Classification spatio-temporelle des localisations d'activité des utilisateurs de cartes à puce en transport en commun

LI HE[1,2,3], MARTIN TRÉPANIER[1,2,3], BRUNO AGARD[1,2,3]

[1] École Polytechnique de Montréal
Département de mathématiques et génie industriel, CP 6079, succursale Centre-Ville, Montréal, Québec, Canada
he.li@polymtl.ca, mtrepanier@polymtl.ca, bruno.agard@polymtl.ca

[2] Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT)
[3] Laboratoire en Intelligence des Données

***Résumé* – Les données des cartes à puce du système de transport en commun sont utiles pour comprendre le comportement des utilisateurs du transport en commun. De nombreuses recherches pertinentes ont été menées concernant : (1) l'utilisation de données de carte à puce, (2) les techniques de fouille de données et (3) l'utilisation de la fouille de données avec des données de carte à puce. Dans ces recherches, la classification des comportements des utilisateurs est basée sur des déplacements dans lesquels les classifications temporelles et spatiales sont considérées comme des processus séparés. Dans le présent article, nous développons une méthode, basée sur les comportements quotidiens des utilisateurs, 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 déformation temporelle dynamique (dynamic time warping ou DTW), sur la classification hiérarchique et sur l'échantillonnage. Un graphique spatio-temporel en 3 dimensions montre l'efficacité de l'algorithme.**

***Mots clés* – Transport en commun · Données de cartes à puce · Déformation temporelle dynamique · Classification spatio-temporelle · Lieux d'activité**
***Keywords* – Public transit · Smart card data · Dynamic time warping · Spatio-temporal classification · Activity locations**

1 INTRODUCTION

Les données provenant des systèmes de collecte de tarifs par carte à puce sont très utiles pour les planificateurs de transport en commun. Ces données aident à mieux connaître le comportement des utilisateurs de carte à puce dans le réseau de transport. Cette connaissance est ensuite utile pour améliorer le service de transport en commun [Pelletier et al., 2011]. De nombreux efforts ont été déployés en utilisant la fouille de données pour classifier les transactions des utilisateurs. Certaines méthodologies ont notamment été proposées pour classifier les comportements temporels et spatiaux des utilisateurs de cartes à puce en utilisant diverses métriques de distance et méthodes de classification. Dans cet article, nous présentons une méthode permettant de classifier les utilisateurs de transport en commun en fonction de l'heure et du lieu de leurs déplacements pendant la journée.

Cet article sera organisé comme suit. Dans la prochaine partie, la revue de la littérature se concentrera sur les travaux pertinents, principalement les méthodes de fouille de données qui seront utilisées. Ensuite, la problématique et l'objectif de cet article seront présentés. Pour résoudre le problème de la classification des comportements spatio-temporels, une méthodologie est développée dans la partie 4. Un cas d'étude sera présenté en section 5. Les résultats et leurs analyses seront présentées dans la section 6. Finalement, une conclusion qui contient la contribution, les limitations et perspectives sera présentée.

2 REVUE DE LITTERATURE

2.1 Utilisation des données de la carte à puce

Au fil des ans, plusieurs travaux ont été réalisés avec les données de cartes à puce dans les transports en commun. En termes de préparation et de complétion des données, des articles pertinents introduisent la description des données de carte à puce [Trépanier et al., 2004], l'enrichissement des données, y compris une méthode d'estimation de la destination [Trépanier et al., 2007]. Ces travaux sont complétés, en utilisant l'estimation (de la densité) du noyau [He et Trépanier, 2015], et la précision de la méthode est améliorée [He et al., 2015]). D'autres méthodes ont été développées, basées sur la détection de transferts [Chu et Chapleau, 2008] et des inférences du but des déplacements [Lee & Hickman, 2013]. Ces recherches constituent la base de l'analyse des comportements des usagers du transport en commun.

En termes de détection du comportement des utilisateurs de cartes à puce, les données peuvent être utilisées pour catégoriser les utilisateurs par informations temporelles (heure de la transaction, durée du déplacement, délai, etc.) [Morency et al., 2007 ; Bunker et al., 2018], par informations géographiques (origine-destination, trajectoire, etc.) [Shi et al., 2014], par choix de mode [Kurauchi et al., 2014; Viggiano et al., 2017] et également par la personnalité des passagers (comme la fidélité de l'utilisateur) [Imaz et al., 2015]), par la caractérisation des réseaux [Sun et al., 2016], par l'analyse des facteurs externes qui influent sur l'utilisation du réseau [Briand et al., 2017] et par la prévision dans l'utilisation des données [Ceapa et al., 2012]. Les méthodes aident notamment à estimer et à comprendre le changement de comportement [Asakura et al., 2012] pour diverses stratégies d'amélioration de la fiabilité des services de transport en commun [Diab & El-Geneidy, 2013]. Ces recherches visent à comprendre l'analyse des comportements

des usagers du transport en commun, et contribuent à améliorer le service de transport en commun. Par exemple, des travaux ont été menés sur l'optimisation des horaires des services de transport en commun [Nishiuchi et al., 2018], sur l'optimisation des arrêts de bus [El-Geneidy & Surprenant-Legault, 2010], etc. Le nombre de transactions par carte à puce pouvant atteindre plusieurs millions pour une ville typique, il est pertinent d'utiliser des techniques de fouille adaptées pour pouvoir analyser les données de manière significative.

2.2 Techniques de fouille de données

De nombreuses techniques de fouille de données peuvent être utilisées. En ce qui concerne les méthodes de segmentation, d'un côté, il existe un choix de méthodes parmi les algorithmes de partition [Chevalier et al., 2013], les algorithmes hiérarchiques [Rokach et al., 2005] et les algorithmes basés sur la densité (Kriegel et al., 2011). De l'autre côté, plusieurs métriques peuvent être utilisées pour évaluer la dissimilarité de deux vecteurs, notamment la distance euclidienne [Deza et al., 2009], la distance de Manhattan [Black, 2006], la distance de corrélation croisée [Mori et al., 2016], et la distance de déformation temporelle dynamique [Giorgino, 2009].

La figure 1 illustre la méthode de déformation temporelle dynamique. La déformation temporelle dynamique est une technique populaire de comparaison de séries temporelles, fournissant à la fois une mesure de distance insensible à la compression et aux étirements locaux qui déforme de manière optimale l'une des deux séries d'entrées sur l'autre [Giorgino, 2009]. Nous pouvons définir formellement le problème de déformation temporelle dynamique comme la minimisation dynamique sur des trajets de déformation potentiels en fonction de la distance cumulée pour chaque trajet, où *d* est une mesure de distance entre deux éléments de série temporelle. On ajuste le dernier moment de la série temporelle *B* au dernier moment de la série temporelle *A* afin que la distance cumulée entre *A* et *B* soit minimale (Fig. 1 (a)).

Pour obtenir une distance cumulée minimale, un point d'une série temporelle peut être renvoyé au point temporel suivant (moment). Par exemple, le point de grille (*M* -1, *N* -1) peut être renvoyé à (*M*, *N* -1), (*M* - 1, *N*), (*M*, *N*) pour calculer chaque distance (Fig. 1 (b)). Il faut alors calculer tous les chemins possibles des points de la grille (1, 1) à (6, 6), et trouver le chemin avec la distance cumulée minimale. Dans cette grille ci-dessus, la distance de DTW est 7 (Fig. 1 (c)).

2.3 Utilisation de la fouille de données avec les données de carte à puce

Une question qui intéresse beaucoup les chercheurs du transport en commun consiste à diviser les passagers en groupes en fonction de leurs comportements de déplacements. La technique classique de fouille de données (k-means et classification hiérarchique) a été utilisée pour classifier le comportement général des utilisateurs sur une période de 12 semaines [Agard et al., 2006]. D'autres travaux ont été réalisés sur la base de k-means [Morency et al., 2006], de réseaux de neurones [Ma et al., 2013] et de DBSCAN (Density-based spatial clustering of applications with noise) [Kieu et al., 2014] afin d'identifier les passagers réguliers ou de proposer une classification en fonction de leurs comportements. De plus, une méthode de classification peut également être développée en vue d'analyser la qualité du niveau de service de transport collectif [de Oña et al., 2015].

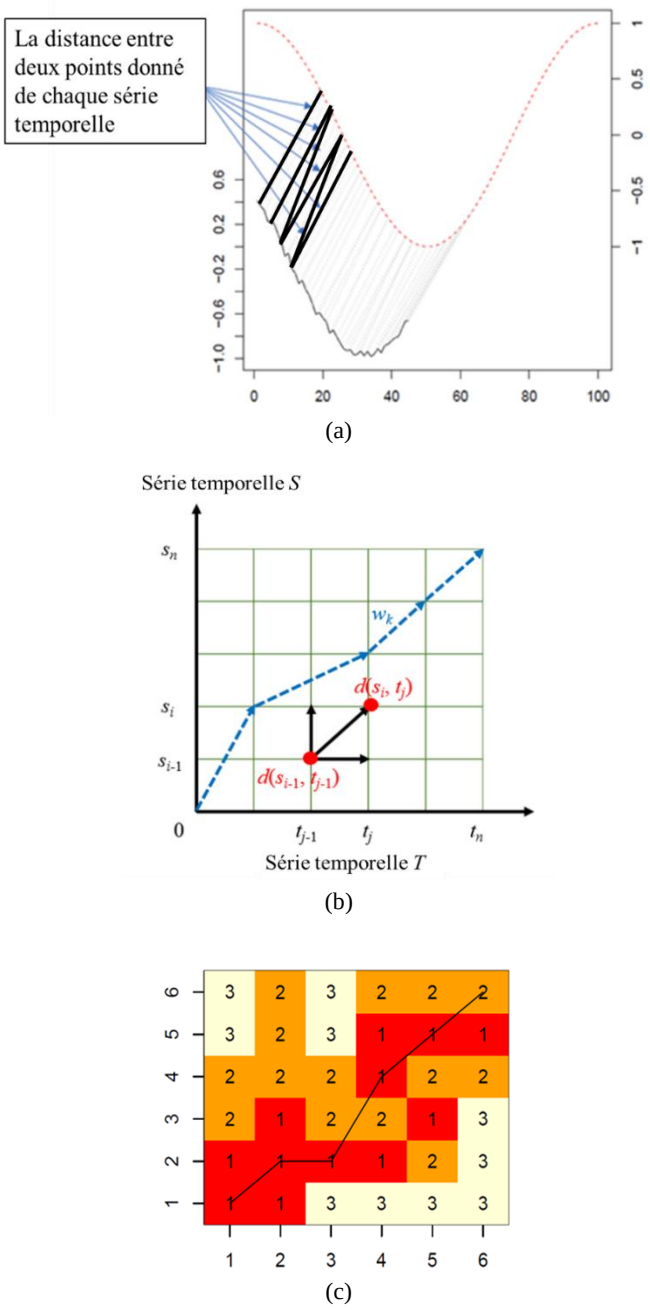


Fig.1 Méthode de déformation temporelle dynamique

Il est également très intéressant d'analyser les comportements des utilisateurs temporellement et spatialement, en se basant sur la méthode de fouille de données temporelle [Ghaemi et al., 2017] et sur la méthode de fouille de données spatiale [Ghaemi et al., 2015]. Dans ce cas-ci, les comportements temporels et spatiaux de l'utilisateur de la carte de transport en commun ont été analysés séparément.

Finalement, pour vérifier l'efficacité des algorithmes de classification spatio-temporelle, un tracé spatio-temporel en 3-dimensions [Farber et al., 2015] permet de montrer le profil de chaque groupe. Comme le montre la figure 2, ce tracé 3D montre l'emplacement d'un utilisateur (ou d'un groupe d'utilisateurs) pendant une journée. Cela permet non seulement de montrer la différence entre les comportements des utilisateurs du transport en commun, mais aussi celle des groupes d'utilisateurs.

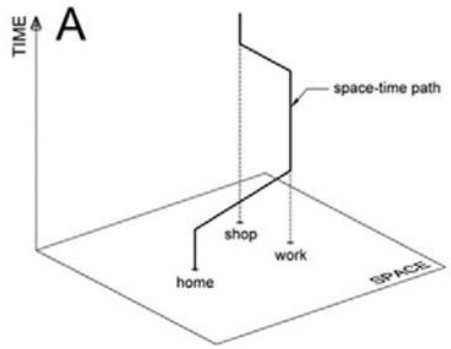


Fig.2 Exemple de tracé spatio-temporel [Farber et al., 2015]

2.4 Limitation des méthodes actuelles

Les documents actuels présentent des méthodologies pertinentes pour la détection des comportements des utilisateurs de cartes à puce, la diversité des méthodes de classification et l'application des méthodes d'exploration de données sur les données de cartes à puce. Toutefois, comme les comportements des utilisateurs peuvent être considérés comme des séries temporelles, peu d'articles présentent une classification des séries temporelles afin de découvrir les comportements temporels et spatiaux des passagers. Comparée à l'analyse dans laquelle les comportements des utilisateurs sont traités séparément à chaque instant, la classification des séries temporelles devrait contenir plus d'informations sur les caractéristiques de l'utilisateur, cependant, la classification des séries chronologiques est un problème particulier en raison des limites de la méthode de classification classique. La recherche actuelle est encore souvent basée sur la liste des transactions de chaque utilisateur de carte à puce au lieu de la série temporelle de comportements quotidiens. Par exemple, lors de la classification à l'aide de k-means, l'algorithme considère uniquement la valeur des éléments vectoriels et non la position de ces éléments dans le vecteur. L'intérêt pour les planificateurs en transport est de considérer l'heure du jour dans la séquence d'embarquement. Pour résoudre ce problème, une méthode de classification temporelle a été développée dans [He et al., 2018], basée sur la métrique de distance de corrélation croisée (cross correlation distance).

Même si une classification temporelle permet de classifier les comportements temporels des utilisateurs en groupes, peu d'articles présentent une méthode pertinente en vue de classifier les comportements quotidiens spatiaux ou spatio-temporels des utilisateurs de cartes à puce en transport public. Dans cet article, ces problèmes seront résolus en reconstruisant la distance de déformation temporelle dynamique et en appliquant une méthode de classification hiérarchique et une méthode d'échantillonnage. Le résultat permettra aux autorités de transport d'offrir un meilleur service pour répondre aux besoins quotidiens des passagers.

3 PROBLEMATIQUE ET OBJECTIF

3.1 Problématique

En raison de sa nature, un chemin de voyage de transport en commun est caractérisé à la fois par le moment de la journée où les activités d'embarquement ont eu lieu et par le lieu où elles se sont passées. La manière la plus prometteuse pour classifier les utilisateurs serait de considérer simultanément l'espace et le temps. Dans cet article, les comportements des utilisateurs seront traités comme une série temporelle de localisations spatiales. La technique de classification tiendra donc compte de l'espace et du temps en même temps, en utilisant une métrique de dissimilarité spécifique.

Dans nos travaux précédents, la distance de corrélation croisée et la distance de déformation temporelle dynamique ont été intégrées à la classification hiérarchique pour créer des méthodes de classification de séries temporelles [He and al., 2017], maintenant, nous proposons d'intégrer la dimension spatiale.

3.2 Objectif

L'objectif de cet article est de proposer une méthodologie permettant de classifier les comportements spatio-temporels des utilisateurs à l'aide d'algorithmes de classification et de métriques de distance pertinents. Le comportement de l'usager est ici représenté par la séquence des emplacements des arrêts de bus chaque heure. Pour illustrer la méthode, la figure 3 présente un exemple des comportements quotidiens de 3 utilisateurs :

• le premier utilisateur quitte son domicile pour arriver à l'école à 06:30 et part pour son domicile à 16:00;

• le deuxième utilisateur quitte son domicile pour arriver au travail à 07:30 et rentre chez lui à 18:00.

• le troisième part également pour arriver au travail à 06:30, mais avant de rentrer chez lui à 18:00, l'utilisateur s'est rendu au supermarché à 16:00.

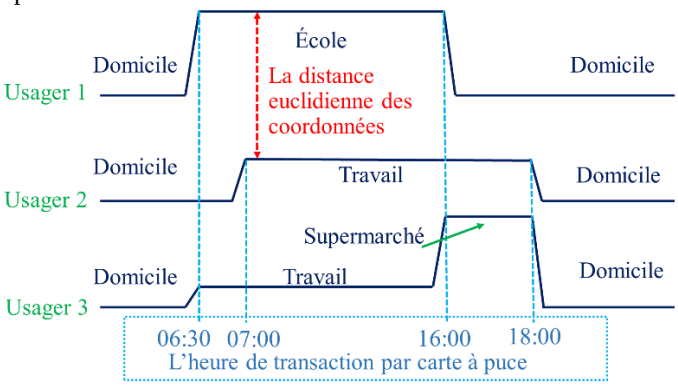


Fig. 3 Petit exemple montrant 3 comportements d'utilisateurs à classer

L'objectif de la classification spatio-temporelle est de classifier ces profils quotidiens en termes de temps et de lieu simultanément, afin de les séparer en différents groupes.

Dans la classification spatio-temporelle, lors de la mesure de la dissimilarité du profil de deux utilisateurs, on considère non seulement le temps de transaction par carte à puce, mais également la distance réelle entre les arrêts de bus, servant de proxy pour la localisation de l'utilisateur pendant la journée (distance euclidienne entre école de l'utilisateur 1 et travail de l'utilisateur 2, par exemple). L'objectif est de disposer d'une mesure de dissimilarité prenant en compte les deux dimensions (espace et temps) afin de procéder à classifier.

4 METHODOLOGIE

La Fig. 4 montre la méthodologie développée pour mettre en œuvre la métrique de dissimilarité proposée et les méthodes de classification. La figure montre le nombre d'enregistrements de données utilisés dans l'étude de cas, décrits dans la section suivante. La méthodologie comporte 3 grandes sections décrites ci-après.

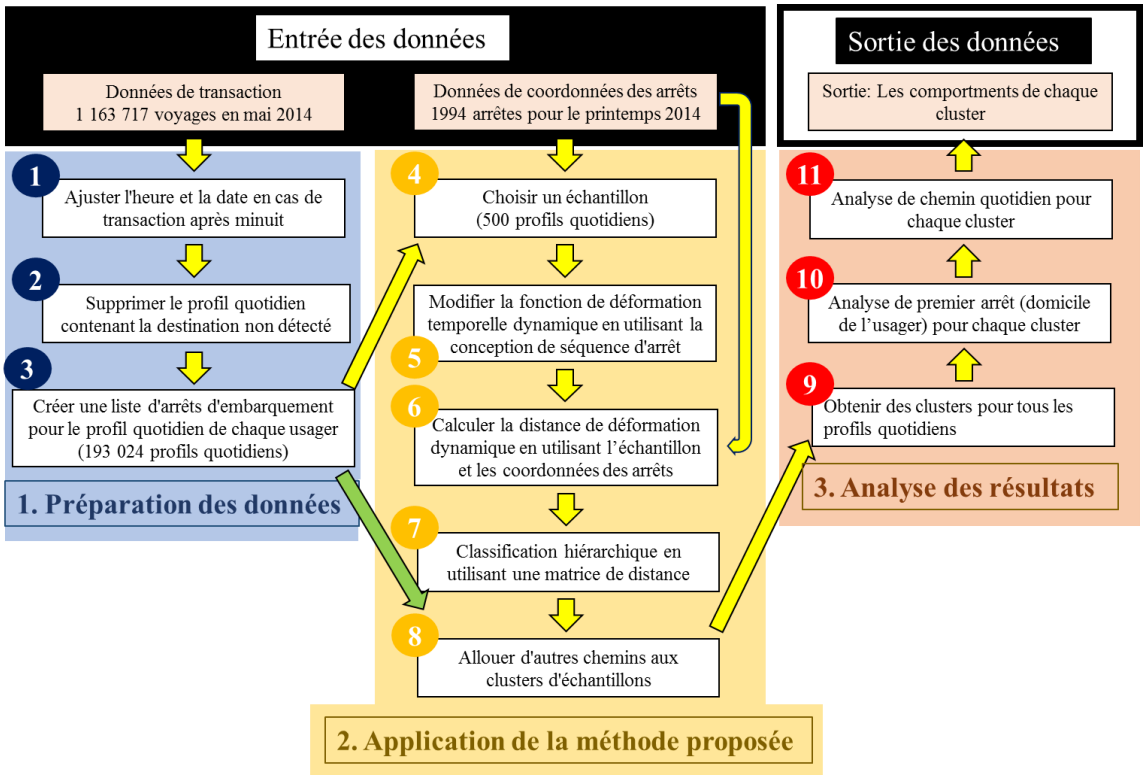


Fig. 4 Méthode proposée

4.1 Préparation des données

Les transactions par carte à puce sont formatées et prétraitées. Les trajectoires qui ont eu lieu après minuit sont ajustées de sorte que la trajectoire reste dans le même trajet utilisateur, en utilisant un système de plus de 24 heures (étape 1 de la Fig. 4). Par exemple, un voyage survenu à 1 heure du matin le lendemain est remplacé par la 25^{e} heure le même jour.

Puis, pour la classification des trajectoires, nous devons utiliser les destinations des transactions par carte à puce. Les données de carte à puce utilisées dans ce cas nécessitent des informations de débarquement (destinations). Les destinations ont été estimées à l'aide de la méthode proposée par [He et Trépanier, 2015]. Les transactions pour lesquelles il n'est pas possible d'estimer de destinations sont enlevées (étape 2, Fig.4).

Lors de la dernière étape de préparation des données, une liste des arrêts de bus est créée pour chaque carte et pour chaque jour. Elle présente la séquence d'heures des arrêts où se trouve l'utilisateur pendant la journée (étape 3 Fig 4). La Fig. 5 présente trois méthodes pour construire la série temporelle dans ce cas. L'idée principale est de relier tous les arrêts à un moment donné, jusqu'à la fin de la journée. Les balles 2 et 3 de la figure présentent des méthodes choisies dans cet article pour construire des séries chronologiques pour la classification spatiale et spatio-temporelle. Le Tableau 1 présente les caractéristiques de chaque approche, nous utilisons les deux dernières.

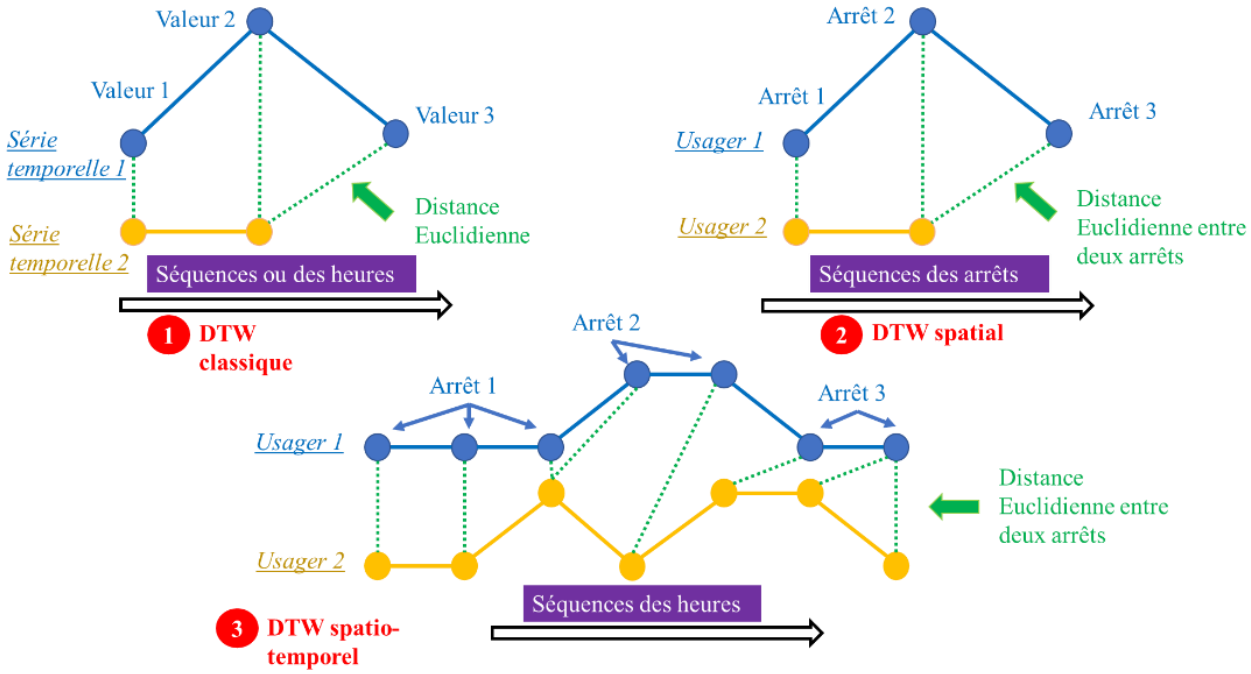


Fig.5 Comparaison des trois méthodes DTW (déformation temporelle dynamique)

Tableau 1 Conception de trois types de DTW

Conception	DTW Classique	DTW Spatial	DTW Spatio-temporelle
Objet à traiter	Séries temporelles	Trajectoires d'utilisateur dans le profil d'une journée (Séquence d'arrêts)	Localisation-heure d'utilisateur dans le profil d'une journée (Séquence d'arrêts à une heure donnée (moment))
Point	Point du temps (moment)	Arrêt	Arrêt à chaque moment donné
Séquence (Série temporelle)	Série temporelle	Séquence d'arrêt (inégale par rapport au temps)	Séquence d'arrêt (inégale par rapport au temps)
Distance entre le point de la grille	Peut être défini comme distance euclidienne, distance de Manhattan, etc.	Distance entre deux arrêts donnés (uniquement distance euclidienne)	Distance entre deux arrêts donnés (uniquement distance euclidienne)
Distance euclidienne	Au sens du temps (X : temps ; Y : valeur en x)	Au sens de la géographie (X: longitude; Y: latitude)	Au sens de la géographie (X: longitude; Y: latitude)

4.2 Application de la méthode proposée

La classification de plus de cent mille profils d'utilisateurs quotidiens est un processus qui prend du temps. Le temps de calcul (lorsque cela est réalisable) est beaucoup trop long et la quantité de mémoire informatique nécessaire va exploser en raison de la taille de la matrice de dissimilarité. Pour faire la classification, nous proposons d'utiliser une approche d'échantillonnage. Nos tests ont montré qu'un échantillon de 500 profils quotidiens (plus de 100 000) est suffisant ici. Cette section explique les étapes 4 à 8 de la méthodologie.

La figure 6 montre le processus d'échantillonnage global [He et al., 2019]. Au début, toutes les observations sont présentées à la Fig. 6 (a). Les points rouges sur la figure 6 (b) sont les points choisis au hasard. Ensuite, nous appliquons des algorithmes de déformation temporelle dynamique et de classification hiérarchique à ces points d'échantillonnage. La Fig. 6 (c) présente les groupes créés dans cet exemple. Nous avons utilisé le dendrogramme indiquant la distance entre les observations pour couper un certain nombre de groupes adaptés aux besoins de l'analyse.

Nous calculons ensuite la distance entre tous les autres points et tous les points d'un groupe d'échantillon, puis nous les attribuons au groupe le plus proche. Enfin, nous pouvons obtenir les groupes pour tous les points (séries temporelles), comme illustré à la Fig. 6 (d).

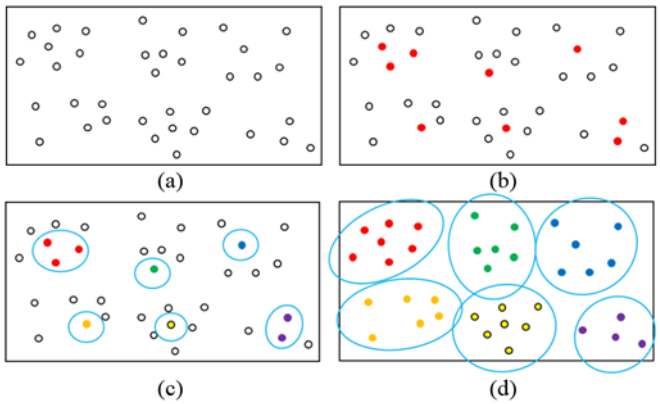


Fig. 6 Méthode d'échantillonnage

4.3 Analyse des résultats

Sur la base des résultats obtenus, nous analysons les comportements des utilisateurs de cartes à puce en ce qui concerne les arrêts d'embarquement, le profil quotidien et le trajectoire espace-temps pour chaque groupe (étapes 9-10-11 de la figure 4).

5 CAS D'ETUDE

Les données sont fournies par la Société de transport de l'Outaouais (STO), une société de transport desservant 280 000 habitants à Gatineau, au Québec. La STO est une chef de file canadienne dans l'utilisation des données de cartes à puce pour le transport en commun. Ce système est utilisé depuis 2001 et une proportion importante (plus de 80%) des utilisateurs de la STO possède une carte à puce [Pelletier et al., 2011].

Dans cette étude, toutes les données de transactions de la semaine de mai 2014 ont été utilisées pour tester la méthode de classification spatiale proposée. Cet ensemble de données contient 1 163 717 voyages.

La méthode est programmée en python, ce qui permet de traiter une base de données assez volumineuse.

Lors de l'implémentation, le nombre de groupes est déterminé en coupant des branches de dendrogramme. La Fig. 7 montre le dendrogramme de l'algorithme de classification spatiale. Nous l'avons découpé en 10 groupes :

• Nous avons essayé d'obtenir des groupes de taille égale autant que possible, même si cela n'est pas obligatoire (les comportements des utilisateurs peuvent ne pas être équilibrés de manière uniforme). Nous pouvons comparer plus de comportements différents si nous obtenons plus de groupes.

• Dans ce cas, si nous augmentons le nombre de groupes de 10 à 11, il y aura un groupe dont la taille est trop petite. Ensuite, après le processus d'allocation, cette taille de groupe sera négligeable par rapport aux autres groupes. Pour l'analyse, nous préférons conserver une taille de groupe plus grande.

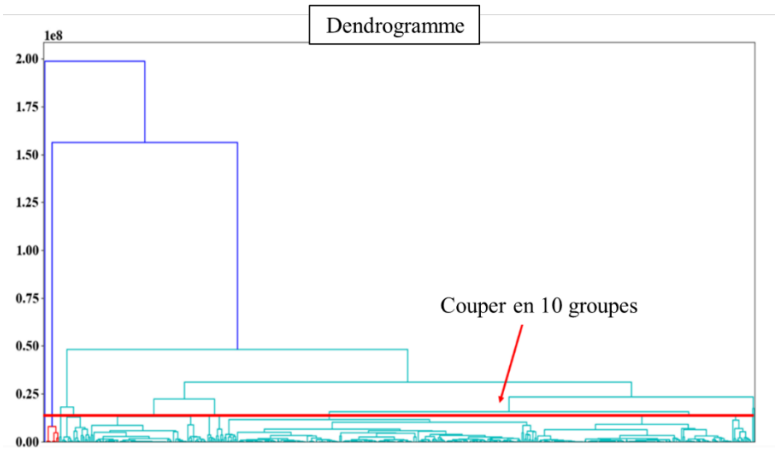


Fig.7 Dendrogramme de la classification hiérarchique de l'algorithme de classification spatiale

6 RESULTATS ET ANALYSES

6.1 Résultats

Un extrait des résultats de la classification spatiale est présenté dans le Tableau 2. Pour chaque combinaison « numéro de carte à puce + date » (carte-jour), une liste d'arrêts est générée et un groupe est obtenu. Par exemple, l'utilisateur 1000309 quitte son domicile pour le travail ou l'étude à l'arrêt 5022 et part pour son domicile à l'arrêt 2604. Basé sur l'algorithme de la classification spatiale, le comportement de cet utilisateur-jour devrait être regroupé dans le cluster 7.

Tableau 2 Résultat de la classification spatiale

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

6.2 Analyse par arrêt d'embarquement

La Fig. 8 montre l'analyse par premier arrêt d'embarquement de la classification spatiale. Chaque couleur représente un groupe et les points représentent uniquement le premier arrêt d'embarquement. En général, les groupes sont regroupés en fonction de l'emplacement (coordonnées). Cependant, dans certains cas, le cas est plus compliqué. Par exemple, dans la zone « Aylmer », les couleurs orange et verte sont mélangées, car les destinations de ces deux groupes sont différentes, même si les origines sont similaires. Dans ce cas, les destinations du groupe vert sont Ottawa, alors que celles du groupe orange sont Hull ou Gatineau. C'est un avantage de la méthode proposée par rapport aux méthodes classiques.

6.3 Analyse par trajectoire quotidienne

La Fig. 9 montre la trajectoire quotidienne de chaque groupe obtenu par classification spatiale. En observant les couleurs, nous pouvons avoir un aperçu des caractéristiques de chaque groupe. Par exemple, les utilisateurs du groupe cyan vivent à Buckingham et vont travailler à Ottawa. Peut-être y vont-ils directement ou ont-ils un transfert à Gatineau. Si nous voulons faire la distinction entre ces deux comportements (correspondance ou non), nous pouvons découper le dendrogramme en plusieurs groupes. Ceci est un avantage de la méthode proposée par rapport aux méthodes classiques.
Cette séparation des deux comportements aide à caractériser la demande. Sur la base de ce résultat, nous pourrions suggérer aux autorités des transports en commun de mettre en place de nouvelles lignes ou d'améliorer le service d'autobus, afin que les personnes puissent se déplacer directement et facilement de Buckingham à Ottawa.

6.4 Analyse par trajectoire spatio-temporelle

Sur la base du résultat de la classification spatio-temporelle, une trajectoire spatio-temporelle 3D de chaque groupe est tracée. La Fig. 10 (a) montre tous les profils individuellement, et sur la Fig. 10 (b), le chemin moyen pour chaque groupe. Le chiffre de l'axe Z correspond à l'heure du jour (la 25e heure correspond à une transaction de 1 heure du matin).
Dans la Fig. 10 (b), même si les utilisateurs du groupe vert habitent plus près de leur lieu de travail que le groupe bleu clair (de l'est au centre-ville), le groupe vert quitte la maison plus tôt et rentre plus tard que celui de bleu clair. Cela peut être dû à une ligne de bus express qui relie l'origine et la destination du groupe bleu clair. Par conséquent, il est possible de suggérer aux autorités de transport en commun d'implémenter une ligne de bus express pour desservir les utilisateurs du groupe vert afin qu'ils puissent gagner du temps lors de leurs déplacements.
Nous pouvons également constater que le comportement du groupe vert clair est stable pendant les heures de travail (de 9 h 30 à 15 h 00, l'emplacement du groupe vert clair ne change pas beaucoup). Cela signifie que ces utilisateurs se déplacent localement. Il est possible de suggérer aux autorités de transport en commun de mettre en place une ligne de bus spéciale pour ces utilisateurs. Cette nouvelle ligne de bus reliera l'origine et la destination du groupe vert clair, et ne sera exploitée qu'aux heures de pointe, mais elle peut très bien répondre à la demande de ce groupe.

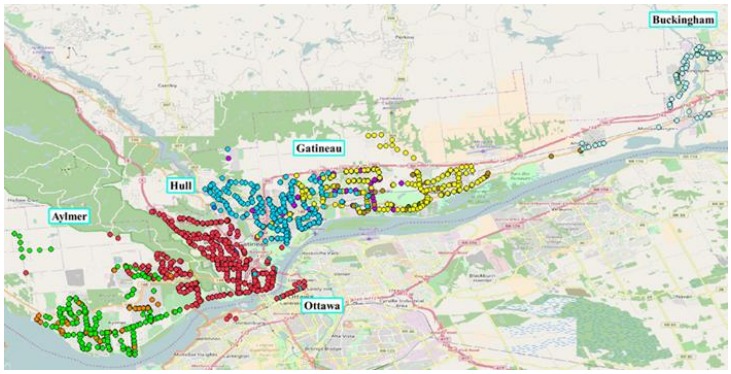


Fig. 8 Analyse par arrêt d'embarquement

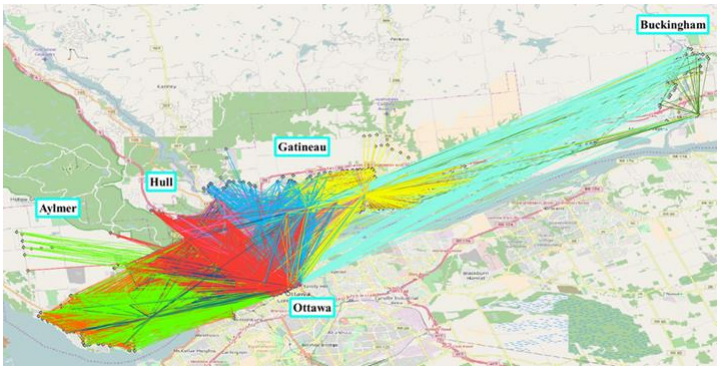
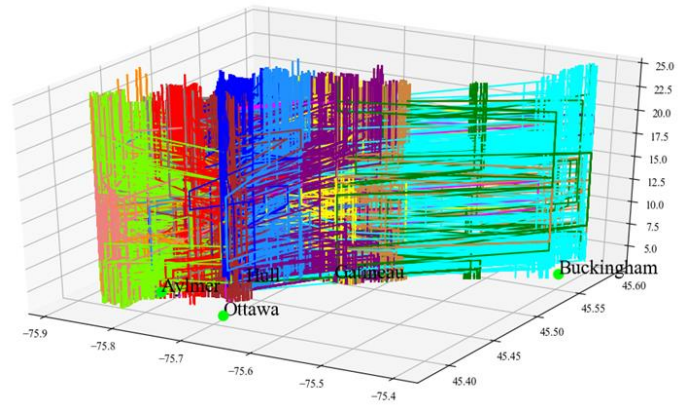
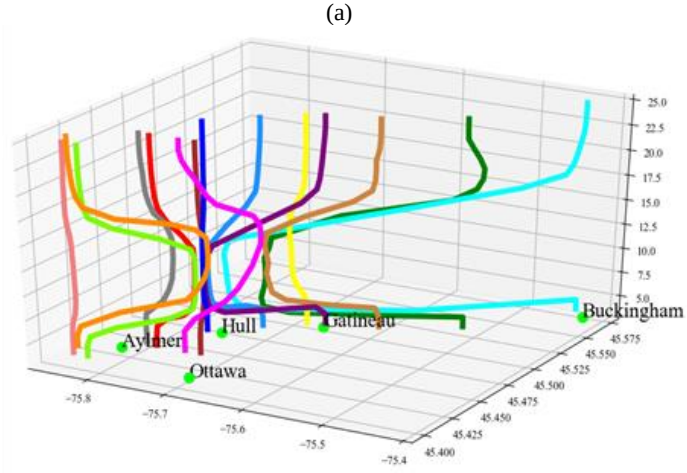


Fig. 9 Analyse par trajectoire quotidienne



(a)



(b)

Fig. 10 Analyse par trajectoire spatio-temporelle

7 CONCLUSION

7.1 Contribution

Dans cet article, une nouvelle méthodologie basée sur la déformation temporelle dynamique, la classification hiérarchique et la méthode d'échantillonnage est proposée pour classifier les comportements spatio-temporels des utilisateurs de cartes à puce en transport en commun. Le résultat montre que les comportements peuvent être bien distingués. En fonction des résultats, il est possible de suggérer des améliorations à l'autorité des transports en commun afin de mieux desservir les passagers de groupes spécifiques.

7.2 Limites

Premièrement, l'algorithme de déformation temporelle dynamique est quadratique, le temps de calcul est donc long. Deuxièmement, le critère de séparation est la distance, ainsi, différents comportements peuvent rester dans le même groupe, car leur dissimilarité des autres facteurs ne sont pas prises en compte (par exemple, l'objectif de voyage n'est pas un critère ici). D'autres limitations proviennent des données: l'estimation des destinations peut ne pas être parfaite (elle n'est pas validée), ce qui peut entraver les résultats de la méthode de classification.

7.3 Perspectives

À l'avenir, il est proposé de réaliser certains travaux pour améliorer cette nouvelle méthode. Premièrement, nous jugeons la qualité de la classification en observant la trajectoire quotidienne et la trajectoire spatio-temporelle. Une méthode quantitative est nécessaire pour mesurer la dissimilarité entre chaque groupe, afin de prouver mathématiquement que la méthode proposée fonctionne bien. Deuxièmement, des travaux devraient être faits pour améliorer le temps de calcul, pour surmonter les limites de la méthode de déformation temporelle dynamique. En outre, plus de suggestions peuvent être proposées aux autorités de transport en commun afin de mieux répondre à la demande des utilisateurs d'un groupe spécifique.

8 REMERCIEMENTS

Les auteurs désirent remercier la Société de transport de l'Outaouais pour leur collaboration au projet et la fourniture des données. Les auteurs soulignent également le support financier de Thalès et du Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG, projet RDC 446107-12).

9 REFERENCES

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.

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

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

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.

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.

Chevalier, F., et al. (2013). La classification. Université de Rennes. https://docplayer.fr/13650741-La-classification-2012-2013-fabien-chevalier-jerome-le-bellac.html

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.

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.

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.

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

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.

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. (2017). A visual segmentation method for temporal smart card data. *Transportmetrica A: Transport Science*, 13(5), 381-404.

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

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., 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*.

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.

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.

Kieu, L. M., Bhaskar, A., and Chung, E. (2014). "Transit passenger segmentation using travel regularity mined from Smart Card transactions data." Proc., Transportation Research Board 93rd Annual Meeting, National Research Council, Washington, DC

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.

Kurauchi, F., Schmöcker, J. D., Shimamoto, H., & Hassan, S. M. (2014). Variability of commuters' bus line choice: an analysis of oyster card data. *Public Transport*, 6(1-2), 21-34.

Lee, S. G., & Hickman, M. (2014). Trip purpose inference using automated fare collection data. *Public Transport*, 6(1-2), 1-20.

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.

Morency, C., Trépanier, M., & Agard, B. (2006, September). 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., Trépanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3), 193-203.

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance measures for time series in R: The TSdist package. *R Journal*, 8(2), 451-459.

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.

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, Boston, MA.

Shi, X., and L. Hangfei. 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.

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.

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.

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*.

Trépanier, et al. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1), 1-14.