Proposition de communication pour les
2e Rencontres Francophones Transport Mobilité (RFTM)
Montréal, 11-13 juin 2019

Titre : Méthode d'échantillonnage pour la classification des comportements des usagers de transport collectif à partir de données de carte à puce

1er choix de session
(Session 44) Données massives en transport public : traitement et valorisation des données
2e choix de session
(Session 10) La donnée : son appropriation, son exploitation et son partage : enjeux pour les transports intelligents de demain

Auteur(s) :
Li HE[1,2,3], M.Sc. A., étudiant Ph.D., he.li@polymtl.ca
Martin TRÉPANIER[1,2,3], Ph.D., Professeur titulaire, mtrepanier@polymtl.ca
Bruno, AGARD[1,2,3], Ph.D., Professeur titulaire, bruno.agard@polymtl.ca
[1] École Polytechnique de Montréal
[2] Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT)
[3] Laboratoire en Intelligence des Données

Mots-clés :
Transport en commun, données de carte à puce, échantillonnage, classification, comportement spatio-temporel.

Résumé :

Les données provenant de systèmes de perception par cartes à puce sont très utiles pour les planificateurs de transport en commun, car elles permettent de mieux connaître les comportements des utilisateurs [Pelletier et al., 2011]. Sur la base de fouille des données, plusieurs chercheurs ont proposé des méthodes de classification du comportement des usagers, en vue de mieux connaître leurs caractéristiques d'utilisation des réseaux, principalement axées sur les fréquences et les moments d'utilisation. Parmi les méthodes, citons le k-means et la classification hiérarchique [Agard et al., 2006], les réseaux de neurones [Ma et al., 2013], le DBSCAN (Density-based spatial clustering of applications with noise) [Kieu et al., 2014], etc.

Des travaux précédents ont démontré la nécessité de raffiner les méthodes de traitement de données provenant de systèmes de perception par cartes à puce, tant pour la fouille de données temporelle [Ghaemi et al., 2017] que sur la fouille de données spatiales [Ghaemi et al., 2015]. En effet, dans le cas des méthodes de classification, le nombre d'observations, qui atteint plusieurs centaines de milliers, est beaucoup trop élevé pour que les données puissent être traitées directement par les outils de classification. Il faut donc proposer une méthode basée sur un échantillonnage. Cette communication présente une telle méthode, et discute des niveaux d'échantillonnage requis pour en arriver à des résultats satisfaisants, appliqués à des données de cartes à puce, qui ont leurs caractéristiques propres.

Dans cette recherche, les comportements des utilisateurs de transport collectif sont traités comme une série temporelle de localisations spatiales découlant des transactions de cartes à

puce. Pour ce faire, une méthode générale de classification de séries temporelles a été proposée en vue de mieux comprendre les comportements temporels de différents groupes [He et al., 2018]. Ici, nous proposons un raffinement de cette méthode, en y ajoutant de l'échantillonnage de données.

Les transactions de cartes à puces sont caractérisées par un numéro de carte (encrypté, anonyme), une date et une heure de transaction, ainsi que l'arrêt d'autobus ou la station de métro où s'est effectuée la transaction. De ces transactions, nous dérivons un vecteur quotidien de 24 valeurs qui représentent l'utilisation (approche temporelle) ou la localisation (approche spatio-temporelle) de la carte à chacune des heures de la journée. Ce vecteur est obtenu à partir d'un traitement séquentiel des transactions effectuées par une carte, chaque arrêt de montée faisant foi de l'utilisation ou de la localisation de la carte lors des heures précédentes, depuis la dernière transaction. Donc, pour chaque carte, nous obtenons une série de « cartes-jours », chacune étant un point d'observation pour notre méthode de classification. Cela permet de classer ces « cartes-jours » dans des groupes montrant des comportements similaires.

La Fig. 1 montre l'intégration de l'échantillonnage dans la méthode de classification. Au début, toutes les observations sont présentes (Fig. 1a). Des observations sont choisies au hasard pour constituer l'échantillon, sur la Fig. 1b). Ces points sont utilisés pour créer des groupes ayant des comportements similaires (Fig. 1c). Nous calculons ensuite la distance entre tous les autres points et ces groupes afin de les attribuer au groupe le plus proche (Fig. 1d). Un des défis ici est de déterminer quelle est la taille d'échantillon qui permet d'obtenir des groupes significatifs, et ce sans trop demander de ressources de calcul.

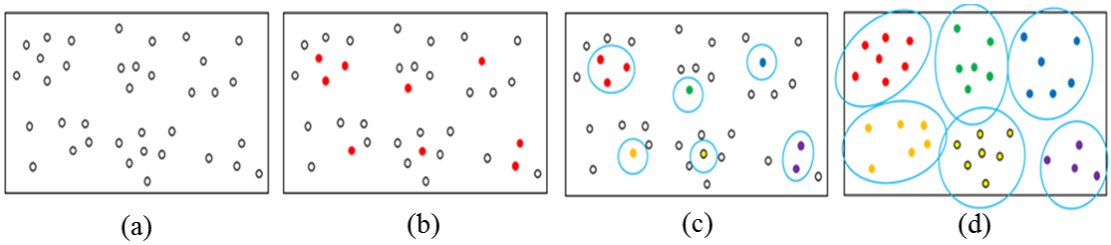


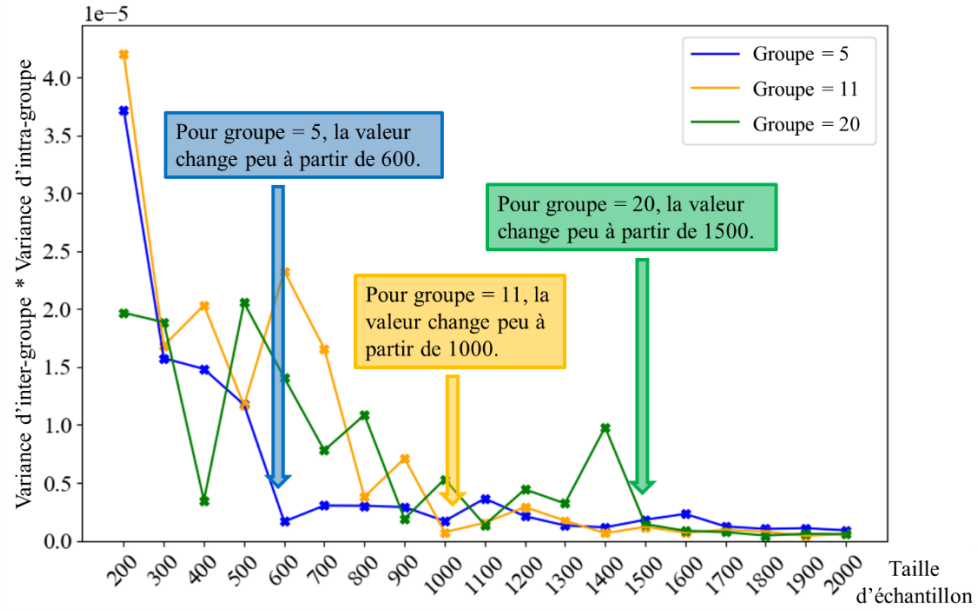
Fig. 1 Méthode d'échantillonnage

Nous proposons une méthode permettant de mesurer l'efficacité de diverses tailles d'échantillons en fonction du nombre de groupes. Cette approche préconise l'utilisation d'indicateurs tels que la variance des distances entre les groupes (dissimilarité intergroupes) et la variance des distances entre les individus à l'intérieur de chaque groupe (intragroupes). Les indicateurs sont obtenus sur la base de 10 tirages d'échantillons. Idéalement, les deux indicateurs doivent être les plus bas possibles, c'est-à-dire que l'échantillon doit être suffisamment grand pour que le tirage aléatoire ne vienne pas perturber le choix des groupes. L'objectif est de déterminer une taille d'échantillon.

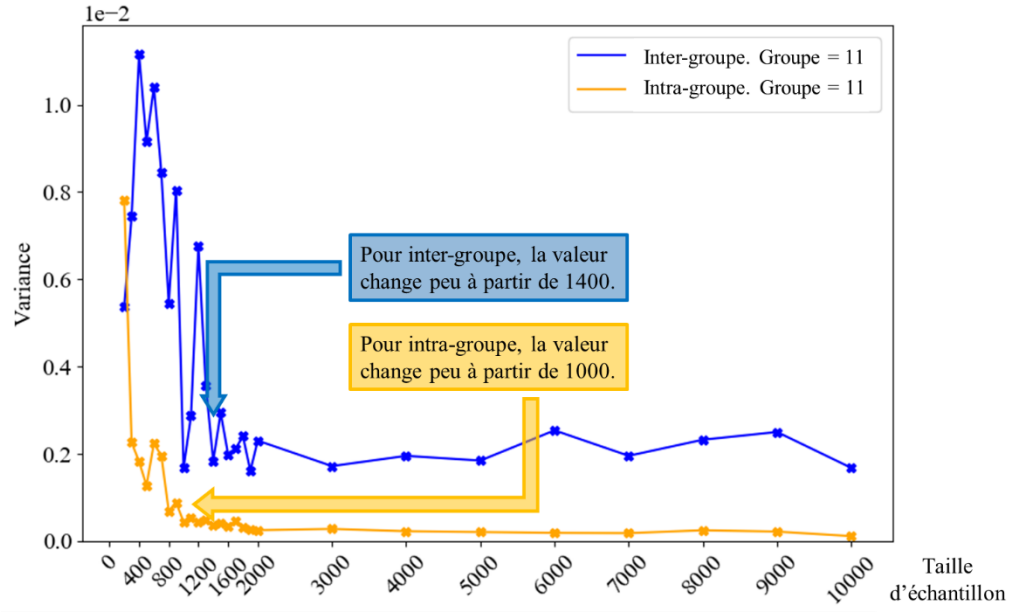
Pour expérimenter notre méthode, nous utilisons un mois de données de la Société de transport de l'Outaouais (STO), qui dessert la ville de Gatineau, au Québec (300 000 habitants). Le jeu de données contient environ 950 000 transactions de cartes à puces réalisées par environ 30 000 cartes, ce qui donne un total de 753 501 vecteurs « cartes-jours ».

Voici les étapes de l'approche proposée :

1) Pour une taille d'échantillon donnée (supposons 1000) et un nombre de groupe donné (supposons 11), prendre un échantillon de 1000 observations et les classer en 11 groupes.
2) Calculer la distance inter-groupe et intra-groupe.
3) Répéter 10 fois les étapes 1 et 2.
4) À partir des résultats des 10 essais, calculer la variance de distance intergroupe et celle intragroupe. Ensuite, calculer le produit de ces deux variances.
5) Répéter les étapes 1 à 4 pour plusieurs d'échantillon.
6) Répéter les étapes 1 à 5 pour quelques nombres de groupes.
7) Analyser les résultats.



(a)



(b)

Fig. 2 Mesure de l'efficacité de méthode d'échantillonnage (classification temporelle)

La Fig. 2 présente les résultats appliqués à la classification temporelle des comportements des usagers pour le cas d'étude. La figure présente une certaine stabilisation autour d'une taille

d'échantillon de 1000, et ce même si le nombre d'observations est plus de 750 000. Ce taux à 0,1% semble très faible, mais il est suffisant, probablement à cause de la nature des données, qui représentent des comportements d'usagers qui ne sont pas si différents que cela. Finalement : les horizons temporel et spatial des usagers sont contraints par le territoire et par les heures d'opération du réseau. Une autre constatation : un nombre de groupes plus élevé requiert une plus grande taille d'échantillon avant de se stabiliser, ce qui est intuitif. La stabilisation intragroupe semble survenir pour des tailles d'échantillon plus basses que pour la variance de la distance entre les groupes.

Dans la communication, nous allons également présenter les résultats en ce qui concerne la classification spatio-temporelle, tenant compte de l'espace et du temps.

Références

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

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.

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.

He, L., Trépanier, M., & Agard B. (2017). Evaluating the impacts on users' temporal patterns of a bus-rapid transit 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.)

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

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.

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.