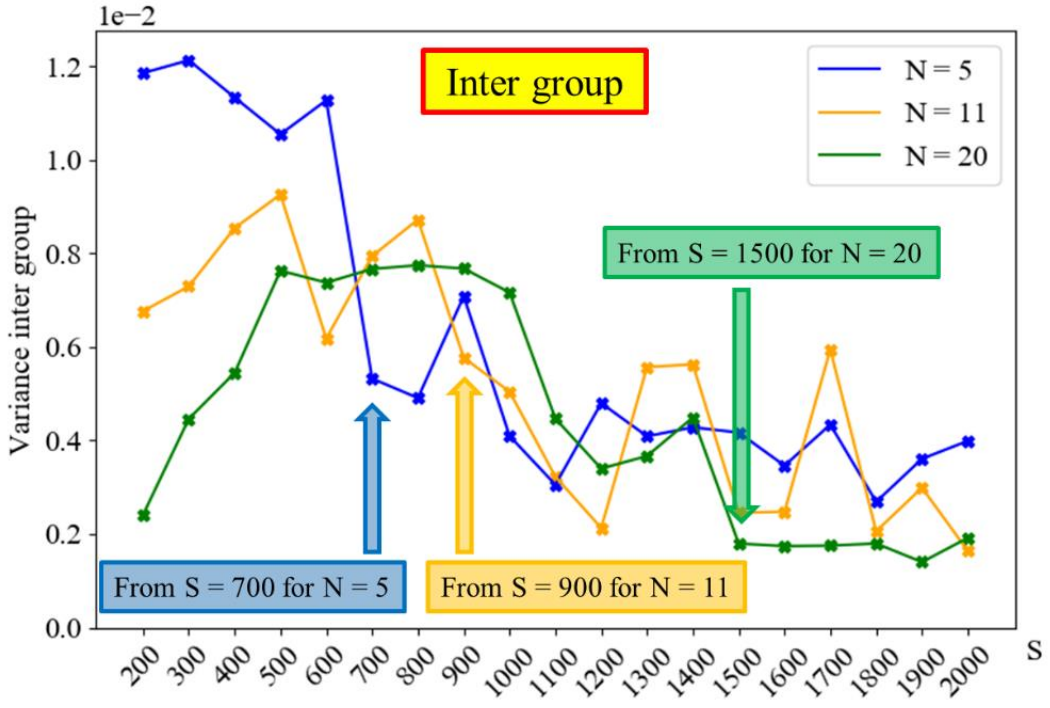


Figure 5-4 (b)

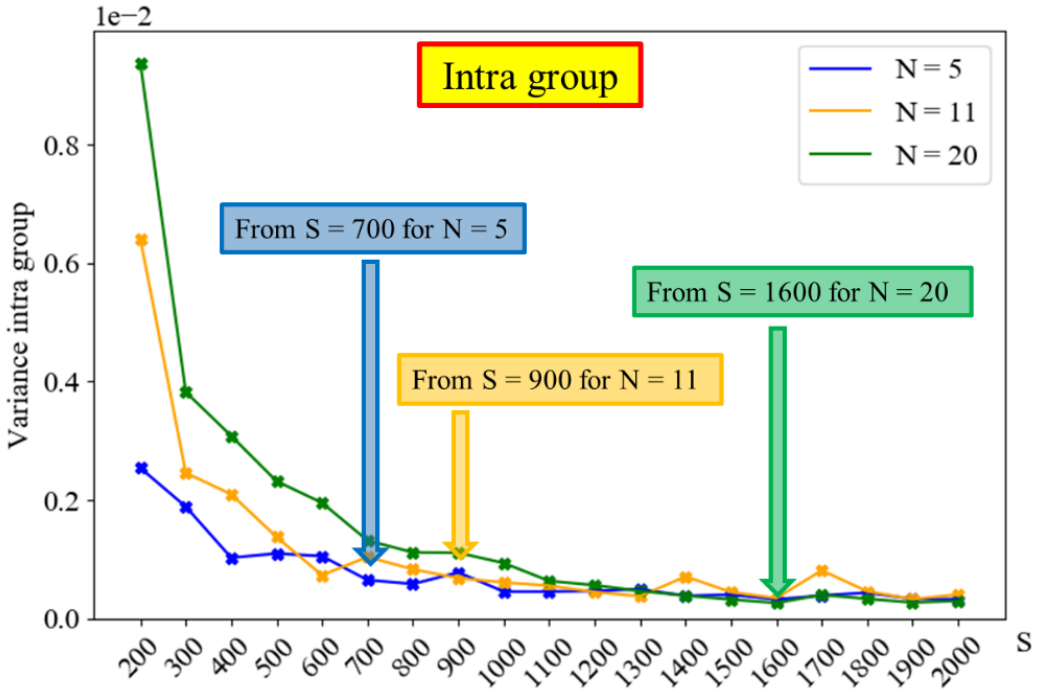


Figure 5-4 (c)

Figure 5-4: Variance of (a) inter-group distance and intra-group distance combined (b) inter-group distance (c) intra-group distance (D=20)

We have to question whether the two combined variances can represent the tendency of both individual variances. Figure 5-4 (b) and (c) illustrate the variances of inter-group distances and intra-group distances for N = 5, 11 and 20, respectively.

Unlike the almost monotonically decreasing function of the intra-group variances, the inter-group variances tend to demonstrate more complexity. First, it increases with the sample size. Then, the indicator stays high for a period determined by the number of groups (for N = 20, the indicator values stay high for longer than for N = 5). Finally, it decreases rapidly and reaches a stable range.

A possible explanation for this may be that in the beginning, the classification algorithm did not work well because the sample size was too small. An extreme case is to classify 6 elements into 6 groups. In this case, each group contains only one element. This would result in the inter-group distance being the same and therefore the variance of the inter-group distance would be 0. Then, as the sample size increases, the variance of the inter-group distances increases because the classification algorithm is not able to efficiently cluster the samples. In the end, the sufficient

sample sizes are determined and other groups of similar sizes are created, thus rapidly decreasing the variance of inter-group distances.

In terms of choosing the sample size, we can see that the individual variances for the inter-group distances and intra-group distances show little change after S = 700, in the case of N=5. This matches the results observed for the variance of intra- and inter-distances combined (Figure 5-4 (a)). Therefore, 700 was chosen for sample size for all three indicators (inter-group, intra-group and the combination of both). In the same way, S = 700 was chosen as the minimum sample size for N = 11. For N = 20, results show that the groups obtained become stable from around a sample size of 1 500 or 1 600.

5.6.2 Sensitivity analysis of the number of draws (D)

Because we are proposing a sampling process, we tested the sensitivity of the number of random draws to see if this method performs well in all cases. In addition to D=20 as presented in Figure 5-4 (a), D=10 and D=50 are tested for 11 groups (N=11), as presented in Figure 5-5.

The comparison demonstrates that the results obtained from the different values for D are almost the same. For D = 10 and 50, the value starts to stabilize at around S=1 000. However, it becomes stable at S=900 when D = 20.

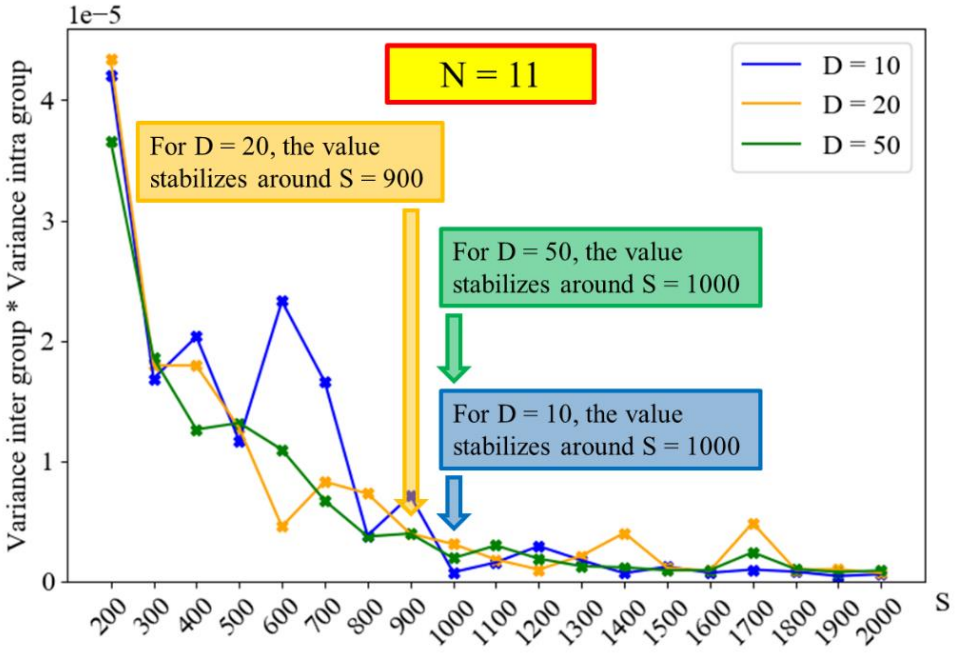


Figure 5-5: Variance analysis by the number of draws

It is surprising to see that a sample size of 2 000 is large enough to create clusters from 335,745 entries. We believe that this is due to the type of data and application used in this study. In fact, not like other cases, the type of public transit users' behaviors may be limited, people don't have too much choice, then it is able to represent all of profiles behavior even with less than 1% of total profiles. Ultimately, temporal patterns remain relatively stable over time, especially for the same card. The calculation of N = 5 and N = 20 supports the result of N =11.

5.6.3 Resulting Temporal Profiles

We applied the proposed method to the dataset using 11 groups, a sample size of 2 000 and 20 draws (N=11, S=2000, D=20). The daily temporal patterns of 6 of the 11 groups are presented in Figure 5-6. It shows that Cluster 2 and Cluster 3 display pendular behaviours related to commuters, however not at the same time. Clusters 6, 7 and 11 are characterized by single surges during peak hours. Cluster 9 regroups the people that are using the system in between the peak periods of the day. This type of analysis will help the transit authority identify the different types of customers in order to establish differentiated fares or service levels. Using this method also decreases the computational time. For example, for the dissimilarity matrix process, the computation time of 1 000 user-day profiles is 1% of 10 000 user-day profiles. In the case study, instead of having to use a 333,745 X 335,745 distance matrix ($1{,}1 \times 10^{11}$ entries), it used a 2 000 X 2 000 matrix (4×10^{6} entries).

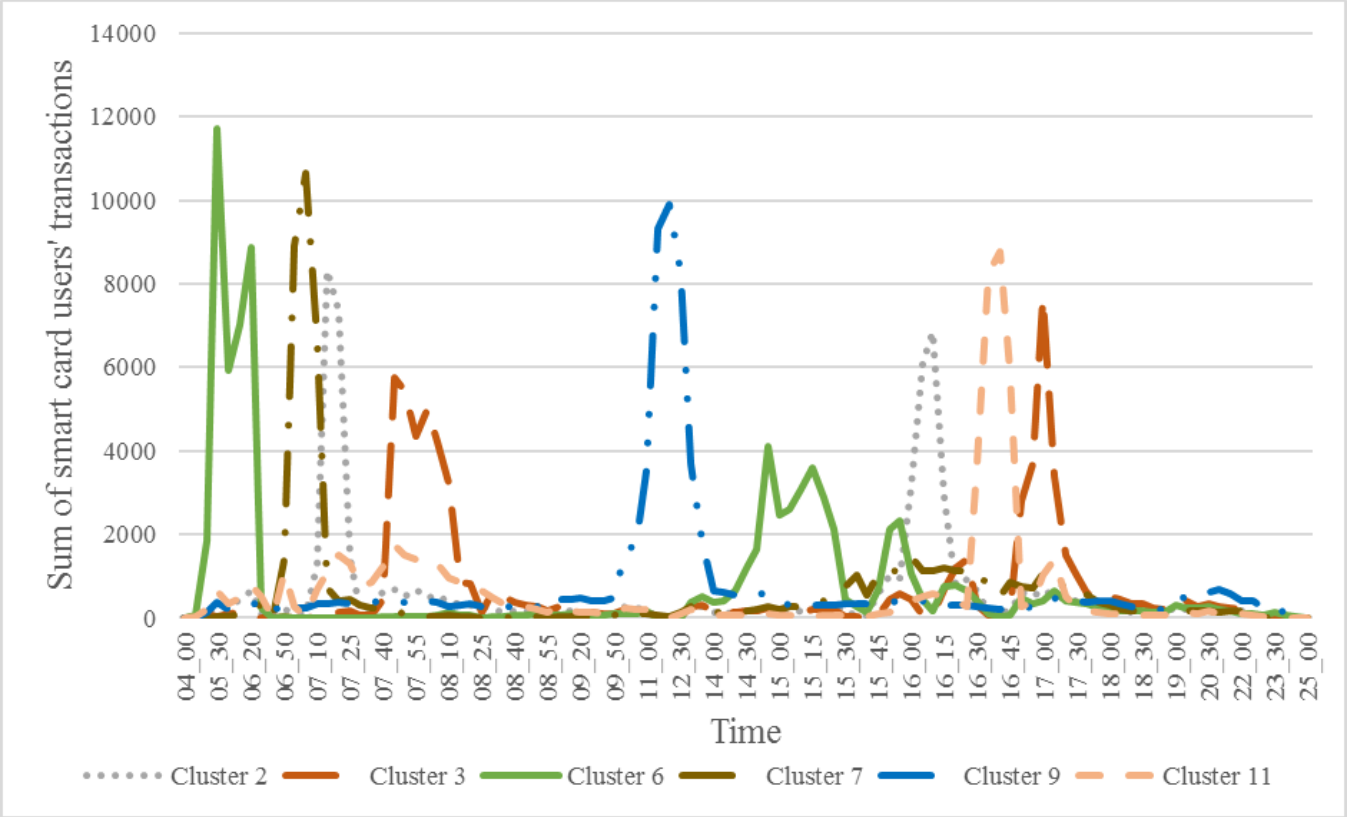


Figure 5-6: Resulting temporal profiles for some groups (N=11, S=2000, D=20), using STO data from Sep. and Nov. 2013

5.7 Conclusion

In this paper, we proposed a framework that combines cross-correlation distance, hierarchical clustering and a sampling method in order to characterize the temporal profiles of public transit travelers using data from smart card transactions. Applying this framework to the *Société de transport de l'Outaouais* transit network helped classify 333,745 user-days. We conducted a sensitivity analysis on the main parameters of this approach in order to test its validity with the dataset, analysing the number of groups, sample sizes and number of random draws for the sample.

The main limitation of this work is that the method for determining sampling efficacy is based on smart card transaction dataset of one city. Transit users from the STO may have their own unique characteristics, and the sample size found here may not apply to data from another city. Therefore, we believe that by using our methodology, the appropriate values of N, S and D can also be determined for other datasets.

We expect that this framework can also be applied to clustering the spatio-temporal profiles of transit users, simultaneously studying their location and time of use (He et al., 2019). We also look forward to improving computational times by proposing a strategy for calculating the distances between sample points and the remaining points, which is a time-consuming practice that actually requires less time than calculating the distance matrices between all points. Concerning sampling processes, dimensionality reduction, performed as a preliminary step on the dataset, could help mitigating the problem where an abnormal point could be chosen as a sample.

5.8 Acknowledgments

The authors wish to acknowledge the support of the *Société de transport de l'Outaouais (STO*) for providing data, and the Thales group and the Natural Science and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

CHAPITRE 6 ARTICLE 3: SPACE-TIME CLASSIFICATION OF PUBLIC TRANSIT SMART CARD USERS' ACTIVITY LOCATIONS FROM SMART CARD DATA [1]

6.1 Abstract

Smart card data from public transit systems has proven to be useful to understand the behaviours of public transit users. Relevant research has been done concerning: (1) the utilization of smart card data (2) data mining techniques (3) the utilization of data mining in smart card data. In prior research, the classification of user behaviour has been based on trips when temporal and spatial classifications are considered to be separate processes. Therefore, it is of interest to develop a method based on users' daily behaviours that takes into account both spatial and temporal behaviours at the same time. In this work, a methodology is developed to classify smart card users' behaviours based on dynamic time warping, hierarchical clustering and the sampling method. A three-dimensional space-time prism plot demonstrates the efficiency of the algorithm.

Keywords: Public transit, smart card data, dynamic time warping, spatiotemporal classification, activity locations

6.2 Introduction

Data from smart card fare collection systems is very useful for public transit planners (Pelletier and al., 2011). Smart card data from a public transit system assists in understanding smart card users' behaviours. This knowledge is helpful in improving the services provided by a public transit authority. Many efforts have been made using data mining to classify users' transactions. In particular, some methodologies were proposed to classify smart card users' temporal and spatial behaviours by using diverse distance metrics and the classification method. In this paper, we

[1] Soumis à *Public Transport* le 12 décembre 2018. Auteurs : Li He, Martin Trépanier, Bruno Agard.

present a method to classify public transit users according to the time and the location of their trips during the day.

This article will be organized as follows. In the next section, a literature review will focus on relevant work; mainly, the data mining methods that will be used. Then, the pragmatics and the objectives of this paper will be introduced. To solve the problem of classifying spatiotemporal behaviours, a methodology is developed in part 4. After, the "Implementation" section will introduce the cases studied and important takeaways when testing the algorithms. Then, the results and their analyses will be in part 6. The end of the article presents a conclusion that contains the contributions, limitations and perspectives of this work.

6.3 Literature review

6.3.1 Utilization of smart card data

Over the years, work has been done with smart card data in the public transit sector. Data from a smart card system enables a better understanding of users' behaviours and offers users a better service through the development of a public transit optimisation method (Pelletier et al., 2011).

In terms of data preparation and completion, relevant articles introduce the description of smart card data (Trépanier et al., 2004) and enriching the data, including a destination estimation method (Trépanier et al., 2007), an unlinked trips destination estimation using kernel density estimation (He and Trépanier, 2015), a method to improve the accuracy of the destination estimation method (He et al., 2015), and so forth. Furthermore, some methods based on transfer detection (Chu and Chapleau, 2008) and trip purpose inferences (Lee & Hickman, 2013) have been developed. These research works construct the base of public transit user behaviour analysis.

In terms of smart card user behaviour detection, smart card data can be used to analyse user behaviour; for example, characterizing users from temporal information (such as transaction time, travel duration, delay, etc.), (Morency et al., 2007; Bunker et al., 2018), by spatial information (Origin-Destination, trajectory, etc.) (Shi et al., 2014), by mode choice (Kurauchi et al., 2014; Viggiano et al., 2017) and also by the personalities of passengers (such as user loyalty) (Imaz et al., 2015), network characterization (Sun et al., 2016), analysis of external factors that influence the utilization of the network (Briand et al., 2017), and data prediction (Ceapa et al., 2012).

In particular, methods help estimate and understand a change in behaviour (Asakura et al., 2012) of various improvement strategies on transit service reliability (Diab & El-Geneidy, 2013). This research aims to understand and analyse public transit user behaviour. A better understanding of user behaviour helps improve the services provided by public transit. For example, work has been done concerning the optimization of public transit timetables (Nishiuchi et al., 2018), bus stop optimisation (El-Geneidy & Surprenant-Legault, 2010), and so forth.

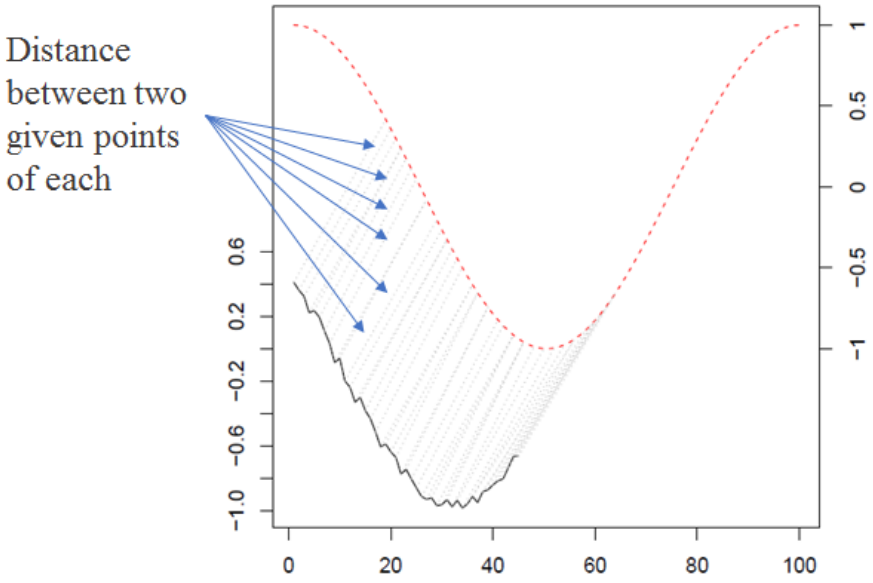
The amount of smart card transactions can be in the multi-millions for a typical city; it is therefore relevant to use data mining techniques to be able to analyse data in a meaningful way.

6.3.2 Data mining techniques

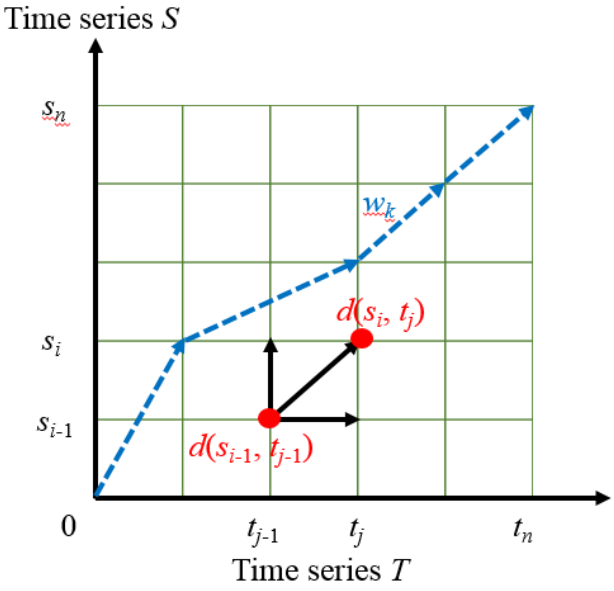
Many data mining techniques can be used to process data. Two elements must be foreseen. On the one hand, there is a range of methods, including partition algorithms (Chevalier et al., 2013), hierarchical algorithms (Rokach et al., 2005), and algorithms based on density (Kriegel et al., 2011). On the other hand, several metrics can be used to evaluate the dissimilarity of two vectors, including Euclidean distance (Deza et al., 2009), Manhattan distance (Black, 2006), cross correlation distance (Mori et al., 2016), and dynamic time warping distance (Giorgino, 2009).

Figure 6-1 illustrates the dynamic time warping method. Dynamic time warping is a popular technique for comparing time series, providing both a distance measure that is insensitive to local compression and stretches and the warping which optimally deforms one of the two input series onto the other (Giorgino, 2009). We can formally define the dynamic time warping problem minimization over potential warping paths based on the cumulative distance for each path, where d is a distance measure between two time-series elements. Warping the last moment of time series *B* to the last moment of time series *A* allows the cumulative distance between A and *B* to be minimal (Figure 6-1(a)).

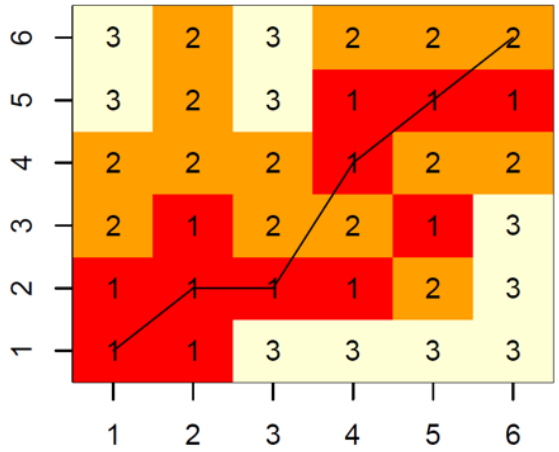
To obtain a minimum cumulative distance, the time series can be wrapped to the next time point (moment). For example, grid point (*M* -1, *N* -1) can be wrapped to (*M*, *N* -1), (*M* - 1, *N*), (*M*, *N*) to compute each distance (Figure 6-1(b)). Then, calculate all of the possible paths from grid points (1, 1) to (6, 6) to find the path with the minimum cumulative distance. In the grid above, the distance of DTW is 7 (Figure 6-1(c)).



(a)



(b)



(c)

Figure 6-1: Dynamic time warping method

Futhermore, the work from Faroqi et al., 2019 provides a very interesting review and comparative analysis of different approaches to perform sequential two-step spatial-temporal (S-T), sequential two-step temporal-spatial (T-S), and combined one-step spatiotemporal (ST) clustering of smart card data.

6.3.3 Utilization of data mining in smart card data

An issue of great interest to public transit researchers involves partitioning passengers into groups based on their trips. The classical data mining technique (k-means and hierarchical clustering) has been used to classify users' general behaviour over a period of 12 weeks (Agard et al., 2006). Some other works have been done based on k-means (Morency et al., 2006), neural networks (Ma et al., 2013) and DBSCAN (Density-based spatial clustering of applications with noise) (Kieu et al., 2014), which were bound to identify regular passengers, or propose clustering according to their behaviours. Moreover, classification methods can also be developed to analyse the quality levels of transit service (de Oña et al., 2015).

It is also of great interest to analyse users' behaviours temporally and spatially, based on the temporal data mining method (Ghaemi et al., 2017) and spatial data mining method (Ghaemi et al., 2015); the public transit card user's temporal patterns and spatial patterns have been analysed separately.

At the end, to verify the efficiency of space-time clustering algorithms, a space-time prism 3D plot (Farber et al., 2015) helps show the profile of each cluster. As illustrated in Figure 6-2, this 3D plot shows a user's (or a user group's) location for one day. It illustrates not only the difference in a public transit user's individual behaviour, but also the difference in a user group.

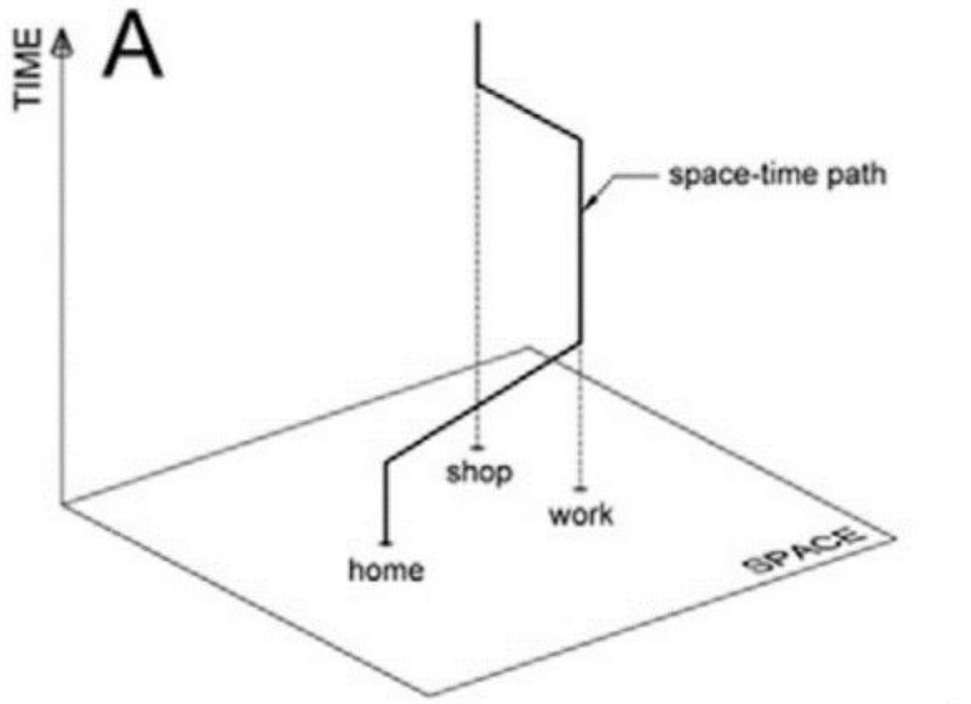


Figure 6-2: Example of space-time prism (Farber et al., 2015)

6.3.4 Limitations of the current methods

Papers have presented pertinent methodology on public transit smart card users' behaviour detection, the diversity of the classification method, and the application of data mining methods to smart card data. However, as users' behaviours can be considered time series, few articles present time series classification as a means to discover passengers' temporal and spatial behaviours. Compared to an analysis in which users' behaviours are treated separately at each time point, time series classification should have contained more information about users' characteristics.

However, time series classification is a special issue because of the limitation of the classical classification method, and the research is based on each individual smart card user's transactions instead of a daily behaviour time series. For example, when clustering using k-means, the algorithm considers only the value of vector elements, not the position of these elements in the vector. The interest in transportation planners is to consider the time of the day in the boarding sequence. To solve this problem, a temporal classification method has been developed based on cross correlation distance metrics (He et al., 2018).

Even though a temporal classification enables a classification of a user's temporal behaviours into groups, few articles present a pertinent method on how to classify public transit smart card users' daily behaviours spatially or spatio-temporally. In this article, these problems would be resolved by reconstructing dynamic time warping distance and an application of the sampling hierarchical clustering method. The results would enable transit authorities to offer better service and to satisfy the daily requirements of their passengers.

6.4 Problematic and objective

6.4.1 Problematic

By nature, a public transit path is characterized by both the time of day when a boarding activity occurs and the location where it occurred. The most intuitive way of clustering users would be to consider space and time at the same time. In this article, user behaviour will be treated as a time series of spatial locations. The classification technique will therefore take into account space and time at the same time, using a specific dissimilarity metric.

In our previous works, cross correlation distance and dynamic time warping distance have been integrated with hierarchical clustering to create time series segmentation methods [He and al., 2018]. Now, we propose integrating the spatial dimension.

6.4.2 Objective

The aim of this paper is to propose a methodology to classify users' spatiotemporal behaviours using pertinent classification algorithms and distance metrics. The behaviour is composed of the sequence of bus stop locations at each hour. To demonstrate the method, Figure 6-3 presents an example of three of the users' daily behaviours:

- The first user leaves home at 06:30 to go to school and returns home at 16:00;
- The second user leaves home at 07:30 to go to work and returns home at 18:00.
- The third one also leaves at 06:30 to go to work, but before going home at 18:00, the user goes to the supermarket at 16:00.

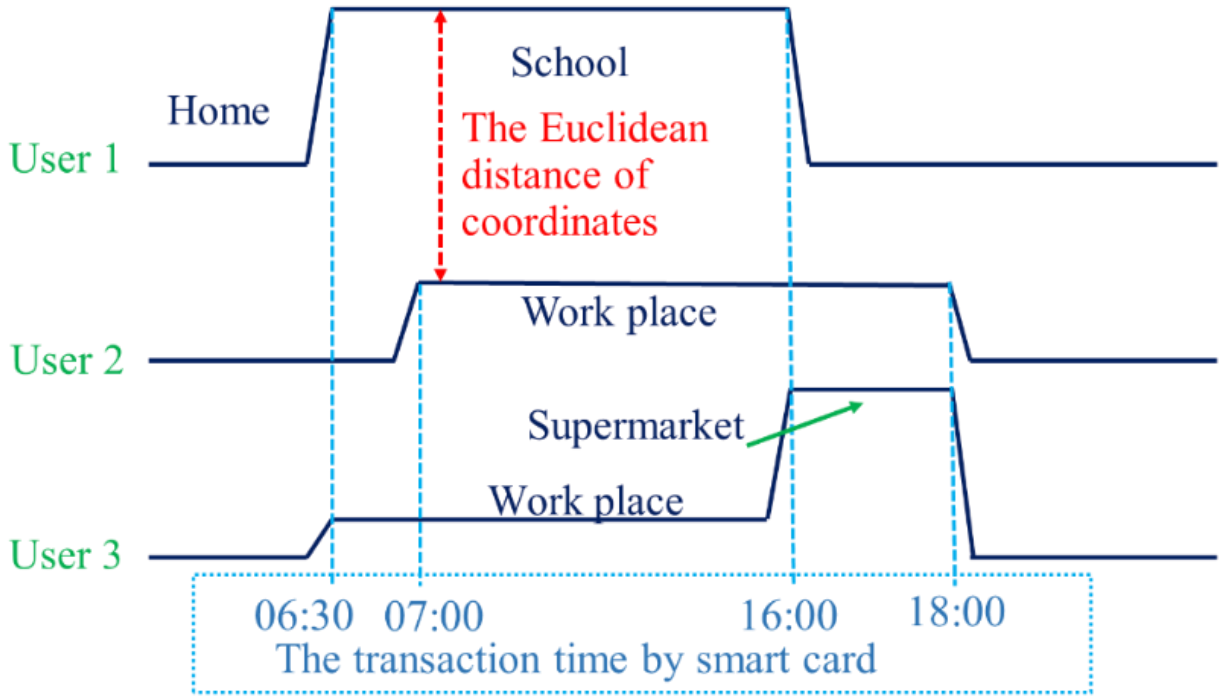


Figure 6-3: Brief example showing three user behaviours to be classified

The objective of spatiotemporal classification is to group these daily profiles in terms of time and location in order to separate them into a few clusters. In this case, if we measure the behaviours of users 1 and 2, which are "more similar" than user 3, then a group will be created with users 1 and 2, and user 3 will be in another group.

In the spatiotemporal classification, when measuring the dissimilarity of two users' profiles, we consider not only the time of the smart card transaction, but also the real distance between bus stops, serving as proxies for the user's location during the day (the Euclidean distance between school of user 1 and work of user 2, for example). The objective is to have a measure of dissimilarity that takes the two dimensions (space and time) into account in order to proceed to clustering.

6.5 Methodology

Figure 6-4 shows the methodology developed to put the proposed dissimilarity metric and clustering methods in action. The figure shows the number of records for data that were used in the case study, which are described in the next section. The methodology contains seven steps that are described hereafter.

6.5.1 Preparation of the data

First, the smart card transactions are formatted and pre-processed. The trips that occurred after midnight are adjusted so that the trip remains in the same user journey, using a 24+ hour system (step 1 in Figure 6-4). For example, a trip that occurred at 1:00 AM the next day is changed to 25:00 the same day.

Secondly, for trajectory classification, we have to use the destinations of the smart card transactions. Smart card data used for this paper do have a tap-out, so the destinations were estimated using the method proposed in (He and Trépanier 2015). Therefore, the transactions that do not contain destinations (destinations that are not estimated) are removed (step 2 in Figure 6-4).

In the final step of data preparation, for each card and for each day, a list of bus stops is created, showing the hourly sequence of stops where the user is located during the day (step 3 in Figure 6-4). Figure 6-5 presents three methods to build the time series in this case.

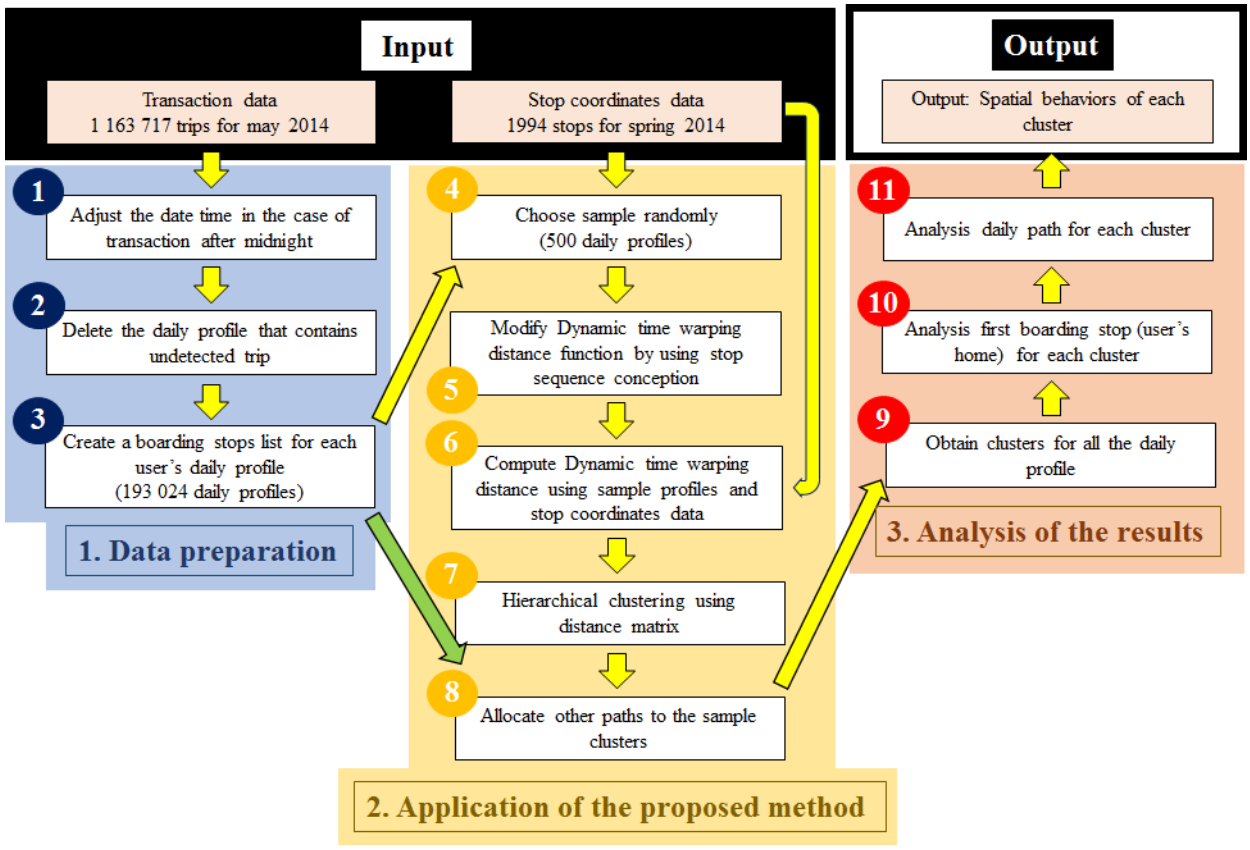


Figure 6-4: Proposed method

6.5.2 Application of the proposed method

The main idea is to link all of the stops in a sequence of given moments (…Stop 1 at 11:00, stop 2 at 12:00, stop 3 at 13:00 …until the end of the day). Bullets 2 and 3 of the figure present the method chosen in this paper to build time series for spatial and spatiotemporal classification. Table 6-1 presents the characteristics of each approach. In this paper, we use the latter two approaches.

Suppose Ts_1 Ts_2 and Ts_3 the time series for classical, spatial and spatio-temporal DTW respectively. V represents whatever value of a time series, while L represents the locations of a stop (where a smart user stays). Then N , S , H be the length of a normal time series, the number of stops visited by a user, and all the hours in a day. The following 3 functions explain the construction of time series for classical DTW (function 1), spatial DTW (function 2) and spatio-temporal DTW (function 3).

$$Ts_1 = (V_1, V_2, \dots V_N) \quad (1)$$

$$Ts_2 = (L_1, L_2, \dots L_S) \quad (2)$$

$$Ts_3 = (L_1, L_2, \dots L_H) \quad (3)$$

Note that in this research, DTW is used to measure the dissimilarity between two time series, because DTW is better adapted to the objective in this paper than other time series metric such as cross correlation distance. The parameter ‘lag’ in cross correlation distance can better represent the case if a user’s transaction is earlier or later, therefore, cross correlation distance is more used in smart card user’s temporal behavior classification. However, the parameter ‘window’ is more likely a limit to value change (in our cases, location change). Therefore, calculate of DTW is more adapted to measure spatial (location) classification.

The clustering of more than a hundred thousand users’ daily profiles is a time-consuming process. The calculation time (when feasible) is way too long, and the amount of computer memory needed would be far too much because of the size of the dissimilarity matrix. To do the clustering, we propose using a sampling approach. Our tests showed that a sample of 500 daily profiles (over 100,000) is sufficient here. This section explains steps 4 to 8 of the methodology.

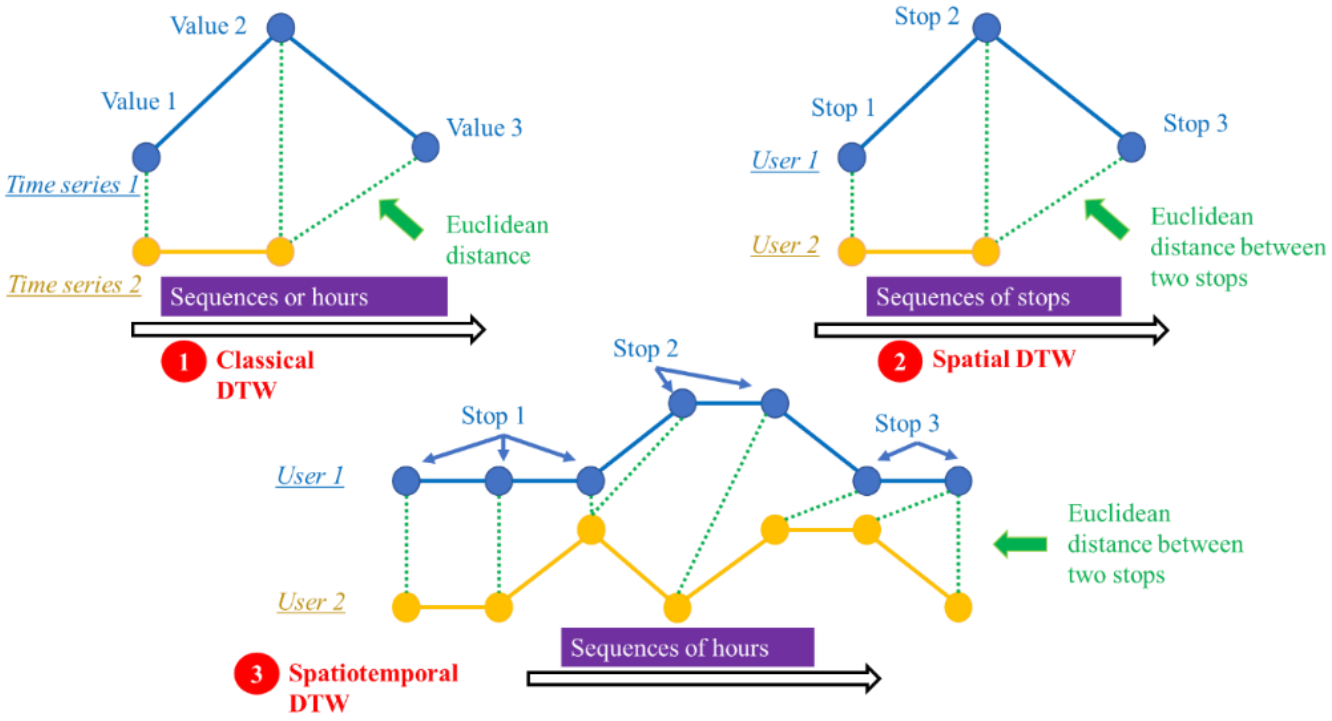


Figure 6-5: Comparison of the three DTW methods

Figure 6-6 shows the overall sampling process (He et al., 2018). At first, all observations are provided in Figure 6-6(a). The red points in Figure 6-6(b) are the randomly selected points. Then, we apply dynamic time warping and hierarchical clustering algorithms to these sample points. Figure 6-6(c) presents the clusters created in this example. We used the dendogram showing the distance between observations to cut a number of groups suitable to the needs of the analysis. We then calculate the distance between any other point and all the points of a sample group, and then allocate them to the nearest group. Finally, we obtain the groups for all of the points (time series), as illustrated in Figure 6-6(d).

6.5.3 Analysis of the results

Based on the results obtained, we analyze smart card users' behaviours, specifically looking at the boarding stops, daily profiles and space-time path for each cluster (steps 9-10-11 in Figure 4).

Table 6-1: Conception of three types of DTW

Conception	Classical DTW	Spatial DTW	Spatiotemporal DTW
Object to be treated	Time series	User path in daily profile (Stop sequences)	User path-hour in daily profile (Stop sequences at every given moment)
Point	Time point (moment)	Stop	Stop at every given moment
Sequence (time series)	Time sequence	Stop sequence (Uneven relation to time)	Stop sequence (Uneven relation to time)
Distance between grid points	Can be defined as Euclidean distance, Manhattan distance, etc.	Distance between two given stops (Only Euclidean distance)	Distance between two given stops (Only Euclidean distance)
Euclidean distance	In sense of time (X: time; Y: value in x)	In the sense of geography (X: longitude; Y: latitude)	In the sense of geography (X: longitude; Y: latitude)

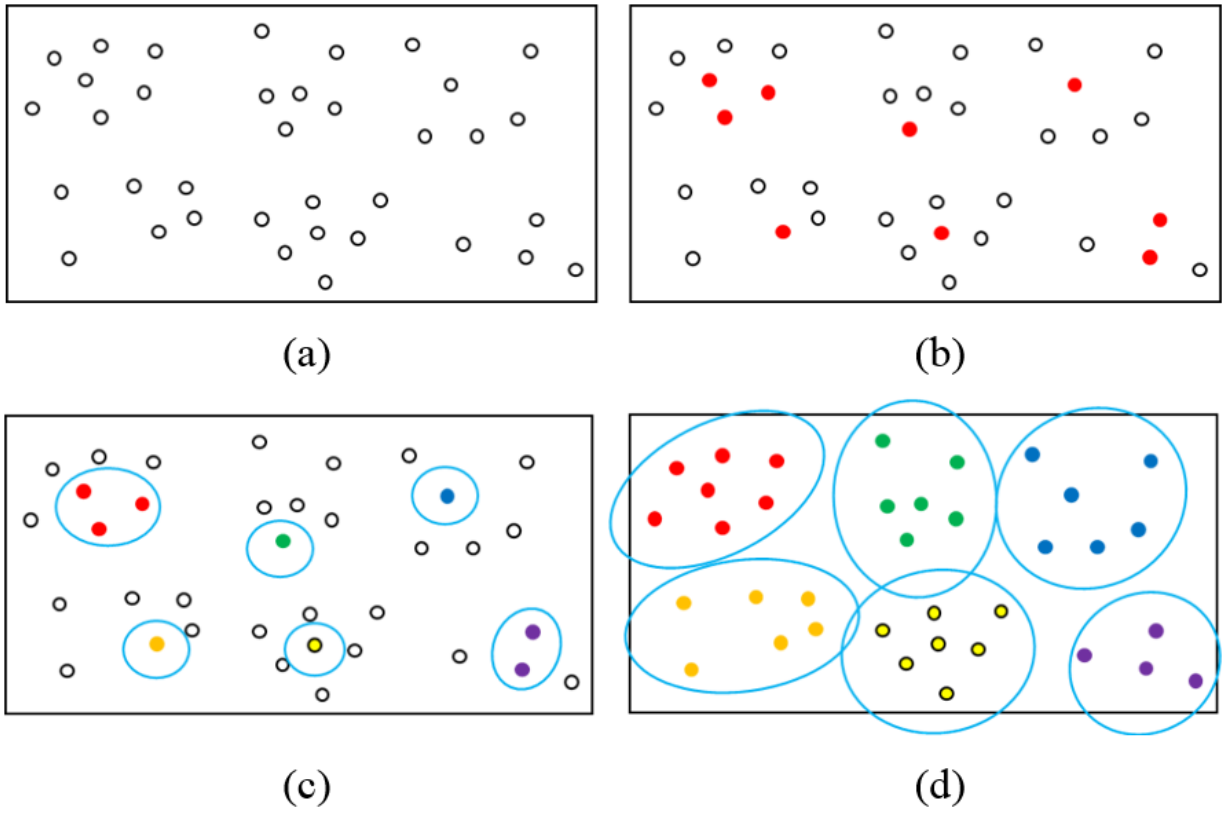


Figure 6-6: Allocation method

6.6 Implementation

The dataset was provided by the *Société de Transport de l'Outatouai*s (STO), a transit authority serving the 280,000 inhabitants of Gatineau, Quebec. The STO authority is a Canadian leader in using fare collection with public transit smart cards. This system has been in use since 2001, and a substantial proportion (over 80%) of STO users have a smart card [Pelletier et al., 2011].

In this study, all of the weekday transaction data from May 2014 has been used to test the proposed method of spatial classification. This dataset contains 1,163,717 trips.

The method is programmed in Python, which enables us to deal with such a large database.

During implementation, the number of clusters should have been determined by cutting dendrogram branches. Figure 6-7 shows the dendrogram of a spatial classification algorithm. We cut it into 10 clusters because:

We attempted to obtain as even-sized clusters as possible, even though this is not mandatory (user behaviours may not be balanced evenly). We can compare more different behaviours if we were to obtain more clusters.

In this case, if we increased the number of clusters from 10 to 11, there would be a cluster with too small of a size. Then, after the allocation process, this cluster size would be negligible compared to other clusters. In the analysis, we preferred to keep a larger cluster size. Especially, after get the result, a check has been done, to avoid mixing between intra-zone and inter-zone behavior.

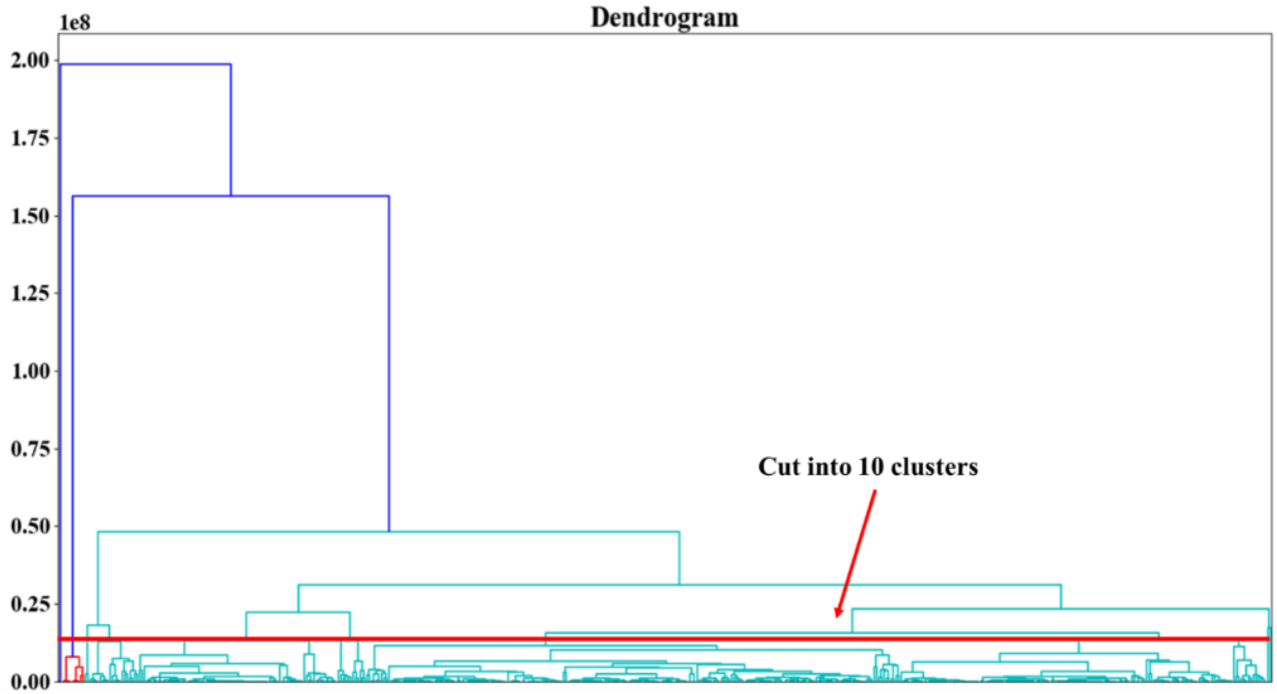


Figure 6-7: Dendrogram of hierarchical clustering of spatial classification algorithm

6.7 Results and analysis

6.7.1 Results

An excerpt of the results of the spatial classification is presented in Table 6-2. For each combination of "smart card number + date" (card-day), a stop list is generated, and a cluster is obtained.

We could find that for the combination "1292322417029248_2014-05-20", many trips were in this daily profile. One of the advantages of dynamic time warping is that it can deal with a different

number of trips during the day. We can also find that for the user “1000309”, the user’s spatial behaviours have not been changed, even though there is a minor difference in the boarding stop.

6.7.2 Analysis according to boarding stop

Figure 6-8 shows the analysis by the first boarding stop for the spatial classification. Every colour represents a cluster and dots represent the first boarding stop only. In general, the clusters are grouped by the location (coordinates); however, there are some places where the case is more complicated. For example, in the “Aylmer” area, the orange and green colours are mixed because the destinations of these two clusters are different even though the origins are similar. In this case, the destinations of green clusters are located in Ottawa, but those in the orange clusters are located in Hull or Gatineau. This is an advantage of the proposed method compared to the classical ones.

Table 6-2: Spatial classification results

Daily_profile	Stop_list	Cluster
1185321492030080_2014-05-01	['2060', '5034']	7
1188606196918144_2014-05-05	['1425', '5030']	5
1162476560982656_2014-05-13	['8071', '2618', '8030']	8
1144962089103488_2014-05-22	['2822', '1377']	6
1256806531407488_2014-05-30	['2390', '2427', '2108']	7
1243736397129600_2014-05-23	['4631', '5030', '3307']	2
1159327275886208_2014-05-27	['4442', '8101', '2724', '3991']	4
1173514901724800_2014-05-12	['3991', '4772']	1
1214358820824960_2014-05-26	['8101', '2318']	10
1292322417029248_2014-05-20	['8101', '3501', '3496', '9735', '5022', '3991']	8
1000309_2014-05-02	['5022', '2604']	7
1000309_2014-05-06	['5022', '2604']	7
1000309_2014-05-15	['5022', '2604']	7
1000309_2014-05-16	['5022', '2604']	7
1000309_2014-05-28	['5022', '2625']	7

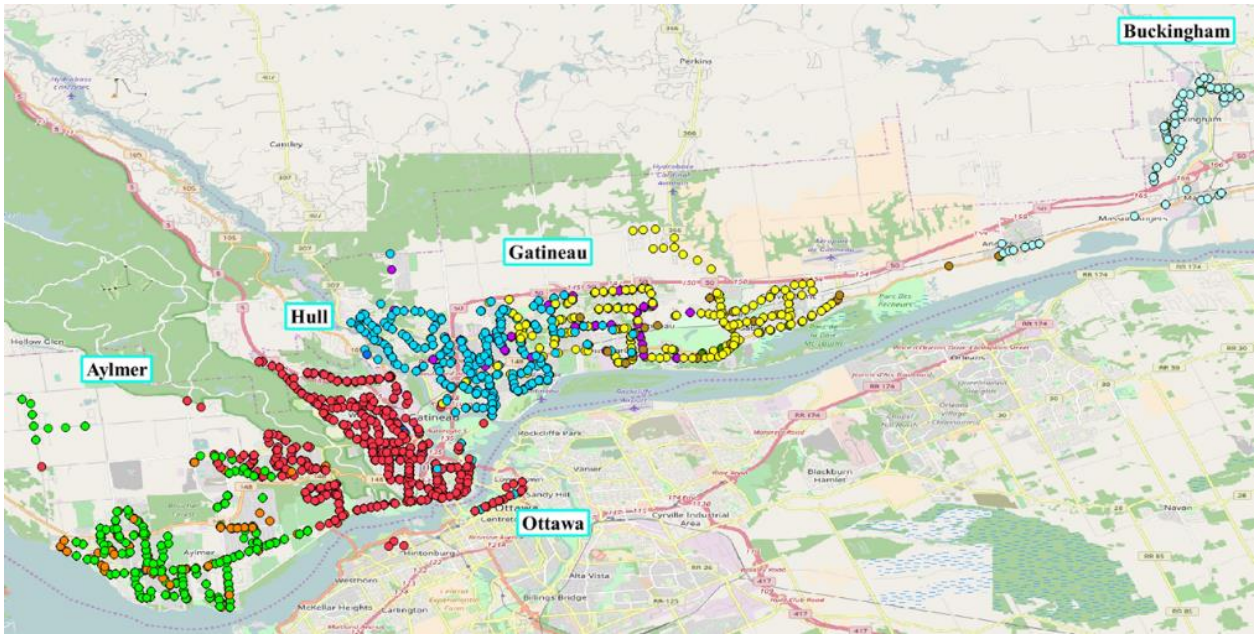


Figure 6-8: Analysis by first boarding stop

6.7.3 Analysis by daily trajectory

Figure 6-9 shows the daily trajectory of each cluster obtained through spatial classification. By watching the colours, we can see an overview the characteristics of each cluster. For example, the users from the cyan cluster live in Buckingham, and they go to work in Ottawa (the trajectory contains only the first and last stop of a transaction). Maybe they go there directly, or maybe they have a transfer in Gatineau. If we want to distinguish between these two behaviours (whether they transfer or not), we can cut the dendrogram into more clusters. This is an advantage of the proposed method compared to the classical ones.

This separation of the two behaviours helps characterize the demand. Based on this result, we may suggest to the public transit authority to implement new lines or enhance the bus service so that the people can travel directly and easily from Buckingham to Ottawa.

6.7.4 Analysis by space-time path

Based on the spatiotemporal classification result, a 3D space-time path prism of each cluster is plotted. Figure 6-10(a) shows all profiles individually, and Figure 6-10(b) shows the average path for each cluster. The Z axis of each figure is the hour within the day (25^{th} hour is for a 1 a.m. transaction).

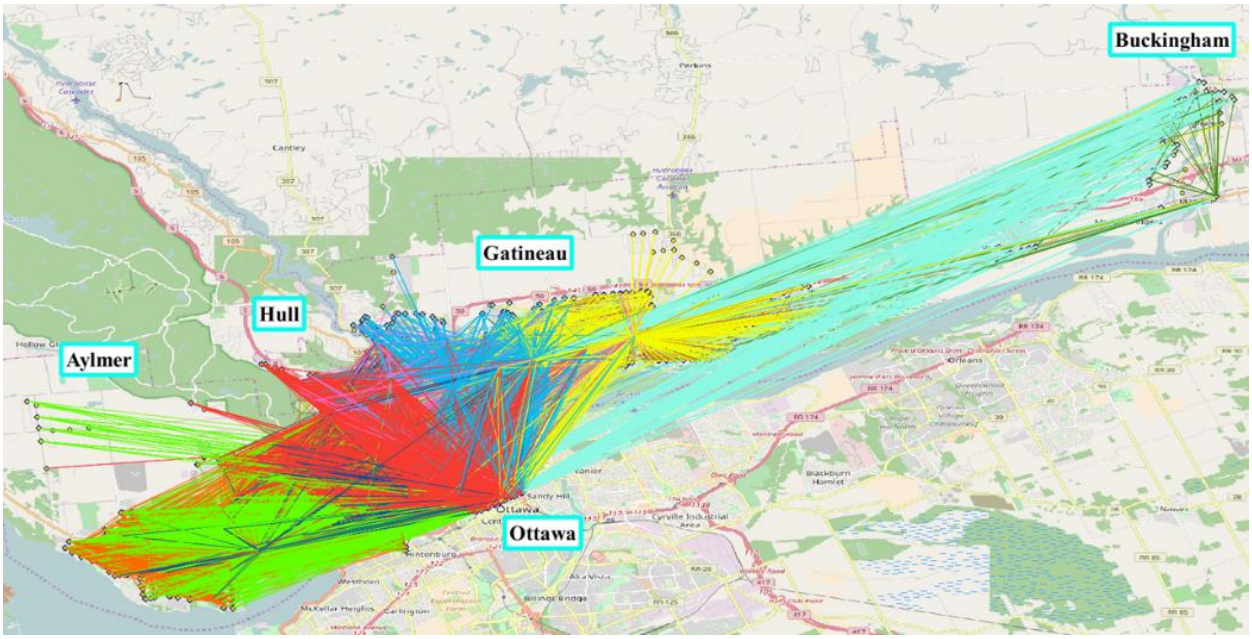
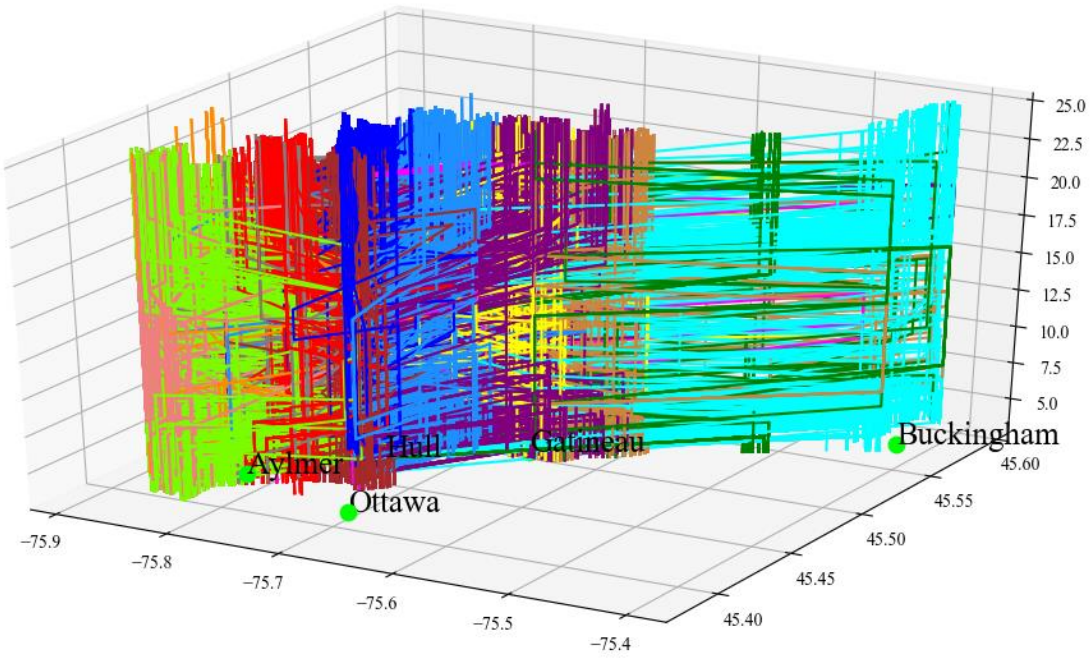


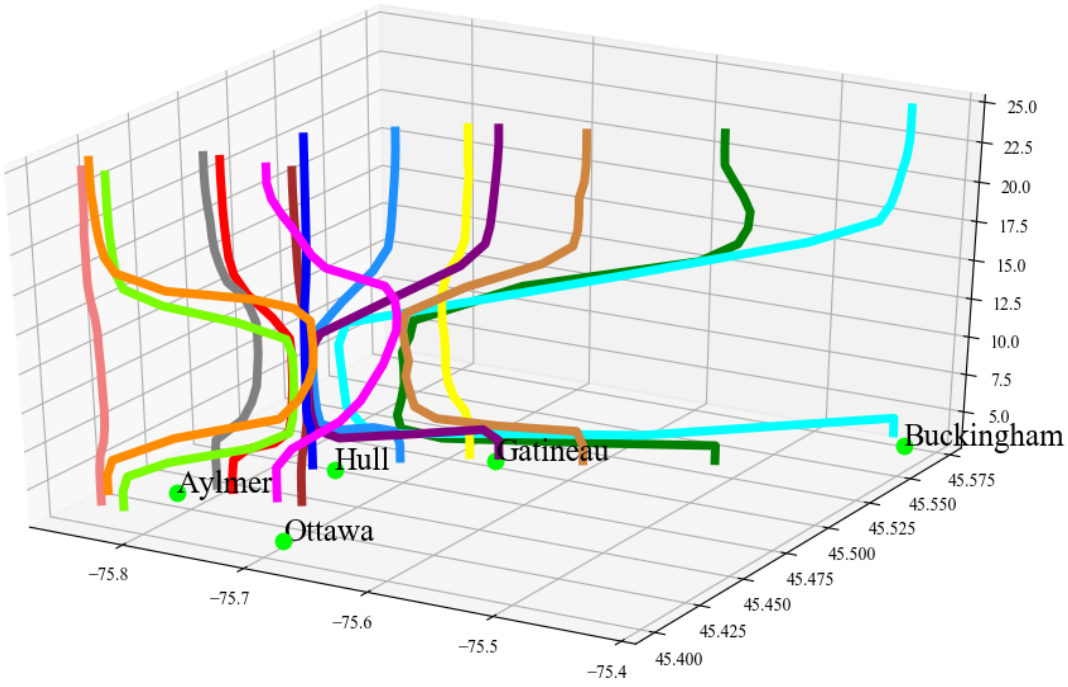
Figure 6-9: Analysis by daily trajectory

In Figure 6-10(b), even though users of the green cluster live closer to their work location than those of the light blue cluster (both from east of downtown), the green cluster leaves home earlier and returns home later than the light blue cluster. This may be due to an express bus line that links the origin and destination of the light blue cluster. Therefore, it would be possible to suggest to the public transit authority to implement an express bus line to serve the users in the green cluster so that they could save more time when commuting.

It is also possible to find that that the behaviour of the light green cluster is stable during work hours (during 9:30 – 15:00, the location of the light green cluster does not change a lot). That means these users only travel locally. It would be possible to suggest to the public transit authority to implement a special bus line for these users. This new bus line could link the origin and destination of the light green cluster, and this would only operate during peak hours, but it could still respond well to the demand of this cluster.



(a)



(b)

Figure 6-10: Space-time prism of (a) each user (b) average for each cluster

6.8 Conclusion

6.8.1 Contribution

In this paper, a new methodology based on dynamic time warping, hierarchical clustering and the sampling method is proposed to classify public transit smart card users' spatiotemporal behaviours. The results demonstrate that the behaviours can be distinguished well. Based on the results, it is possible to suggest enhancements to the public transit authority to better serve customers from specific clusters.

6.8.2 Limitations

First, the dynamic time warping algorithm is quadratic; therefore, the computation time is long. Secondly, the separation criterion is based on distance, so different behaviors can still stay in the same cluster because their dissimilarity of other factors is not considered (for example, the purpose of travel is not a dissimilarity criterion in this case). Other limitations come from the data: the estimation of destinations may not be perfect (it was not validated) and therefore this might hamper the results of the clustering method.

6.8.3 Perspective

In the future, some works are proposed to improve this new method. Firstly, at this time, we judge the quality of the classification by watching the daily trajectory and the space-time path plot. A quantitative method is needed to measure the dissimilarity between each cluster to prove that the proposed method works mathematically. Secondly, a sensitivity analysis needs to be performed to determine size of sample. Thirdly, some work should be done to reduce computation time of the dynamic time warping method. Furthermore, more suggestions could be made to the public transit authority to better respond to user demand in a specific cluster.

6.9 Acknowledgements

The authors wish to acknowledge the support of the Société de transport de l'Outaouais (STO) for providing data, the Thales group and the National Science and Engineering Research Council of Canada (NSERC project RDCPJ 446107-12) for funding.

CHAPITRE 7 ARTICLE 4: COMPARING TRANSIT USER BEHAVIOR OF TWO CITIES USING SMART CARD DATA [1]

7.1 Abstract

Public transit users behavior varies all around the world. An interesting problem is to be able to compare behaviors between them. Existing research works propose some approaches that were suitable for specific cities. However, to identify behaviors from different cities, we need a common method, based on compatible datasets. To solve this problem, this article introduces a method, based on time series metrics, hierarchical clustering algorithm and sampling method to compare public transit smart card users behavior of Gatineau (a North American city) and those of Santiago (a South American city). The result shows that 66.24% of daily behaviors of users can be classified differently in the two cities. The analysis of results demonstrates that users behavior of Gatineau are more concentrated in the morning (7:00 peak) and come to home earlier (15:00 – 17:00) compared to those of Santiago (5:00 - 8:00, 18:00 - 21:00 respectively). The analysis provides a possibility of offering better service for the public transit authority of both cities, and allows transportation researchers to analyze if a method developed in a city can be transferred to another city.

Key words: public transit, travel behavior, smart card, time series metrics, hierarchical clustering algorithm

7.2 Introduction

Data collected from an automatic collection system (in this case, smart card data) can be used to understand characteristics of public transit card users (Pelletier et al., 2011). Research has been done to exploit the potential information from smart card data. An algorithm has been developed based to estimate the alighting location given a smart card user's boarding location (Trépanier et

[1] Soumis à *Journal of Transport Geography* le 30 avril 2019. Auteurs : Li He, Martin Trépanier, Bruno Agard, Marcela Munizaga, Benjamin bustos.

al., 2007). This algorithm has been improved to estimate the unlinked trips (He & Trépanier, 2015) and has been calibrated to improve the accuracy of estimation (He et al., 2015).

Applying data mining technique to public transit user behavior is of great interest for transportation researchers (Jou et al., 2007). It helps them better understand smart card user patterns and assess the demand of public transit (Hoh et al., 2006). Classification of smart card users' daily behavior series has been explored based on temporal analysis (Ghaemi et al., 2016) and spatial analysis (Ghaemi et al., 2015). Different time series metrics have been tested and compared, to find a most suitable method for transportation case (He et al., 2018). To deal with big data from smart card systems, a sampling method is also introduced (He et al., 2017). The clusters obtained by data mining can help to obtain a better knowledge of the passenger activity during a certain period (Briand et al., 2017).

Comparing travel behaviors from different cities requires not only compatible raw data from different cities, but also novel methods to process this data, as the standard methods designed for analyzing the data from specific cities may not be well suited for the task. We solve this problem utilizing an algorithm that combines time series metrics (cross correlation distance, dynamic time warping distance), hierarchical clustering algorithm and sampling method. This algorithm allows to compare transit smart card users behavior of Gatineau, Canada from those of Santiago, Chile.

The content of the paper goes as it follows. The literature review first emphasizes the importance of smart card data classification in public transport planning and presents the current data mining methods, including time series metrics and hierarchical clustering. Then, the proposed method is presented in the methodology section, and a pedagogical example shows how we choose the metric. Next, the results of the experiments are presented and analyzed, based on one week (from Monday to Thursday) of data from a Canadian and a Chilean public transport operator respectively. Further discussions and perspectives are presented in the conclusion.

7.3 Literature review

7.3.1 Public transit smart card users' behavior classification

Classification of public transit smart card users' behaviors is widely used. A variety of methods are used for different objectives. Clustering approaches can be used such as the Hierarchical

Ascendant Classification (HAC) or k-means algorithm, to show that transit users can be quickly divided into four major behavior groups, regardless of the type of ticket they use (Agard et al., 2006). DBSCAN can be used to select regular passengers considering the dissimilarity of a sequence (last alighting stop, first boarding stop of the day, boarding time) (Kieu et al., 2014). Based on boarding stops, then boarding time, neural networks can also be used to classify regular passengers (Ma et al., 2013). Naïve Bayesian networks can be used to classify the trip purpose based on time and location of origin and destination (Kusakabe, 2014). Considering the duration of activity and land use, a method is developed using a continuous hidden Markov model to subgroup users into eight groups (Han et al., 2016).

Spatio-temporal classification can also be done by using a time series relevant technique. Based on temporal analysis (Ghaemi et al., 2016) and spatial analysis (Ghaemi et al., 2015), one can discover and analyze the public transit card user's temporal patterns and spatial patterns. Cross correlation distance and dynamic time warping are implemented to classify smart card users' daily transaction time profile (daily time profile), and the cross correlation distance has been proven more pertinent (He et al., 2018). In addition, considering that hierarchical clustering is quadratic, a sampling method can be used to deal with millions of transactions (He et al., 2017). All these methods are developed by using data from only one city, or only one public transit authority.

7.3.2 Time series metrics

In this section we introduce two time series metrics: cross correlation distance and dynamic time warping distance.

Cross correlation distance is based on the correlation between two time series. The similarity between two time series is measured by shifting one time series to find a maximum cross-correlation with another time series. The cross correlation between two time series at lag k is calculated as (Mori et al., 2016)*:*

$$CC_k(X,Y) = \frac{\sum_{i=0}^{N-1-k}(x_i-\bar{x})(y_{i+k}-\bar{y})}{\sqrt{(x_i-\bar{x})^2}\sqrt{(y_{i+k}-\bar{y})^2}} \qquad (1)$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the series. Based on this, the distance measure is defined as:

$$CCD(X,Y) = \sqrt{\frac{(1-CC_0(X,Y)^2)}{\sum_{k=1}^{max} CC_k(X,Y)^2}} \qquad (2)$$

Figure 7-1 illustrates the dynamic time warping method. Dynamic time warping is a popular technique for comparing time series, providing both a distance measure that is not sensitive to local compression and stretches the warping, which optimally deforms one of the two input series onto the other (Giorgino, 2009). We can formally define the dynamic time warping problem minimization over potential warping paths based on the cumulative distance for each path, where *d* is a distance measure between two time-series elements. Warping the last moment of time series *T* to the last moment of time series *S*, in order that the cumulative distance between *S* and *T* is minimum (Figure 7-1(a)).

To obtain a minimum cumulative distance, the time series can be wrapped to the next time point (moment). For example, grid point (i -1, j -1) can be wrapped to (i, j -1), (i - 1, j), (i, j) to compute each distance (Figure 7-1(b)). Then, we calculate all the possible paths from grid points (1, 1) to (6, 6) for finding the path with minimum cumulative distance. In this grid above, the distance of DTW (Dynamic time warping) is 7 (Figure 7-1(c)).

7.3.3 Hierarchical Algorithm

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis that builds a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types (Rokach & Maimon, 2005): agglomerative and divisive. For the first one, each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. This is a "bottom-up" approach. For the second, all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. It is a "top-down" approach. The results of hierarchical clustering are usually presented in a dendrogram. Compared to partitioning algorithms, hierarchical clustering is available for a variety of distances, but it cannot deal easily with large datasets, due to high computational costs.

7.3.4 Sampling Method

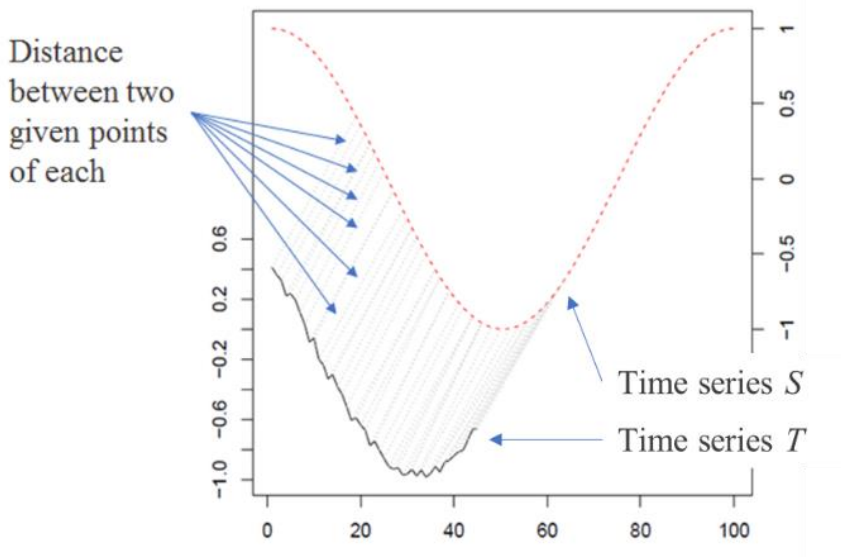
A sampling method is implemented. At first, all observations are provided. The sampling is done as following (He et al., 2017):

1) Firstly, choose random points as samples. These sample points should be as evenly distributed as possible.

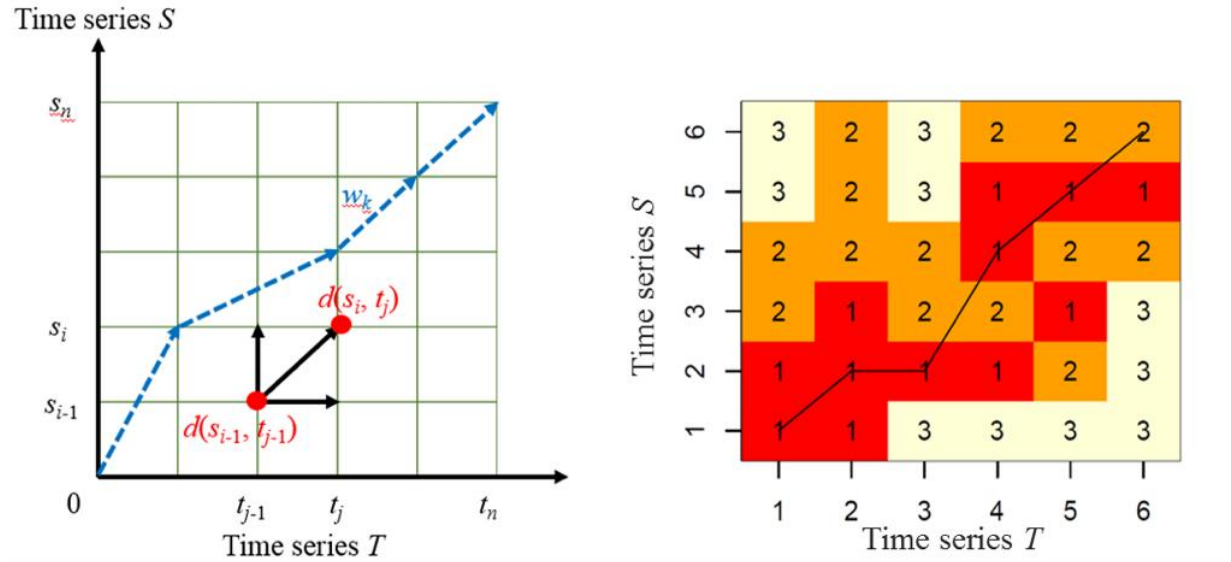
2) Secondly, apply distance metric and a hierarchical clustering algorithm to these sample points.

3) Thirdly, calculate the distance between any other point and all the points of a sample group, and allocate all the other points to the nearest group.

4) Finally, obtain the groups for all the points (time series).



(a)



(b) (c)

Figure 7-1: Dynamic time warping method (Giorgino, 2009)

7.3.5 Synthesis of literature review

In conclusion, cross correlation distance and dynamic time warping are metrics to measure dissimilarity between public transit smart card users' daily behaviors, and to obtain a matrix of distance between any two daily behaviors. By using this distance matrix, the hierarchical clustering algorithm allows us to obtain some clusters, in which each cluster represents a kind of behavior. Because hierarchical clustering does not work with big data, a sampling method is developed to deal with the data of a whole city.

However, most researches study the classification of behaviors of one city, or comparation of behaviors from different cities (Devillaine et al., 2012), but few address the recognition of those of different cities by using classification methods. This may due to two reasons. First, the confidentiality of data needs more collaboration of different cities. Second, considering the different characteristics of ridership between two cities, one method that was used to recognize different kinds of behaviors within a city may not work when we need to recognize behaviors from two cities. For example, transaction time may be used to classify users within a city, however, if the transaction time of two cities are similar, but travel time are different, then the transaction time is no longer helpful, but travel time will be helpful to recognize behaviors from different cities. The different behaviors between cities need a different attempt of methodology other than the case of only one city. Considering these factors, in the following part, we propose a methodology for analyzing smart card data from different cities.

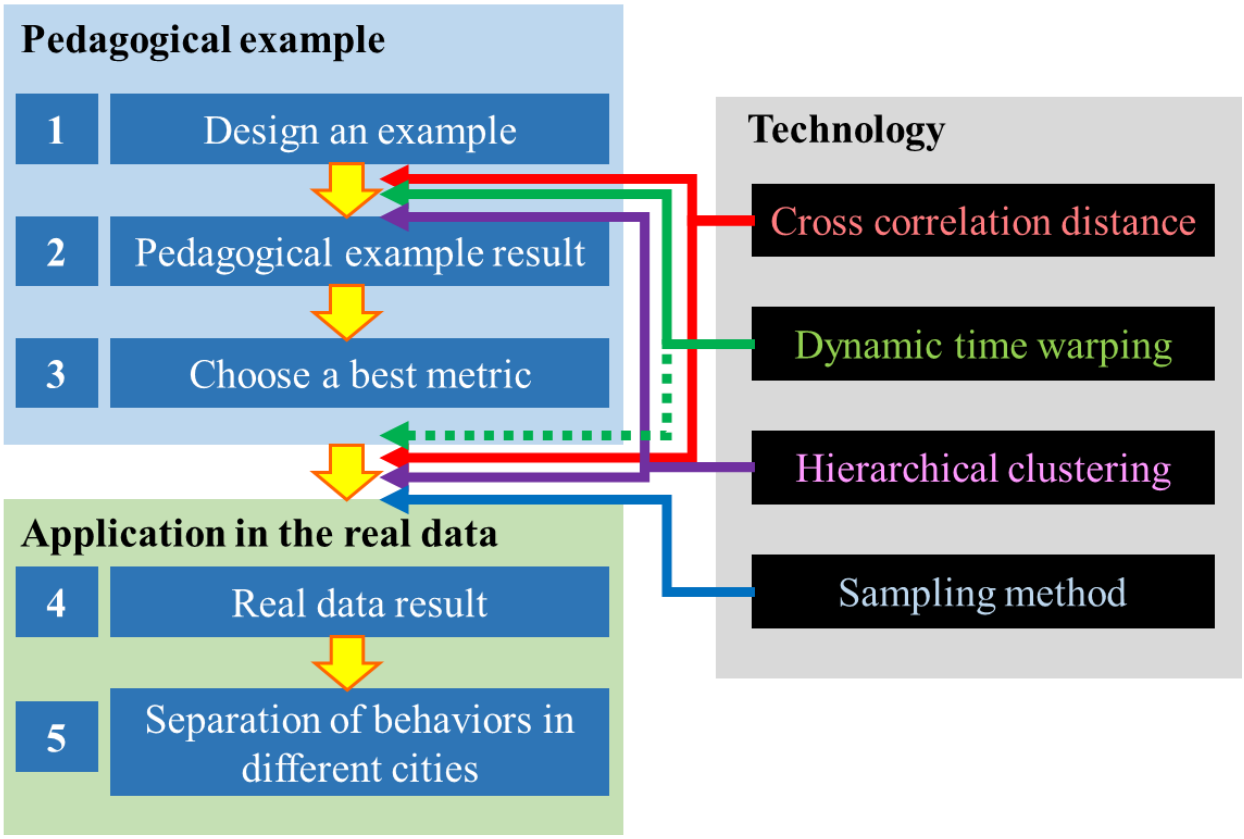


Figure 7-2: Methodology

7.4 Methodology

Figure 7-2 shows how the methodology consists of three parts: pedagogical example, application in the real data, and technology.

(1) Pedagogical example: This is an example with simplified series data. It represents some typical smart card users' behaviors, and also yields fast results. Even though cross correlation distance has been proven more pertinent when clustering smart card users' daily transaction time series (He et al., 2018), we check if in our case this is also true. Therefore, it is necessary to design a good example, and then analyze the pedagogical example result to get a best metric when clustering (Step 1 – 3 in the Figure 7-2).

(2) Application in the real data: The best metric(s) will be applied to the real data. Based on the result, one needs to check whether the behaviors of different cities are well recognized (Step 4 - 5 in the Figure 7-2).

(3) Technology: We use cross correlation distance and dynamic time warping, respectively, to calculate the dissimilarity between any two user's daily transaction time series. In fact, as expected, cross correlation is also in this case more pertinent. Therefore, dynamic time warping is no longer applied when dealing with real data (dotted line in the Figure 7-2). We apply hierarchical clustering with the dissimilarity matrix obtained by cross correlation; the result will be the clusters of smart card users' behaviors. A sampling method is applied when dealing with real data, because hierarchical clustering is a quadratic method and the computation time is long.

7.4.1 Pedagogical Example Design

Table 7-1 shows a pedagogical example. We assume ten users' daily transaction time series for each city, as presented in the first column of the Figure 7-3. For every daily profile, we verify if there is a transaction during each hour. This check covers the hours 5:00 to 21:00. During each hour, if there is at least one transaction, a "1" is put in that cell. For the daily profile "Gatineau_regular_1", this user has a transaction at 8:00 – 9:00, and at 15:00 – 16:00. For this pedagogical example, we define a variety of travel types, as presented in the last column of the Table 7-1. The "regular", "early" (transaction in the morning), "shopping" (before returning home), "dinner" (after returning home) behaviors are assumed for Gatineau, while the "regular", "early", "lunch" (at the break time at noon), "late" (come back home) behaviors are assumed for Santiago.

7.4.2 Choice of the best metric

In this paper, the best metric means a metric that allows us to better separate user daily profile according to their characteristics. The hierarchical clustering algorithm produces dendrogram, which needs to be cut. Figure 7-3(a) shows the dendrogram by cross correlation. In this case, all the daily profiles in the pedagogical example are tested. In this dendrogram, each color of leaves represents a cluster, and it can be cut into 5 clusters. The bottom of the dendrogram shows the daily profiles in two colors, and each color represents a city. For the right half clusters, the smart card users' behaviors can be well recognized. Each cluster represents or almost represents one city. However, for the left half of the dendrogram, the separation is not good.

Considering the result of left half of that dendrogram, a farther test is done to explore if dynamic time warping works based on the current result. This test is to check if dynamic time warping can

consider travel duration and distinguish users' behaviors from the clusters in which behaviors from different cities are mixed. The result of reclassification of the first cluster (green leaves in the Figure 7-3(a)) is shown in the Figure 7-3(b). The dendrogram can be cut at the height of 0.6 to obtain 2 clusters. Behavior in the first cluster (left blue leaf) is from Gatineau ("Gatineau_dinner_2"), while behaviors in the second cluster (right blue leaf) are mixed, and the behaviors from different cities cannot be distinguished. This result shows that the dynamic time warping does not work well, and the behaviors of two cities cannot be distinguished anymore.

At the end of the section, considering the efficiency and low computation cost compared to dynamic time warping, cross correlation is chosen to be the metric to test real data.

Table 7-1: Pedagogical example

Name	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Note
Gatineau_regular_1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	Gatineau regular
Gatineau_regular_2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	
Gatineau_regular_3	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	
Gatineau_regular_4	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
Gatineau_early_1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Gatineau early
Gatineau_early_2	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
Gatineau_shopping_1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	Gatineau shopping
Gatineau_shopping_2	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	
Gatineau_dinner_1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	Gatineau dinner
Gatineau_dinner_2	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	
Santiago_regular_1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	Santiago regular
Santiago_regular_2	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	
Santiago_regular_3	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	
Santiago_early_1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	Santiago early
Santiago_early_2	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	
Santiago_early_3	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	
Santiago_lunch_1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Santiago lunch
Santiago_lunch_2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	
Santiago_late_1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	Santiago late
Santiago_late_2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	

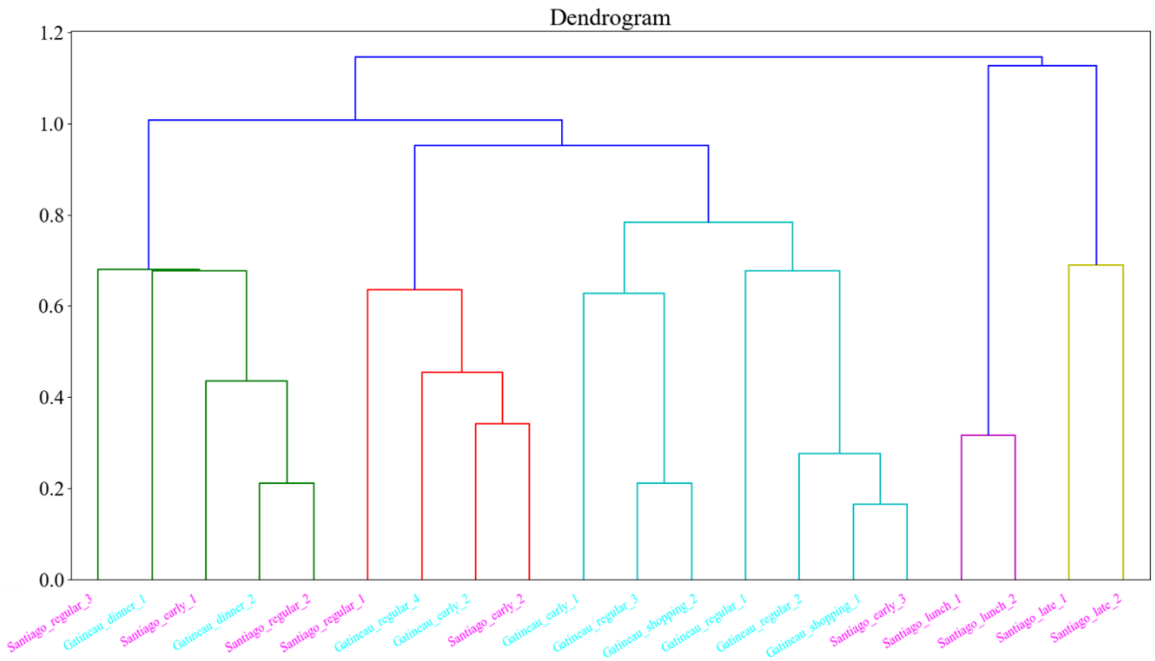


Figure 7-3 (a)

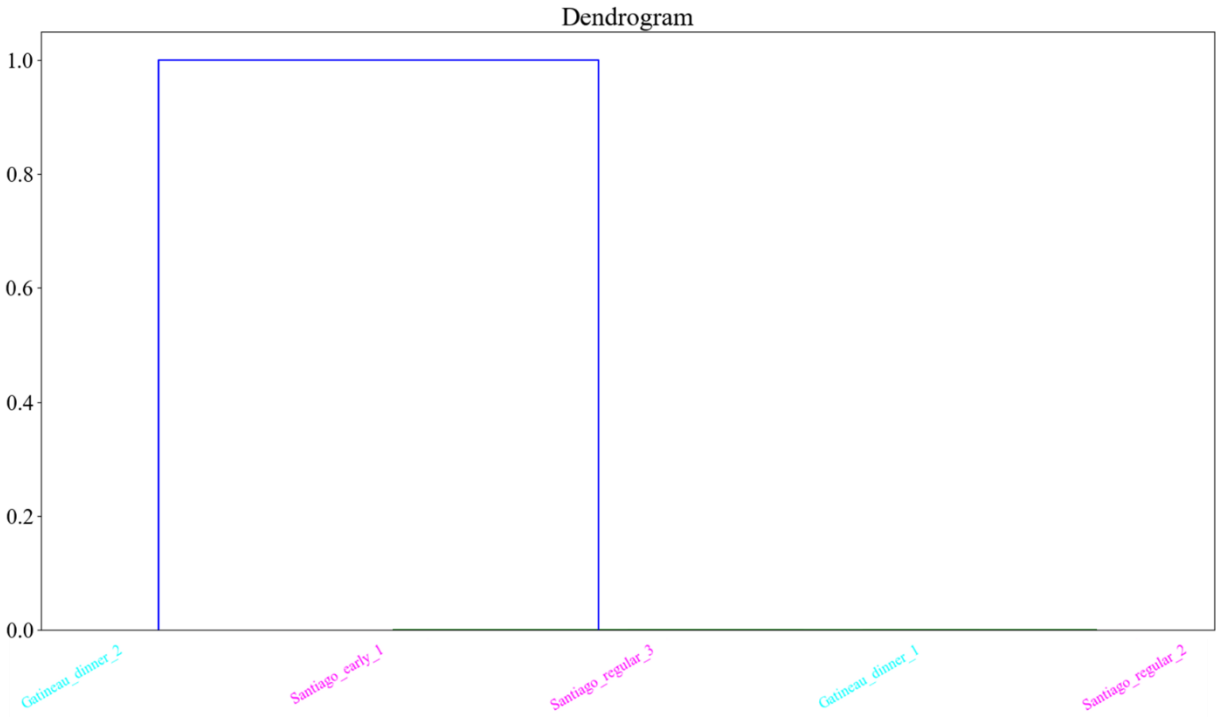


Figure 7-3 (b)

Figure 7-3: Dendrogram (a) by cross correlation distance (b) by dynamic time warping distance

7.5 Experiments

7.5.1 Case Study

The first dataset is provided by the Société de Transport de l'Outatouais (STO), a transit authority serving the 280,000 inhabitants of Gatineau, Quebec. The STO is a Canadian leader in user transit using smart card fare collection and operates a bus-only service. This system has been in use since 2001, and a substantial proportion (over 80%) of STO users has a smart card (Agard et al., 2006).

The second dataset comes from the public transport system Transantiago, based in Santiago de Chile since 2007. It is a multimodal integrated system (bus and metro) that serves a population of 6.6 million inhabitants. Overall, the system has over 6,500 buses operating daily in a network that contains 68 km of segregated busways, 150 km of reserved streets or exclusive bus lanes and over 11,000 bus stops. The integrated Metro network has 5 lines, 104 km of rails and 108 stations, and it is currently expanding (Gschwender et al., 2016; Amaya et al., 2018)

The size of Santiago data is much larger than Gatineau. The transaction number of Santiago accounts for about 5 million each day, while that of Gatineau accounts for only about 53,000. Gatineau and Santiago are located in northern and southern hemispheres respectively. Because seasons are opposite, we use Gatineau's November data and Santiago's August data as matching months.

This data contains a variety of information, including the information related to transaction, line and stop.

- Transaction: card id, boarding date, boarding time, boarding stop, boarding area, mode (subway, bus, train), vehicle (number of vehicles), etc.
- Line (only for Gatineau): the sequence of stops given a line and direction.
- Stop: the coordinates X and Y given a stop.

The information used in this paper is: card id, boarding date, boarding time.

7.5.2 Implementation

Figure 7-4 shows the implementation processes:

a) All the data from both cities are mixed. In this algorithm, all the transactions of Gatineau from 2013-11-03 (Monday) to 2013-11-06 (Thursday) are selected (97164 Card_date). Then the transactions of Santiago from 2017-07-31 (Monday) to 2017-08-03 (Thursday) are sampled so that the amount of "card_date" will be the same as Gatineau (step 1 in the Figure 7-4).

b) A 0-1 table is made by checking if there is a transaction during a certain period (Step 3 in the Figure 7-4). The duration of this period called time interval, and it should be set (Step 2 in the Figure 7-4).

When dealing with Gatineau data, time interval varies from 5 min to 30 min, depending on whether it happens during peak hours. However, in this case it is better not to preset the conception of "peak hours"; the interval of time in this algorithm will be fixed. Then, to reduce the computation cost, the interval of time is finally set to be "1 hour".

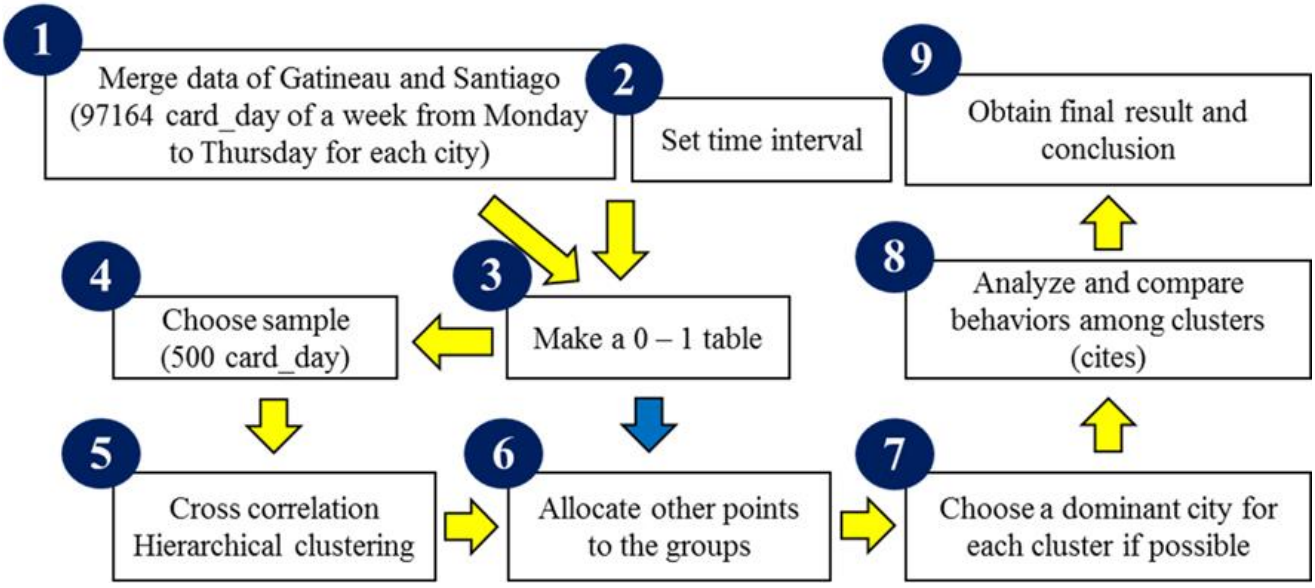


Figure 7-4: Implementation

c) 500 profiles (card_day) is selected as the sample, and the others will be non-sampled data (step 4 in the Figure 7-4)

d) Cross correlation distance and hierarchical clustering (step 5 in the Figure 7-4) is applied to regroup all the sample data profiles according to their dissimilarities.

e) Non-sample data observations are assigned to their nearest cluster. The data from each user is used to compute the distance to the closest cluster (the dissimilarity is also measured the average of cross correlation distance to all the points of a cluster), the cluster of smart card users' daily profile is obtained (step 6 in the Figure 7-4).

f) In each cluster, if the profile number of City A is 50% more than City B in a cluster, the behaviors of this cluster can represent that of City A (step 7 in the Figure 7-4).

g) Then, the behaviors of clusters representing each city will be compared and analysis. The output of this step will be the result (the step 8 in the Figure 7-4).

Table 7-2: Results of recognition

Cluster	Gatineau	Santiago	Behaviors			
			Morning	Noon	Afternoon	Evening
1	5059	11913	10:00 - 12:00			(19:00 - 21:00)
2	366	295	6:00 - 8:00		16:00 - 17:00	
3	30690	17334	7:00		15:00 - 17:00	
4	11233	9996	6:00		16:00 - 17:00	
5	5313	5035	5:00 - 6:00	13:00 - 14:00		
6	2353	8335	8:00	13:00 - 14:00		19:00 - 20:00
7	17121	8746	(6:00)		15:00	
8	12227	14396	8:00		16:00 - 18:00	
9	2822	11134	6:00		18:00 -19:00	

Table 7-3: Accuracy of recognition of the algorithm

		Real		Accuracy of recognition
		Gatineau	Santiago	
Estimated (recognized)	Gatineau	47811	26080	64.70%
	Santiago	10234	31382	75.41%
	Total	-	-	70.06%

7.6 Results

7.6.1 Results of recognition

Table 7-2 shows the number of profiles of each city for each cluster, and the behaviors (transaction time) for each cluster. As mentioned, in a cluster, if the number of cards from city A is 50% more than city B, the behavior of this cluster is associated to city A. The behaviors in green background are the behaviors representing Gatineau, while the behaviors in yellow background are the behaviors representing Santiago. The behaviors by green and yellow background can be recognized (either Gatineau, either Santiago), and each behavior belong to a cluster. Therefore, some clusters can be recognized by a city, and they are labeled by red background. For the cluster without background, in these clusters, behaviors from two cities are mixed.

The behaviors that can be differentiated (red background) account for 66.24% of the observations. For the other clusters, the behaviors of Gatineau and Santiago are about the same.

The accuracy of recognition is an important indicator to measure the efficiency of proposed algorithm. As shown in the Table 7-3, the accuracy of recognition of each city is calculated. For example, for all the behaviors that estimated (recognized) to be Santiago's, there are 10,234 behaviors that are real Gatineau's, while there are 31,382 that are real Santiago's. The accuracy of recognition of Santiago estimation is 75.41%, and the total accuracy of recognition for both cities is (Gatineau 64.70% + Santiago 75.41%) = 70.06%.

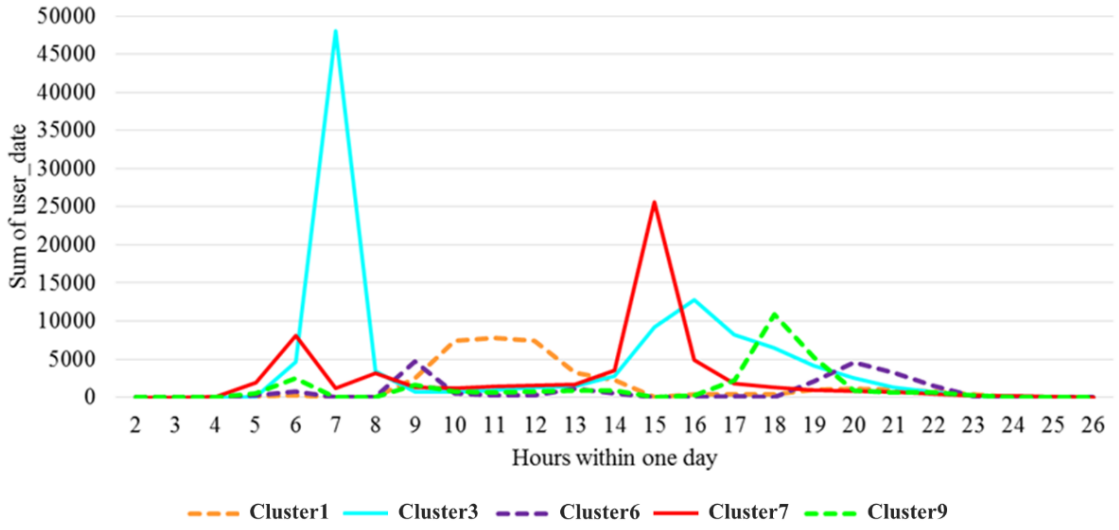


Figure 7-5: Behaviors of clusters which can be recognized

Figure 7-5 presents the behaviors that can be recognized. Combining the conclusion of behaviors which has been presented in the Table 7-2, we can compare the difference of behaviors between two cities.

1) Morning peak hours are more concentrated in Gatineau compared to Santiago. So, if a transaction is carried out at 7 am, it is more likely that this transaction is happened in Gatineau (cluster 3).
2) Gatineau users return home earlier. Therefore, if a transaction is carried out at 3 pm, it is more likely that this transaction happened in Gatineau (cluster 6).
3) During the period from 10 am to noon, Santiago users dominate this period. So if a transaction is carried out during this period, it is more likely that this transaction is happened in Santiago (cluster 1).
4) Two typical types of users' behaviors in Santiago are easy to recognize. First one (cluster 6): leaves home at 8:00, has lunch from 13:00 to 14:00, then returns home from 19:00 – 20:00. The other one (cluster 9): leaves home at 6:00, then returns home from 18:00 – 19:00.

In conclusion, users' behaviors in Gatineau are more concentrated (7:00 only) and earlier (15:00 – 17:00) compared to those in Santiago (5:00 – 8:00, 18:00 – 21:00 respectively).

7.6.2 Exploration of users' behaviors difference

It is of interest to explore why there exists such difference between users' behaviors from these cities. Some experience of two country help us to understand the difference.

To explain why behaviors in Santiago are later, the reasons could be rule and habit. Firstly, we compare the opening hours of both cities. In weekdays, the opening hours of Les Promenades Gatineau (a Canada's National Capital Region's major shopping center located in Gatineau) is from 9:00/9:30 to 20:00/21:00, while the opening hours of Mall Costanera Center (the largest shopping center in South America located in Santiago) is from 10:00 to 22:00. Then, such rules impact residences' habits, especially some people in Santiago who have a flexibility schedule. These users can work until evening and have dinner much later than people in Gatineau. Therefore, these reasons make behaviors in Santiago later than Gatineau.

To explain why behaviors in Gatineau are more concentrated, the reason could be the distance travelling to work/study. Users in both cities tend to work/study in a core area. For users in Gatineau, this core area is Ottawa. For most of these travelers, the inter zone travel distance is not that long, and users only take bus for commuting. While for users in Santiago, the core area for working/studying is from Estacion Central to Las Condes (a narrow corridor about 12 km), and the travel time is diverse. For users living far from this core area, they must leave home very early. Many of them leave home from 5:00 to 6:00. However, for users living near this core area, they have more time to sleep and leave home after 8:00. Therefore, the diverse travel distance makes behaviors in Santiago less concentrated than Gatineau.

7.6.3 Potential application

At the end, it is also possible to apply the result of the algorithm to public transit planning operations and research. For the public transit authority, they may offer a better service according to their cluster's characteristics. For example, during more peak hours, they may offer more vehicles at 7:00 – 8:00 in Gatineau, while offering less vehicles in Santiago but the duration (6:00 – 9:00) lasts longer than in Gatineau.

Transportation researchers may check if a methodology developed by using the data of a city can be transferred to another city. It is possible to determine how many vehicle service hours of transit should be assigned to a zone during a given time horizon (7-9 am, 11-2 pm, 4-7 pm etc.). Schedules setting and fare policies can be compared within a cluster between two cities. In this cluster, users from two cities share same behaviors. Each result of comparison helps us to adjust schedule or fare policy of a city.

For example, in the cluster 8 of the Table 7-2, users from both cities share a same behaviour (transaction(s) in the morning during 8:00 – 8:59 and transaction(s) in the afternoon during 16:00 – 18:59).

- For Schedules setting, peak routes of Gatineau, which are routes providing service during peak periods, could be checked if they are transferable to Santiago. Some more analysis of characteristics similarity can be done such as Origin-Destination. If cases from two cities are similar, then peak routes can be considered as a suggestion to improve service of Santiago public transit during certain hours, especially for that cluster of smart card users.

- For fare policy, modified ticket price, which means the ticket price of Santiago subway inflated during peak hours, could be transferable to Gatineau, some more analysis of characteristics similarity can be done such as users' elasticity (how they respond to fare policy) and objective (the authority needs attract or limit passengers). If cases from two cities are similar, then modified ticket price can be considered as a suggestion to improve service of Gatineau public transit during certain hours, especially for that cluster of smart card users.

If we want to apply methods developed for a cluster to another cluster, the method should be calibrated in order to shift peak hours and lengthen peak hour duration, from the characteristics of a cluster to anther cluster. This research of transferable calibration would be of interest in the future.

7.7 Conclusion

7.7.1 Contribution

Considering the difference of behaviors between different cities, this article introduces a methodology based on the time series metrics, hierarchical clustering and sampling method, to recognize them. A pedagogical example is designed to get a best metric, while the application of the algorithm to the real data shows 66.24% smart card users' daily behaviors can be recognized, and the total accuracy of recognition for both cities is 70.06%. It is easier to recognize behaviors from Santiago because the accuracy of recognition reaches 75.41%. The analysis of result demonstrates that users' behaviors of Gatineau are more concentrated and earlier compared to those of Santiago.

7.7.2 Limitation

One third of users' behaviors are not recognized due to the same behaviors of the two cities. However, a finer difference between them has not been detected. The cross correlation distance is more likely to distinguish the difference of users' transactions time, which has been done in the article. While the dynamic time warping distance may be used to distinguish the difference of travel duration, which has not been carried out in the article. A modification of dynamic time warping distance or another time series metrics may be applied to solve this problem. Therefore, a finer separation by any other method should be considered, to obtain a better result.

7.7.3 Perspective

The results of the algorithm are expected to light out the influence of external variables. Socio-economic factors could explain the belonging to clusters, for example to help exploring the impact of the level of income on smart card users' behaviors. It is of great interest to test if the impact of income is more important than the impact of city on the users' behaviors. At this time, socio-demographics attributes of smart card users are not available, but a method could be used to derive them from the estimated home location.

This algorithm offers a variety of potential applications, especially for optimizing schedules of public transit. A relevant method may be developed based on the proposed method in this article. Furthermore, if the schedule method is applied, the algorithm in this article may recognize the behaviors before and after the application of a new schedule.

7.8 Acknowledgements

The authors wish to acknowledge the continuous support of the *Société de transport de l'Outaouais* (STO) and *Transantiago de Chile* for providing relevant data, the Thales group and the Natural Science and Engineering Research Council of Canada (NSERC) for providing the funding. The authors also wish to acknowledge the support of the Fonds de Recherche du Québec – Nature et technologies (FRQNT) for providing an international internship scholarship, partially funded by Millennium Institute for Foundational Research on Data (IMFD).

CHAPITRE 8 RÉSULTATS COMPLÉMENTAIRES

Les résultats complémentaires se concentrent sur deux parties : la section 8.1 illustre les recherches plus avancées sur les résultats de classification spatio-temporelle et la section 8.2 propose une méthode de classification spatiale sur la base de densité.

8.1 Analyse de résultats de classification spatio-temporelle

8.1.1 Proportion de métro

En utilisant la méthode de classification spatiale indiquée à Chapitre 6, les données de cartes à puce de transport en commun de Santiago du Chili sont testées.

Cette base de données provient du système de transport public Transantiago, basé à Santiago du Chili depuis 2007. Il s'agit d'un système intégré multimodal (bus et métro) desservant une population de 6,6 millions d'habitants. Globalement, le réseau compte plus de 6 500 bus fonctionnant quotidiennement dans un réseau comprenant 68 km de voies de bus séparées, 150 km de rues réservées ou voies de bus exclusives et plus de 11 000 arrêts de bus. Le réseau de métro intégré compte 5 lignes, 104 km de voies ferrées et 108 gares et est en pleine expansion (Gschwender et al., 2016; Amaya et al., 2018). Le tarif du métro est un peu plus élevé que les autres modes (800 CLP (*Chilean peso*) vs 700 CLP en heure de pointe, avril 2019).

Figure 8-1 présenter le résultat de la classification spatiale (la méthode pour obtenir cette figure est la même que pour la Figure 6-9). Cette figure a pour but de présenter tous les clusters par la classification spatiale, représentés par toutes les trajectoires des usager-jours. Ensuite, la Figure 8-2 est une représentation simplifiée des déplacements de la Figure 8-1. Par exemple, pour le cluster rouge, ce sont les habitants du sud-ouest qui travaillent au centre-ville ou à l'est de Santiago. Ensuite, le Tableau 8-1 montre la proportion d'utilisation du métro, soit le ratio entre le nombre de transactions de métro et le nombre de transactions total.

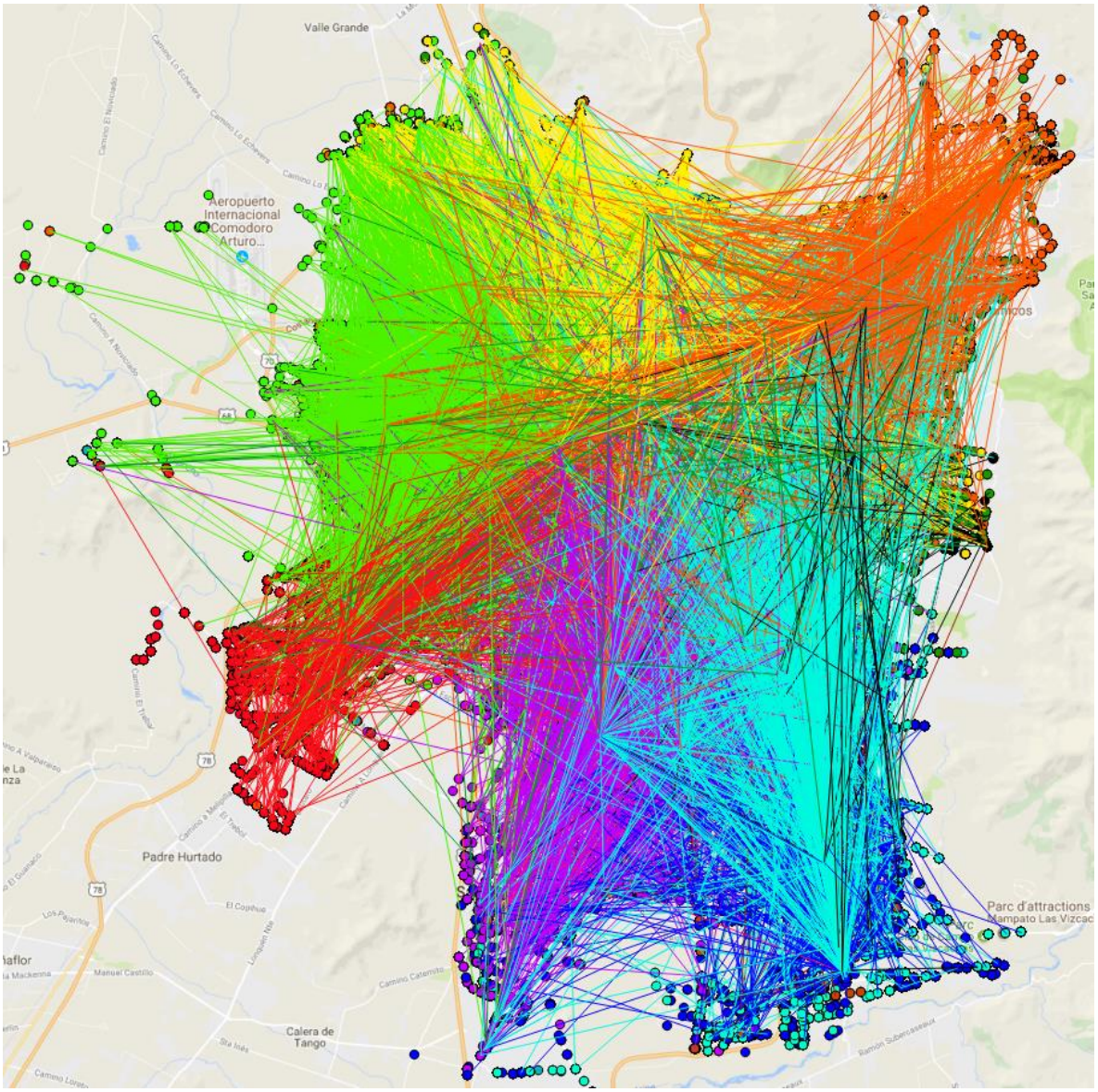


Figure 8-1: Trajectoires quotidiennes de chaque cluster

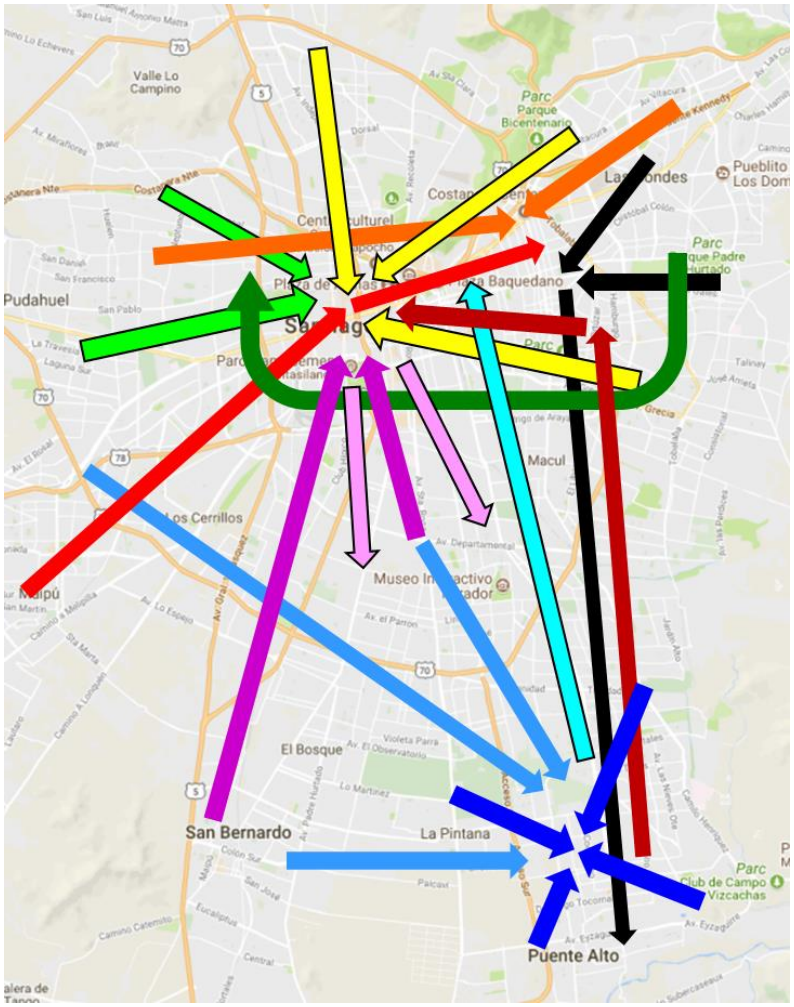


Figure 8-2: Représentation simplifiés des trajectoires quotidiennes de chaque cluster

Tableau 8-1: Proportion d'utilisation du métro par cluster

Cluster	Color	Home	Work / study	Transfer via	Proportion	Subway proportion	Note
1	Black	East / north east	South east	La Reina	1,51%	36,64%	
2	Dodgerblue	South / South west	South east		0,38%	33,84%	Nearly null
3	Red	South west	Downtown		5,20%	34,67%	
4	Lime	West	Downtown		17,02%	36,76%	
5	Pink	Downtown	Mid south		3,34%	55,21%	
6	Orange	North east west	Las condes		9,23%	40,43%	The home location are diverse
7	Yellow	North East	Downtown		31,25%	49,36%	The home location are diverse
8	Dark green	East	West / Mid west		3,67%	38,03%	
9	Blue	South east	South east		5,32%	31,58%	Only intra zone trips
10	Cyan	South east South east	Downtown Mid north east	(Maybe) via south	5,93%	39,52%	
11	Purple	South South	Downtown North / mid north east	Downtown	12,57%	43,56%	Seprated from South to mid north east directly
12	Maroon	Far south east	Downtown	Las condes	4,56%	66,89%	

La proportion d'utilisation du métro varie entre différents clusters. Les explications sur la proportion d'utilisation du métro dans certains groupes sont les suivantes:

1. La proportion d'utilisation du métro du groupe 12 est élevée (66,89%), car lors du transfert du centre-ville à Las Condes, le métro est le moyen le plus efficace pour éviter la congestion.

2. La proportion d'utilisation du métro du groupe 5 est élevée (55,21%), car le réseau de métro est bien couvert dans cette zone, ligne 2 et ligne 5. De plus, la direction (quitter le centre-ville) est opposée aux mouvements principaux de population, ce qui signifie que le métro est moins encombré pendant les heures de pointe, les passagers préfèrent donc utiliser le métro. Ces usagers habitent au centre-ville, et ont normalement des revenus plus importants.

3. La proportion d'utilisation du métro du groupe 2 est faible (33,84%), car il n'existe pas de métro directement du sud-ouest au sud-est; par conséquent, il est essentiel de prendre un bus.

4. La proportion d'utilisation du métro du groupe 9 est faible (31,58%), car pour les trajectoires intra-zone, les distances des trajectoires ne sont pas si longues. Le tarif d'un bus sera moins cher.

8.1.2 Analyse sur la coupe transversale de la trajectoire espace-temps

Basée sur la méthode pour obtenir la Figure 6-10(a), la Figure 8-3 est le résultat de la classification spatio-temporelle avec des données de Santiago. Cette figure a pour but de présenter tous les clusters par la classification spatio-temporelle, représentés par toutes les trajectoires des usager-jours. Cette figure montre les trajectoires spatio-temporelles de 200 000 usagers le 31/07/2017 (les axes X et Y sont des coordonnées (carte), et l'axe Z correspond au temps / heure d'une journée).

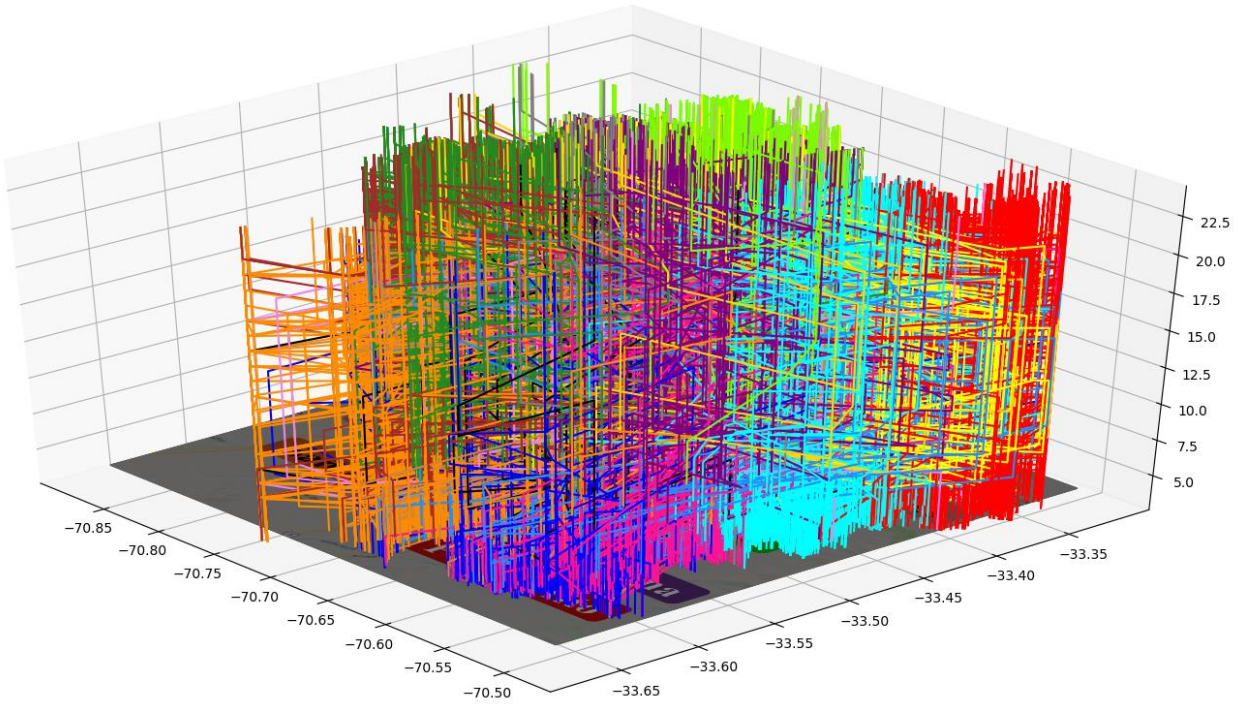


Figure 8-3: Trajectoire spatio-temporel de chaque cluster

Dans certain, Figure 8-3 est difficile à lire. Par conséquent, pour simplifier l'analyse et pour mieux voir les comportements de chaque cluster, la Figure 8-4 montre le chemin d'accès espace-temps de 100 usagers de cartes. Chaque trajectoire représente le comportement spatio-temporel d'un usager. Par exemple, pour la trajectoire la plus droite (rouge), c'est un usager qui habite à Los Condes, qui a une transaction à 9 h pour se rendre au travail ou à l'étude. Il reste là pendant quelques heures puis il a une autre transaction à 15 h pour se rendre au domicile. Basé sur la méthode de la classification spatio-temporelle proposée, cet usager fait partie du groupe rouge.

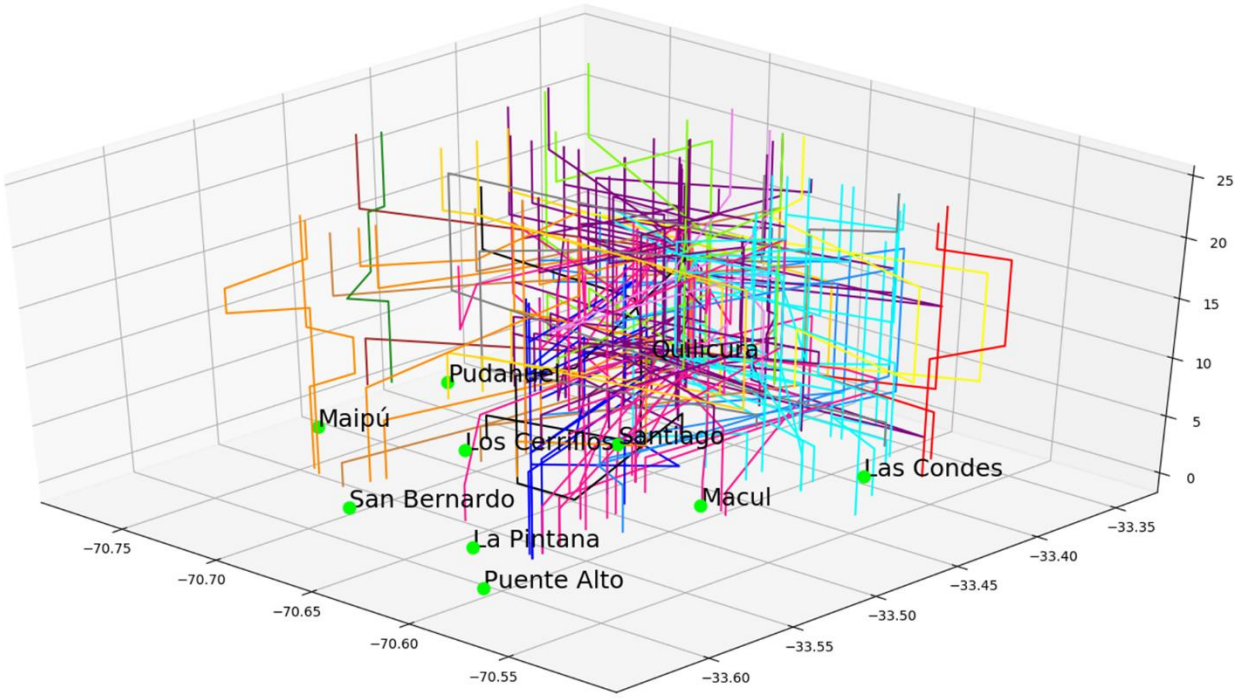


Figure 8-4: Trajectoire spatio-temporel de 100 usagers

Nous pouvons obtenir les mêmes informations de différentes façons. Dans la Figure 8-5, de cet angle, nous pouvons voir des informations spatiales pour chaque cluster. Cette fois, le nombre de groupes (16) dépasse la classification spatiale (12 groupes). Les comportements intra-zone des zones ouest (groupe vert proche de Maipú) apparaissent cette fois, au lieu du résultat de la classification spatiale (Figure 8-2), où les comportements intra-zone et inter-zone des zones ouest sont mélangés.

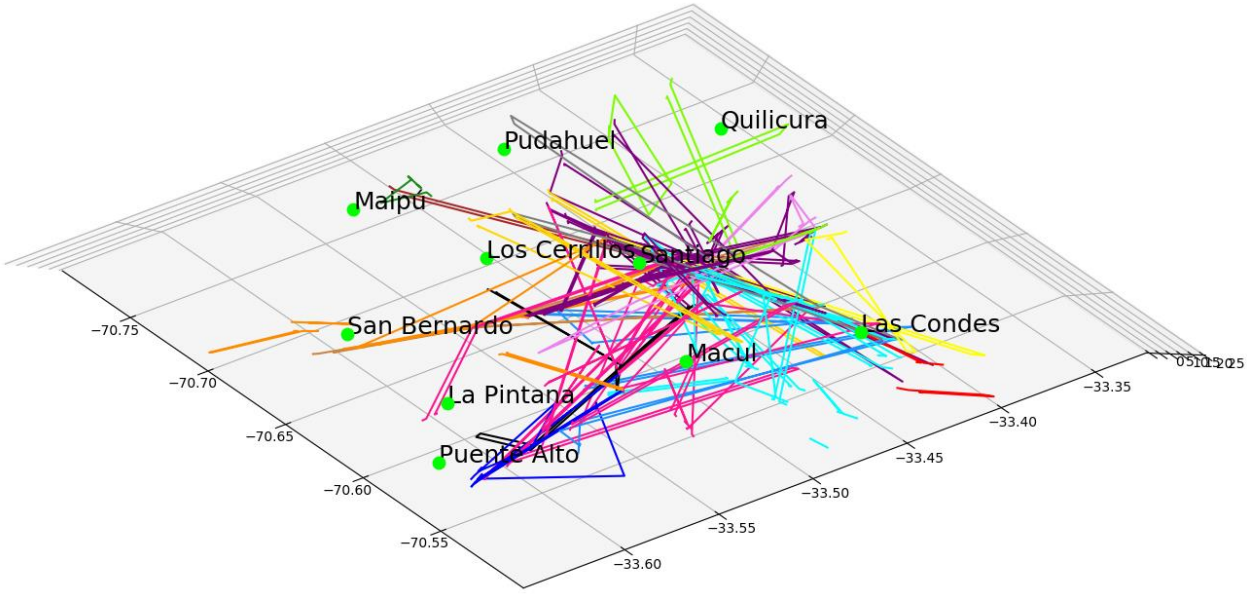


Figure 8-5: Trajectoire spatio-temporel: information spatiale

Dans la Figure 8-6, de ce côté, nous pouvons identifier les heures de pointe pour le matin (7h00 - 10h00) et pour l'après-midi (15h00 - 20h00). Par exemple, pour définir le lieu de travail (ou d'étude) d'une manière systématiquement, une solution pourrait être de calculer la densité des usagers à certains moments.

Supposons que la plupart des usagers sont à leur lieu de travail à 11h00, puis, nous coupons les trajectoires espace-temps pour obtenir une carte. Cette carte pourrait afficher tous les emplacements des usagers à 11h00. Ensuite, en calculant la densité d'usagers dans toutes les zones et en choisissant les zones de densité maximale, nous pouvons déterminer les lieux de travail. De même façon, les trajectoires peuvent être coupées à 3h00 pour calculer la densité d'usagers et mesurer de mesurer leur lieu de domicile potentiel.

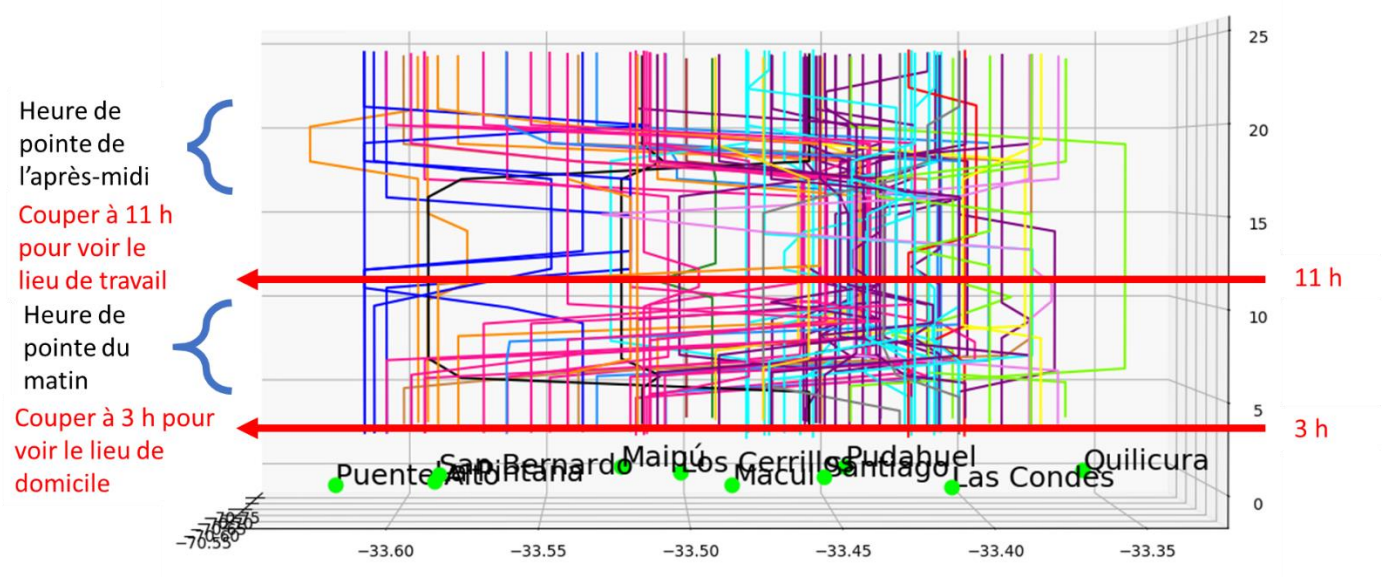


Figure 8-6: Définition du lieu de travail / domicile en coupant le chemin spatio-temporel

La Figure 8-7 illustre un échantillon de 100 usagers auxquels on applique cette méthode de densité. En utilisant une carte de chaleur, on présente la densité de domicile à 3h du matin et la densité de travail (ou étude). Les zones jaunes sont celles où les usagers sont le plus concentrés, tandis que les zones bleues sont celles où les usagers sont le moins concentrés. Il est facile de constater que le lieu de travail est plus concentré que le lieu de domicile.

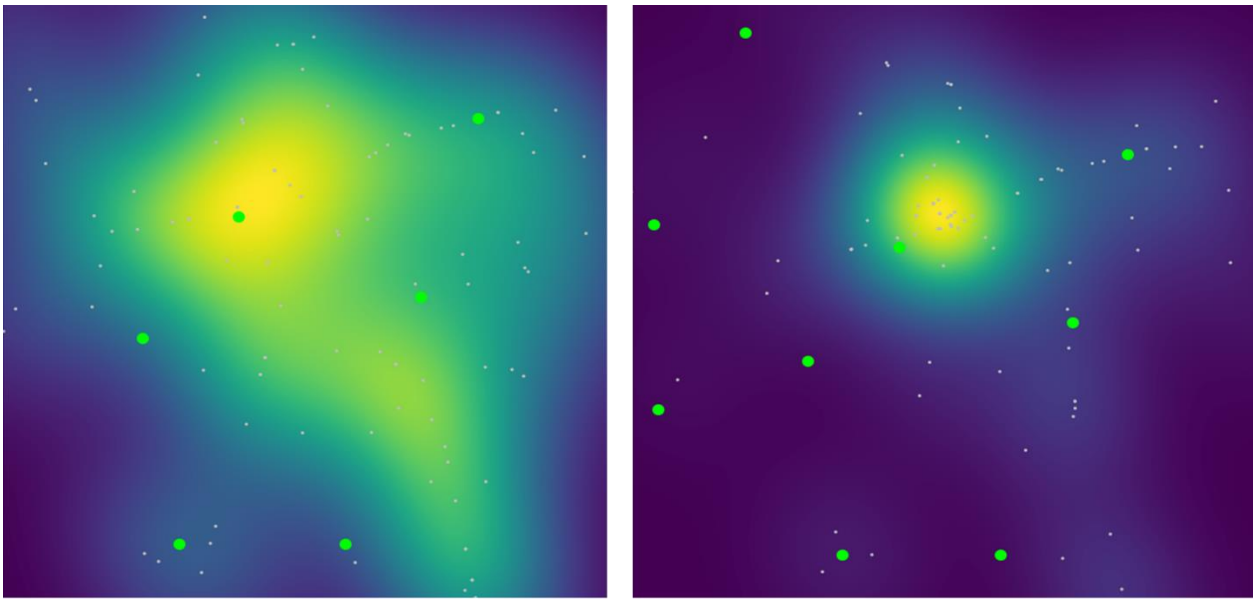
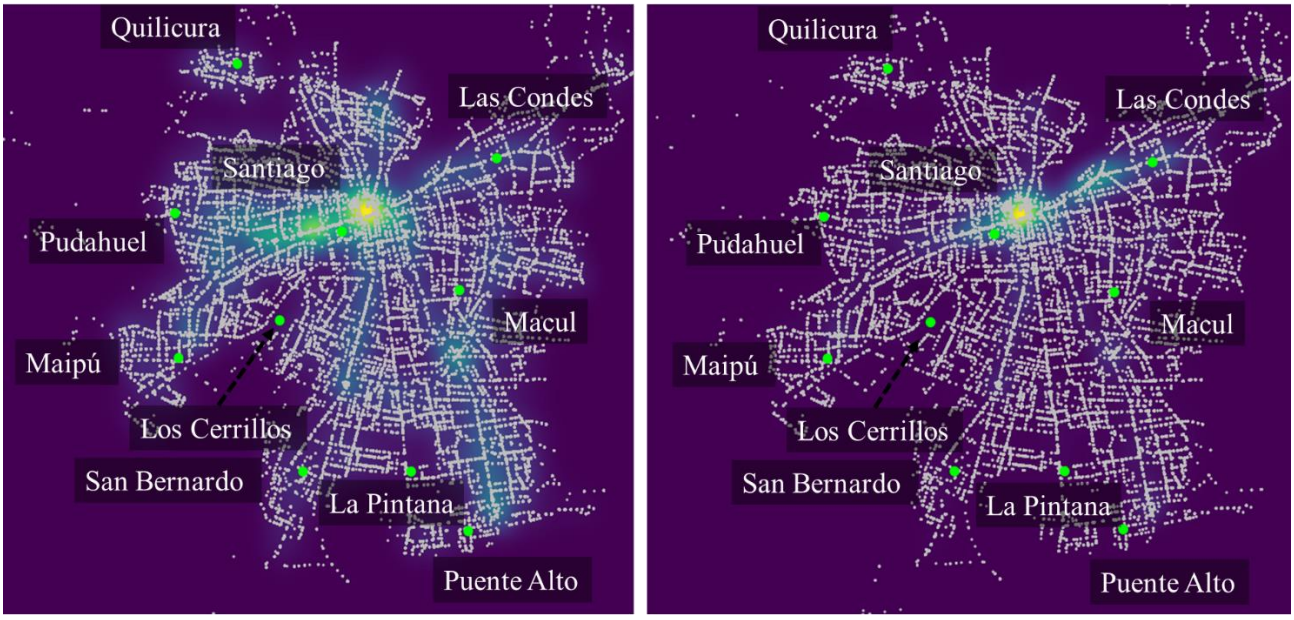


Figure 8-7: Lieu de domicile / travail de 100 usagers

Ensuite, cette méthode est testée avec tous les usagers pendant une journée. La Figure 8-8 présente le lieu de domicile à gauche et le lieu de travail (ou étude) à droite pour tous les usagers. Concernant le lieu de domicile, les usagers se concentrent sur quelques couloirs : le couloir d'est (Las Condes) à ouest (Maipú), le couloir de centre-ville à mid-sud, et le couloir d'est à sud-est (de Macul à Puente Alto).

Concernant le lieu de travail, il est beaucoup plus concentré que le lieu de domicile. Les usagers travaillent plutôt au centre-ville, sinon, il est aussi possible qu'ils travaillent dans un couloir du centre-ville à Las Condes. Notons qu'il s'agit seulement des usagers de transport en commun, et ces lieux peuvent être influencés par l'offre de service.

Enfin, en analysant le cas de tous les usagers, nous pouvons confirmer ce que nous pensions: le lieu de travail est plus concentré que le lieu de domicile.



3 a.m. 11 a.m.

Figure 8-8: Lieu de domicile / travail de tous les usagers

8.1.3 Avantage de l'algorithme

L'algorithme de la classification spatio-temporelle peut prendre en compte la durée dans différentes zones. Dans la Figure 8-9, il y a 3 comportements quotidiens comme suit. Tous les usagers concernés vivent dans les zones ouest. Le premier usager travaille à Las Condes. Le deuxième

travaille à Santiago. Pour le troisième usager, il / elle travaille à Santiago, mais visite Las Condes deux fois avec une durée limitée. Même si le troisième utilisateur visite Las Condes, son lieu d'activité principal est Santiago. L'algorithme développé peut prendre en compte ces informations et regrouper l'usager 3 dans un groupe avec l'usager 2 et non avec l'usager 1. Il s'agit d'un avantage par rapport à l'algorithme de classification spatiale uniquement.

Si l'on a seulement un déplacement interzone de faible durée, le comportement de ce profil d'usager sera considéré comme un comportement intra-zone. La Figure 8-10 compare les comportements intra-zone et inter-zone pour les usagers des zones ouest.

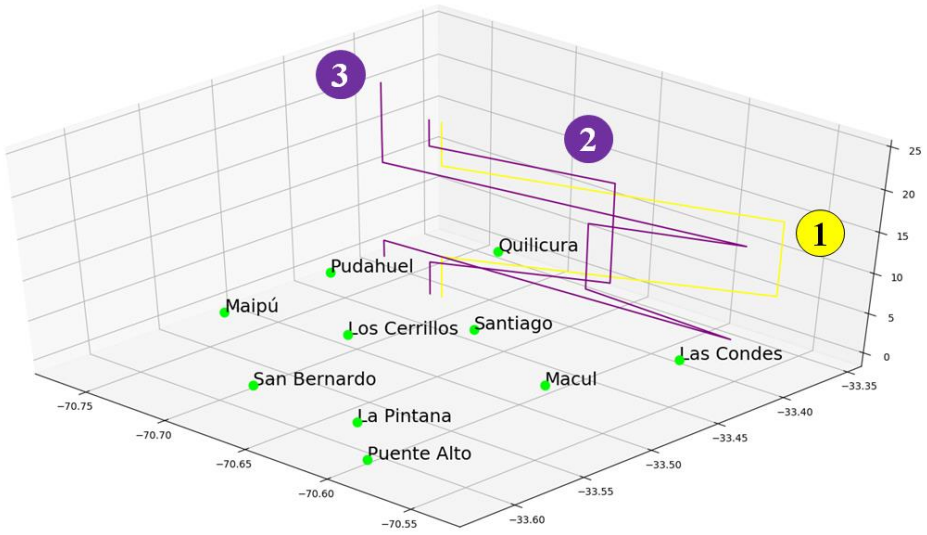


Figure 8-9: Prise en compte du temps de séjours

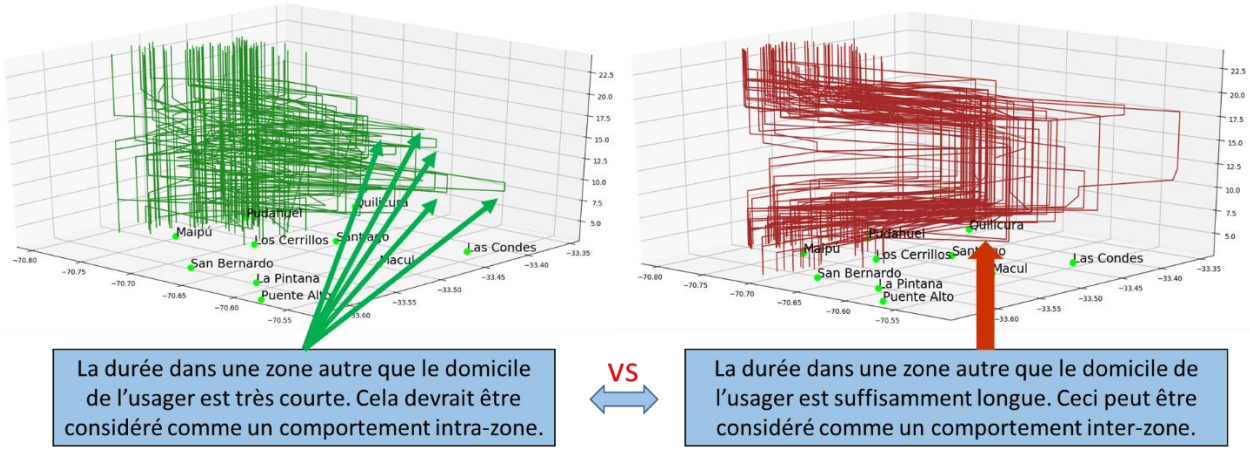


Figure 8-10: Reconnaîssance du type de déplacement : inter-zone ou intra-zone

Par conséquent, par rapport à la classification spatiale, la classification spatio-temporelle permet d'éliminer le bruit : si un usager reste peu de temps à une destination, cette destination peut être éliminée du comportement du groupe.

8.1.4 Analyse sur la moyenne de la trajectoire espace-temps

L'emplacement moyen de chaque heure est calculé pour chaque groupe. Cela permet de mieux comprendre les comportements de chaque cluster. Comme présenté à la Figure 8-11.

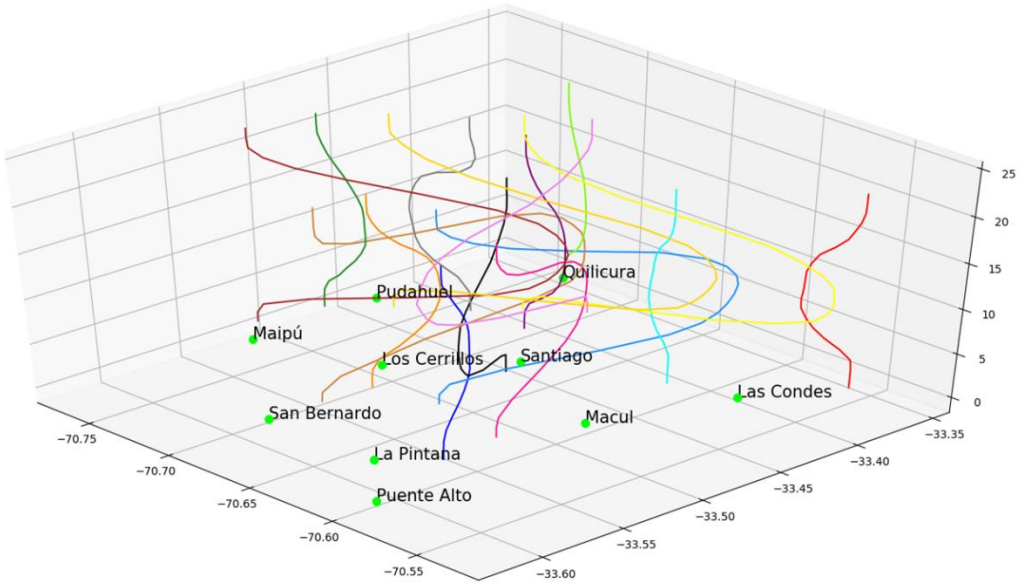


Figure 8-11: Moyen de trajectoire espace-temps

Il est intéressant de comparer l'heure de départ entre les usagers qui réalisent des trajets longs et courts. La Figure 8-12 compare les usagers de longues distances (du sud à Santiago (cluster brun) et du sud à Las Condes (cluster bleu)) et de courtes distances (intra-zone de Las Condes (cluster rouge)). L'analyse montre que l'heure de départ des groupes de longues distances est à partir de 4 heures du matin, tandis que celle des groupes de courtes distances est de 2 heures plus tard (6 heures du matin).

De plus, l'emplacement moyen du cluster rouge n'a pratiquement pas changé, passant de 11 heures à 15 heures. Cela signifie que les usagers de ce groupe travaillent pendant presque toutes ces heures. Tandis que pour les groupes de longues distances de trajet (tel que le groupe bleu), l'emplacement moyen change toutes les heures (les emplacements moyens de 10h, 11h, 12h… sont tous

différents). Cela pourrait supposer que la planification du transport en commun est plus facile pour les utilisateurs intra-zone que pour les usagers effectuant de longs trajets.

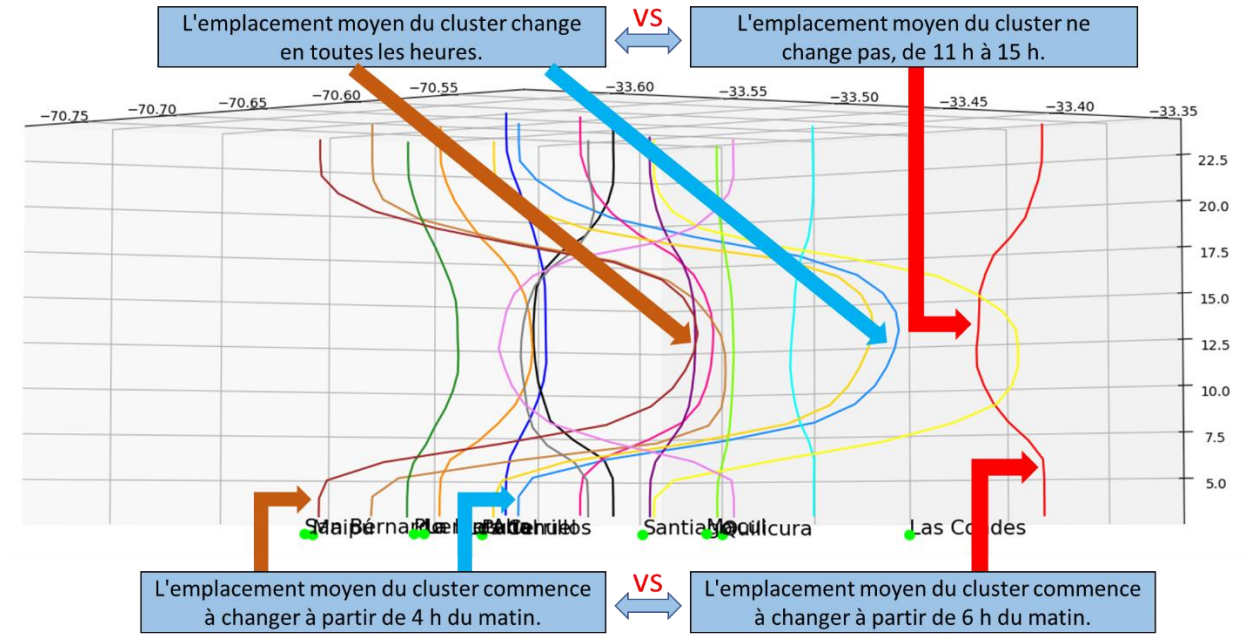


Figure 8-12: Heure de départ et durée du travail

8.1.5 Analyse sur la déviation de trajectoire espace-temps

La Figure 8-13 représente l'écart type d'emplacement pour chaque cluster. Il est intéressant de voir le premier écart-type (68%), le deuxième écart-type (95%) et le troisième écart-type (99,7%) d'un groupe, comme présentés à la Figure 8-14.

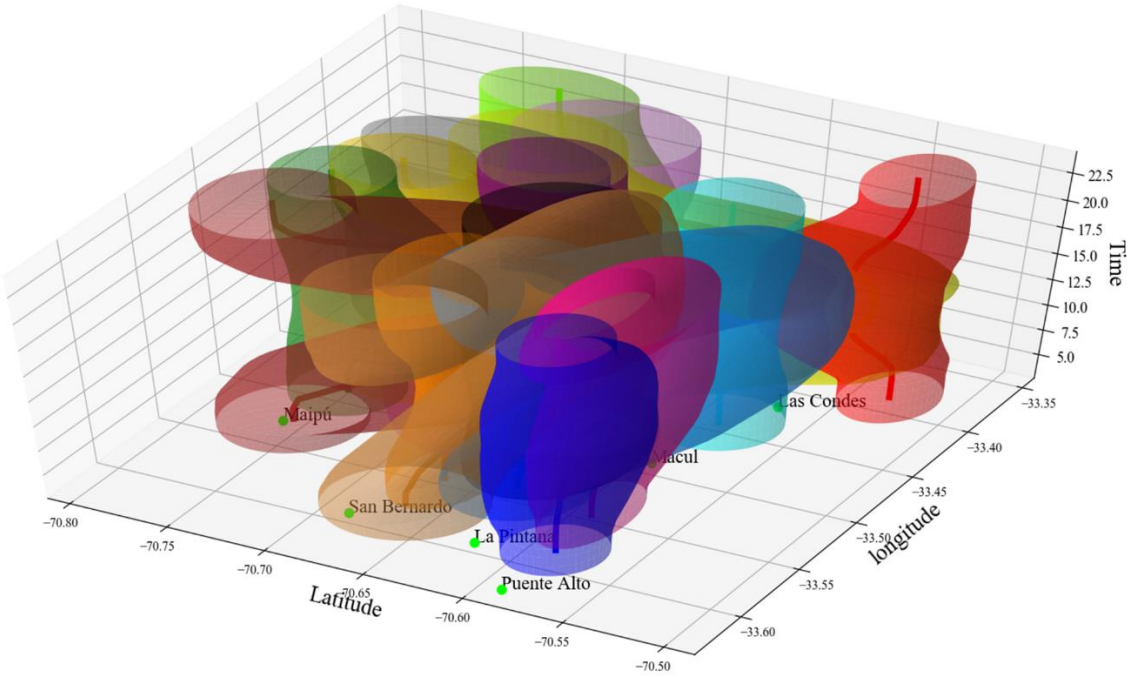


Figure 8-13: 1 écart type de localisation de chaque cluster

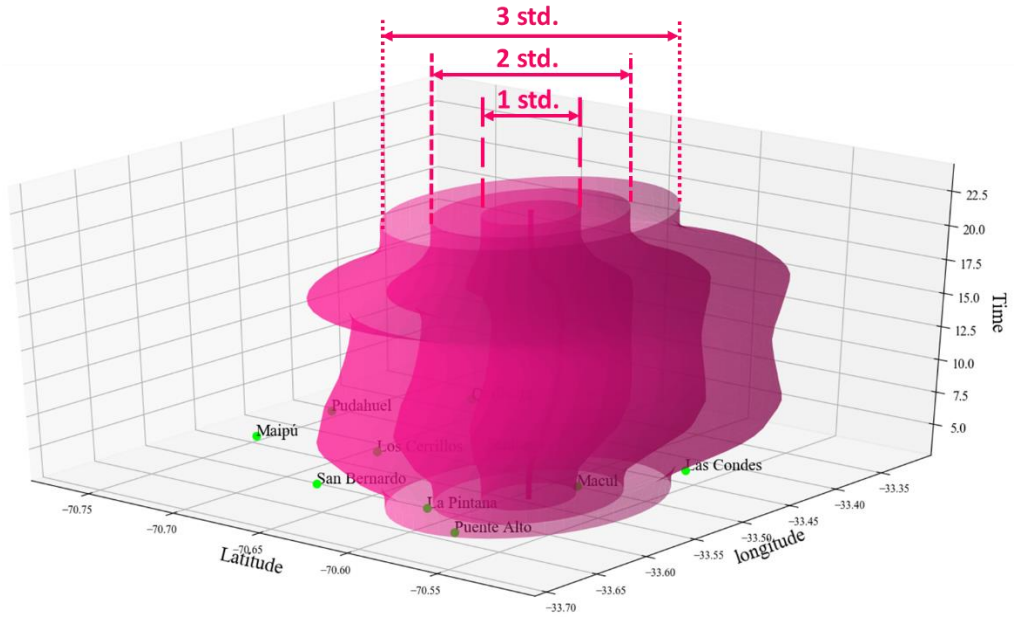


Figure 8-14: 3 écarts type de localisation d'un cluster

Pour les usagers à longue distance, l'écart type des heures de pointe est généralement très grande, car les comportements des usagers ne sont pas concentrés. Par exemple, certains usagers sont déjà arrivés à leur lieu de travail, tandis que d'autres sont encore au domicile. L'écart type des heures de pointe du matin sont parfois inférieures aux heures de pointe de l'après-midi, ce qui signifie que

les comportements des usagers aux heures de pointe du matin sont plus concentrés. La Figure 8-15 compare les zones de différents écarts type et les trajectoires espace-temps réels.

(a) Pour la zone inférieure au premier écart type, c'est la zone principale, la plupart des usagers apparaissent dans cette zone.

(b) Pour la zone entre le premier écart type et le deuxième écart type, le temps de séjour d'usager dans cette zone est « normal » (souvent supérieur à 5 heures, c'est-à-dire ils se rendent là pour travail ou étude).

(c) Pour la zone entre le deuxième et le troisième écart type, quelques usagers apparaissent, et le temps de séjour des usagers dans cette zone est inférieure à 5 heures.

(d) Pour la zone à l'extérieur de trois écart type, peu d'usagers apparaissent et le temps que l'utilisateur reste dans cette zone est inférieur à 2 heures.

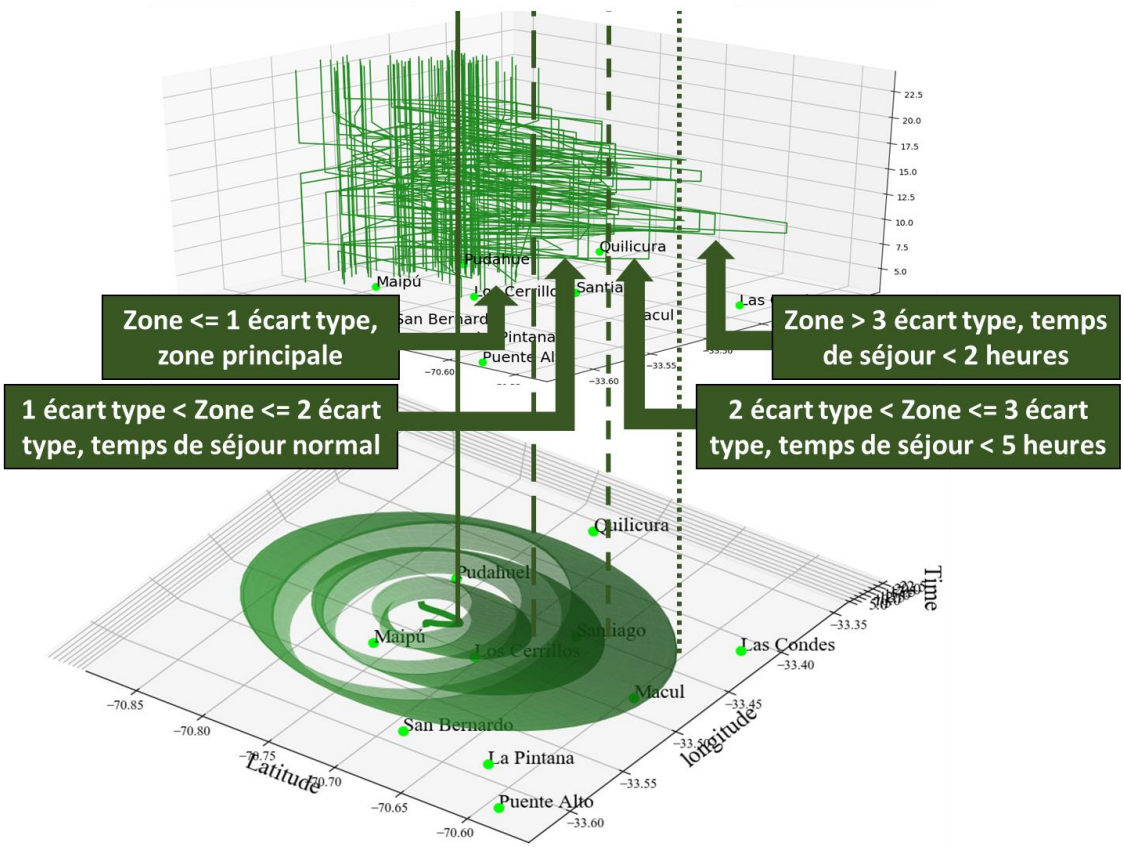


Figure 8-15: Comparaison entre les écarts type et le comportement des usagers

Sur la base de l'analyse précédente, il est possible de suggérer la planification du transport en commun spécifiquement pour un cluster en utilisant 3 écarts type. La Figure 8-16 est un petit exemple:

(a) Pour la zone inférieure au premier écart type, nous proposons des bus de grande capacité car la plupart des usagers se déplacent dans cette zone.

(b) Pour la zone entre le premier écart type et le deuxième écart type (mid-sud à Santiago et à Los Cerrillos), il est recommandé d'offrir des autobus de capacité normale parce que la plupart des usagers restent longtemps dans cette zone. Nous pouvons inférer qu'ils travaillent ou étudient là-bas, ainsi ils seraient des voyageurs réguliers.

(c) Pour la zone entre le deuxième et le troisième écart type, (mid-sud à Macul, à Providencia, au nord-ouest de Santiago), nous n'avons pas besoin de les considérer spécialement. Mais nous pouvons réserver les possibilités de transfert vers ces zones, car certains utilisateurs y restent moins de 5 heures.

(d) Pour la zone à l'extérieur de trois écart type (centre-sud jusqu'à Las Condes, jusqu'à Pudahuel), nous n'avons pas besoin de les prendre en compte, et nous n'avons pas besoin d'étudier les possibilités de transfert spécialement, car peu d'usagers visitent, et il n'y a que moins de 2 heures s'ils visitent.

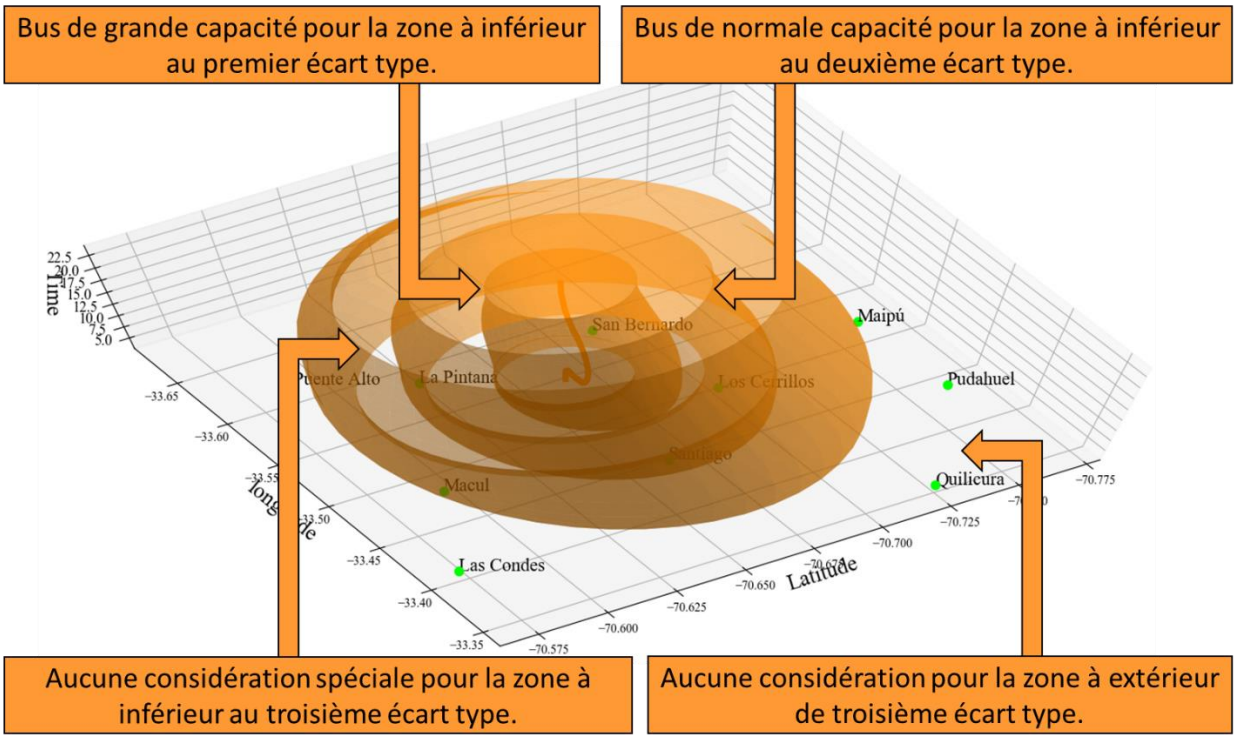


Figure 8-16: L'écart type aide à la planification du transport en commun

8.2 Classification des zones basée sur la densité

8.2.1 Zonage des arrondissements

Le résultat de la classification spatio-temporelle peut être utilisé pour distinguer des arrondissements. Comme mentionné à la Figure 8-6, les trajectoires spatio-temporelles peuvent être coupées à 3 heures du matin afin de localiser le domicile de chaque usager (obtenir la localisation de chaque usager à 3 heure du matin, et considérer cette localisation comme la localisation de domicile de cet usager).

Nous calculons la densité d'usagers de chaque cluster à l'aide de l'estimation du noyau, puis nous choisissons le cluster avec la densité maximale pour représenter une zone. Le résultat est représenté sur la Figure 8-17.

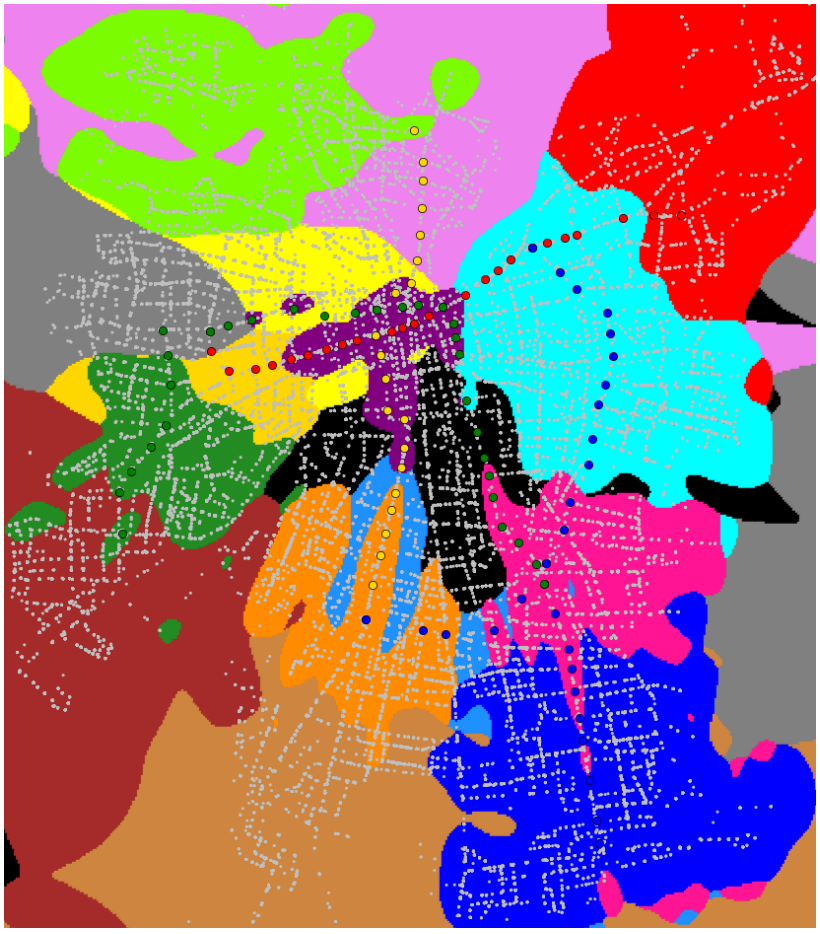


Figure 8-17: Zonage des arrondissements basé sur la densité de domicile de chaque groupe de classification spatio-temporelle

La densité des groupes d'usagers (de comportements spatio-temporels) divise la ville en 16 « arrondissements ». Il est également intéressant de voir que :

(1) Les terminaux de métros sont souvent proches de la limite de deux arrondissements. Par exemple, le terminal nord de la ligne 2 (ligne jaune) se trouve près de la frontière des arrondissements vert et violet. Cela signifie que les comportements des usagers de cartes à puce sont modifiés au-delà de la zone couverte par le métro. Cela s'est également produit dans le terminal sud-ouest de la ligne 5 (ligne verte), le terminal est de la ligne 1 (ligne rouge).

(2) Les lignes de métro peuvent prolonger un arrondissement. Par exemple, le groupe rose (sud-est) comprend deux zones très étroites avec des stations de métro. Les comportements des utilisateurs deviennent plus similaires à cause du métro. En particulier, pour certaines lignes telles que le terminal sud-ouest de la ligne 5 (ligne verte), la zone située entre deux stations appartient à un autre groupe, mais les deux stations appartiennent à un même groupe.

8.2.2 Classification des zones basé sur la densité d'heure de première transaction

La conception de l'estimation par noyau s'applique lorsque nous développons l'algorithme de la classification des zones sur la base de la densité. Dans l'estimation par noyau, la bande passante est la largeur du noyau de convolution utilisé. Dans notre cas d'étude, la bande passante mesure si une transaction à un arrêt a un impact sur un arrêt plus loin. Par exemple, si la bande passante est illimitée, une transaction d'arrêt à Santiago pourrait avoir une incidence sur la densité de transactions de Gatineau.

Lorsque vous choisissez une bande passante large, vous pouvez voir les comportements des usagers plus agrégés, tandis que vous choisissez une faible bande passante, les comportements des usagers sont plus désagrégés. Dans la Figure 8-18, la largeur de bande est fixée à 0.3.

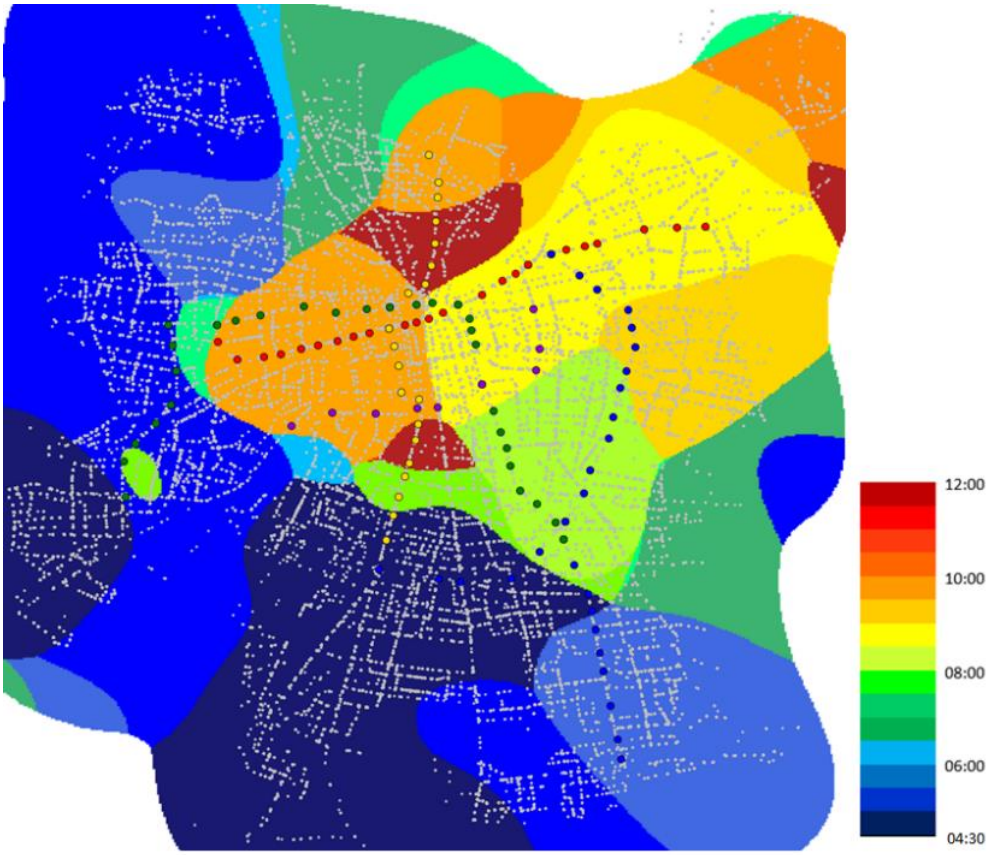


Figure 8-18: Classification des zones d'heure de première transaction – façon agrégée

Ensuite, la bande passante est définie sur une valeur inférieure: 0.1, de sorte que la Figure 8-19 a été obtenue.

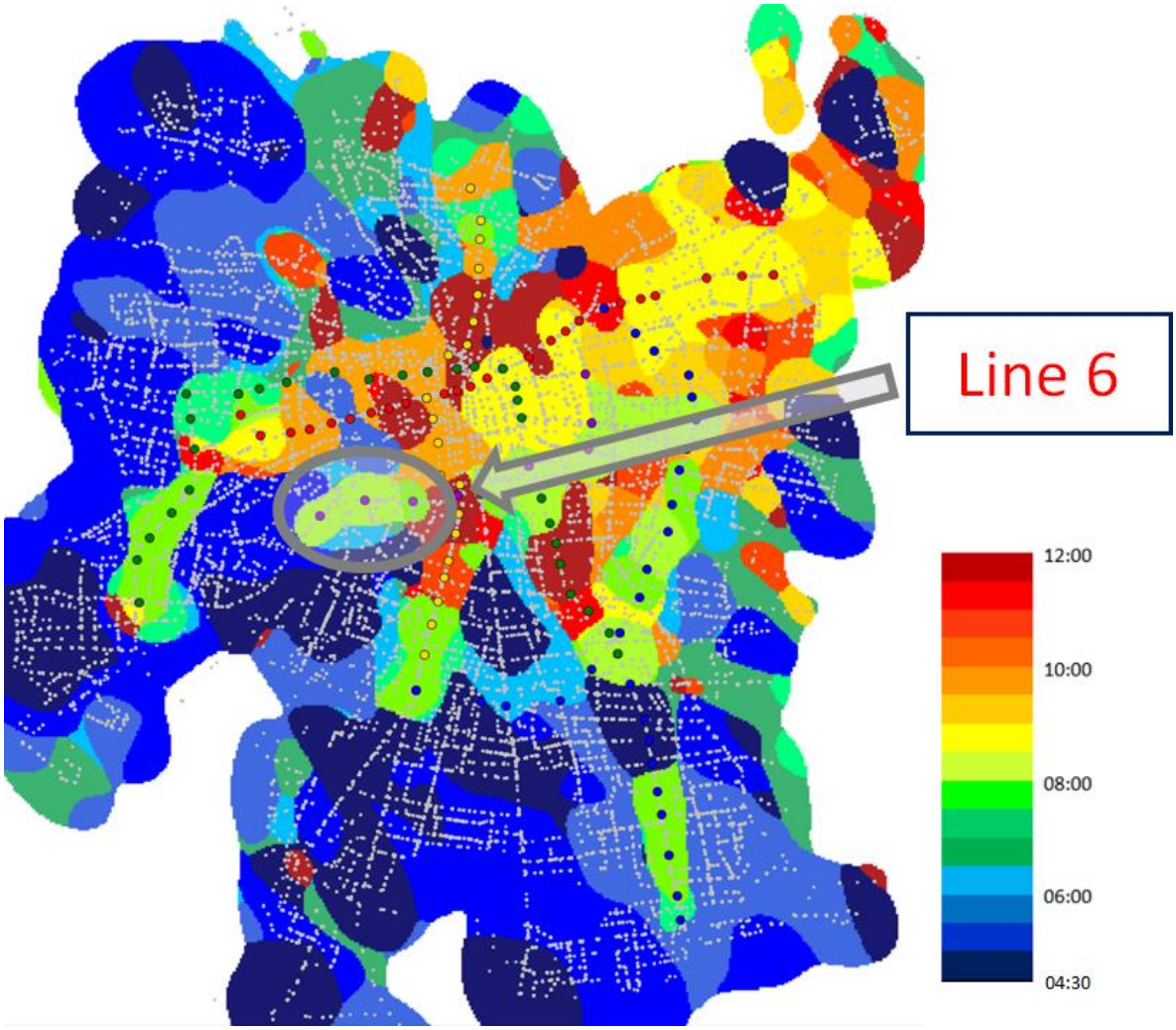


Figure 8-19: Classification des zones d'heure de première transaction – façon désagrégée

Sur la Figure 8-19, nous pouvons voir les comportements des usagers de façon plus désagrégée. En particulier, dans les zones pauvres, les usagers qui habitent proche du métro quittent leur domicile plus tard que ceux qui vivent loin du métro. Dans la région riche, les usagers qui habitent proches du métro quittent leur domicile plus tôt que ceux qui vivent loin du métro.

Une constatation importante est la suivante: lorsque l'on examine les trois stations de métro les plus à l'est de la ligne 6, on peut constater que l'heure de départ dominante pour les usagers situés à proximité de ces trois stations est 1 à 2 heures plus tard que celle qui habite loin de ces trois stations. Cela peut être perçu comme un impact positif de la mise en œuvre de la ligne 6.

8.2.3 Mesure du changement de comportements des usagers

Basé sur la Figure 8-19, de la même façon, nous pourrions faire une classification des zones basée sur les groupes de classification temporelle de STO. De cette façon, on peut faire une comparaison avant et après l'implémentation du BRT (Rapibus) et nous pourrions savoir le temps de première transaction des zones autour le BRT.

La première carte de la Figure 8-20 montre le cas avec une bande passante = 0.005. Le paramètre bande passante représente si une transaction a un impact sur un arrêt très loin et si nous augmentons le paramètre bande passante (0.01), on pourrait avoir un résultat plus agrégé, comme la deuxième carte de la Figure 8-20.

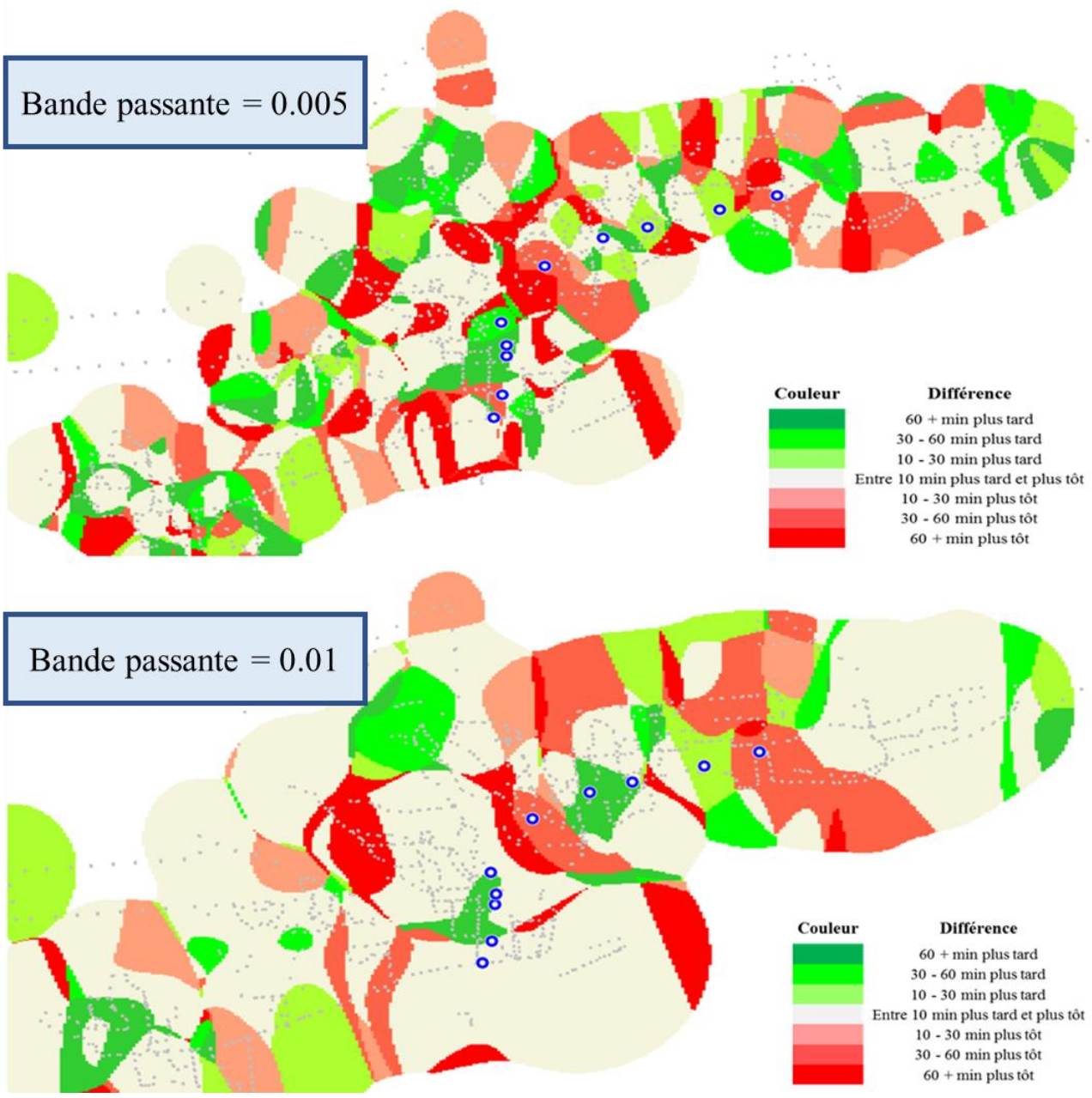


Figure 8-20: Différence d'heure de première transaction avant et après l'implémentation de Rapibus

Nous pouvons voir que pour les deux cas, les passagers faisant parties de certains arrêts voyagent plus tard. Cependant, beaucoup d'autres arrêts sont dans les zones où l'heure de première transaction dominant est plus tôt. Le changement d'heure de première transaction d'un jour ne change pas d'une manière significative après l'implantation du Rapibus.

CHAPITRE 9 DISCUSSION GÉNÉRALE

La problématique générale de la thèse est de la classification spatio-temporelle du comportement des usagers de transports en commun avec des données provenant d'un système de perception automatique, en vue de mieux connaître les demandes spatio-temporelles des voyageurs pour que les autorités des transports collectifs puissent offrir un meilleur service aux citoyens d'une ville. Pour cela, nous avons proposé 6 travaux inter-reliés, mais résolus un par un de manière indépendante dans les chapitres 4 à 8. La Figure 9-1 illustre les relations entre les objectifs, méthodes, articles et contributions de la thèse. Les sections 9.1 à 9.4 ont pour but d'expliquer ces relations indiquées à la Figure 9-1 un par un.

9.1 Relation entre les articles

Dans le premier article, nous avons résolu deux problèmes: le choix d'une meilleure métrique (les métriques classiques ne fonctionnent pas avec les séries temporelles) et le choix d'une méthode pertinente pour une petite taille des données. À cause du manque d'une métrique traditionnelle qui peux bien adapter la classification des séries temporelles, deux métriques, soit CCD et DTW, ont été comparées pour mesurer la dissimilarité des comportements des usagers. Avec cette méthode, des étapes de base ont été établies pour la classification temporelle. Il manquait cependant une méthode pour l'applications à de grandes tailles des données. Les données réelles ne pouvaient pas être testées à ce moment-là. En outre, cet article ne touchait que la classification temporelle, et les informations spatiales des usagers n'étaient pas traitées.

Dans le deuxième article, nous avons résolu le problème de l'application de la méthode de classification temporelle avec des données réelles. Dans cet article, une méthode d'échantillonnage a été proposée et des analyses sur la taille d'échantillonnage, le nombre de tirages et le nombre de groupes ont prouvé le bon fonctionnement de la méthode proposée. Elle a donc résolu la limitation du premier article. À ce moment-là, l'algorithme de classification temporelle a été développé, et il restait encore l'algorithme de classification spatiale à développer.

Dans le troisième article, nous avons résolu deux problèmes: la classification spatiale et la classification spatio-temporelle. Elles répondent à la préoccupation du deuxième article. Dans cet article, l'algorithme de DTW a été ajusté pour qu'il adapte des comportements spatiaux et spatio-temporelles, tandis que CCD n'étais pas disponible dans ce cas-ci. Avec l'algorithme temporel

développé, trois algorithmes ont été proposés, avec qui nous pouvons regrouper les comportements des usagers en groupes. Cependant, ces algorithmes ne s'appliquent qu'à une seule ville.

Dans le quatrième article, nous avons essayé d'appliquer l'algorithme à une autre ville. Les résultats démontrent que les algorithmes développés fonctionnent pour des villes différentes. Elles répondent à la préoccupation du troisième article. La métrique CCD avait été prouvée plus pertinente dans le premier article, pourtant, CCD et DTW ont été comparées encore une fois parce que nous voulons voir si DTW fonctionne pour certains cas lors de la classification temporelle. À la fin de cette étape, les comportements des usagers de carte à puce peuvent être bien regroupés en certains groupes. Cependant, il faudrait avoir une méthode de connaître les demandes des voyageurs à partir de ces groupes obtenus, en vue d'avoir les méthodes d'application pour améliorer le service de transports en commun.

Dans la partie ‘‘contributions supplémentaires’’ nous avons présenté une série de visualisations des résultats qui permettent de mieux connaître les demandes et de reconnaître la différence avant et après l'implémentation d'une infrastructure. Avec ces connaissances, des suggestions en vue d'améliorer le service de transports en commun ont été proposées. Elles répondent à la perspective du quatrième article.

9.2 Relation entre les objectifs

Après l'introduction de la relation entre six contributions (quatre articles + deux contributions supplémentaires), il y a aussi des relations solides entre les autres éléments de la thèse, par exemple, les relations de la méthode avec les contributions, etc.

Concernant les objectifs, au début de la recherche, l'objectif général est la classification spatio-temporelle des comportements des usagers en transport en commun. Pour développer les étapes de recherche, nous avons divisé cet objectif général en trois parties : classification temporelle, classification spatiale et classification spatio-temporelle. Ces trois objectifs ont été réalisés un par un. Après avoir développé les algorithmes de la classification spatio-temporelle, nous avons fait une extension d'objectifs. Les nouveaux objectifs (objectif 4 à 6) nous posent les questions : si l'algorithme développé avec les données d'une ville pourrait facilement s'appliquer à une autre ville? Et, comment interpréter les résultats de la classification spatio-temporelle? Cette extension d'objectifs nous permet de penser la compatibilité et l'utilisé des algorithmes proposés.

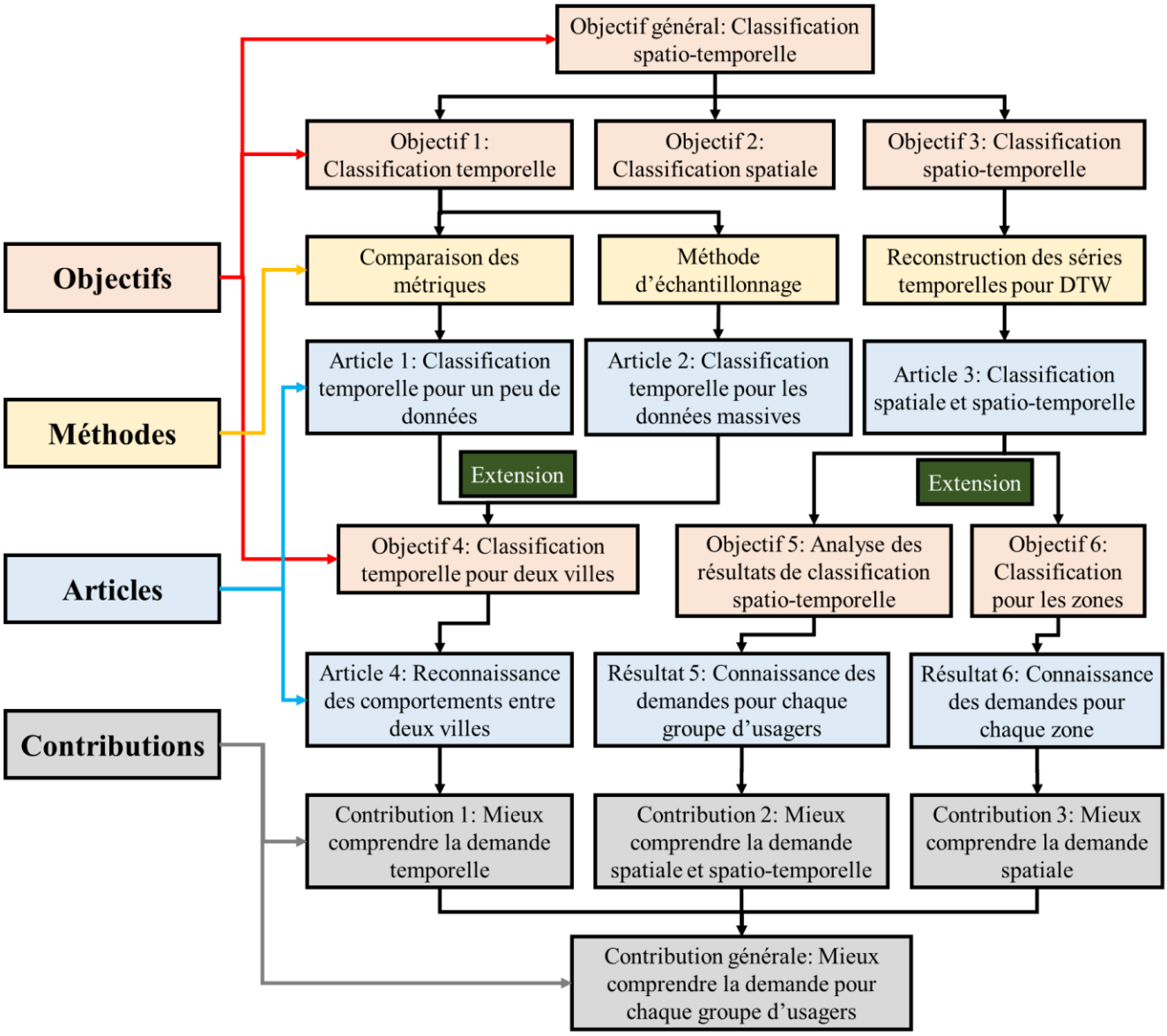


Figure 9-1: Relation entre les objectifs, méthodes, articles et contributions de la thèse

9.3 Relation entre les méthodes

Concernant les méthodes, la comparaison des métriques se fait dans la plupart algorithmes (sauf dans la contribution de la classification sur la base de densité), même si nous n'utilisons qu'une métrique dans chaque cas. Pour la classification temporelle, CCD est meilleure parce que CCD interprète mieux des comportements de délais de transactions par rapport à DTW, que ce soit une (article 1) ou deux villes (article 4). Pour la classification spatiale et spatio-temporelle, DTW est prouvée mieux que CCD, parce que DTW peut interpréter le changement de localisations des

usagers de transport en commun. Notons qu'il y a une nuance de la méthode DTW entre l'article 1 et 4. Dans l'article 1, l'implémentation de DTW est simplement une comparaison mathématique avec CCD, pourtant, dans l'article 4, l'implémentation est plutôt pratique, parce que dans ce cas-là, CCD compare le délai de transactions entre deux usagers, et le DTW compare le nombre de correspondance pendant une certaine période.

En outre, la méthode d'échantillonnage est développée dans l'article 2, mais cette méthode s'applique partout dans la thèse. Tous les algorithmes ont besoin de traiter les données massives de carte à puce (pour toute la ville pendant une certaine période). Avec les résultats de l'article 2, nous profitions d'un échantillon avec un nombre limité de profils de transactions des usagers.

9.4 Relation entre les contributions

Concernant les contributions, en suivant le processus du côté de la classification temporelle (Chapitre 4 et 5), l'algorithme de la classification temporelle pour une ville et l'algorithme de reconnaissance des comportements pour deux villes (Chapitre 7) sont développés, ceci nous permet de mieux comprendre la demande temporelle pour une ville, ou de mieux comprendre la différence de la demande temporelle entre deux villes.

En suivant le processus du côté de la classification spatio-temporelle, l'algorithme de la classification spatiale et spatio-temporelles, des méthodes de visualisation des résultats de la classification spatiale et spatio-temporelles sont développés (Chapitre 8). Cela nous permet non seulement de regrouper les comportements des usagers de cartes à puce en transport en commun en peu de groupes, mais aussi de trouver les différentes caractéristiques pour chaque cluster, en vue de mieux comprendre la différente demande spatiale et spatio-temporelle de chaque cluster.

En suivant le processus du côté de la classification spatio-temporelle (Chapitre 6), une branche est la classification des zones. Nous n'utilisons plus l'algorithme hiérarchique dans cette section, mais le résultat peut être considéré comme une classification spatiale, et ainsi il nous permet de mieux comprendre la demande spatiale des usagers.

En combinant ces trois sous-contributions, nous obtenons la contribution générale de la thèse. Ces des méthodes, ensembles, nous permettent de mieux comprendre les comportements spatio-temporels de chaque groupe d'usagers, pour que nous puissions leur fournir un service de transport en commun plus personnalisé.

Au total, nous avons donc résolus différents morceaux cohérents d'une même problématique, cependant, il resterait encore à améliorer l'application dans la planification des transports avec des stratégies plus fines, accélérer la vitesse de calcul en optimisant le processus de choix de la métrique et le calcul de DTW, prendre en compte les variables externes en vue de mieux comprendre comment les comportements de chaque groupe répondent aux changements de chaque variable externe.

CHAPITRE 10 CONCLUSION ET RECOMMANDATIONS

10.1 Contributions

D’abord, une analyse des profils quotidiens des usagers de cartes à puce dans les transports en commun nécessite une méthode permettant de classifier les séries temporelles. Bien qu’il existe des méthodes de mesure de similarité de séries temporelles standards, peu de recherches ont développé une méthode de classification de séries temporelles représentant un profil d’usager pendant une journée. En raison des limitations des métriques de distance traditionnelles, une méthode a été conçue en combinant une métrique de série temporelle et un clustering hiérarchique. Les résultats montrent que la corrélation croisée est mieux adaptée à la classification du profil temporel des données de cartes à puce du transport en commun. Le test utilisant les données d’une société de transport en commun de taille moyenne montre une classification claire des profils de transaction quotidiens des usagers de cartes. Cela permet une meilleure information sur chaque sous-groupe d'usagers, puis une meilleure adéquation entre l’offre et la demande du système de transports.

Ensuite, nous avons proposé une méthode en combinant la distance de corrélation croisée, la classification hiérarchique et une méthode d'échantillonnage afin de classifier les profils temporels des voyageurs en transport en commun utilisant des données de transactions par cartes à puce. L’application de cette méthode aux réseaux de transport en commun de la Société de transport de l’Outaouais a permis de classifier 333 745 jours-usagers. Nous avons également effectué une analyse de sensibilité sur les principaux paramètres de cette approche afin de tester sa validité avec l'ensemble de données, en analysant le nombre de groupes, la taille de l'échantillon et le nombre de tirages aléatoires pour l'échantillon.

Par la suite, une nouvelle méthodologie basée sur la déformation dynamique temporelle, la classification hiérarchique et la méthode d’échantillonnage est proposée pour classifier les comportements spatio-temporels des usagers de cartes à puce en transport en commun. Les résultats démontrent que les comportements peuvent être bien distingués. Sur la base des résultats, il est possible de suggérer des améliorations au service de transport en commun afin de mieux desservir les usagers de groupes spécifiques.

Ensuite, considérant la différence de comportements entre différentes villes, une recherche introduit une méthodologie basée sur les métriques de séries temporelles, la classification hiérarchique et la méthode d'échantillonnage, pour les reconnaître. Un exemple pédagogique est conçu pour obtenir une meilleure métrique, et l'application de l'algorithme aux données réelles permet de reconnaître 66,24% des comportements quotidiens des utilisateurs de cartes à puce et une précision totale de reconnaissance de 70,06% pour les deux villes. Il est plus facile de reconnaître les comportements de Santiago car l'exactitude de la reconnaissance atteint 75,41%. L'analyse des résultats montre que les comportements des usagers de Gatineau sont plus concentrés et plus tôt que ceux de Santiago.

À la fin, des analyses complémentaires sur les résultats de classification spatio-temporelle et une méthode de classification basée sur la densité sont présentés. Ces méthodes fournissent toutes des résultats de visualisation. Cela permet de mieux traduire et valoriser les données.

10.2 Limitations

En ce qui concerne les limitations et les perspectives. La première limite est sur le temps de calcul: (1) Le temps de calcul lors de l'essai de différents paramètres (décalage pour corrélation croisée et fenêtres pour DTW) est long si nous cherchons le meilleur paramètre parmi toutes les possibilités. Par exemple, pour calculer une matrice de dissimilarité entre 1000 usager-jour, le temps de calcul est environ 10 minutes. (2) L'algorithme de déformation dynamique temporelle est quadratique; par conséquent, le temps de calcul est long. Par exemple, pour calculer une matrice de dissimilarité entre 1000 usager-jour, le temps de calcul est environ 12 minutes.

La deuxième limite est le choix des paramètres pour traiter les problèmes de transport. Dans ce cas, le CCD est approprié car le délai du temps de transaction d'un utilisateur de carte à puce est comparable au paramètre "lag" du CCD. Toutefois, afin de traiter d'autres séries temporelles sur les problèmes de transports, d'autres distances devraient être essayées et appliquées.

Troisièmement, concernant la méthode d'échantillonnage, la principale limite de ce travail est que la méthode de détermination de l'efficacité de l'échantillonnage n'est basée que sur une ville (seulement avec les données de la STO). Les usagers des transports en commun de la STO peuvent avoir leurs propres caractéristiques et la taille de l'échantillon indiquée ici peut ne pas s'appliquer à des données provenant d'une autre ville. Par conséquent, nous pensons qu'en utilisant notre

méthodologie, les valeurs appropriées de N, S et D devraient également être déterminées pour d'autres expérimentations de données.

Quatrièmement, le critère de classification étant basé sur la distance, différents comportements peuvent rester dans le même groupe car leur différence, entre autres facteurs, n'est pas prise en compte (par exemple, l'objet du voyage n'est pas un critère de dissimilarité dans ce cas). Les données présentent d'autres limites: l'estimation des destinations peut ne pas être parfaite (elle n'a pas été validée), ce qui pourrait entraver les résultats de la méthode de classification.

Cinquièmement, concernant la méthode de reconnaissance des comportements entre des villes, un tiers des comportements des usagers ne sont pas reconnus en raison des mêmes comportements entre les deux villes. Une différence plus fine entre eux n'a pas été détectée. La distance de corrélation croisée est plus susceptible de distinguer la différence de temps de transaction des utilisateurs, ce qui a été fait dans l'article. La distance de déformation dynamique temporelle pourrait être utilisée pour distinguer la différence de durée de trajet, ce qui n'a pas été réalisée dans l'article.

10.3 Perspectives

Pour résoudre le problème du temps de calcul, un nouvel algorithme pourrait être développé afin d'éviter certains cas dans lesquels le calcul du CCD entre certains vecteurs est annulé en supposant que la distance de ces deux vecteurs est trop grande. Il faudrait aussi travailler à réduire le temps de calcul de la méthode de déformation temporelle dynamique.

Afin de traiter d'autres séries temporelles sur les problèmes de transports, d'autres distances peuvent être appliquées. Par exemple, la distance de transformation de Fourier basée sur l'analyse des fluctuations pourrait être testée pour expliquer les fluctuations des transactions, etc. Un autre exemple : une modification de la distance de déformation temporelle dynamique ou une autre métrique de série temporelle peut être appliquée, en vue de résoudre le problème de reconnaissance des comportements entre des villes. Par conséquent, une classification plus fine pour toute autre méthode doit être envisagée pour obtenir un meilleur résultat.

Dans le futur, des travaux sont proposés pour améliorer la méthode de classification spatiale et spatio-temporelle. Premièrement, à ce stade, nous jugeons la qualité de la classification en observant la trajectoire quotidienne et le tracé du chemin spatio-temporel. Une méthode

quantitative est nécessaire pour mesurer la dissimilarité entre chaque groupe afin de prouver que la méthode proposée fonctionne de manière mathématique.

Concernant la méthode de reconnaissance des comportements entre des villes, les résultats de l'algorithme pourraient permettre d'identifier l'influence des variables externes. Des facteurs socio-économiques pourraient expliquer l'appartenance à des grappes, par exemple pour aider à explorer l'impact du niveau de revenu sur le comportement des usagers de cartes à puce. Il est très intéressant de vérifier si l'impact du revenu est plus important que l'impact de la ville sur le comportement des utilisateurs. À ce moment, les attributs socio-démographiques des usagers de cartes à puce ne sont pas disponibles, mais une méthode pourrait être utilisée pour les déduire de l'emplacement du domicile estimé.

Concernant la performance de l'algorithme, pour l'instant, la visualisation des résultats est utilisée pour déterminer la performance en voyant les distributions des clusters. Cependant, nous espérons de développer une méthode en vue de mesurer la performance d'une façon quantitative. La métrique idéale sera la plus simple possible.

En outre, davantage de suggestions pourraient être faites aux autorités de transport en commun afin de mieux répondre à la demande des usagers dans un cluster spécifique. Les algorithmes offrent diverses applications potentielles, notamment pour optimiser les horaires et les trajectoires de transport en commun. Des méthodes appropriées peuvent être développées sur la base de la méthode proposée dans cette thèse.

BIBLIOGRAPHIE

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behavior from smart card data. *IFAC Proceedings Volumes*, *39*(3), 399-404.

Agard, B., Partovi Nia, V., & Trépanier, M. (2013). *Assessing public transport travel behaviour from smart card data with advanced data mining techniques.* Paper presented at the World Conference on Transport Research.

Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y., 2015. Time-series clustering–A decade review. Information Systems, 53, pp.16-38.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications* (Vol. 27): ACM.

Amaya, M., Cruzat, R., & Munizaga, M. A. (2018). Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. *Journal of Transport Geography*, *66*, 330-339.

Ankerst, M., Breunig, M., Kriegel, H. P., Ng, R. T., & Sander, J. (2008). Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD* (Vol. 99).

Arana, P., Cabezudo, S., & Peñalba, M. (2014). Influence of weather conditions on transit ridership: A statistical study using data from Smartcards. *Transportation research part A: policy and practice*, *59*, 1-12.

Asakura, Y., Iryo, T., Nakajima, Y., & Kusakabe, T. (2012). Estimation of behavioural change of railway passengers using smart card data. *Public Transport*, *4*(1), 1-16.

Bagchi, M., & White, P. (2004). What role for smart-card data from bus systems? *Municipal Engineer*, *157*(1), 39-46.

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006, June). A comparative study for outlier detection techniques in data mining. In *Cybernetics and Intelligent Systems, 2006 IEEE Conference on* (pp. 1-6). IEEE.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer Berlin Heidelberg.

Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In *KDD workshop* (Vol. 10, No. 16, pp. 359-370).

Black, P. E. (2006). Manhattan distance. Dictionary of algorithms and data structures. *http://xlinux. nist. gov/dads//*.

Bordagaray, M., dell'Olio, L., Ibeas, A., & Cecín, P. (2014). Modelling user perception of bus transit quality considering user and service heterogeneity. *Transportmetrica A: Transport Science*, 10(8), 705-721.

Bradley, P. S., Mangasarian, O. L., & Street, W. N. (1997). Clustering via concave minimization. In *Advances in neural information processing systems* (pp. 368-374). MIT Press.

Briand, A. S., Côme, E., Trépanier, M., & Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behavior using smart card data. *Transportation Research Part C: Emerging Technologies*, *79*, 274-289.

Brockwell, P. J., Davis, R. A., & Calder, M. V. (2002). *Introduction to time series and forecasting* (Vol. 2). New York: springer.

Bunker, J. M. (2018). High volume bus stop upstream average waiting time for working capacity and quality of service. *Public Transport*, *10*(2), 311-333.

Cats, O., Wang, Q., & Zhao, Y. (2015). Identification and classification of public transport activity centres in Stockholm using passenger flows data. *Journal of Transport Geography*, *48*, 10-22.

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

Chang, H., Park, D., Lee, S., Lee, H., & Baek, S. (2010). Dynamic multi-interval bus travel time prediction using bus transit data. *Transportmetrica*, 6(1), 19-38.

Chen, C. F., Chang, Y. H., & Chang, Y. W. (2009). Seasonal ARIMA forecasting of inbound air travel arrivals to Taiwan. *Transportmetrica*, *5*(2), 125-140.

Chen, Y., Mahmassani, H. S., & Hong, Z. (2015). Data mining and pattern matching for dynamic origin–destination demand estimation: Improving online network traffic prediction. *Transportation Research Record: Journal of the Transportation Research Board*, (2497), 23-34.

Chen, Y., Tang, S., Bouguila, N., Wang, C., Du, J. and Li, H., 2018. A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data. *Pattern Recognition*, 83, pp.375-387.

Chevalier, F., Le Blanc, J. (2013). La classification. Faculté des sciences économiques. Université de Rennes. *https://docplayer.fr/13650741-La-classification-2012-2013-fabien-chevalier-jerome-le-bellac.html*

Chu, K. A. (2015). Two-year worth of smart card transaction data–extracting longitudinal observations for the understanding of travel behaviour. Transportation Research Procedia, 11, 365-380.

Chu, K.A., & Chapleau, R. (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (2063), 63-72.

Cui, Q., Wei, Y. M., Li, Y., & Li, W. X. (2017). Exploring the differences in the airport competitiveness formation mechanism: evidence from 45 Chinese airports during 2010–2014. *Transportmetrica B: Transport Dynamics*, *5*(3), 325-341.

Das, S., & Pandit, D. (2015). Determination of level-of-service scale values for quantitative bus transit service attributes based on user perception. Transportmetrica A: Transport Science, 11(1), 1-21.

de Oña, R., & de Oña, J. (2015). Analysis of transit quality of service through segmentation and classification tree techniques. *Transportmetrica A: Transport Science*, *11*(5), 365-387.

Del Castillo, J. M., & Benitez, F. G. (2013). Determining a public transport satisfaction index from user surveys. *Transportmetrica A*: Transport Science, 9(8), 713-741.

Devillaine, F., Munizaga, M., & Trépanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, (2276), 48-55.

Deza, M. M., & Deza, E. (2009). Encyclopedia of distances. In *Encyclopedia of Distances* (pp. 1-583). Springer Berlin Heidelberg.

Diab, E. I., & El-Geneidy, A. M. (2013). Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. *Public Transport*, *4*(3), 209-231.

Ding, R., Wang, Q., Dang, Y., Fu, Q., Zhang, H. and Zhang, D., 2015. Yading: fast clustering of large-scale time series data. *Proceedings of the VLDB Endowment*, 8(5), pp.473-484.

E. I., & El-Geneidy, A. M. (2013). Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. *Public Transport*, *4*(3), 209-231.

El-Geneidy, A. M., & Surprenant-Legault, J. (2010). Limited-stop bus service: an evaluation of an implementation strategy. *Public Transport*, *2*(4), 291-306.

El Mahrsi, Mohamed & Côme, Etienne & Baro, Johanna & Oukhellou, Latifa. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. *The 3rd International Workshop on Urban Computing (UrbComp 2014)*

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of transport geography*, *49*, 26-38.

Faroqi, H., Mesbah, M. and Kim, J., 2019. Comparing Sequential with Combined Spatiotemporal Clustering of Passenger Trips in the Public Transit Network Using Smart Card Data. *Mathematical Problems in Engineering*, 2019.

Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., & Bento, C. (2014). *Catch me if you can: Predicting mobility patterns of public transport users.* Paper presented at the Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference.

Ghaemi, M. S., Agard, B., Nia, V. P., & Trépanier, M. (2015). Challenges in Spatial-Temporal Data Analysis Targeting Public Transport. *IFAC-PapersOnLine*, *48*(3), 442-447.

Ghaemi, M. S., Agard, B., Trépanier, M., & Partovi Nia, V. (2016). A Visual Segmentation Method for Temporal Smart Card Data. *Transportmetrica A: Transport Science*, *13*(5), 381-404.

Gilpin, S., Qian, B. and Davidson, I., 2013, October. Efficient hierarchical clustering of large high dimensional datasets. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (pp. 1371-1380). ACM.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software, 31*(7), 1-24.

Giraud, A. (2016). *Outils de visualisation de données de cartes à puce pour une société de transport collectif*. École Polytechnique de Montréal. Mémoire de maîtrise.

Gschwender, A., Munizaga, M., & Simonetti, C. (2016). Using smart card and GPS data for policy and planning: The case of Transantiago. *Research in Transportation Economics, 59*, 242-249.

Guha, S., Rastogi, R., & Shim, K. (1998). *CURE: an efficient clustering algorithm for large databases.* Paper presented at the ACM Sigmod Record.

Han, G. and Sohn, K. (2016). Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transportation Research Part B: Methodological*, 83: 121-135.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics, 151*(1-2), 304-318.

He, L. (2014). *Contributions à l'amélioration d'un algorithme d'estimation des destinations des déplacements unitaires dérivées des validations d'un système de perception par carte à puce.* École Polytechnique de Montréal. Mémoire de maîtrise.

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: Transport Science*, 1-20.

He, L., Agard, B., & Trépanier, M. (2018b). Space-time classification of public transit smart card users' activity locations from smart card data. *Conference on Advanced Systems in Public Transport and TransitData 2018*, paper 62.

He, L., & Trépanier, M. (2015). Estimating the Destination of Unlinked Trips in Transit Smart Card Fare Data. *Transportation Research Record: Journal of the Transportation Research Board*, (2535), 97-104.

He, L., Trépanier, M., Agard B., Munizaga M., Bustos B. (2019). Comparing transit user behavior of two cities using smart card data. *Annual Meeting of the Transportation Research Board*, Washington, DC. No. 19-05564.

He, L., Trépanier, M., Hickman, M., & Nassir, N. (2015). Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems (No. CIRRELT-2015-52). *CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport= Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation*.

He, L., Trépanier, M., & Agard, B. (2017). Evaluating the Impacts of a Bus-Rapid Transit on Users' Temporal Patterns Using Cross Correlation Distance and Sampled Hierarchical Clustering Applied to Smart Card Data. *Annual Meeting of the Transportation Research Board*, Washington, DC. No. 17-03711.

Huang, Z., & Ng, M. K. (2003). A note on k-modes clustering. *Journal of Classification*, *20*(2), 257-261.

Imaz, A., Habib, K. M. N., Shalaby, A., & Idris, A. O. (2015). Investigating the factors affecting transit user loyalty. *Public Transport*, *7*(1), 39-60.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Joh, C. H., Timmermans, H. J. P., & Arentze, T. A. (2006). Measuring and predicting adaptation behavior in multidimensional activity-travel patterns. *Transportmetrica*, *2*(2), 153-173.

Jou, R. C., Lam, S. H., & Wu, P. H. (2007). Acceptance tendencies and commuters' behavior under different road pricing schemes. *Transportmetrica*, 3(3), 213-230.

Kang, H.Y., Kim, J.S. and Li, K.J., 2009, March. Similarity measures for trajectory of moving objects in cellular space. In Proceedings of the 2009 ACM symposium on Applied Computing (pp. 1325-1330). ACM.

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, *32*(8), 68-75.

Ketabi, R., Alipour, B. and Helmy, A., 2018, November. Playing with matches: vehicular mobility through analysis of trip similarity and matching. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 544-547). ACM.

Kieu, L. M., Bhaskar, A., & Chung, E. (2014). Transit passenger segmentation using travel regularity mined from Smart Card transactions data. In *Transportation Research Board 93rd Annual Meeting*, 12-16 January 2014, Washington, D.C.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 231-240.

Kruskall, J. B. (1983). The symmetric time warping algorithm: From continuous to discrete. In *Time warps, string edits and macromolecules*. Addison-Wesley.

Kurauchi, F., & Schmöcker, J. D. (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. *Transportation Research Part C: Emerging Technologies*, *46*, 179-191.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics*, *24*(5), 719-720.

Langlois, G. G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, *64*, 1-16.

Lathia, N., Froehlich, J., & Capra, L. (2010, December). Mining public transport usage for personalised intelligent transport systems. In *2010 IEEE International Conference on Data Mining* (pp. 887-892). IEEE.

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, *6*(1-2), 1-20.

Lee, W. H., Tseng, S. S., Shieh, J. L., & Chen, H. H. (2011). Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1047-1056.

Legara, E. F. T., & Monterola, C. P. (2018). Inferring passenger types from commuter eigentravel matrices. *Transportmetrica B: transport dynamics, 6*(3), 230-250.

Lhermitte, S., Verbesselt, J., Verstraeten, W. W., & Coppin, P. (2011). A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, *115*(12), 3129-3152.

Li, H., & Chen, X. (2016). Unifying Time Reference of Smart Card Data Using Dynamic Time Warping. *Procedia Engineering*, *137*, 513-522.

Li, Y. T., Schmöcker, J. D., & Fujii, S. (2015). Demand adaptation towards new transport modes: the case of high-speed rail in Taiwan. *Transportmetrica B*: Transport Dynamics, 3(1), 27-43.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857-1874.

Liao, W. K., Liu, Y., & Choudhary, A. (2004, April). A grid-based clustering algorithm using adaptive mesh refinement. In *7th workshop on mining scientific and engineering datasets of SIAM international conference on data mining* (Vol. 22, pp. 61-69).

Liu, Y., & Cheng, T. (2018). Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science*, 1-28.

Ma, X., Wang, Y., McCormack, E., & Wang, Y. (2016). Understanding Freight Trip-Chaining Behavior Using a Spatial Data-Mining Approach with GPS Data. *Transportation Research Record: Journal of the Transportation Research Board*, (2596), 44-54.

Ma, X., Wu, Y. J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies, 36*, 1-12.

Meyer, D., Buchta, C., & Meyer, M. D. (2017). Package 'proxy'.

Mirkes, E. M. K-means and K-medoids Applet, University of Leicester, 2011. *Online: http://www. math. le. ac. uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids. html (Dec 2013)*.

Mohamed, K., Côme, E., Baro, J., & Oukhellou, L. (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data. *UrbComp,(Seattle, WA, USA)*.

Mohamed, K., Côme, E., Oukhellou, L., & Verleysen, M. (2017). Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems*, *18*(3), 712-728.

Morency, C., Trépanier, M., & Agard, B. (2006). Analysing the variability of transit users behaviour with smart card data. In *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE* (pp. 44-49). IEEE.

Morency, C., Trepanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, *14*(3), 193-203.

Morency, C., Trépanier, M., Agard, B., Martin, B., & Quashie, J. (2007, September). Car sharing system: what transaction datasets reveal on users' behaviors. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE* (pp. 284-289). IEEE.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance Measures for Time Series in R: The TSdist Package. *R JOURNAL*, 8(2), 451-459.

Naboulsi, D., Fiore, M., Ribot, S. and Stanica, R., 2015. Large-scale mobile traffic analysis: a survey. IEEE Communications Surveys & Tutorials, 18(1), pp.124-161.

Nishiuchi, H., Kobayashi, Y., Todoroki, T., & Kawasaki, T. (2018). Impact analysis of reductions in tram services in rural areas in Japan using smart card data. *Public Transport, 10*(2), 291-309.

Nishiuchi, H., King, J., & Todoroki, T. (2013). Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. *International Journal of Intelligent Transportation Systems Research*, *11*(1), 1-10.

Nuzzolo, A., & Comi, A. (2016). Advanced public transport and intelligent transport systems: new modelling challenges. *Transportmetrica A: Transport Science*, *12*(8), 674-699.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications*, *36*(2), 3336-3341.

Parzani, C., Leclercq, L., Benoumechiara, N., & Villegas, D. (2017). Clustering route choices methodology for network performance analysis. *Transportmetrica B: Transport Dynamics*, *5*(2), 191-210.

Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557-568.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer US.

Sedgwick, P. (2012). Pearson's correlation coefficient. *Bmj*, *345*(7).

Shi, X., and Lin. H. (2014). The Analysis of Bus Commuters' Travel Characteristics Using Smart Card Data: The Case of Shenzhen, China. Presented at *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., (No. 14-2571), 2014.

Spurr, T., Chapleau, R., & Piché, D. (2014). *Discovery and Partial Correction of Travel Survey Bias Using Subway Smart Card Transactions.* Paper presented at the Transportation Research Board 93rd Annual Meeting. 2405, 56-67.

Srimani, P., Mahesh, S., & Bhyratae, S. A. (2013). *Improvement of Traditional K-means algorithm through the regulation of distance metric parameters.* Paper presented at the Intelligent Systems and Control (ISCO), 2013 7th International Conference

Subbiah, K. (2011). *Partitioning Methods in Data Mining*.

Sun, Y., Shi, J., & Schonfeld, P. M. (2016). Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: a case study of Shanghai Metro. Public Transport, 8(3), 341-363.

Tao, S., Rohde, D., & Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *Journal of Transport Geography*, *41*, 21-36.

Ten Holt, G. A., Reinders, M. J., & Hendriks, E. A. (2007, June). Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging. The Netherlands*. (Vol. 300, pp. 23-32).

Tranchant, N. (2005). Analyse des déplacements d'usagers à partir de données de cartes à puce.

Trépanier, M., Barj, S., Dufour, C., & Poilpré, R. (2004). Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain. *Congrès de l'Association des transports du Canada*. Québec. 10–14.

Trépanier, M., & Chapleau, R. (2001). Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination. *Canadian Journal of Civil Engineering, 28*(1), 48-58.

Trépanier, M., & Morency, C. (2010). *Assessing transit loyalty with smart card data*. Paper presented at the 12th World Conference on Transport Research, Lisbon, Portugal.

Trépanier M., Morency, C., & Agard, B. (2009). Calculation of transit performance measures using smartcard data. *Journal of Public Transportation, 12*(1), 5.

Trépanier, Morency, C., Agard, B., Descoimps, E., & Marcotte, J. (2012). *Using smart card data to assess the impacts of weather on public transport user behavior*. Paper presented at the Conference on Advanced Systems for Public Transit-CASPT12, Santiago, Chile.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, *11*(1), 1-14.

Vicente, P., & Reis, E. (2016). Profiling public transport users through perceptions about public transport providers and satisfaction with the public transport service. *Public Transport*, *8*(3), 387-403.

Viggiano, C., Koutsopoulos, H. N., Wilson, N. H., & Attanucci, J. (2017). Journey-based characterization of multi-modal public transportation networks. *Public Transport*, *9*(1-2), 437-461.

Wang, W., Yang, J., & Muntz, R. (1997, August). STING: A statistical information grid approach to spatial data mining. In *VLDB* (Vol. 97, pp. 186-195).

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Yang, C., Yan, F., & Ukkusuri, S. V. (2018). Unraveling traveler mobility patterns and predicting user behavior in the Shenzhen metro system. *Transportmetrica A: Transport Science*, *14*(7), 576-597.

Yap, M., Cats, O., & van Arem, B. (2018). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science*, 1-20.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, *17*(10), 977-987.

Yuan, Y. and Raubal, M., 2014. Measuring similarity of mobile phone user trajectories–a Spatio-temporal Edit Distance method. International Journal of Geographical Information Science, 28(3), pp.496-520.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996, June). BIRCH: an efficient data clustering method for very large databases. In *ACM Sigmod Record* (Vol. 25, No. 2, pp. 103-114). ACM.

Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *6*(3), 29.

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. *Transportation Research Part C: Emerging Technologies*, *75*, 17-29.