Évaluation des impacts de l'implantation d'un service de bus rapide à partir de données de cartes à puce

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

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

[2] Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT)

***Résumé* – Bien qu'il ait déjà été démontré à de nombreuses reprises que les données provenant de systèmes de paiement par cartes à puce puissent être utilisées pour avoir une meilleure compréhension des comportements des usagers de transport collectif, ces données restent difficiles à analyser. Dans cet article, nous proposons une méthode basée sur la corrélation croisée des profils d'utilisation pour évaluer l'impact de l'implantation d'un service de bus rapide à la Société de transport de l'Outaouais (Canada). Près de 2 millions de transactions de cartes à puce ont été analysées pour montrer que 56,5% des usagers ont modifié leur comportement d'utilisation du réseau de transport collectif suite à l'implantation du nouveau service.**

***Abstract* - Smart card data has proven to be useful to characterize public transit ridership and travel behavior. However, this data is still hard to analyze. In this study, we propose a cross correlation distance method to evaluate the impact of the implementation of a new bus-rapid transit (BRT) at the Société de transport de l'Outaouais, Canada. The experiment shows that 56.50% users have changed their pattern in weekdays, so introduction of BRT may represent a large impact to users.**

***Mots clés* – transport collectif, fouille de données, service de bus rapide, carte à puce.**
***Keywords* – public transport, data mining, bus-rapid transit, smart card.**

1 INTRODUCTION

La mise en œuvre d'un nouveau système rapide par bus (SRB) dans un réseau de bus existant peut avoir un impact énorme sur le comportement de voyage des utilisateurs. Alors que de nombreux utilisateurs peuvent profiter de l'augmentation du niveau de service du SRB, d'autres pourraient être désavantagés par les nouvelles configurations du réseau. Bien que certains impacts puissent être évalués au niveau agrégé, par exemple, les méthodes visant à analyser le comportement des voyageurs basées sur chaque transaction des cartes à puce sont bien développées (Ghaemi et al., 2015), peu d'études proposent une méthode pour évaluer les changements de comportement sur une base de l'utilisateur individuel. L'analyse des changements de comportement, basé sur l'utilisateur individuel, permettra d'enrichir la compréhension du phénomène, de manière plus fine que le traitement des données agrégées.

Des données sont générées lors qu'un voyageur passe sa carte à puce sur une machine de perception installée dans un bus. Les 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 utilisons ces données pour faire notre recherche. Dans cet article, nous présentons une approche, au niveau de l'utilisateur, pour évaluer individuellement les impacts de la mise en place d'un système rapide par bus à partir de l'analyse des données de carte à puce. L'utilisation de données de cartes à puce à partir d'un système de perception automatique donne une évaluation des changements dans le comportement de déplacement de chaque utilisateur avant et après la mise en œuvre du SRB. L'efficacité et utilité des données de carte à puce pour analyser le comportement de déplacement des usagers du transport a déjà été prouvée [Pelletier et al., 2011]. Une technique d'exploration des données basée sur la distance de corrélation croisée, l'échantillonnage et la segmentation hiérarchique est utilisée ici pour segmenter les utilisateurs de transport en commun ayant une distribution similaire de l'heure de départ (de transactions par carte à puce). Nous proposons également une méthode comparative pour mesurer les changements de comportement introduits par le nouveau système.

L'article est organisé de la manière suivante. La section suivante (La Section 2) se concentre sur l'état de l'art des méthodes de fouille de données, entre autres, les métriques (pour les méthodes de segmentation de données). Dans la section 3, compte tenu des limites des méthodes actuelles, nous avons conçu un algorithme en combinant la distance de corrélation croisée, la segmentation hiérarchique, une méthode d'échantillonnage et d'affectation. Dans le cadre de sa mise en œuvre, nous introduisons les détails de la pratique. Enfin, dans la section 4, nous présentons les résultats sur des données de cartes à puce dans les transports en commun, lors de l'introduction d'un service SRB, à partir d'une étude de cas sur un réseau de transport en commun (STO).

2 REVUE DE LITTÉRATURE

La fouille de données permet de caractériser la demande des voyageurs qui utilisent le système de transport. Différentes métriques seront présentées pour l'analyse des données disponibles. Tout d'abord la distance euclidienne, pour sa popularité et sa simplicité, mais aussi des outils pour l'analyse des séries temporelles : la déformation temporelle dynamique et la corrélation croisée.

2.1 *Méthodes de fouille de données*

Les méthodes de segmentation de données ont pour but de partager un ensemble d'observations en clusters. Une bonne méthode de segmentation va produire des clusters de haute qualité avec une grande similarité intra-classe et une faible similarité inter-classe. Plusieurs grandes approches de segmentation existent [Subbiah, 2011]:

1. Les algorithmes de partitionnement construisent différentes partitions et les évaluent. Il contient essentiellement deux méthodes heuristiques.
   - k-means: chaque groupe est représenté par sa valeur moyenne.
   - k-medoids ou PAM (partitionnement autour de medoids): chaque groupe est représenté par l'un des objets du cluster.
2. Les algorithmes hiérarchiques créent une décomposition hiérarchique de l'ensemble des données (ou objets) en utilisant certains critères d'agglomération ou de division.
3. Les méthodes basées sur la densité utilisent des fonctions de connectivité et de densité (DBSCAN, OPTICS)
4. D'autres méthodes de segmentation existent comme celles basées sur une grille (une structure de granularité de niveaux multiples: STING, CLIQUE) et celles basées sur des modèles (un modèle est supposé pour chacun des clusters et l'idée est de trouver le meilleur ajustement à ce modèle).

Au fil des ans, plusieurs auteurs ont proposé d'utiliser des techniques d'exploration de données pour analyser les données de transaction par carte à puce. Une recherche [Morency et al., 2007] a montré la variabilité du comportement de déplacements d'utilisateurs dans un réseau de bus en utilisant la technique des k-means. Cependant, la méthode k-means est de calculer une valeur moyenne pour chaque groupe, en se basant sur le « vecteur » de chaque individu, malheureusement ceci n'est pas adapté aux séries temporelles. Des travaux plus récents utilisent DBSCAN [Kieu et al., 2015] [Ma et al., 2013] pour évaluer le comportement de déplacement des utilisateurs de cartes à puce. Li & Chen [Li et al., 2016] ont utilisés la déformation temporelle dynamique (DTW) pour unifier la référence temporelle des transactions par carte à puce, cependant la complexité des calculs de DTW n'est pas adaptée aux données de grandes échelles. Une recherche [El Mahrsi et al., 2014] a également utilisé un mélange de techniques pour analyser les données de la ville de Rennes, en Franc. Dans cette recherche, les auteurs se basent sur chaque transaction, mais non sur chaque profil de déplacement des voyageurs. Cependant, il y a encore des défis associés à la caractérisation des comportements basés sur la distribution temporelle parce que les méthodes de calcul de la distance réelle entre les observations ne sont pas bien adaptées à la nature des analyses qui sont requises par les autorités de transport en commun [Ghaemi et al., 2015].

2.2 *Distance euclidienne*

Diverses mesures de distance existent pour mesurer la (dis) similarité entre deux instances (vecteurs liés aux observations). Dans cette partie, nous comparons trois types de distances: la distance euclidienne, la distance de déformation temporelle dynamique (DTW) et la distance de corrélation croisée.

La distance euclidienne mesure la distance directement entre deux points dans l'espace euclidien [Deza et al., 2009]. Soit x_i et v_j chaque être un vecteur P-dimensionnel. La distance euclidienne est calculée comme [Liao, 2005]:

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

La distance euclidienne est largement utilisée dans de nombreux domaines d'application. En particulier, dans le cas du système utilisé à Gatineau (Canada) qui contient de très grands ensembles de données : environ 600 000 entrées sont collectées chaque mois. Les techniques d'exploration de données ont été utilisées pour analyser ces données avec des résultats intéressants [Agard et al., 2006]. Néanmoins, la distance euclidienne utilisée dans ces études n'est pas bien adaptée pour analyser les séries chronologiques.

2.3 *Déformation temporelle dynamique*

La déformation temporelle dynamique est une technique populaire pour comparer les séries chronologiques, fournissant une mesure de distance insensible à la compression et aux étirements locaux. La distance utilisée se base sur un ajustement optimal de l'une des deux séries d'entrée sur l'autre [Giorgino, 2009]. La méthode pour calculer la déformation dynamique entre deux séries S et T est la suivante [Berndt et al., 1994]:

$$S = s_1, s_2, \dots s_i, \dots, s_n \qquad (2)$$

$$T = t_1, t_2, \dots t, \dots, t_n \qquad (3)$$

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

$$W = w_1, w_2, \dots w_i, \dots, w_n \qquad (4)$$

2.4 *Corrélation croisée des distances*

La distance de corrélation croisée est basée sur la corrélation entre les 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 U et al., 2016] :

$$CC_k(X,Y) = \frac{\sum_{i=0}^{N-1-k}(x_i-\bar{x})(y_{i+k}-\bar{y})}{\sqrt{(x_i-\bar{x})^2}\sqrt{(y_{i+k}-\bar{y})^2}} \qquad (8)$$

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

$$d_{CC}(X,Y) = \sqrt{\frac{(1-CC_0(X,Y))}{\sum_{k=1}^{max} CC_k(X,Y)}} \qquad (9)$$

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.

Dans cette étude, l'intérêt porte sur le changement de l'heure de départ des voyageurs. Le changement temporel correspond le décalage dans la méthode corrélation croisée. Par conséquent, nous préférons utiliser la corrélation croisée au lieu de DTW ou la distance euclidienne en raison de la nature de la distribution horaire de transactions par carte à puce pour un usager dans la journée. Ces distributions sont caractérisées par quelques observations à travers la journée.

3 CONCEPTION DE L'ALGORITHME

Après cette discussion sur les métriques de distance utilisées pour évaluer la similarité entre les observations, cette section présente deux éléments supplémentaires nécessaires pour compléter l'algorithme (HCA et d'échantillonnage), et présente ensuite la mise en œuvre.

3.1 *Regroupement hiérarchique*

La classification (segmentation) hiérarchique (également appelée analyse de segmentation hiérarchique) est une méthode de segmentation qui cherche à construire une hiérarchie de clusters. Les stratégies pour la segmentation hiérarchique se répartissent généralement en deux types [Rokach et Oded, 2005]:

1. Agglomérative: Chaque observation commence dans son propre cluster, et des paires de clusters sont fusionnées comme on se déplace vers le haut la hiérarchie.
2. Divisive: Toutes les observations commencent dans un cluster unique, et on effectue une séparation de manière récursive en se déplaçant vers le bas de la hiérarchie.

En général, les fusions et les scissions sont déterminées d'une manière gourmande en calcul. Les résultats de la classification hiérarchique sont habituellement présentés dans un dendrogramme. Tout d'abord, nous avons besoin de calculer une matrice de distance entre toutes les observations. Ensuite, l'algorithme de classification hiérarchique est utilisé pour traiter cette matrice de distance afin de fusionner toutes les données en clusters. Enfin, nous obtenons un dendrogramme à partir duquel nous pouvons couper des branches pour obtenir les clusters désirés. Avec de grands ensembles de données, la matrice de distances peut être impossible de calculer, en particulier lorsque le calcul de la distance entre deux éléments prend du temps (ce qui est le cas de corrélation croisée). La situation envisagée dans cette étude est dans ce cas, nous utilisons donc une méthode d'échantillonnage pour seconder la segmentation.

3.2 Échantillonnage et affectation

La combinaison de la distance de corrélation croisée et de l'algorithme de classification hiérarchique est faite pour segmenter les données. Cependant, dans la pratique, la distance de corrélation croisée ne peut guère être appliquée à une très grande base de données, en raison de l'étendue en temps de calcul. Dans notre cas, nous utilisons une méthode d'échantillonnage et affectation pour régler le problème.

La figure 1 montre le processus d'échantillonnage et affectation utilisé. Dans un premier temps, nous avons toutes les observations (a). L'échantillonnage se fait comme suit. Tout d'abord, choisir des points aléatoires comme échantillon. Ces points d'échantillonnage doivent être aussi uniformément distribués que possible. Les points rouges sur la figure 1 (b) sont les points sélectionnés.

Deuxièmement, nous appliquons la distance de corrélation croisée et de l'algorithme de classification hiérarchique à ces points d'échantillonnage. Sur la figure 1 (c), on voit des clusters fabriqués à partir de cet échantillon.

Troisièmement, en calculant la distance entre chaque autre point et les points du groupe d'échantillons, nous attribuons tous les points au groupe le plus proche en fonction de la distance de corrélation croisée moyenne de tous les clusters. Nous avons déterminé le groupe d'un point en minimisant la distance moyenne entre ce point à tous les points du groupe déterminé. Enfin, on obtient les groupes pour tous les points (séries chronologiques), comme représentés sur la figure 1 (d).

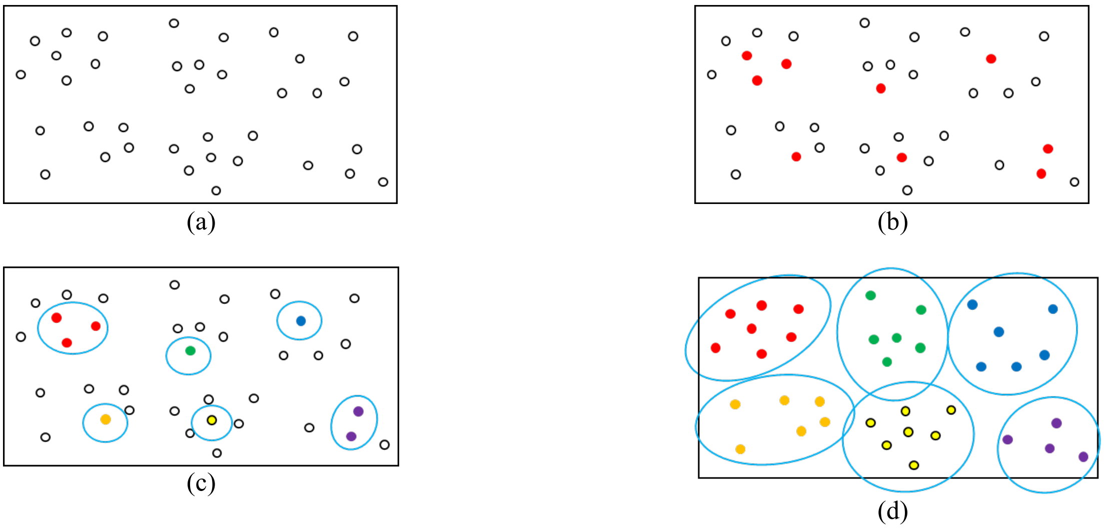


Figure 1. Méthode d'échantillonnage et affectation

3.3 Mise en œuvre

La figure 2 montre les 8 étapes pour la mise en œuvre de l'algorithme développé. Classification est basée sur la répartition temporelle des transactions de carte à puce pendant la journée. Une observation est créée pour chaque utilisateur, chaque jour où il s'est déplacé (utilisateur-jour).

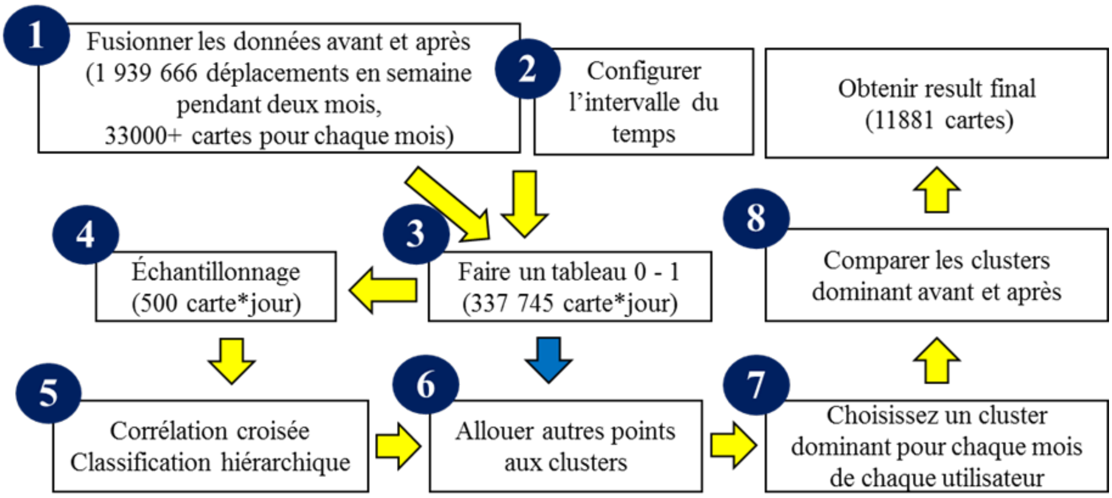


Figure 2. Mise en œuvre d'algorithme

Étape 1. Fusionner ensemble toutes les transactions avant et après l'introduction de la SRB (totalement 1 939 666 déplacements en semaine, de plus de 33 000 cartes pour chaque mois). Cette étape consiste à faire en sorte que les clusters soient cohérents pour les deux situations.

Étape 2. Définir l'intervalle de temps. L'heure de transaction doit être classifiée avant de calculer la distance. Par exemple, 09: 21 et 09 :28 peuvent appartenir à la période 09 :20 - 09 :29 et 09 :32 à la période 09:30 - 09:39. Dans ce cas, l'intervalle de temps est de 10 minutes. Différents intervalles sont fixés pour les heures de pointe pour éviter d'avoir trop d'éléments dans les vecteurs.

Le tableau 1 montre les intervalles appliqués sur un jour. Ainsi de 00 :00 à 5 :59, aux heures creuses, les intervalles sont de 30 minutes, alors qu'aux heures de pointe, ils sont de 5 minutes.

Tableau 1. Les intervalles de temps pour la distribution quotidienne des transactions

Heure	Type de période	Intervalle
00:00 - 05:59	Hors heure de pointe	30
06:00 - 06:59	Heure régulière	10
07:00 - 08:59	Heure de pointe	5
09:00 - 09:59	Heure régulière	10
10:00 - 13:59	Hors heure de pointe	30
14:00 - 14:59	Heure régulière	10
15:00 - 15:59	Heure de pointe	5
16:00 - 17:59	Heure régulière	10
18:00 - 23:59	Hors heure de pointe	30

Étape 3. Faire un tableau qui contient la répartition temporelle des transactions par carte à puce pour chaque utilisateur-jour. Les ID uniques sont créés par concaténation du numéro de la carte avec la date. Un total de 35, 886 utilisateurs-jour sont créés.

Tableau 2. Exemple de données d'utilisateurs-jour (tableau 0-1)

Combinaison	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	..

Le tableau 2 montre un tout petit extrait de la base de données construite. On peut y lire, par exemple, que l'individu 1150296033731200, en date du 4 septembre 2013, a validé une transaction dans le réseau, entre 6h20 et 6h29.

Étape 4. Un échantillon qui contient 500 cartes-jour a été choisi. Pour choisir un échantillon aléatoire approprié, il faut notamment éviter de mettre un profil de l'utilisateur de temps deux fois dans l'échantillon. L'échantillon contient donc 500 usagers de carte à puce, et 500 jours différents. C'est avec cet échantillon que nous calculons la distance de corrélation croisée et nous faisons la classification hiérarchique.

Étape 5. Calculer les corrélations croisées et l'algorithme de classification hiérarchique sur l'échantillon. Analyser le dendrogramme obtenu par l'algorithme de classification hiérarchique, décider du nombre de groupes.

Étape 6. Affecter les profils d'utilisateur-jour restants au cluster le plus proche.

Étape 7. Identifier le cluster dominant pour chaque utilisateur (carte), avant et après l'introduction du SRB respectivement. Nous définissons comme dominant le cluster avec la fréquence d'apparition maximale (si aucun cluster ne dépasse 2 apparitions, la carte n'a pas de cluster dominant).

Étape 8. Comparer le cluster dominant avant et après. Trois cas peuvent se produire: l'utilisateur ne change pas de cluster dominant (code 0), l'utilisateur change de cluster dominant (code 1) ou le groupe dominant n'existe pas avant ou après (code 2).

Les étapes 5 et 6 sont implémentées dans le logiciel R (https://www.r-project.org). D'autres étapes sont implémentées en Python (https://www.python.org). Premièrement, avec les données brutes (les données de transaction de Rapibus), nous calculons le tableau 0 -1 et choisissons un échantillon en utilisant Python. Deuxièmement, avec cet échantillon, la distance de corrélation croisée est calculée et la classification hiérarchique est faite en utilisant R, en vue d'obtenir les groupes. Dans chaque groupe, nous avons les profils de l'heure de départ d'un jour des voyageurs. Troisièmement, nous comparons le changement de comportement (heure de départ) en utilisant Python.

4 ÉTUDE DE CAS ET RÉSULTATS

4.1 Système d'information

L'ensemble des données sont fournies par la Société de Transport de l'Outatouais (STO), une autorité de transit desservant les 280, 000 habitants de Gatineau, au Québec. L'autorité de la STO est un leader canadien en utilisant le système de la collection intelligente par la carte à puce. Ce système a été utilisé depuis 2001, et actuellement une grande proportion des utilisateurs de la STO a une carte à puce [Morency et al., 2007].

Le tableau 3 montre un extrait de l'ensemble de données de carte à puce brutes; il contient l'information de déplacement de l'utilisateur. Outre l'identification de la carte (un numéro anonyme), il y a le code de billet (catégories tarifaires), la date et l'heure de la transaction, le numéro de la ligne (route) et la direction (Nord, Sud, Est, Oeust). Toutes les transactions sont effectuées sur un réseau de bus; l'emplacement de la transaction est également disponible.

Tableau 3. Extrait de l'ensemble de données de carte à puce brutes

Id carte	Type de carte	Date	Heure
1150629967111800	140	2013-09-03	65232
1273590714804090	110	2014-09-02	71909
1273590714804090	110	2014-09-02	154607

Ligne	Direction	Jour de la semaine	Id arrêt
44	Sud	2	1140
224	Sud	1	2801
224	Nord	1	2610

Pour la suite de l'étude, nous avons sélectionné des données en semaine parce que les profils d'utilisation des travailleurs sont plus faciles à identifier dans une étude préliminaire. Nous avons choisi deux mois similaires pour évaluer la différence de comportement des usagers: septembre 2013 et septembre 2014. Le système SRB a été introduit en octobre 2013. Cette base de données contient 1 939 666 déplacements en semaine pour ces deux mois, 33 016 cartes pour septembre 2013 et 33 089 cartes pour septembre 2014. Cependant, toutes les cartes ne sont pas présentes sur les deux périodes. Enfin, nous continuons à 760 210 déplacements liés aux 11 881 cartes qui ont enregistré des voyages au cours des deux mois.

4.2 Nombre de groupes

Après le calcul de la distance de corrélation croisée et l'exécution de l'algorithme hiérarchique, nous obtenons un dendrogramme présenté dans la Figure 3. Ce dendrogramme permet de déterminer le nombre pertinent de clusters pour cette étude. Le principe est d'obtenir, en même temps, un grand nombre de clusters pour une bonne précision, mais pas trop grand pour l'interprétation, aussi nous voulons assurer la cardinalité des clusters sont similaires dans chaque cluster.

Nous avons d'abord testé 25 clusters. Les cases plus basses à la Figure 3 présentent le résultat. Il y avait de nombreux groupes contenants peu d'usagers. Dans ce cas, lors de l'analyse sur le changement de comportement, un usager peut facilement changer de groupe, même s'il change peu de comportement réel. Par conséquent, nous avons donc réduit progressivement le nombre de clusters sur le dendrogramme. Un bon compromis est de garder 11 groupes. Si le nombre de groupes est trop réduit, en tenant compte la prochaine étape qui va allouer les autres points aux clusters, il pourrait y avoir trop d'éléments dans un même groupe. Dans ce cas, lors de l'analyse sur le changement de comportement, il sera difficile d'identifier un usager qui change de groupe, même si son comportement réel évolue plus grandement. Finalement, nous choisissons 11 groupes. La somme des transactions de chaque groupe (cf. Figure 4) illustre la pertinence de ce nombre de groupe, les comportements sont bien séparés.

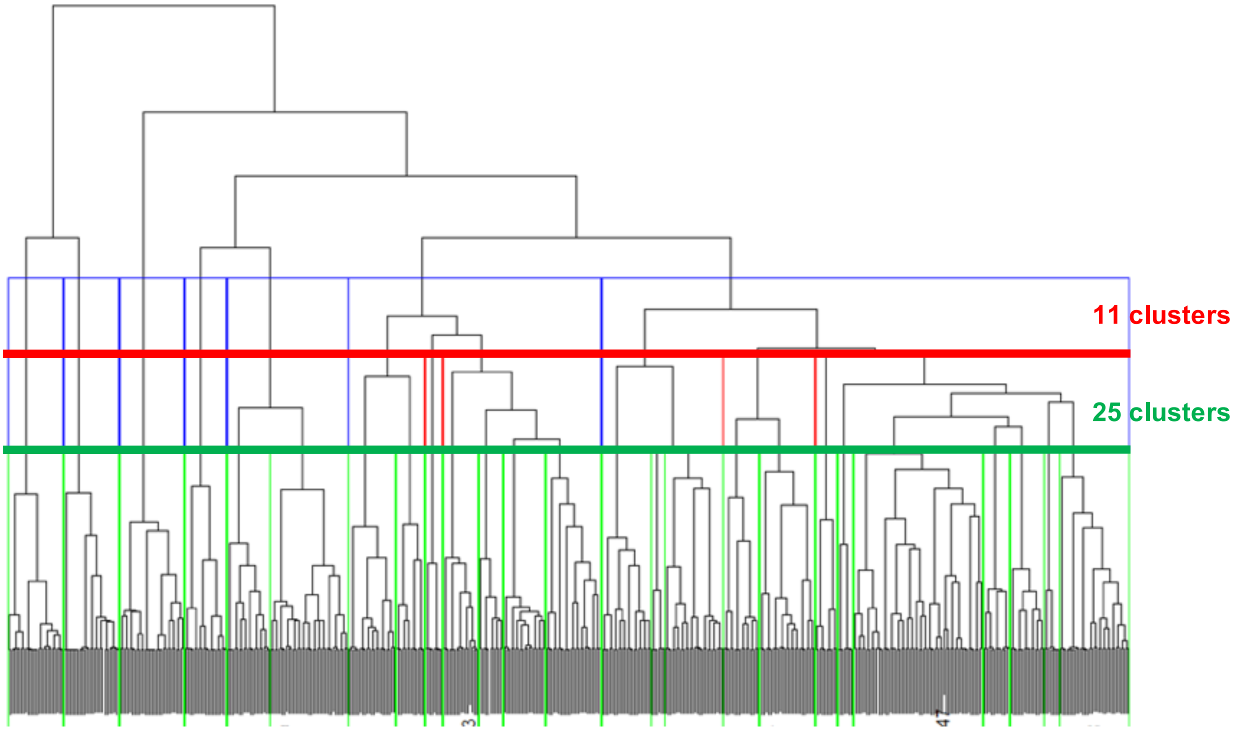


Figure 3 Dendrogramme des points d'échantillonnage

Nous déterminons alors le tableau de changements de cluster dominant (tableau 4). Le tableau identifie les groupes dominants avant et après l'introduction de la BRT. Un code indique si le cluster reste le même (le code 0), a changé (code 1) ou n'a pas été trouvé avant et après (code 2). Code 2 signifie que la carte est là avant et après, mais ne sont pas liés à un profil dominant (trop de clusters différents pour l'individu). Le temps de calcul dure environ 12h dans notre cas STO.

Table 4 Exemple de tableau de différence de cluster dominant

Carte	Cluster 2013-09	Cluster 2014-09	Code
706209	7	7	0
706259	8	11	1
706188	0	0	2
...	...	...	...

Ensuite, nous calculons la somme des transactions de chaque cluster. Ce résultat est représenté sur la Figure 4. Sur les 11 clusters, seulement 6 sont représentés dans cette figure. Cette figure montre comment le profil quotidien des utilisateurs peut être séparé. Par exemple, pour le groupe 3, les utilisateurs ont généralement une transaction d'embarquement 07:25 - 07:40 le matin, et une autre transaction d'embarquement entre 17 :00 - 17 :30 dans l'après-midi. Pour le groupe 6, les utilisateurs ont généralement une transaction d'embarquement 05 :30 - 06 :20 le matin, et une autre transaction d'embarquement entre 14:30 - 15 :45 dans l'après-midi, etc. Comme prévu, la méthode que nous avons développée peut clairement séparer les séries chronologiques clusters.

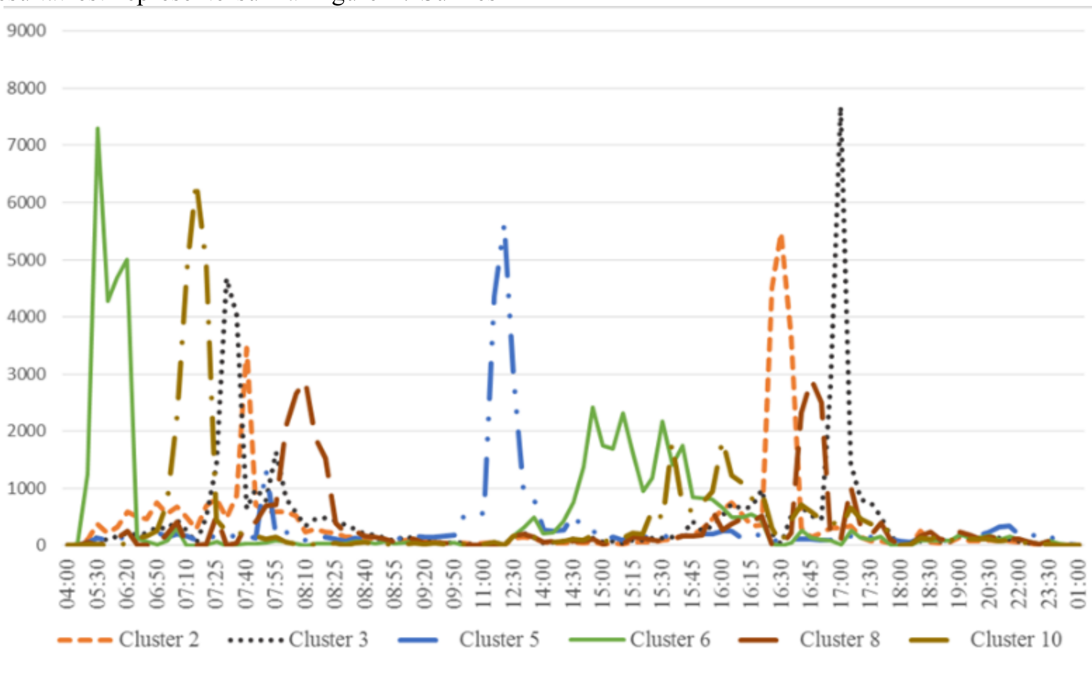


Figure 4. Somme des transactions de chaque groupe pendant les périodes du temps d'un jour

4.3 Comparaison générale

Après avoir comparé le groupe dominant avant et après du SRB, on obtient le résultat présenté dans la Figure 5. Le pourcentage des cartes pour lesquelles le groupe dominant change est 60,75%, et le pourcentage dont le temps de profils ne change pas est de 39,74%. En outre, il y a 3,76% de cartes pour lesquelles il n'y a pas de profil de temps avant et après.

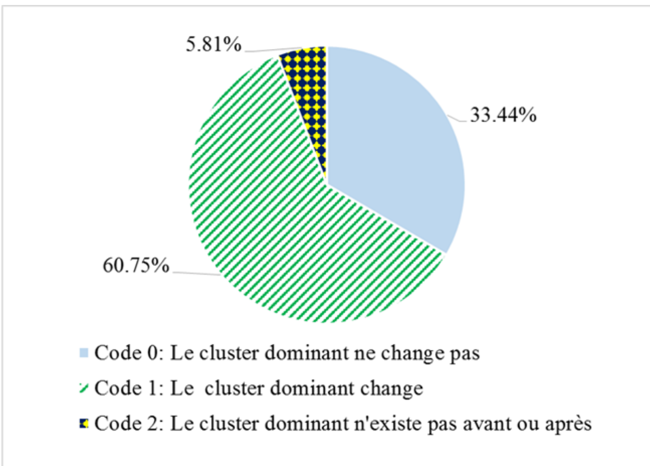


Figure 5. Résultat de différence de profil du temps

La Figure 6 montre la matrice de changement de cluster avant et après pour les 11 881 cartes. Premièrement, cela montre toutes les combinaisons de groupes qui peuvent être observées. Les utilisateurs peuvent changer de tout cluster à un autre. Nous pouvons voir que dans chaque groupe l'option dominante est de rester dans le même groupe, mais aussi de nombreux utilisateurs peuvent avoir un comportement différent avant et après, aussi certains clusters sont plus volatils que d'autres. Par exemple, la colonne verticale (2013) représente les groupes avant la mise en œuvre de SRB, et la colonne horizontale représente les groupes après la mise en œuvre du Rapibus. Par exemple, le premier nombre de la diagonale montre que 401 utilisateurs, dans le groupe 0 (pas de groupe dominant pendant cette période) avant la mise en œuvre de SRB, n'ont pas changé de comportement (groupe 0) après la mise en œuvre du SRB.

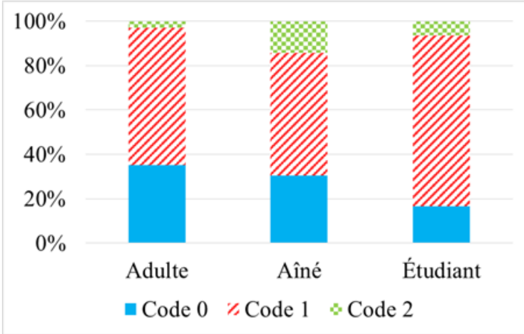
Après la mise en œuvre du SRB, seulement 20,7% des membres du cluster 1 restent, tandis que pour le cluster 6 le pourcentage monte à 69,2%. Le plus haut mouvement est pour le cluster 4, qui a perdu 11,9% de ses membres pour le cluster 8.

		2014 (après)												
		0	1	2	3	4	5	6	7	8	9	10	11	Total
2103(avant)	0	401	47	59	84	77	85	58	80	57	89	67	34	1138
	1	56	60	7	3	17	26	18	17	11	35	10	14	274
	2	68	6	75	49	21	3	9	43	38	5	47	6	370
	3	67		33	157	59	6	10	37	69	8	67	17	530
	4	95	12	18	61	199	16	6	17	73	24	24	33	578
	5	79	27	9	9	19	127	13	9	9	59	15	12	387
	6	55	21	15	7	7	14	557	74	6	11	22	1	790
	7	98	11	54	33	16	12	120	562	17	25	98	10	1056
	8	72	8	58	49	64	9	21	31	126	6	21	9	474
	9	103	29	5	6	21	63	17	18	5	154	13	18	452
	10	87	13	42	78	16	14	17	109	18	10	249	12	665
	11	35	15	4	15	26	4	4	12	13	7	7	40	182
	Total	1216	249	379	551	542	379	850	1009	442	433	640	206	6896

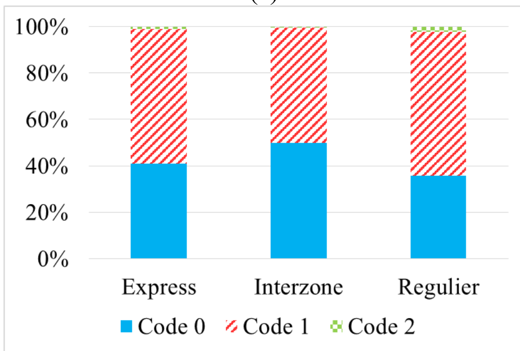
Figure 6. Matrice des clusters d'utilisateurs avant et après (NA = pas de cluster dominant)

4.4 Comparaison par type de carte

Il est intéressant d'observer les changements de cluster entre les différents détenteurs de tarifs. La Figure 7 montre la répartition des codes de différents types de cartes. Dans la partie (a), nous voyons que les adultes et les aînés ont été moins touchés. Dans la Figure 7 (b), nous voyons que les détenteurs de tarif d'express et interzone sont moins touchés que le tarif régulier. Ceci peut être expliqué par le fait que peu de routes express et interzone utilisent le nouveau corridor du Rapibus. La figure montre également que, contrairement aux étudiants et aux aînés, les utilisateurs express et interzone gardent des comportements dominants.



(a)



(b)

Figure 7. Analyse du changement par type de carte

4.5 Comparaison par secteur

Nous divisons le territoire de la STO en trois parties: Aylmer, Hull et Gatineau. Comme le montre la Figure8, Aylmer est une partie d'un territoire non desservi par le SRB, tandis que Hull et Gatineau sont traversés par le corridor SRB. Nous déterminons arbitrairement l'emplacement de la maison d'un utilisateur par la dominante de ses premières transactions d'un jour. Sur la carte, la couleur des arrêts est liée à la quantité d'utilisateurs qui changent leur comportement. Un bleu arrêt (foncé) signifie que plus de comportements d'utilisateurs sont en train de changer à cet arrêt, un (clair) arrêt jaune signifie que les comportements ne changent pas. Sur la carte, on voit une légère différence entre Aylmer (avec moins de changements) et d'autres utilisateurs.

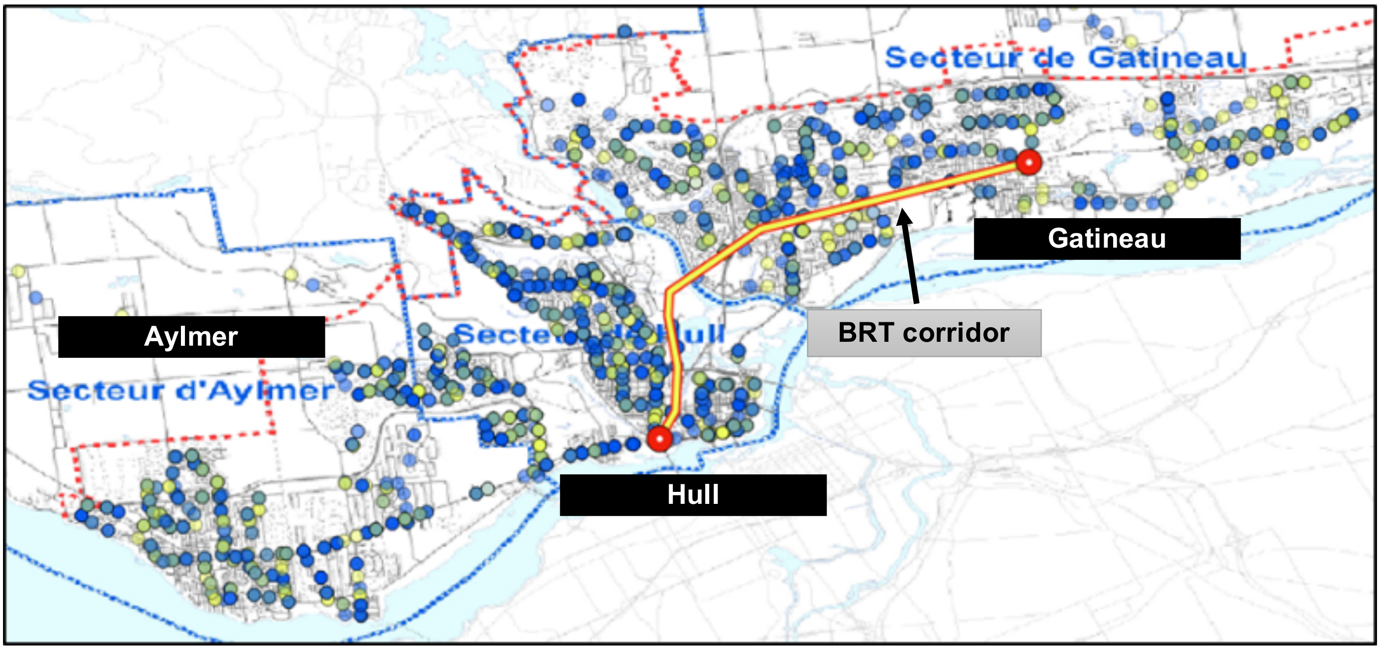


Figure 8. Répartition dû au changement de profil du temps par secteur

La Figure 9 montre le changement d'utilisation par secteur. Même si la différence d'échelle est de petite taille, nous voyons un peu plus de changements dans les secteurs de Hull et de Gatineau. Cela peut montrer que les détenteurs de cartes des secteurs desservis par le SRB sont plus touchés, comme prévu.

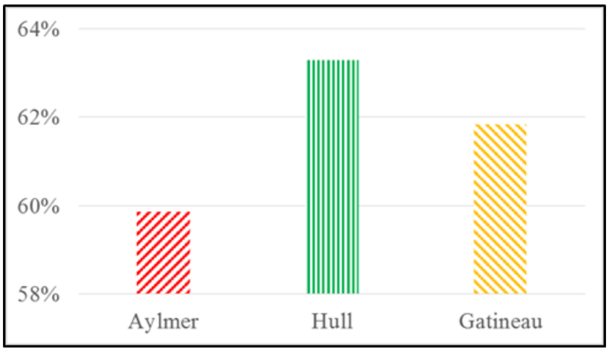


Figure 9. Proportion de codes 1 (le cluster dominant a changé) par secteur

5 Conclusion

5.1 Contributions

Dans cet article, nous avons développé un algorithme qui combine la distance de corrélation croisée, la classification hiérarchique et une méthode d'échantillonnage pour caractériser le comportement de déplacement des utilisateurs du transport en commun, en utilisant des données de transaction par carte à puce. L'application de l'approche à l'étude de la mise en œuvre d'un SRB dans le réseau de la Société de transport de l'Outaouais à Gatineau, Canada, a permis d'évaluer les changements de comportement de déplacement qui ont eu lieu suite à la mise en place d'un SRB. Nous avons observé que 60,75% des cartes ont été associés à un comportement de déplacement de changement de temps.

5.2 Limites

La principale limitation de ce travail est que les changements de comportement ont pu se produire en raison d'autres facteurs exogènes autres que la mise en œuvre du SRB lui-même. Par conséquent, nous avons essayé de tester les utilisateurs qui ne sont pas situés dans le secteur où le SRB a directement touché, mais nous ne pouvons pas être sûrs que ce soit une mesure de contrôle suffisant. Une autre limitation est liée à la technique d'échantillonnage qui est utilisée pour réduire le temps de calcul. L'échantillonnage peut ne pas refléter la situation réelle. Toutes les variations ne peuvent pas être liées à l'introduction de la SRB, aussi sur une sorte longue période de temps (un an), certains utilisateurs peuvent également avoir changé leurs habitudes pour des raisons personnelles.

5.3 Perspectives

Les premières perspectives de recherche sont liées à aborder les limites actuelles. Nous prévoyons d'utiliser des périodes prolongées de données avant et après la mise en œuvre d'un SRB pour tenter de purger les autres facteurs exogènes qui peuvent affecter le comportement de mobilité observé. L'utilisation d'ordinateurs plus puissants et l'introduction de techniques de calcul le temps de coupe sont également envisagés. A plus long terme, nous souhaitons développer un modèle pour être en mesure de prévoir l'évolution du déplacement des transports liés aux changements de service, en fusionnant les données des deux côtés de la demande et de l'offre.

6 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).

7 Références

Ghaemi M.S., Agard B., Trépanier M., Partovi N.V. (2015), Challenges in spatial-temporal data analysis in the public transport domain, IFAC - Proceedings, 48(3), 442,447, INCOM-IFAC Conference, Ottawa, Canada.

Agard B., Morency C., Trépanier M. (2006). Mining public transport user behavior from smart card data In: The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM), Saint-Étienne, France, May 17–19.

Berndt, Donald J., Clifford J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. KDD workshop. Vol. 10. No. 16..

Deza M. M., Deza E. (2009). Encyclopedia of distances. Springer Berlin Heidelberg, 583 p.

El Mahrsi M.K., Côme E., Baro J., Oukhellou L. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. In 3rd International Workshop on Urban Computing (SigKDD).

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

Kieu L.-M., Bhaskar A., Chung E. (2015), A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data, Transportation Research Part C: Emerging Technologies, Volume 58, Part B, Pages 193-207.

Li H., Chen X. (2016). Unifying Time Reference of Smart Card Data Using Dynamic Time Warping, Procedia Engineering, Volume 137, Pages 513-522.

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

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, Volume 36, November 2013, Pages 1-12.

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. Retrieved from https://cran.r-project.org/.

Pelletier M.-P., Trépanier M., Morency C. (2011). Smart Card Data Use in Public Transit: A Literature Review. Transportation Research Part C, Vol. 19, pp. 557–568.

Rokach, Lior, and Oded Maimon (2005). Clustering methods. In Data mining and knowledge discovery handbook. Springer US. 321-352.

Subbiah, K. (2011). Partitioning Methods in Data Mining [PowerPoint slides]. Retrieved from http://www.authorstream.com/Presentation/msusuresh-1133119-partitioning-methods/