EPI sciences

Les Cahiers Scientifiques du Transport / Scientific Papers in Transportation

Développement d'indicateurs de mesure de la variabilité d'utilisation du transport en commun à partir de données de cartes à puce

Élodie Deschaintres

Polytechnique Montréal, Département des génies civil, géologique et des mines

Catherine Morency

Polytechnique Montréal, Département des génies civil, géologique et des mines ; Chaire Mobilité, CRC Mobilité des personnes, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport (CIRRELT)

Martin Trépanier

Polytechnique Montréal, Département de mathématiques et de génie industriel ; Chaire Mobilité, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport (CIRRELT)

L'achalandage du transport en commun varie dans le temps, au sein du comportement d'un même usager, mais aussi d'un usager à l'autre. Ces variations rendent difficile l'ajustement des services et des modèles de prévision de la demande, conduisant potentiellement à des coûts d'opération supplémentaires et à une affectation non optimale des véhicules sur le réseau. Toutefois, la disponibilité croissante de données longitudinales et individualisées permet désormais de mieux comprendre la variabilité des comportements de déplacement. Cet article bénéficie ainsi de données de cartes à puce provenant du système OPUS de Montréal, Canada. Plus de 429 millions de validations, réalisées par près de 2 millions de cartes, sont exploitées pour étudier la variabilité d'utilisation du transport en commun à un niveau totalement désagrégé sur une période d'un an. Quatre indicateurs sont proposés afin de mesurer plusieurs types de variations : la dispersion des déplacements parmi les usagers, la variabilité de la fréquence d'utilisation, la variance temporelle du nombre de déplacements par mois et la diversité spatiale des lieux d'embarquement. Ces indicateurs sont appliqués pour comparer la régularité de dix groupes de cartes distincts définis en fonction de leur composition tarifaire. Face aux limites des tests statistiques traditionnels, sensibles à la taille de l'échantillon, la notion de taille d'effet est introduite pour mieux quantifier l'importance des différences observées entre les groupes. Les résultats révèlent qu'il existe une relation entre l'utilisation du transport en commun et le type de titre utilisé. Les utilisateurs d'abonnements annuels ou mensuels sont en moyenne très réguliers et fréquents, alors que les utilisateurs de carnets de tickets sont des usagers plus occasionnels. De plus, la variabilité d'usage tend à augmenter avec le nombre de titres différents utilisés durant l'année.

Mots-clés : transport en commun, cartes à puce, variabilité, indicateurs, taille d'effet

Deschaintres, É., Morency, C. et Trépanier, M. (2021). Développement d'indicateurs de mesure de la variabilité d'utilisation du transport en commun à partir de données de cartes à puce. *Les Cahiers scientifiques du transport,* n° 76-77, p. 99-136, https://doi.org/10.46298/cst-9845.

Development of Indicators to Measure Transit Use Variability Based on Smart Card Data

Transit ridership varies over time, within the behaviour of a same user, and from one user to another. These variations make it difficult to adjust service and demand forecasting models, potentially leading to additional operating costs and a non-optimal allocation of vehicles on the network. However, the growing availability of longitudinal and individualized data now allows to better understand travel behaviour. In particular, this paper benefits from smart card data from the OPUS system of Montreal, Canada. More than 429 million validations, made by nearly 2 million cards, are mined to investigate transit use variability at a totally disaggregated level over a one-year period. Four indicators are proposed to measure several types of variations: trip dispersion among the users, variability of the frequency of use, temporal variance of the monthly number of trips and spatial diversity of the boarding locations. These indicators are applied to evaluate the regularity of 10 distinct groups of cards defined by their fare composition. Because of the sensitivity of statistical tests on sample size, an effect size analysis is provided to better quantify the magnitude of the differences between the groups. The results reveal relationships between transit use and the type of product used. Both annual and monthly pass users are found to be regular and frequent passengers, whereas ticket book users are rather occasional travellers. Moreover, variability tends to increase with the number of different products used during the year.

Keywords: public transit, smart cards, variability, indicators, effect size

Classification JEL : C55, L91, N70, O18, R41, R42, R48

Du fait de leur commodité à la fois pour les clients et les opérateurs, les systèmes automatisés de perception des tarifs par carte à puce (*SCFAFC - Smart Card Automated Fare Collection Systems* en anglais) sont de plus en plus utilisés dans les réseaux de transport en commun. Ils permettent notamment de réduire le temps d'embarquement des passagers, la durée d'immobilisation des véhicules aux arrêts et la charge de travail des conducteurs (Chira-Chavala et Coifman, 1996). Ces systèmes reçoivent également beaucoup d'intérêt de la part des chercheurs et des planificateurs car ils permettent de collecter de grandes quantités de données longitudinales et individualisées.

Ces gros ensembles de données peuvent notamment être exploités pour mieux comprendre les variations continues de la demande de transport à long terme. En effet, ils permettent de suivre l'utilisation du transport en commun dans le temps et dans l'espace à un niveau totalement désagrégé (chaque carte étant associée à un identifiant unique). L'étude spatio-temporelle des comportements de mobilité contribue à améliorer la modélisation de la demande afin d'obtenir des prévisions d'achalandage plus précises. Une plus grande prise en compte des fluctuations de l'utilisation du transport en commun dans les modèles permettrait notamment d'optimiser l'adéquation offre/demande, réduisant ainsi les coûts d'opération et rendant plus efficace l'affection des véhicules sur le réseau. L'analyse désagrégée des déplacements en transport en commun peut également être utile dans un objectif de personnalisation des services : de nouveaux titres de transport peuvent être créés pour mieux répondre à certains types de comportements.

Dans cette perspective, cet article tire profit d'une année entière (2016) de données de cartes à puce provenant de la Société de Transport de Montréal (STM), Canada. Le système de perception tarifaire montréalais, appelé le système OPUS, a été implanté dans toute la région métropolitaine en 2008 pour desservir plus de 4 millions d'habitants (17 % d'entre eux utilisant le transport en commun au moins une fois par jour d'après l'enquête Origine-Destination de 2013). Plusieurs compagnies opèrent ce système mais la STM régit la partie centrale de la région. Cette entreprise publique gère ainsi 4 lignes de métro, d'une longueur totale de 71 kilomètres, ainsi qu'un réseau de bus composé d'environ 220 lignes régulières réparties sur une superficie de 500 km^2. Approximativement 1,3 million de déplacements sont enregistrés chaque jour par ce système. La carte OPUS, adoptée par plus de 90 % des usagers du transport en commun, fournit donc un bon portait de l'utilisation du réseau.

S'appuyant sur ces données, cet article présente un cadre méthodologique et des outils pour mesurer la variabilité d'utilisation du transport en commun à un niveau totalement désagrégé (Deschaintres, 2018). La méthode proposée est mise en application à partir de la politique tarifaire du transport en commun de Montréal afin d'étudier la relation entre la variabilité d'usage de ce mode et sa tarification. Une segmentation tarifaire est d'abord réalisée afin de construire dix blocs de cartes en fonction du nombre et des types de titres de transport utilisés pendant l'année. Différents indicateurs, couvrant plusieurs dimensions de la mobilité individuelle, sont ensuite calculés afin de quantifier quatre types de variabilités dans chaque groupe de cartes. La variabilité d'usage du réseau à moyen-long terme (sur une année entière) est ainsi comparée entre les différents groupes. Les différences observées entre les groupes sont finalement vérifiées par des tests statistiques et des indices de taille d'effet. La méthode de validation statistique proposée, originale dans le domaine du transport, permet de se libérer de l'influence de la taille de l'échantillon et a donc un fort potentiel d'application dans l'ère du Big Data.

Cet article est composé de plusieurs sections. Tout d'abord, une revue synthétique de la littérature scientifique permet d'introduire quelques notions liées aux données de cartes à puce et à la mesure de la variabilité des comportements de déplacement. Le jeu de données utilisé est ensuite décrit, puis la méthodologie et les indicateurs proposés sont définis. Les résultats de la segmentation tarifaire des cartes et de l'application des indicateurs sont alors rapportés, analysés et supportés par des tests statistiques et des analyses de taille d'effet. En conclusion, les résultats et les perspectives de recherche sont discutées.

Revue de littérature

Utilité des données de cartes à puce en transport

Au-delà de leurs fonctions primaires de collecte des revenus et de prévention de la fraude, les systèmes par cartes à puce fournissent de grandes quantités de données pour les opérateurs et les chercheurs. Ces données ont plusieurs avantages sur les données des enquêtes traditionnelles. Premièrement, elles sont disponibles sur de plus longues périodes de temps puisque leur collecte n'est limitée ni par l'équipement, ni par la durée de vie d'une quelconque batterie, ni par la fatigue du répondant (Spurr et al., 2015). De plus, elles contiennent des informations plus précises au niveau temporel, mais aussi au niveau spatial si les lieux de validation (embarquement et/ou débarquement) sont géolocalisés (El Mahrsi et al., 2017). Une fois le système bien installé, elles sont également moins coûteuses à recueillir (Trépanier, 2012) et leur collecte passive permet de supprimer les biais de déclaration des personnes enquêtées (Bagchi et White, 2004).

Toutefois, ces données sont souvent dites partielles car certaines informations (entre autres, le motif de déplacement, la destination et l'origine exactes, les temps d'accès, les propriétés du voyageur, etc.) sont manquantes (Spurr et al., 2015). C'est pourquoi Bagchi et White (2004) ont écrit qu'elles devaient compléter et non remplacer les données des enquêtes traditionnelles. La protection de la vie privée est un autre enjeu qui rend difficile la mise en correspondance de ces données avec les caractéristiques sociodémographiques des individus (El Mahrsi et al., 2017 ; Trépanier, 2012). En outre, la qualité de ces données peut être altérée par des pannes d'équipements ou des erreurs de manipulation (White et al., 2010).

Néanmoins, de nombreux travaux ont été réalisés afin de dépasser ces limites. Plusieurs ont été recensés par Pelletier et al. (2011). Les auteurs décrivent l'utilisation des données de cartes à puce en transport en commun à trois niveaux : stratégique (planification à long terme), tactique (ajustement des services et développement du réseau) et opérationnel (statistiques d'achalandage et indicateurs de performance). Le sujet développé dans cet article fait partie des études longitudinales sur les comportements de mobilité ; il relève donc à la fois des niveaux stratégique et tactique. En effet, il s'agit ici de mieux comprendre l'utilisation du transport en commun dans le but de planifier à long terme et de mieux prédire la demande, pour ensuite améliorer l'ajustement de l'offre à cette demande et développer des services en fonction des besoins identifiés.

Variations des comportements de mobilité

L'analyse longitudinale des comportements d'utilisation des systèmes de transport permet notamment de déceler leur variabilité. D'une part, il existe des variations entre les usagers (variabilité

interpersonnelle). Cette variabilité s'explique par la grande diversité des individus, qui possèdent chacun leurs propres habitudes de déplacement. D'autre part, des variations sont visibles au sein du comportement de chaque usager (variabilité intrapersonnelle), qui évolue dans le temps en fonction de ses contraintes quotidiennes ou d'événements ponctuels. Par ailleurs, ces deux types de variabilité peuvent être définies selon plusieurs attributs, principalement d'ordre temporel ou spatial. En effet, on observe principalement des variations individuelles dans la fréquence d'utilisation, la plage horaire ou le lieu d'embarquement des usagers (Morency et al., 2007).

À titre d'exemple, Morency et al. (2006) ont pointé des différences significatives entre deux cartes à puce en termes de nombre d'embarquements par jour et d'heure d'embarquement. Inversement, les auteurs ont mis en évidence une certaine régularité intrapersonnelle puisque, après application d'un algorithme de classification, les mêmes jours de la même carte se sont retrouvés concentrés dans les mêmes groupes de journées types. La régularité spatiale des comportements de mobilité a aussi été démontrée par Barabási et al. (2008). En étudiant les trajectoires de 100 000 utilisateurs anonymes de téléphones cellulaires, les chercheurs ont trouvé une forte probabilité de retourner aux mêmes endroits. Cependant, à l'aide de données de cartes à puce, Nishiuchi et al. (2013) ont construit des profils spatio-temporels à une échelle individuelle et ont révélé que les comportements pouvaient varier selon le type de passagers analysés.

Mesure de la variabilité des comportements – indicateurs

La mesure de la variabilité (ou à l'inverse, de la régularité) des comportements individuels de mobilité est un sujet de recherche exploré depuis de nombreuses années. Jones et Clarke (1988) sont parmi les premiers à avoir reconnu la nécessité d'examiner les habitudes de déplacement des usagers sur plusieurs jours plutôt que sur un jour typique. Ils ont ainsi développé différentes méthodes (graphiques et numériques) pour évaluer ces variations, mais celles-ci se sont avérées difficiles à automatiser. Par la suite, de nombreux indicateurs ont été proposés dans la littérature. Des auteurs ont parfois associé les notions de *régularité* et d'*intensité* d'utilisation du transport en commun, suggérant que les usagers les plus réguliers sont aussi les plus fréquents et les plus actifs sur le réseau. Des indicateurs comme le taux de mobilité ou d'activité des usagers ont alors été calculés (Morency et al., 2006).

Des indicateurs basés sur la fréquence des déplacements faits aux mêmes arrêts et aux mêmes moments de la journée ont également été souvent estimés (Huang et al., 2015 ; Morency et al., 2007). On ne s'intéresse plus ici à la fréquence de tous les déplacements de l'usager, mais seulement des déplacements qui sont similaires. Il est en effet assez courant de définir la régularité de la mobilité par la répétition des mêmes attributs de déplacement. Manley et al. (2016) ont par exemple utilisé l'algorithme de partitionnement DBSCAN (*density-based spatial clustering of applications with noise*)

pour déterminer des groupes d'heures habituelles de déplacement pour chaque individu et ils ont ensuite évalué leur régularité par la proportion des déplacements appartenant à ces groupes. De même, Raux et al. (2012) ont mesuré la régularité des comportements en identifiant des nœuds (*core stops*), définis comme des « déplacements, classés selon les caractéristiques de quatre attributs (activité, mode, heure d'arrivée, lieu), se produisant au moins trois jours différents de la semaine » [notre traduction]. En outre, Schlich et Axhausen (2003) ont recensé des indicateurs calculés à partir d'une matrice de contingence mesurant la co-occurrence des mêmes caractéristiques de déplacement.

De manière plus anecdotique, d'autres familles d'indicateurs peuvent être repérées dans la littérature. Des indicateurs basés sur la périodicité des comportements, souvent mesurée à l'aide de la transformation de Fourrier (Eagle et Pentland, 2006 ; Kim et Kotz, 2007), sont parfois appliqués afin d'identifier des périodes de temps typiques entre deux occurrences d'un même événement. Williams et al. (2012) ont par exemple évalué la variabilité de l'intervalle de temps entre deux visites consécutives au même endroit. Des indicateurs basés sur l'allocation du temps (Schlich et Axhausen, 2003) ou sur la diversité spatiale des lieux visités, cette dernière étant généralement mesurée par le nombre d'arrêts différents empruntés (Morency et al., 2007) ou par un indice d'entropie (Briand et al., 2017), sont d'autres types d'indicateurs pouvant être cités.

Par ailleurs, la mesure de la variabilité des comportements peut également reposer sur le calcul d'une variance. Très tôt, Pas et Koppelman (1987) ont capturé la variabilité individuelle quotidienne de différents segments de la population par l'estimation d'une variance intrapersonnelle moyenne du nombre de déplacements par jour par individu. En utilisant des prismes spatio-temporels, Kitamura et al. (2006) ont quant à eux modélisé la variance de l'heure de départ du premier déplacement de la journée. Ils ont aussi décomposé cette variance en plusieurs parties, en différenciant d'abord les variations systématiques dues aux contraintes de temps et les variations aléatoires, puis en les redissociant en une variabilité intrapersonnelle et une variabilité interpersonnelle. De manière similaire, Raux et al. (2016) se sont appuyés sur la variance de trois attributs de mobilité et l'ont divisée en une variance interpersonnelle et une variance intrapersonnelle. Les trois attributs considérés par les auteurs sont la fréquence des déplacements par jour, l'utilisation quotidienne du temps et la séquence d'activités journalières.

Mesure de la variabilité des comportements – autres méthodes

L'analyse de séquences d'activités ou de lieux visités est généralement réalisée grâce à la méthode d'alignement des séquences (*SAM - Sequential Alignment Method* en anglais), qui permet de prendre en compte l'ordre des événements. Les travaux de Wilson (1998), Joh et al. (2002) ou encore ceux de Moiseeva et al. (2014) peuvent être consultés pour avoir plus de détails et des exemples d'application

de cette méthode dans le cadre de la mesure de la variabilité des comportements de mobilité. La variabilité interpersonnelle est explorée en comparant les séquences de plusieurs individus, tandis que la variabilité intrapersonnelle est étudiée en comparant les séquences d'un seul utilisateur. Une autre manière d'inclure le caractère séquentiel des déplacements est de calculer un taux d'entropie comme l'ont fait Goulet-Langlois et al. (2017). Leurs résultats confirment que la fréquence et l'ordre des événements sont essentiels pour aborder la question de la variabilité d'usage en transport.

De plus, la segmentation des individus en plusieurs groupes d'usagers ayant un comportement similaire peut servir à mieux comprendre la variabilité interpersonnelle. En effet, cette méthode permet d'identifier des différences entre les utilisateurs et de les synthétiser en un nombre limité de groupes de comportements typiques. La richesse des données de cartes à puce permet de décrire la mobilité individuelle de multiples façons et de nombreux algorithmes peuvent être appliqués. Parmi les plus typiques se trouvent celui des K-moyennes, l'algorithme hiérarchique (Agard et al., 2006) ou encore le DBSCAN (Kieu et al., 2014). Ma et al. (2013) ont quant à eux développé une méthode plus complexe basée sur les réseaux de neurones, tandis que Briand et al. (2017) ont présenté un modèle de mélanges gaussiens qui permet une représentation continue du temps. Par ailleurs, Joh et Timmermans (2011) ont proposé une approche heuristique pour appliquer la méthode SAM dans un contexte de classification à de larges bases de données comme celles des données de cartes à puce.

Enfin, la variabilité des comportements de mobilité peut être modélisée, indirectement par ses effets sur les attributs de déplacement dans un modèle d'équations structurelles (Roorda et Ruiz, 2008), ou directement en tant que variable dépendante dissociée en plusieurs composantes dans un modèle multiniveaux (Xianyu et al., 2017).

Mesure de la variabilité des comportements – enjeux et limites

En résumé, il existe de nombreuses méthodes pour étudier la variabilité des comportements de mobilité. La principale difficulté rencontrée pour réaliser des études comme celles susmentionnées réside dans la collecte de données longitudinales et individualisées. Les enquêtes de déplacements longue durée et les carnets de déplacements tels que Mobidrive (Kitamura et al., 2006 ; Schlich et Axhausen, 2003) ont ces propriétés, mais le coût et le fardeau de telles collectes conduisent généralement à des échantillons de petite taille. Pourtant, d'après les résultats de Schlich et Axhausen (2003), l'analyse de la variabilité des comportements requiert des données longitudinales s'étendant sur au moins deux semaines. C'est pourquoi les données de cartes à puce, exploitées dans cet article, sont grandement appropriées. Malgré leur disponibilité croissante, ces données ont jusqu'alors été peu exploitées dans la mesure de la variabilité des comportements à l'aide d'indicateurs. Leur grande taille est notamment un important défi à relever.

Par ailleurs, les études décrites précédemment se sont principalement concentrées sur la variabilité quotidienne des comportements alors que des variations pourraient aussi être visibles à d'autres niveaux, par exemple à l'échelle hebdomadaire ou mensuelle. Ainsi, l'originalité de cet article se trouve également dans l'analyse d'une plus longue période de données ; ce travail décrit la variabilité d'utilisation du transport en commun sur une année entière avec une résolution mensuelle. Cette résolution implique des variations d'usage visibles à moyen-long terme à l'échelle du mois, incluant des variations dues à des facteurs externes (périodes de congé) ou à d'autres phénomènes saisonniers. De plus, aucun auteur n'a jusqu'alors fourni un lot d'indicateurs simples et généralisables pour caractériser différents types de changements (impliquant différentes dimensions) dans les comportements de mobilité. C'est là le but de cet article : proposer des indicateurs reproductibles qui pourront être calculés pour comparer différentes années, villes ou groupes d'utilisateurs à partir de données opérationnelles. En guise d'exemple, des groupes de cartes définis en fonction de leur composition tarifaire seront comparés dans la suite de ce papier. En outre, cet article présente également une procédure de validation statistique des indicateurs développés.

Données et méthodologie

Échantillon de données

L'échantillon de données exploité dans cet article est issu de la base de données transactionnelle de la STM. Il couvre une période d'un an, du 1er janvier 2016 au 31 décembre 2016, et correspond aux validations des cartes à puce OPUS (soit près de 90 % des validations). En quelques chiffres, cet extrait représente 429 millions de validations réalisées par presque 2 millions de cartes. Il est important de préciser ici que deux principaux types de cartes OPUS existent à la STM : des cartes personnalisées, propres à une personne particulière, et des cartes anonymes, qui peuvent être prêtées. Ces dernières peuvent donc être utilisées par plusieurs individus. À l'inverse, deux cartes peuvent être associées à un même usager si celui-ci a renouvelé sa carte au cours de l'année. Par conséquent, des cartes et non des individus sont étudiées dans cet article. Toutes les cartes ayant été validées au moins une fois durant la période d'étude sont incluses dans l'analyse. Le but étant de dresser un portrait exhaustif des usagers (réguliers et occasionnels) de la STM en 2016, aucun filtrage n'a été appliqué en amont afin de supprimer les utilisateurs peu actifs sur le réseau.

Chaque observation contient les informations listées dans le Tableau 1 ci-dessous, c'est-à-dire un identifiant de validation, un identifiant (unique et anonyme) de carte, le code du titre de transport utilisé, l'horodatage de la transaction (date et heure) ainsi que des informations partielles sur le lieu où la carte a été validée (soit un numéro de station pour le métro, soit un numéro de véhicule et de ligne

pour le bus). Aucune information personnelle sur l'usager n'est disponible. De plus, ces informations ne sont rapportées qu'à l'embarquement car aucune validation n'est requise à la sortie du réseau montréalais. La direction du bus est également connue mais peu fiable en raison de possibles erreurs de manipulation des chauffeurs.

Tableau 1 : Information recueillie à chaque validation tarifaire par mode (à l'embarquement)

Information	Métro	Bus
ID validation	√	√
ID carte (anonymisé)	√	√
Code titre	√	√
Date, heure (AA/MM/JJ HH:MM:SS)	√	√
Station/arrêt	√	X
Véhicule	X	√
Ligne	X	√
Direction	X	?

Par ailleurs, ces validations ont été converties en déplacements afin de travailler avec la fréquence réelle d'utilisation du transport en commun. De cette manière, les validations de correspondance entre des lignes ou des modes ne sont pas considérées comme des déplacements à part entière et les détenteurs de cartes à puce ayant effectué beaucoup de correspondances ne seront pas traités comme des utilisateurs plus fréquents du système. Pour réaliser cette transformation, des règles d'affaires basées sur la politique tarifaire de la STM ont été appliquées. Ainsi, une colonne a été ajoutée à la base de données initiale pour spécifier si chaque validation était une première montée ou une correspondance. Seuls les premiers embarquements ont ensuite été sélectionnés pour comptabiliser les déplacements associés à chaque carte. Les résultats de cette conversion (ainsi que toute autre information impliquant des nombres de déplacements) ne seront pas fournis dans cet article pour des raisons de confidentialité.

Cadre général

Le cadre méthodologique de cet article est illustré à la Figure 1. L'idée directrice est de mettre en évidence la variabilité d'utilisation du transport en commun à l'échelle désagrégée en fonction des titres de transport utilisés par les détenteurs de carte durant l'année. Ainsi, on suppose que l'usage de différents titres conduit à différents comportements de mobilité. Cette hypothèse sera validée par les résultats obtenus. La terminologie employée est précisée par l'équation suivante :

$$1\ titre\ de\ transport = 1\ support \times 1\ produit \times 1\ tarif \qquad \text{(Éq. 1)}$$

Un titre de transport est défini comme la combinaison d'un support (ticket papier ou carte à puce), d'un produit (billets unitaires, carnets de tickets, abonnement, etc.) et d'un tarif (ordinaire, réduit, etc.). Environ 2 000 combinaisons sont possibles dans la politique de tarification du transport en commun à Montréal. Cependant, la carte à puce est le seul support considéré ici. Pour simplifier l'analyse, les produits ont été regroupés en six catégories et les tarifs en cinq catégories : celles-ci sont énumérées dans la Figure 1. Ainsi, cet article se limite à trente combinaisons possibles. Nous n'expliquerons pas complètement ici la politique de tarification du transport en commun de Montréal car celle-ci n'est qu'un prétexte pour fournir un exemple d'application de la méthode proposée. Toutefois, le lecteur pourra trouver des informations complémentaires à ce sujet dans le mémoire de Deschaintres (2018).

Plusieurs types de variations peuvent être observées à l'aide des données de cartes à puce. Les différentes dimensions choisies pour étudier la variabilité d'utilisation du transport en commun sont l'intensité, la fréquence ainsi que les attributs temporels et spatiaux de cet usage. Ces quatre dimensions correspondent aux grandes classes de caractéristiques des déplacements les plus couramment analysées dans le domaine du transport. Plus spécifiquement, la distribution des déplacements parmi les différents détenteurs de cartes peut révéler une certaine hétérogénéité dans l'intensité annuelle d'utilisation du transport en commun : la majorité des déplacements de 2016 ont été réalisés par un plus ou moins petit nombre d'usagers. De même, la variabilité de la fréquence d'utilisation du transport en commun sur un an peut être déterminée par l'analyse du nombre de déplacements par mois de chaque usager. L'analyse longitudinale de ce même nombre à travers l'année peut permettre de mesurer des variations temporelles. La variabilité spatiale peut quant à elle être représentée par la diversité des lieux visités durant l'année. Cette analyse spatiale sera néanmoins partielle puisque seuls les lieux d'embarquement sont rapportés dans les données disponibles.

L'objectif de cet article est de proposer un indicateur pour quantifier chacun de ces quatre types de variations. Pour ce faire, les indices de Pareto et de Gini, des mesures statistiques (mode, variance, etc.) et l'entropie de Shannon seront utilisés. Plus de détails sur ces indicateurs sont fournis dans la section suivante. Par ailleurs, ces différents types de fluctuations peuvent être caractérisés selon deux points de vue : interpersonnel (variations entre les usagers) ou intrapersonnel (variations du comportement d'un même usager). Certains indicateurs seront donc dissociés en deux composantes.

Variabilité d'utilisation du transport en commun à l'échelle désagrégée : application à la politique de tarification de Montréal

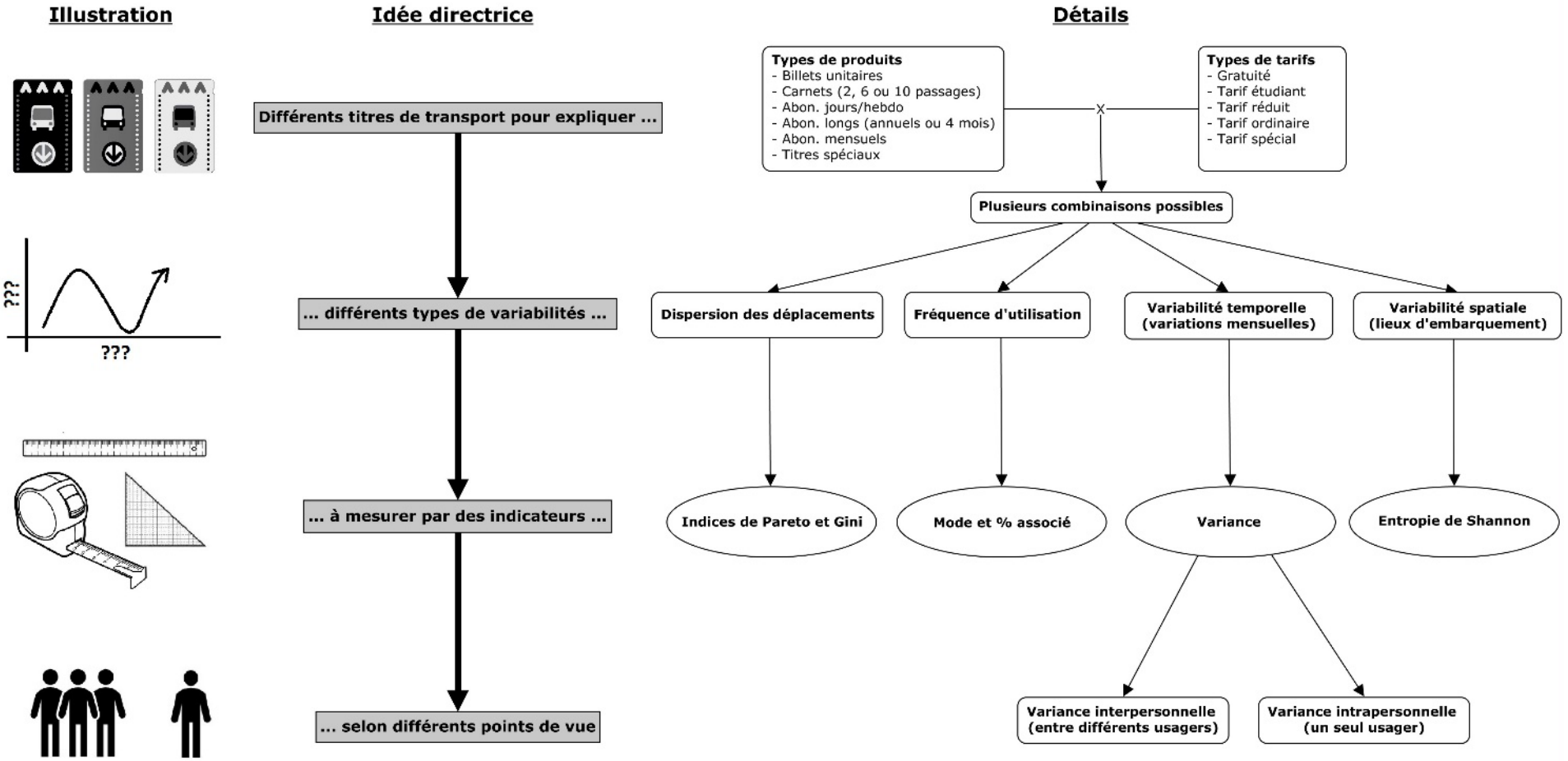

Figure 1 : Cadre méthodologique du présent article

Indicateurs de variabilité

En vue de quantifier la variabilité d'utilisation du transport en commun au niveau désagrégé, un indicateur est associé à chacun des quatre types de variations mentionnées précédemment. Les indicateurs proposés sont construits sur une base mensuelle ou annuelle pour mesurer la variabilité des usagers à travers toute l'année 2016. Toutefois, les indicateurs proposés pourraient aussi être adaptés à d'autres rythmes temporels (journée, semaine, ou saison) en fonction des données disponibles et du type d'analyses souhaitées (à court, moyen ou long terme).

Dispersion des déplacements parmi les usagers

L'indice de Pareto, noté α, est utilisé pour évaluer l'homogénéité de la dispersion des déplacements parmi les usagers. L'emploi de cet indicateur, commun en économie, est ici transposé au transport en distribuant des nombres de déplacements plutôt que des revenus annuels.

La définition de cet indicateur est basée sur les courbes de Lorenz. Deux courbes doivent être distinguées : la courbe de Lorenz L et la courbe complémentaire de Lorenz $\overline{L}$. Celles-ci sont dessinées sur la Figure 2 pour illustrer la distribution des nombres de déplacements réalisés par chaque détenteur de carte OPUS en 2016. La courbe de Lorenz représente la distribution cumulative des intensités annuelles de déplacement (en ordonnée) en fonction de la distribution cumulative des cartes (en abscisse), les intensités étant triées dans l'ordre croissant. La courbe complémentaire est similaire avec les intensités rangées cette fois-ci dans l'ordre décroissant. Le cas où les deux courbes sont sur la ligne $y = x$ correspond au cas d'égalité parfaite entre tous les usagers (c'est-à-dire lorsqu'ils font tous le même nombre de déplacements par an).

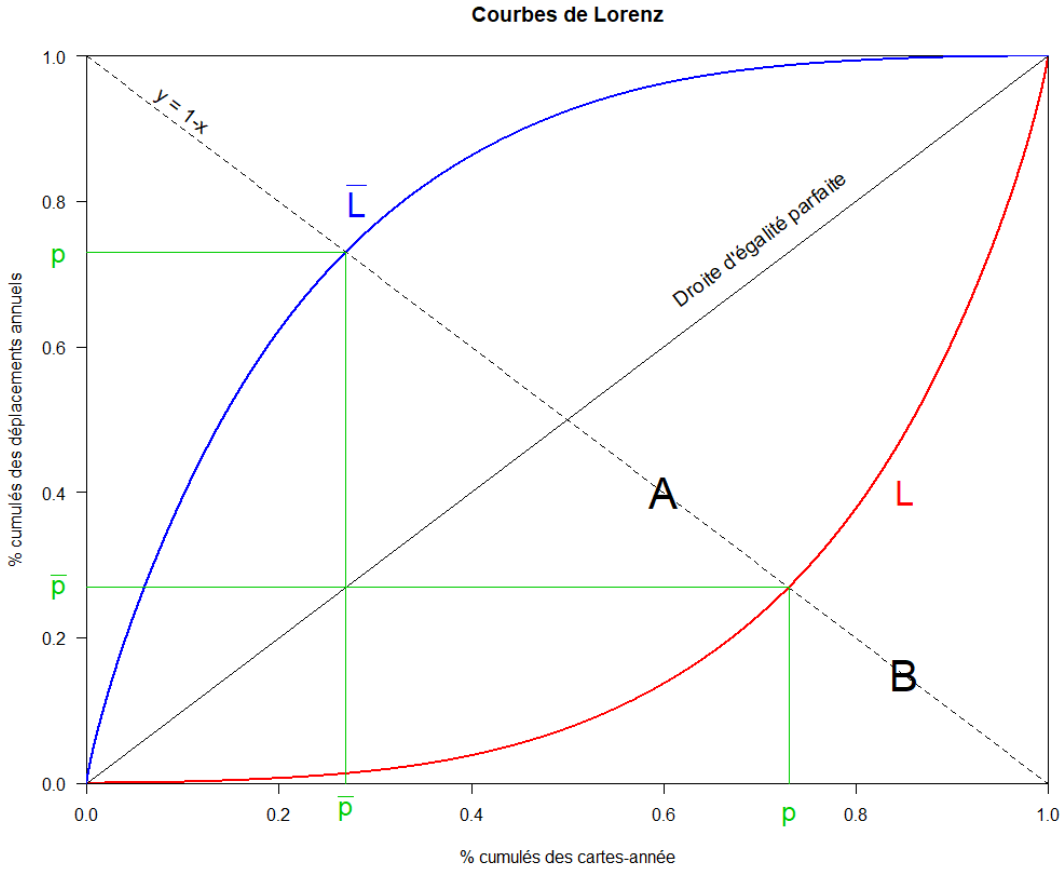

Figure 2 : Courbes de Lorenz et proportions de Pareto appliquées au nombre annuel de déplacements par carte

Ces courbes permettent de déterminer les proportions de Pareto $(p, \bar{p})$ qui sont également positionnées sur la Figure 2. La proportion p, parfois appelée indice k (Ghosh et al., 2014; Inoue et al., 2015), est définie comme le point d'intersection entre la courbe de Lorenz $L(x)$ et la droite $y = 1 - x$. La proportion complémentaire $\bar{p} = L(p)$ est l'ordonnée correspondante, de sorte que $\bar{p} = L(p) = 1 - p$ puis $p + \bar{p} = 1$. De plus, on a $p = \bar{L}(\bar{p}) = 1 - \bar{p}$. Ces relations signifient que p % des usagers les moins fréquents ont réalisé $\bar{p}$ % des déplacements de 2016 et que $\bar{p}$ % des usagers les plus fréquents ont fait p % des déplacements de 2016.

L'indice de Pareto se calcule à partir de ces proportions comme suit :

$$\alpha = \frac{1}{1 - \frac{\log(p)}{\log(1-p)}} = \frac{1}{1 - \frac{\log(p)}{\log(\bar{p})}} = \log_{\frac{p}{\bar{p}}} \frac{1}{\bar{p}} \qquad \text{(Éq. 2)}$$

La démonstration de cette équation a été rappelée par Deschaintres (2018). Ainsi, plus l'indice de Pareto est faible, plus les déplacements sont concentrés dans un petit nombre d'usagers. Une valeur faible reflète donc une plus grande variabilité interpersonnelle de l'intensité annuelle d'utilisation du transport en commun. Inversement, une valeur haute indique une distribution plus homogène des déplacements parmi les usagers. La fameuse règle du « 80-20 » $(p = 0{,}80, \bar{p} = 0{,}20)$, aussi connue

sous le nom de principe de Pareto et selon laquelle 80% des effets proviennent de 20% des causes, conduit à $\alpha \approx 1{,}16$. Dans le cas de la distribution tracée sur la Figure 2, on a $p = 0{,}73$ et $\bar{p} = 0{,}27$ donc $\alpha \approx 1{,}32$. La répartition des intensités annuelles de déplacement analysées dans cet article est donc plus homogène que celle énoncée par le principe de Pareto.

Le coefficient de Gini est une autre mesure de dispersion statistique qui est souvent rapportée avec l'indice de Pareto (Ghosh et al., 2014; Inoue et al., 2015). Sa valeur varie entre 0 et 1 : un indice de 0 souligne une égalité parfaite entre toutes les valeurs de la distribution tandis qu'une valeur de 1 exprime une inégalité totale. Ce coefficient peut être calculé à partir des courbes de Lorenz par

$$G = \frac{A}{A+B} = 2A = 1 - 2B \qquad \text{(Éq. 3)}$$

où A et B sont les aires représentées sur la Figure 2 (A est l'aire entre la ligne d'égalité parfaite et la courbe de Lorenz L, B est l'aire entre la courbe de Lorenz L et les bords du carré unitaire 1×1).

Par ailleurs, l'indice de Pareto α et le coefficient de Gini G sont liés par l'équation suivante :

$$G = 1 - 2\int_0^1 L(F)dF = \frac{1}{2\alpha - 1} \qquad \text{(Éq. 4)}$$

Une augmentation de α entraîne donc une diminution de G.

Variabilité de la fréquence d'utilisation mensuelle

La variabilité de la fréquence d'utilisation du transport en commun est mesurée au niveau mensuel à partir de la distribution des cartes en fonction du nombre moyen de déplacements par mois actif par carte (voir la Figure 6 plus loin). Ce nombre a été calculé pour chaque carte comme le nombre moyen de déplacements par mois en considérant seulement les mois durant lesquels l'usager a réalisé au moins un déplacement, soit :

$$\overline{N}_i = \frac{1}{M}\sum_{m=1}^{12} N_{im} \qquad \text{(Éq. 5)}$$

où $\overline{N}_i$ est le nombre moyen de déplacements par mois actif de l'utilisateur de la carte i, N_{im} le nombre de déplacements réalisés avec la carte i durant le mois m et M le nombre de mois d'activité de la carte i (c'est-à-dire le nombre de mois durant lesquelles elle a été validée au moins une fois). De cette manière, seuls les M mois durant lesquels l'usager était présent sur le réseau sont considérés. Ces mois ne sont pas nécessairement continus sur l'année.

Les valeurs individuelles (par carte) obtenues ont ensuite été regroupées par classes de fréquences, de la classe n° 1 pour les plus faibles fréquences d'utilisation mensuelles à la classe n° 21 pour les plus

grandes fréquences. Le nombre de classes (21) a été choisi arbitrairement de manière à produire le graphique de la Figure 6. Les largeurs d'intervalle de ces classes sont étroites et constantes, à l'exception de la dernière classe qui regroupe les cartes avec un nombre moyen de déplacements par mois actif supérieur à un certain nombre. Les bornes de ces intervalles (en nombre moyen de déplacements par mois actif) ne sont pas précisées ici par souci de confidentialité.

Les proportions de cartes appartenant à chaque classe ont ainsi été déterminées, la classe modale étant la classe la plus fréquente (c'est-à-dire celle avec la plus grande proportion de cartes). Plus formellement, la classe modale notée c_m est telle que $p(c_m) \geq p(c)$ pour toute classe $c \neq c_m$ où $p(c)$ est la proportion de cartes dans la classe c. L'indicateur proposé est simplement la proportion $p(c_m)$ associée à cette classe modale. Plus ce pourcentage est élevé, plus il y a de cartes qui présentent les mêmes fréquences d'utilisation, et donc plus il y a d'usagers qui ont un nombre moyen de déplacements par mois actif similaire. Cet indicateur évalue ainsi la variabilité interpersonnelle de la fréquence d'utilisation mensuelle du transport en commun.

Variabilité temporelle (variations du nombre de déplacements par mois)

La variabilité temporelle est quantifiée par les variations longitudinales du nombre mensuel de déplacements réalisés par chaque utilisateur de carte au cours des douze mois de l'année. L'indicateur proposé est similaire à celui de Raux et al. (2016) : alors que ces auteurs ont calculé la variance du nombre quotidien de déplacements pour mettre en évidence des variations au niveau hebdomadaire, la variabilité au niveau annuel est ici évaluée à partir de la variance du nombre de déplacements par mois. La variance totale (*TSS: total sum of squares*) peut être décomposée en une variance intrapersonnelle (*WPSS: within-person sum of squares*) et une variance interpersonnelle (*BPSS: between-person sum of squares*) comme suit :

$$TSS = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(n_{ij} - \bar{n}\right)^2 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(n_{ij} - \bar{n}_i\right)^2 + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{n}_i - \bar{n})^2 \qquad \text{(Éq. 6)}$$

$$= WPSS + BPSS$$

où I est le nombre total de cartes, J le nombre de mois observés (= 12 ici), n_{ij} le nombre de déplacements faits par l'utilisateur de la carte i durant le mois j, $\bar{n}_i$ le nombre moyen de déplacements par mois de la carte i sur la période J, $\bar{n}$ le nombre moyen de déplacements par mois de toutes les I cartes sur la période J. Les valeurs de $\bar{n}_i$ et $\bar{n}$ sont donc obtenues par :

$$\bar{n}_i = \frac{1}{J}\sum_{j=1}^{J} n_{ij} \quad \text{et} \quad \bar{n} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \qquad \text{(Éq. 7)}$$

Dans la suite, ces différentes variances du nombre de déplacements par mois seront mesurées pour plusieurs groupes de cartes et comparées (la division par le nombre de cartes I de chaque groupe les rendant comparables). Plus la variance totale (TSS) mesurée sera faible, plus les comportements seront considérés comme réguliers au niveau temporel. Les proportions respectives de la variabilité intrapersonnelle ($\frac{BPSS}{TSS} \times 100$) et de la variabilité intrapersonnelle ($\frac{WPSS}{TSS} \times 100$) dans la variabilité totale seront également estimées.

Variabilité spatiale (diversité des lieux d'embarquements)

Finalement, la variabilité spatiale est évaluée par la variation des lieux d'embarquement choisis par l'usager pendant l'année. Inspiré par les travaux de Briand et al. (2017), un indicateur d'entropie individuel est calculé pour chaque carte afin d'estimer cette diversité d'usage spatial. Plus précisément, l'entropie de Shannon définie ci-dessous est utilisée :

$$H_i(X) = -\mathbb{E}\left[\log P\left(X = x_{ij}\right)\right] = -\sum_{j=1}^{n} P_{ij} \log P_{ij} \qquad \text{(Éq. 8)}$$

puis normalisée pour obtenir des valeurs comprises entre 0 et 1:

$$H_i^*(X) = \frac{H_i(X)}{\log(n)} \qquad \text{(Éq. 9)}$$

où H_i est l'indice d'entropie pour l'utilisateur de la carte i, H_i^* l'entropie normalisée correspondante, n le nombre de lieux d'embarquement différents possibles et P_{ij} la probabilité pour l'usager de valider sa carte i au lieu d'embarquement j. La probabilité P_{ij} est en fait la proportion (observée) des validations de la carte i effectuées au lieu d'embarquement j. Il est à noter ici que le nombre de validations et non le nombre de déplacements est utilisé, car un même déplacement peut être composé de plusieurs validations réalisées à différents endroits. De plus, cet indice d'entropie est calculé séparément pour les réseaux de bus et de métro, en considérant les stations ou les lignes d'embarquement selon les données disponibles (donc $n = 68$ stations de métro ou $n = 233$ lignes de bus différentes, incluant des lignes non régulières).

D'après l'équation précédente, l'entropie normalisée H_i^* est proche de 0 lorsque beaucoup de probabilités P_{ij} sont égales à ou proches de 0 (lieux pas ou peu visités) ou de 1 (lieux fréquemment visités), c'est-à-dire lorsque les lieux d'embarquement de l'usager i sont peu diversifiés (l'usager fait presque toutes ses validations aux mêmes endroits). Par conséquent, les deux entropies normalisées calculées pour chaque carte (une avec les validations de métro et l'autre avec les validations de bus) mesurent la variabilité intrapersonnelle spatiale de l'usager à travers l'année. Plus l'entropie est faible,

plus l'usager est dit régulier au niveau spatial. La variabilité interpersonnelle peut aussi être analysée en comparant les entropies individuelles de différents usagers entre elles.

Pour comparer les différents groupes de cartes définis ultérieurement, une entropie spatiale moyenne ainsi que le coefficient de variation associé seront estimés à partir des entropies individuelles des cartes de chaque groupe. Seuls les usagers du groupe ayant utilisé au moins une fois le service de transport collectif (métro ou bus selon le cas) pendant l'année seront pris en compte dans ces calculs, c'est-à-dire que les non-utilisateurs du service (pour lesquels l'entropie individuelle évaluée serait nulle) ne seront pas comptabilisés.

Résumé des indicateurs développés

Le tableau suivant permet de résumer l'ensemble des indicateurs définis précédemment. Le point de vue (interpersonnel ou intrapersonnel) et la résolution temporelle (annuelle ou mensuelle) de chaque indicateur sont mis en évidence dans la colonne « Détails ».

Tableau 2 : Récapitulatif des indicateurs de mesure de la variabilité d'utilisation du transport en commun à un niveau totalement désagrégé (sur une année)

Type de variabilité	**Indicateur**	**Détails** (point de vue en gras, résolution temporelle soulignée)
Dispersion des déplacements	Indices de Pareto et de Gini	Mesure de l'homogénéité de la dispersion des déplacements parmi les usagers, ou mesure de la variabilité **interpersonnelle** de l'intensité annuelle d'utilisation du transport en commun.
Fréquence d'utilisation	Classe modale	Mesure de la variabilité **interpersonnelle** de la fréquence d'utilisation mensuelle du transport en commun (évaluée à partir du nombre moyen de déplacements par mois actif).
Variabilité temporelle	Variance	Mesure des variations du nombre de déplacements par mois de chaque usager sur une année. Cette variabilité est décomposée en une variabilité **interpersonnelle** et une variabilité **intrapersonnelle**.
Variabilité spatiale	Entropie moyenne	Mesure de la diversité des lieux d'embarquement (stations de métro ou lignes de bus) choisis par l'usager pendant l'année, ou mesure de la variabilité **intrapersonnelle** spatiale d'utilisation du réseau. La variabilité **interpersonnelle** peut être évaluée à partir de la comparaison des entropies individuelles.

Analyse des résultats et validation statistique

L'utilité des indicateurs de variabilité proposés précédemment réside dans la possibilité de comparer les comportements d'utilisation du transport en commun de plusieurs groupes d'usagers, villes ou années. Dans cet article, la comparaison portera sur différents groupes de cartes définis par la combinaison des titres de transport validés en 2016. Premièrement, une analyse tarifaire est conduite

de manière à construire ces blocs de cartes en fonction du nombre et des catégories de produits et tarifs utilisés. Les indicateurs décrits dans la section précédente sont ensuite appliqués pour chacune des combinaisons sélectionnées. Finalement, les différences observées entre les groupes sont testées statistiquement.

Segmentation tarifaire

L'objectif de cette segmentation est de construire des groupes de cartes homogènes en taille et dans leur composition tarifaire. Le Tableau 3 ci-dessous fournit tout d'abord la répartition des cartes en fonction du nombre de types de produits et tarifs utilisés durant l'année. Rappelons ici que la classification décrite dans le cadre général de cet article permet six catégories de produits et cinq catégories de tarifs. Les clients de la STM ont néanmoins utilisé un nombre maximum de cinq types de produits et de quatre types de tarifs différents en 2016. Cependant, la grande majorité (63,2 %) d'entre eux n'ont validé qu'un seul type de produit et qu'un seul type de tarif durant toute l'année. Si on considère les produits et les tarifs séparément, 92,5 % des usagers n'ont utilisé qu'un type de tarif et 63,9 % des usagers n'ont utilisé qu'un type de produit en 2016 (la décomposition de ces deux groupes est donnée dans les Figure 3 et Figure 4 respectivement). Ces résultats prouvent que les passagers de la STM sont globalement assez réguliers dans leurs achats tarifaires.

Tableau 3 : Distribution des cartes en fonction du nombre de types de produits et de tarifs différents utilisés en 2016

% de cartes		Nombre de tarifs (5 catégories possibles)				
		1	2	3	4	TOTAL
Nombre de produits (6 catégories possibles)	1	63,2%	0,7%	0,0%	0,0%	63,9%
	2	19,9%	3,8%	0,2%	0,0%	23,9%
	3	7,5%	1,9%	0,1%	0,0%	9,6%
	4	1,8%	0,7%	0,1%	0,0%	2,6%
	5	0,0%	0,0%	0,0%	0,0%	0,1%
	TOTAL	92,5%	7,2%	0,4%	0,0%	100,0%

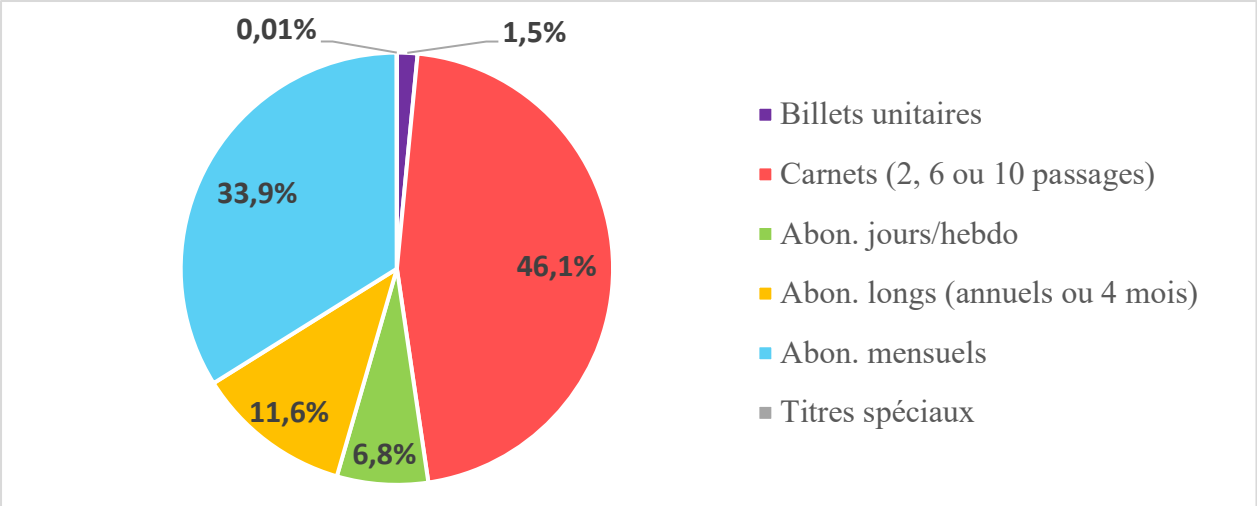

Figure 3 : Distribution des cartes ayant utilisé un seul type de produit durant l'année (63,9 % des cartes totales)

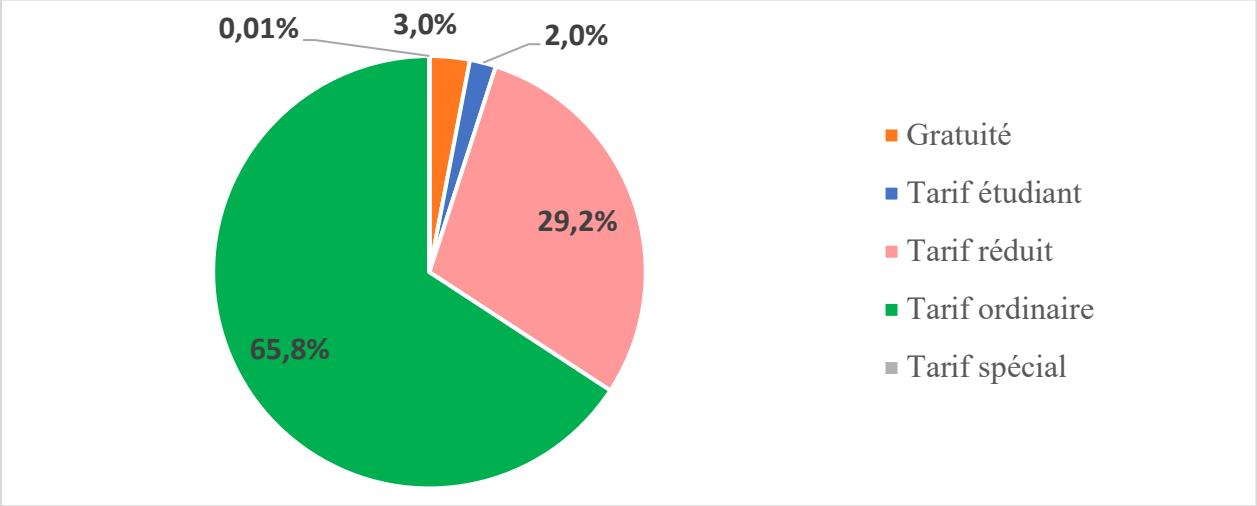

Figure 4 : Distribution des cartes ayant utilisé un seul type de tarif durant l'année (92,5 % des cartes totales)

Les groupes obtenus dans le Tableau 3 étant trop disparates, il a été décidé de préciser la catégorie en plus du nombre de produits ou tarifs utilisés. Les gros blocs de cartes « 1 produit – 1 tarif » (en rouge) et « 2 produits – 1 tarif » (en orange) ont ainsi été désagrégés en plusieurs sous-groupes, alors que les autres blocs (en jaune ou vert) ont été combinés ensemble de manière à rassembler les usagers ayant utilisé plus de trois produits différents et ceux ayant utilisé plus de deux tarifs différents durant l'année. Le critère appliqué consistait à définir des groupes contenant plus de 5 % mais moins de 20 % du total des cartes. Les dix combinaisons nombre*type de produits et de tarifs ainsi sélectionnées sont présentées dans le Tableau 4. Chaque combinaison sera notée COx où $x = 1, \ldots, 10$ dans la suite de cet article.

Tableau 4 : Dix combinaisons de cartes construites en fonction de leur composition tarifaire (nombre et type de produits et de tarifs utilisés durant l'année 2016)

	NOMBRE de		TYPE de		% de cartes
CO	**Produits**	**Tarifs**	**Produits**	**Tarifs**	
1	1	1	Carnets	Ordinaire	17,6%
2	1	1	Abon. mensuels	Ordinaire	10,5%
3	1	1	Carnets	Réduit	9,7%
4	1	1	Abon. mensuels	Réduit	9,1%
5	1	1	Abon. longs (annuels)	Ordinaire	5,4%
6	1	1	Autres		10,9%
7	2	1	Carnets + Abon. mensuels	Ordinaire	5,3%
8	2	1	Autres		14,6%
9	≥ 3	1	Tous		9,4%
10	1 à 5	≥ 2	Tous		7,5%
				TOTAL	100,0%

Les combinaisons les plus populaires correspondent à l'utilisation de produits de types carnets, abonnements mensuels ou abonnements longs, employés séparément (CO1 à CO5) ou de manière complémentaire (CO7). Les tarifs associés (les plus couramment utilisés) sont le tarif ordinaire et le tarif réduit. Précisons que le type de produit « abonnements longs » correspond à des abonnements annuels ou à des abonnements de quatre mois dans la grille tarifaire de la STM. Toutefois, la combinaison 5 (CO5) est composée uniquement d'utilisateurs d'abonnements annuels à tarif ordinaire car les abonnements de quatre mois ne sont disponibles qu'en tarif réduit. Les autres types de produits et de tarifs ayant été moins utilisés par les usagers de la STM en 2016, les combinaisons formées avec ces catégories ont mené à des groupes de moins de 5 % des cartes. Étant donné le critère adopté, ces combinaisons n'ont pas pu être considérées comme des combinaisons indépendantes et elles ont donc été regroupées dans les catégories « Autres » ou « Tous ».

Analyse de la variabilité

Les résultats de l'application des quatre types d'indicateurs dans chaque groupe de cartes sont donnés dans le Tableau 5. Les analyses qui suivent montrent comment de tels résultats peuvent être interprétés et illustrés dans un contexte tarifaire.

Tableau 5 : Calcul des indicateurs de variabilité dans chacune des dix combinaisons de cartes et pour le total des cartes

Type de variabilité	Indicateur	COMBINAISONS DE CARTES*										
		CO1	CO2	CO3	CO4	CO5	CO6	CO7	CO8	CO9	CO10	Total
Dispersion des déplacements	p	74,1%	68,0%	73,9%	67,6%	64,7%	78,4%	64,5%	72,1%	64,5%	66,1%	73,0%
	$\bar{p}$	25,9%	32,0%	26,1%	32,4%	35,3%	21,6%	35,5%	27,9%	35,5%	33,9%	27,0%
	α	1,28	1,51	1,29	1,53	1,72	1,19	1,74	1,35	1,73	1,62	1,32
	G	0,63	0,49	0,63	0,47	0,42	0,72	0,40	0,59	0,41	0,44	0,61
Fréquence d'utilisation	Classe modale	1	8	1	7	8	1	5	1	6	5	1
	% de cartes	62,6%	13,5%	69,3%	12,2%	17,5%	35,8%	14,9%	20,6%	12,5%	13,2%	26,5%
Variabilité temporelle	TSS	31	756	25	600	511	203	448	457	547	466	428
	BPSS	50,4%	56,4%	52,4%	39,5%	64,0%	49,0%	41,4%	54,4%	42,4%	34,1%	55,7%
	WPSS	49,6%	43,6%	47,6%	60,5%	36,0%	51,0%	58,6%	45,6%	57,6%	65,9%	44,3%
Variabilité spatiale MÉTRO	Entropie moyenne	0,29	0,36	0,26	0,37	0,34	0,30	0,37	0,38	0,42	0,40	0,35
	CV	48,4%	35,6%	51,6%	34,2%	35,7%	50,9%	31,5%	34,3%	28,4%	29,5%	40,7%
Variabilité spatiale BUS	Entropie moyenne	0,16	0,25	0,17	0,25	0,22	0,17	0,24	0,24	0,28	0,25	0,22
	CV	74,3%	50,1%	68,7%	43,8%	57,3%	73,0%	49,0%	51,9%	43,1%	50,6%	57,6%

*Ces combinaisons ont été définies dans le Tableau 4

Dans la première partie de ce tableau, les indicateurs proposés permettent de mesurer l'homogénéité de la distribution des déplacements de 2016 parmi les usagers. Cette distribution est représentée sur la Figure 5 pour chaque combinaison de cartes. Les nombres individuels de déplacements par an ayant été rangés dans l'ordre décroissant, les courbes tracées sont des courbes complémentaires de Lorenz. La courbe la plus à gauche du graphique correspond à la combinaison CO6 (cartes avec un seul type de produit et de tarif différent de ceux des combinaisons les plus populaires) : il s'agit donc de la répartition la plus inégale. En effet, la majorité (78,4 %) des déplacements de ce groupe ont été effectués par un petit nombre d'utilisateurs (21,6 % du total). Cette hétérogénéité est également traduite par le faible indice de Pareto α et le haut coefficient de Gini G rapportés dans le Tableau 5. Toutefois, cette tendance peut être expliquée par la diversité des cartes regroupées dans la combinaison CO6 : on y trouve à la fois des utilisateurs d'abonnements longs, annuels ou quatre mois (tarif réduit), et des utilisateurs de billets unitaires. Ces usagers ont donc probablement des intensités d'utilisation assez différentes. Les combinaisons CO1 et CO3 présentent également des indices de Pareto faibles et des coefficients de Gini élevés, témoignant que certains utilisateurs de carnets sont beaucoup plus assidus que d'autres. Cette conclusion semble être indépendante du tarif (ordinaire versus réduit) puisque les deux courbes

de Lorenz associées sont presque superposées sur la Figure 5. Les indices de Pareto (respectivement Gini) obtenus pour les utilisateurs d'abonnements mensuels (CO2 et CO4) sont plus élevés (respectivement faibles), suggérant des comportements d'utilisation plus similaires entre ces usagers. Cette uniformité s'accentue encore pour les utilisateurs d'abonnements annuels (CO5). De même, une plus grande diversité d'utilisation de produits ou de tarifs (CO9 et CO10) entraîne une plus faible variabilité interpersonnelle dans le nombre de déplacements réalisés par an. Le groupe le plus homogène en termes d'intensité d'usage (indice de Pareto le plus haut et coefficient de Gini le plus faible) correspond à la combinaison CO7 (carnets et abonnements mensuels utilisés conjointement avec un tarif ordinaire).

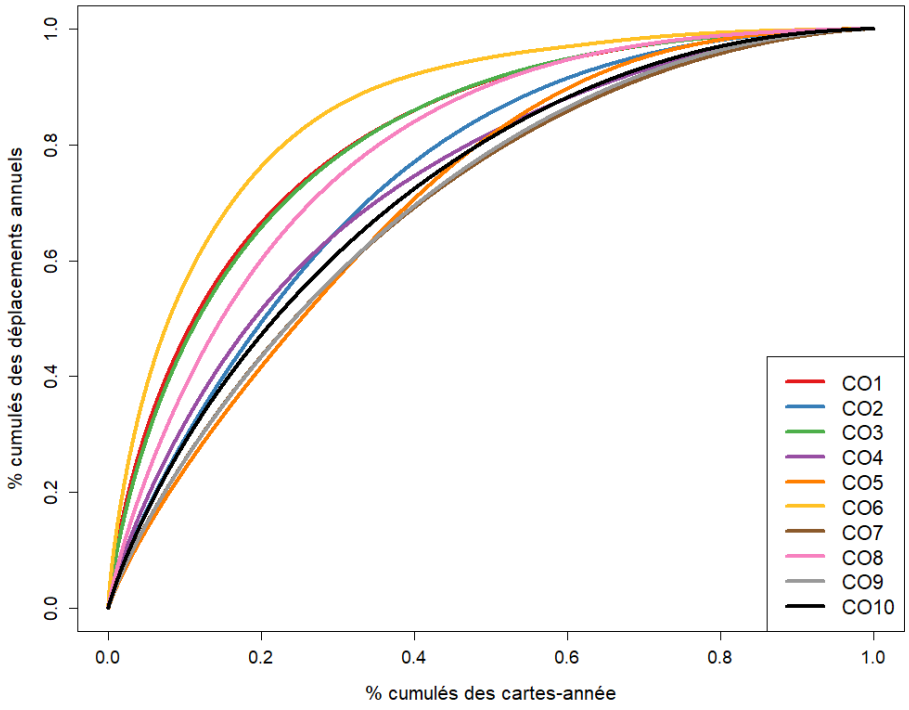

Figure 5 : Courbes de Lorenz (complémentaires) par combinaison de cartes

La variabilité de la fréquence d'utilisation mensuelle du transport en commun est mesurée à l'aide des indicateurs de la deuxième partie du Tableau 5 et illustrée par la Figure 6. Cette figure représente la distribution des cartes de chaque combinaison dans 21 intervalles (rangés dans l'ordre croissant) du nombre moyen de déplacements par mois actif. Trois types de distributions se distinguent : celles qui sont concentrées sur la gauche du graphique, donc dans la classe n° 1 composée de fréquences d'utilisation faibles (CO1, CO3, CO6, CO8 – utilisateurs de carnets ou de produits/tarifs moins populaires), celles qui présentent un pic de cartes dans les classes 7 ou 8 de fréquences élevées (CO2, CO4, CO5 – utilisateurs d'abonnements mensuels ou annuels), et celles qui sont centrées sur les classes 5 ou 6 de fréquences intermédiaires (CO7, CO9, CO10 – utilisateurs de carnets et d'abonnements

mensuels en complémentarité, ou d'une plus grande variété de produits/tarifs). En effet, plus de 60 % des utilisateurs de carnets (CO1 et CO3) ont effectué un nombre moyen de déplacements par mois actif compris dans l'intervalle n° 1. Ce pourcentage très élevé montre que la faible intensité d'utilisation mesurée est commune à la majorité des cartes de ces combinaisons. De même, les combinaisons CO6 et CO8 regroupent des usagers peu fréquents, probablement en raison du grand nombre d'utilisateurs d'abonnements quotidiens, de carnets et de billets unitaires dans ces deux groupes. À l'inverse, les utilisateurs d'abonnements mensuels (CO2) ou d'abonnements annuels (CO5) avec tarif ordinaire correspondant à la classe modale la plus élevée (classe n° 8). La proportion de cartes associée à cette classe est néanmoins plus élevée dans le cas des abonnements annuels (17,5 % contre 13,5 %) : une forte utilisation du transport en commun est donc plus courante chez ces usagers. Les utilisateurs d'abonnements mensuels à tarif réduit (CO4) sont également des usagers fréquents, mais leur nombre moyen de déplacements par mois actif est légèrement inférieur à celui obtenu pour le tarif ordinaire (classe n° 7). Comme on pouvait s'y attendre, l'utilisation combinée d'abonnements mensuels et de carnets (CO7) conduit à des fréquences d'utilisation intermédiaires entre CO1 (carnets) et CO2 (abonnements mensuels). De plus, les classes modales et les pourcentages de cartes correspondants sont comparables lorsque plusieurs types de produits ou de tarifs sont utilisés (CO7, CO9, CO10).

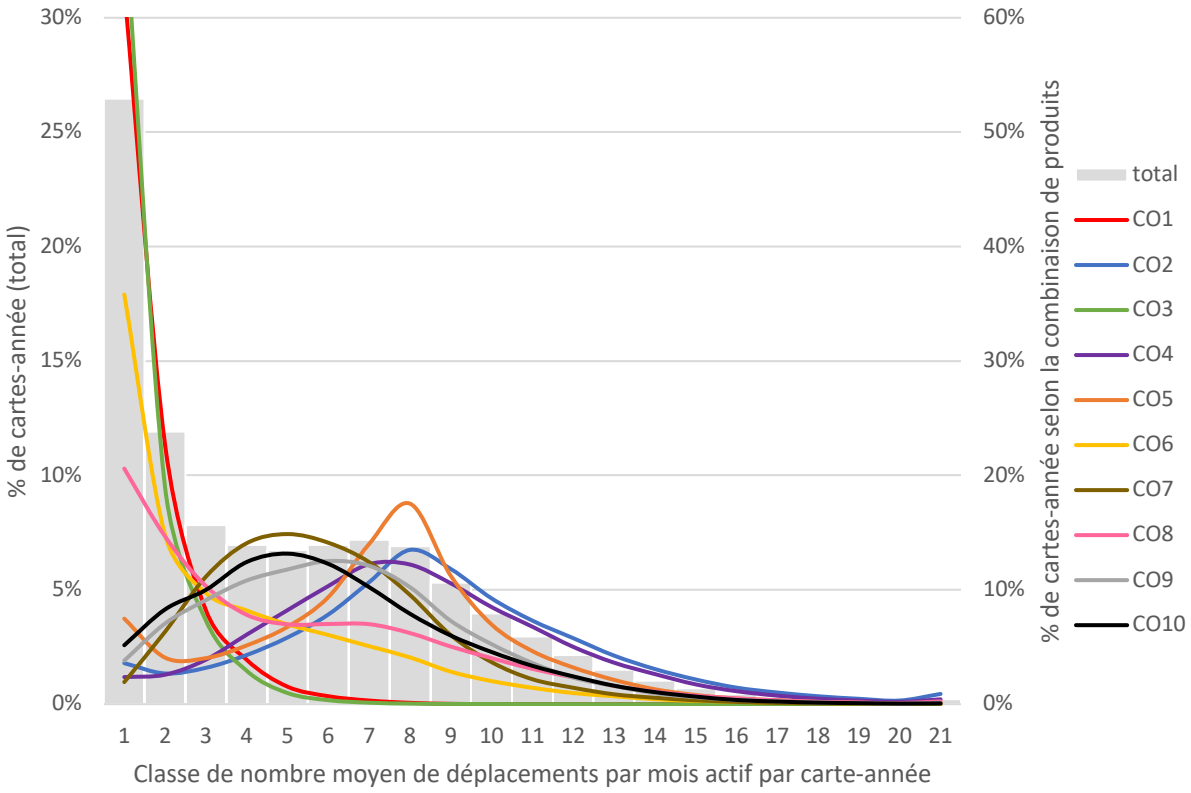

Figure 6 : Distribution fréquentielle des cartes en fonction du nombre moyen de déplacements par mois actif par carte regroupé en 21 classes (pour le total et par combinaison)

Les indicateurs de la troisième partie du Tableau 5 permettent quant à eux de chiffrer la variabilité temporelle du nombre de déplacements par mois par carte : les déviations observées dans le comportement de chaque individu par rapport à la grande moyenne (moyenne mensuelle pour toutes les cartes) ont été sommées sur chacun des mois de l'année de manière à calculer une variance totale dans chaque combinaison de cartes. D'après les résultats obtenus, les utilisateurs les plus variables au niveau temporel (variance totale élevée) sont aussi les plus fréquents. Le coefficient de corrélation poly sérielle (*polyserial correlation coefficient*) calculé entre la variable ordinale « Classe modale » et la variable numérique « TSS » est positif et égal à 0,61. Cette relation semble cohérente : il est plus probable d'observer des variations dans le comportement des usagers qui se déplacent beaucoup que dans celui des usagers qui sont plus occasionnels ou moins mobiles (en transport en commun). Les utilisateurs de carnets (CO1 et CO3) présentent donc les valeurs les plus faibles du fait de leur faible usage du transport en commun. Cependant, même si les utilisateurs d'abonnements annuels (CO5) sont plus fréquents que les utilisateurs d'abonnements mensuels (CO2 et CO4), ils sont aussi plus réguliers au niveau temporel puisque leur variance totale de 511 est inférieure à 756 et 600 respectivement. En outre, la part de la variance interpersonnelle (BPSS) et celle de la variance intrapersonnelle (WPSS) sont assez similaires pour la plupart des combinaisons de cartes étudiées. Néanmoins, la variance interpersonnelle (BPSS) prédomine et explique 64 % de la variance totale des utilisateurs d'abonnements annuels (CO5) : ces usagers sont donc parfois différents les uns des autres, mais le comportement de mobilité d'un même usager tend à rester assez stable au cours de l'année. Au contraire, la variance intrapersonnelle (WPSS) prédomine dans le cas des utilisateurs d'abonnements mensuels à tarif réduit (CO4), indiquant que les habitudes de déplacement de ces usagers changent durant l'année, mais de façon similaire pour tous. C'est également le cas pour les utilisateurs de plus d'un type de produit ou de tarif (CO7, CO9, CO10) : l'hétérogénéité des usagers présents dans ces groupes est donc moins importante que la variabilité longitudinale de leurs comportements individuels. Cette variabilité intrapersonnelle est également manifeste sur la Figure 7, qui permet de visualiser la variation moyenne à travers l'année 2016 du nombre de déplacements par mois (normalisé) pour l'ensemble des individus de chaque combinaison (variabilité intrapersonnelle moyenne par groupe de cartes). Le comportement moyen mensuel associé aux cartes des combinaisons CO4 et CO10 est particulièrement variable. Comme indiqué précédemment, ce sont les groupes de cartes pour lesquels la part de la variance intrapersonnelle WPSS est la plus élevée (60,5 % et 65,9 % respectivement).

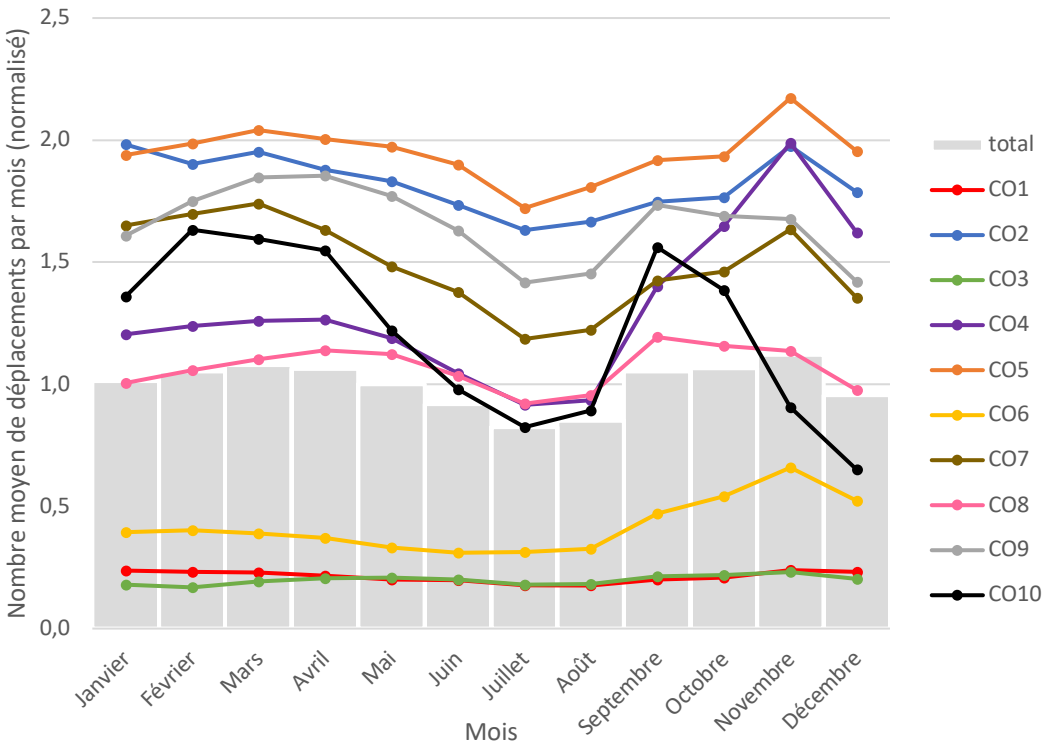

Figure 7 : Nombre moyen de déplacements par mois par carte pour chaque mois de l'année 2016 (pour le total et par combinaison), normalisé avec le nombre moyen de déplacements par mois tous mois confondus

Enfin, la variabilité spatiale est quantifiée à l'aide des indicateurs d'entropie rapportés dans la quatrième partie (pour le métro) et la cinquième partie (pour le bus) du Tableau 5. Une entropie de Shannon a été estimée pour chaque usager puis une moyenne et un coefficient de variation ont été calculés dans chaque combinaison à partir de toutes les valeurs individuelles. Des diagrammes violons ont également été tracés dans la Figure 8 (pour le métro uniquement) afin de visualiser la densité de probabilité, la médiane et l'intervalle interquartile de la distribution des entropies dans chaque groupe. Les valeurs les plus basses des entropies moyennes obtenues pour le métro sont atteintes dans les combinaisons CO1 et CO3. Les utilisateurs de carnets ont donc emprunté un ensemble peu diversifié de stations de métro durant l'année (sur les 68 stations du réseau complet). D'après les définitions proposées précédemment, ils sont caractérisés par une faible variabilité spatiale. Cette surprenante régularité est en fait expliquée par leur faible activité sur le réseau. Cette dernière peut être évaluée à l'aide des deux indicateurs présentés dans le Tableau 6 : le nombre moyen de stations utilisées et l'espace d'action moyen par carte, défini comme l'aire moyenne de l'enveloppe convexe englobant toutes les stations de métro visitées par un usager pendant l'année. Les résultats du calcul de ces deux indicateurs prouvent que les utilisateurs de carnets (CO1 et CO3) ont visité peu de stations en moyenne durant l'année (5,7 et 4,8 respectivement), représentant une petite superficie d'action (14,8 et 12,6 km^2 respectivement). Cette faible utilisation du réseau a généré de nombreuses probabilités P_{ij} égales à 0, d'où une entropie

qui tend vers 0. De plus, les coefficients de variation associés à ces deux combinaisons sont élevés (48,4 % et 51,6 % respectivement), et les diagrammes violons de la Figure 8 révèlent une distribution très hétérogène au sein de chaque groupe. Au contraire, les utilisateurs de plus de trois types de produits ou de plus de deux types de tarifs (CO9 et CO10) cumulent les entropies moyennes les plus hautes, des nombres moyens de stations utilisées et des aires d'actions élevés ainsi que de faibles coefficients de variation (inférieurs à 30 %). Les usagers de ces deux groupes ont donc majoritairement une utilisation plus large et plus variée du réseau au niveau spatial. Par ailleurs, les utilisateurs d'abonnements annuels (CO5) sont caractérisés par une entropie moyenne plus faible que les utilisateurs d'abonnements mensuels (CO2 et CO4) : 0,34 contre 0,36 et 0,37. Ainsi, non seulement les utilisateurs d'abonnements annuels sont plus réguliers au niveau temporel que les utilisateurs d'abonnements mensuels, mais ils sont aussi plus réguliers au niveau spatial. Les mêmes conclusions ou presque peuvent être tirées à partir des entropies estimées avec les validations de bus. Toutefois, les valeurs moyennes sont plus basses en raison du plus grand nombre de lieux d'embarquement possibles ($n = 233$ lignes).

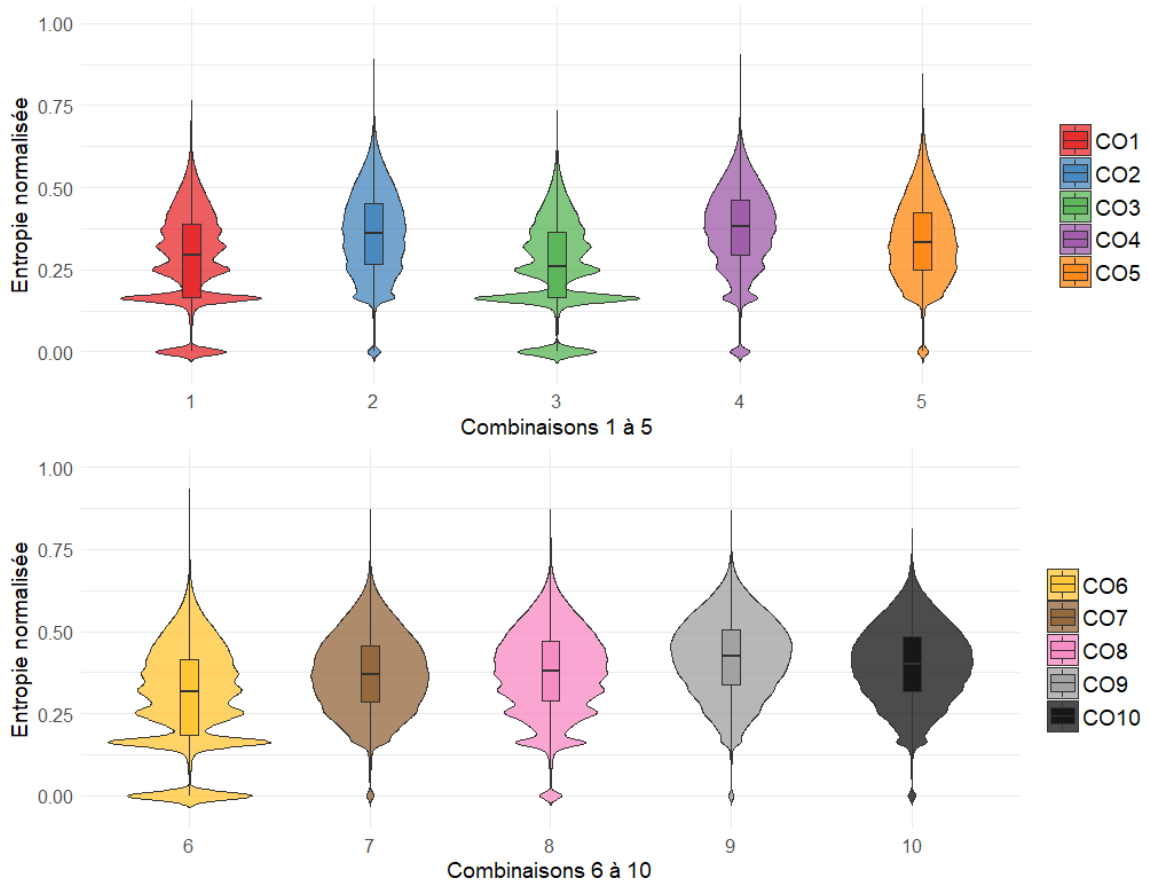

Figure 8 : Diagramme violon des entropies individuelles pour les validations de métro (par combinaison de cartes)

Tableau 6 : Nombre moyen de stations utilisées et espace moyen d'action par carte (pour le total et par combinaison de cartes) – pour les utilisateurs du métro

Combinaison	Nombre moyen de stations utilisées	Espace d'action moyen [km^2]
CO1	5,7	14,8
CO2	12,9	41,6
CO3	4,8	12,6
CO4	11,7	35,4
CO5	13,4	39,2
CO6	7,1	21,0
CO7	13,2	40,3
CO8	11,3	37,5
CO9	16,1	53,8
CO10	13,9	45,5
Total	10,0	31,1

Tests statistiques et taille d'effet

Des tests statistiques sont finalement exécutés afin de vérifier que les différences observées précédemment entre les dix combinaisons de cartes sont significatives. Ces tests sont appliqués, pour chaque indicateur, sur toutes les paires de combinaisons de cartes possibles afin de comparer deux à deux les dix groupes. Des tests bilatéraux sont utilisés de manière à tester une hypothèse nulle d'égalité et non seulement une relation de supériorité ou d'infériorité. De plus, des tests non paramétriques sont choisis car ils ne supposent pas une distribution spécifique des données et ils sont moins sensibles aux valeurs aberrantes que les tests paramétriques (Cleophas et Zwinderman, 2011). Ces tests sont particulièrement adaptés à notre cas puisque les données étudiées ne suivent pas une loi gaussienne et, n'ayant pas été filtrées au préalable, elles contiennent des valeurs extrêmes. Les tests appliqués et les statistiques associées sont répertoriés dans le Tableau 7.

Les différences observées à l'aide de l'indicateur de dispersion des déplacements sont éprouvées avec le test de Kolmogorov-Smirnov. Ce test permet de comparer par paire les distributions cumulées représentées sur la Figure 5 (qui sont ici des courbes complémentaires de Lorenz) en se basant sur la distance maximale mesurée entre ces courbes. Le test d'homogénéité des proportions (test du χ^2) est quant à lui utilisé pour comparer les distributions de la Figure 6 (distributions des cartes dans des classes de nombres moyens de déplacements par mois actif). Ce test contrôle si les proportions de cartes dans les mêmes classes sont égales. De plus, le test d'homogénéité des variances de Fligner-Killeen est réalisé pour tester l'égalité des variances totales du nombre mensuel de déplacements. Plus précisément, la

version asymptotique de la procédure modifiée par Conover et al. (1981) est appliquée : l'approximation asymptotique de la distribution exacte permet alors d'obtenir une statistique Z. Enfin, le test de Wilcoxon Mann-Whitney est employé pour comparer les entropies moyennes de variabilité spatiale. Étant donnés les gros échantillons de données exploités dans cet article, on peut supposer que la statistique U suit asymptotiquement une loi normale et une statistique Z peut donc aussi être calculée (Adjengue, 2014). Le test évalue si les deux distributions comparées ont les mêmes tendances centrales (même centre, même moyenne, même médiane).

Tableau 7 : Tests statistiques appliqués pour chaque indicateur de variabilité

Type de variabilité	Indicateur	Test utilisé	Statistique
Dispersion des déplacements	Indices de Pareto et de Gini	Test de Kolmogorov-Smirnov sur les courbes de Lorenz	d
Fréquence d'utilisation	Classe modale	Test du χ^2 (test d'homogénéité des proportions dans chaque classe)	χ^2
Variabilité temporelle	Variance	Test de Fligner-Killeen (version asymptotique)	Z
Variabilité spatiale	Entropie moyenne	Test de Wilcoxon Mann-Whitney (version asymptotique)	Z

Les résultats de ces tests statistiques sont analysés à partir des valeurs-p calculées. D'après le tableau obtenu (disponible sur demande aux auteurs), toutes les différences observées précédemment (à l'exception de deux paires de combinaisons pour l'indicateur d'entropie) sont significatives ($p < 0,05$ donc l'hypothèse nulle d'égalité a été rejetée à chaque comparaison). Bien que cette conclusion puisse satisfaire de nombreux statisticiens, elle ne permet pas de nuancer les différences rapportées : ont-elles toutes la même importance ? De plus, la taille de l'échantillon utilisé est probablement la cause de ces « trop bons » résultats. Ce problème typique des grosses bases de données a déjà été signalé dans la littérature (Van der Laan et al., 2010). Par conséquent, nous nous sommes intéressés à la notion de taille d'effet, qui permet d'évaluer la taille d'un effet indépendamment de la taille de l'échantillon.

Par définition, la taille d'effet est une grandeur qui permet de quantifier l'ampleur ou l'importance d'un effet, tel que la relation entre une variable dépendante et une variable indépendante ou la différence entre deux populations. Elle ne doit pas être confondue avec la significativité statistique, qui correspond à la probabilité que l'effet mesuré soit dû au hasard ou à l'échantillonnage. Cooper et Hedges (1994) relient ces deux notions par l'équation suivante :

$$Significativité\ statistique = taille\ d'effet \times\ taille\ d'échantillon \qquad \text{(Éq. 10)}$$

Cette relation suppose qu'une grande significativité statistique (ou une faible valeur-p) peut être expliquée par un effet important, mais aussi par un effet moindre mesuré dans un gros échantillon (Cooper et Hedges, 1994). Cela implique également que tout test peut devenir statistiquement

significatif si la taille de l'échantillon utilisé est assez grande. La taille d'effet est au contraire indépendante de la taille de l'échantillon (Coe, 2002 ; Cohen, 1988 ; Fritz et al., 2012).

Depuis les travaux de Cohen (Cohen, 1969, 1988, 1994), la notion de taille d'effet a principalement été utilisée dans les sciences comportementales, notamment en médecine ou psychologie, par exemple pour évaluer l'efficacité d'un traitement : une taille d'effet plus élevée reflète un plus grand impact sur les patients. Coe (2002) a signalé un véritable changement de paradigme : cette mesure permet de passer d'un *does it work?* (est-ce que cela fonctionne ?) à un *how well does it work?* (à quel point cela fonctionne ?). Dans un contexte statistique plus général, Cohen (1988) a défini la taille d'effet comme *the degree to which the null hypothesis is false* (le degré selon lequel l'hypothèse nulle est fausse) ou *the degree of departure from the null hypothesis* (le degré de sortie de l'hypothèse nulle). En d'autres termes, la significativité statistique permet seulement de confirmer ou de rejeter l'hypothèse nulle, alors que la taille d'effet permet de quantifier et de nuancer ce résultat. Cette notion permet donc une meilleure compréhension et une meilleure interprétation d'un effet significatif. C'est pourquoi elle a été officiellement encouragée par l'association américaine de psychologie en 1994 (American Psychological Association, 1994). La taille d'effet reste néanmoins peu rapportée par les chercheurs car ces derniers ont moins d'expérience avec cette notion et ils ont généralement du mal à l'interpréter (Fritz et al., 2012). Initialement appliquée sur de petits échantillons en sciences sociales, la taille d'effet n'a vu son utilité reconnue pour les données massives que très récemment (Khalilzadeh et Tasci, 2017). Ce concept statistique pourrait devenir particulièrement intéressant dans l'ère du Big Data : les bases de données étudiées ne cessent de grossir, conduisant à des résultats toujours plus significatifs, mais non représentatifs de l'importance des effets observés. Les méthodes de validation statistique doivent donc être adaptées à ces données émergentes.

Dans cet article, la notion de taille d'effet est exploitée pour mesurer l'ampleur des différences observées par paire entre les dix combinaisons de cartes sélectionnées. Ces différences ont été trouvées significatives par les tests statistiques appliqués précédemment, mais la taille d'effet va permettre d'apprécier dans quelle mesure les hypothèses nulles d'égalité testées doivent être rejetées. Elle va ainsi quantifier l'importance de ces différences indépendamment de la taille des échantillons comparés. Les indices présentés ci-après permettent d'estimer la taille d'effet dans une population donnée à partir d'un échantillon. Ces indices augmentent avec la discordance entre les deux distributions, variances ou ensembles de proportions testés. Des indices non paramétriques sont utilisés car leurs équivalents paramétriques sont très sensibles au non-respect des hypothèses de normalité (Coe, 2002). De plus, tous ces indices sont rapportés avec un signe positif car l'ordre des deux groupes comparés n'est pas primordial ici. Leur définition est basée sur les statistiques obtenues à partir des tests appliqués précédemment.

Dans le cas de l'indicateur de dispersion des déplacements, la distance du test de Kolmogorov-Smirnov est utilisée comme indice de taille d'effet :

$$d = \max_x |F_A(x) - F_B(x)| \qquad \text{(Éq. 11)}$$

où A et B sont deux échantillons comparés, $F_A(x)$ et $F_B(x)$ leurs fonctions de répartition respectives. Cette statistique n'est en réalité pas un indice de taille d'effet mais elle permet de mieux capturer l'ampleur de la différence entre les deux distributions que la valeur-p. L'indice de taille d'effet associé à l'indicateur de variabilité de la fréquence d'utilisation mensuelle est l'indice V de Cramér défini par l'équation 12. D'après Cohen (1988), il s'agit de l'indice le plus général et le plus facilement interprétable dans le cas des tableaux de contingence de fréquences ou de proportions.

$$V = \sqrt{\frac{\chi^2}{N.t}} \qquad \text{(Éq. 12)}$$

où χ^2 est la statistique du test du khi-deux, $N = n_A + n_B$ la taille totale des deux échantillons comparés et $t = \min[(r-1),(c-1)]$ avec r le nombre de lignes et c le nombre de colonnes du tableau de données utilisé pour chaque comparaison. Dans notre cas, cet indicateur est équivalent au coefficient Φ et à l'indice w de Cohen, car $r = 2$ groupes sont testés à chaque comparaison et nous avons $c = 11$ classes de nombre de déplacements par mois actif avec au moins 5 observations par classe, donc $t = \min(1,10) = 1$. Voir les travaux de Cohen (1988) et Fritz et al. (2012) pour avoir plus de détails sur ces calculs.

$$V = \sqrt{\frac{\chi^2}{N}} = \Phi = w \qquad \text{(Éq. 13)}$$

Pour les deux indicateurs restants (variabilité temporelle et variabilité spatiale), le coefficient de corrélation r défini par Cooper et Hedges (1994) et utilisé par Pallant (2007) est calculé. Ce coefficient est basé sur la statistique Z :

$$r = \sqrt{\frac{Z^2}{N}} = \frac{|Z|}{\sqrt{N}} \qquad \text{(Éq. 14)}$$

avec $N = n_A + n_B$ la taille totale de l'échantillon.

Les valeurs au carré de ces indices de taille d'effet sont peut-être plus connues et mieux comprises par les chercheurs. En effet, le terme Φ^2 ou r^2 peut être interprété comme la proportion de la variance totale (de l'indicateur mesuré dans la population combinée des groupes A et B) qui est expliquée par l'appartenance à un des deux groupes. Autrement dit, appartenir à une combinaison de cartes plutôt

qu'à une autre, c'est-à-dire utiliser un titre de transport plutôt qu'un autre, représente Φ^2% ou r^2% de la variance totale de l'indicateur dans les deux groupes combinés. Par conséquent, la taille d'effet permet ici de quantifier l'influence de l'utilisation d'un certain titre (par rapport à un autre) sur la variabilité des comportements individuels.

Les résultats de l'application de ces indices de taille d'effet sont fournis dans le Tableau 8. Chaque ligne 'COi v COj' correspond à la comparaison d'une paire de combinaisons de cartes, consistant à évaluer la différence entre les cartes des combinaisons i et j pour chaque type de variabilité. Des couleurs ont été ajoutées pour faciliter l'analyse de ces résultats. Celles appliquées pour le premier indicateur sont différentes puisque la statistique d n'est pas un vrai indice de taille d'effet : le gradient de couleurs rouge-blanc-bleu a ainsi été attribué dans l'ordre décroissant des valeurs obtenues. Pour les trois autres indicateurs, les couleurs conférées correspondent au critère de Cohen, défini dans le Tableau 9. L'auteur (Cohen, 1988) présente les adjectifs « petite », « moyenne » et « grande » comme une convention arbitraire caractérisant trois niveaux d'importance de la taille d'effet. Les valeurs représentatives associées à ces trois niveaux ont été déterminées à partir de résultats expérimentaux en recherche comportementale et biologique. L'auteur avertit lui-même les experts que cette échelle peut différer selon le domaine de recherche mais, dans une situation comme la nôtre où aucune référence antérieure ne peut servir de base de comparaison, cette convention reste recommandée. Néanmoins, pour s'affranchir du critère de Cohen, les nombreuses valeurs de taille d'effet calculées dans cet article peuvent également être comparées entre elles (comparaisons relatives) plutôt qu'avec les valeurs représentatives proposées par Cohen (comparaisons absolues).

Tableau 8 : Résultats de l'application des indices de taille d'effet (indicateurs de variabilité)

Test	Dispersion des déplacements	Fréquence d'utilisation	Variabilité temporelle	Variabilité spatiale – métro	Variabilité spatiale – bus
CO1 v CO2	0,11	0,87	0,54	0,24	0,34
CO1 v CO3	0,01	0,07	0,00	0,08	0,05
CO1 v CO4	0,16	0,87	0,24	0,29	0,37
CO1 v CO5	0,17	0,83	0,70	0,15	0,21
CO1 v CO6	0,12	0,44	0,03	0,06	0,06
CO1 v CO7	0,20	0,78	0,64	0,25	0,29
CO1 v CO8	0,04	0,58	0,27	0,30	0,31
CO1 v CO9	0,19	0,79	0,69	0,43	0,44
CO1 v CO10	0,17	0,76	0,48	0,36	0,32
CO2 v CO3	0,11	0,88	0,51	0,33	0,29
CO2 v CO4	0,07	0,10	0,28	0,06	0,02
CO2 v CO5	0,06	0,19	0,03	0,08	0,10
CO2 v CO6	0,22	0,60	0,56	0,18	0,27
CO2 v CO7	0,10	0,40	0,01	0,05	0,02
CO2 v CO8	0,08	0,44	0,27	0,07	0,02

Test	Dispersion des déplacements	Fréquence d'utilisation	Variabilité temporelle	Variabilité spatiale – métro	Variabilité spatiale – bus
CO2 v CO9	0,09	0,33	0,01	0,23	0,12
CO2 v CO10	0,06	0,37	0,13	0,15	0,00
CO3 v CO4	0,17	0,89	0,23	0,39	0,33
CO3 v CO5	0,17	0,84	0,74	0,27	0,19
CO3 v CO6	0,12	0,45	0,05	0,14	0,01
CO3 v CO7	0,20	0,83	0,68	0,38	0,28
CO3 v CO8	0,03	0,57	0,25	0,38	0,26
CO3 v CO9	0,19	0,82	0,68	0,53	0,41
CO3 v CO10	0,17	0,80	0,47	0,47	0,29
CO4 v CO5	0,06	0,19	0,14	0,14	0,13
CO4 v CO6	0,28	0,58	0,24	0,23	0,31
CO4 v CO7	0,06	0,33	0,15	0,02	0,05
CO4 v CO8	0,15	0,41	0,01	0,01	0,05
CO4 v CO9	0,06	0,26	0,17	0,18	0,11
CO4 v CO10	0,03	0,31	0,17	0,09	0,02
CO5 v CO6	0,26	0,49	0,55	0,10	0,17
CO5 v CO7	0,08	0,34	0,10	0,14	0,08
CO5 v CO8	0,13	0,33	0,23	0,14	0,07
CO5 v CO9	0,07	0,23	0,07	0,31	0,22
CO5 v CO10	0,04	0,28	0,16	0,24	0,10
CO6 v CO7	0,32	0,44	0,54	0,21	0,25
CO6 v CO8	0,14	0,20	0,26	0,23	0,24
CO6 v CO9	0,31	0,46	0,56	0,38	0,38
CO6 v CO10	0,28	0,41	0,51	0,31	0,27
CO7 v CO8	0,18	0,32	0,21	0,02	0,00
CO7 v CO9	0,02	0,12	0,11	0,19	0,15
CO7 v CO10	0,04	0,13	0,16	0,11	0,02
CO8 v CO9	0,17	0,30	0,25	0,16	0,15
CO8 v CO10	0,14	0,26	0,16	0,08	0,02
CO9 v CO10	0,03	0,08	0,19	0,09	0,12

Tableau 9 : Critère de Cohen

Valeur	Taille d'effet
0,1	petite
0,3	moyenne
0,5	grande

Dans l'ensemble, les nombreuses petites tailles d'effet rapportées (en vert) dans le Tableau 8 révèlent que la plupart des différences trouvées significatives précédemment ne sont en fait pas très importantes. Cette conclusion confirme que la taille des échantillons manipulés était la principale cause de cette apparente significativité. En effet, les distances de Kolmogorov-Smirnov et les indices de taille d'effet calculés pour ces deux paires de combinaisons étant faibles, ni les cartes des combinaisons CO1

et CO3 (‘CO1 v CO3’) ni celles des combinaisons CO2 et CO4 (‘CO2 v CO4’) ne sont vraiment très distinctes. Par conséquent, le type de tarif (ordinaire versus réduit) a peu d’effet sur la variabilité d’utilisation du transport en commun. De même, les résultats des comparaisons ‘CO2 v CO5’ et ‘CO4 v CO5’ attestent que les indicateurs de variabilité calculés pour CO5 (utilisateurs d’abonnements annuels) sont assez proches de ceux obtenus pour CO2 et CO4 (utilisateurs d’abonnements mensuels) : ces usagers ont donc un niveau de régularité similaire. Les différences mesurées entre les combinaisons CO6 et CO1 ou CO3 (utilisateurs de carnets) sont également peu importantes, à l’exception de celle se rapportant à l’indicateur de variabilité de la fréquence d’utilisation. Cette ressemblance est probablement due au grand nombre d’utilisateurs de billets unitaires et d’abonnements journaliers dans CO6, conduisant à des comportements d’utilisation du transport en commun peu éloignés de ceux qui caractérisent les utilisateurs de carnets. En outre, les faibles tailles d’effet obtenues dans la dernière partie du Tableau 8 prouvent que les utilisateurs d’une plus grande variété de produits ou de tarifs (CO7, CO8 et CO9) présentent des comportements semblables. Inversement, les grandes distances et tailles d’effet mesurées entre les utilisateurs d’abonnements annuels ou mensuels (CO2, CO4, CO5) et les utilisateurs de carnets (CO1 et CO3), notamment pour la fréquence d’utilisation, révèlent une grande dissimilarité entre ces usagers.

Conclusion

S’appuyant sur une année de données de cartes à puce provenant de la Société de Transport de Montréal, cet article a proposé une méthode basée sur quatre types d’indicateurs pour mesurer la variabilité d’utilisation du transport en commun à une échelle totalement désagrégée. Ces indicateurs ont été mis en pratique afin de comparer 10 groupes de cartes définis en fonction des titres de transport utilisés au cours de l’année. Des tests statistiques ont finalement été appliqués pour prouver la puissance des indicateurs développés à mettre en évidence des différences significatives entre les groupes comparés. Toutefois, la grande taille des échantillons a limité l’interprétation des résultats obtenus. La notion de taille d’effet a alors été présentée et a permis de dépasser ce problème en quantifiant non plus la significativité mais l’importance des différences observées. Il a été démontré qu’une telle approche statistique pouvait être particulièrement pertinente pour valider des résultats obtenus à partir de données massives.

Pour résumer les résultats de cet article, les utilisateurs d’abonnements annuels ou mensuels ont des comportements de mobilité très différents de ceux des utilisateurs de carnets de tickets, quel que soit le tarif payé (ordinaire ou réduit). Les utilisateurs d’abonnements annuels ou mensuels sont les clients les plus fiables de la STM puisqu’ils sont à la fois fréquents et réguliers. Les utilisateurs d’abonnements annuels tendent d’ailleurs à être plus constants aux niveaux temporel et spatial que les utilisateurs

d'abonnements mensuels, mais les analyses de taille d'effet ont montré que cette différence n'était pas très importante. À l'inverse, les utilisateurs de carnets sont des usagers occasionnels du transport en commun (faible fréquence d'utilisation) et une certaine hétérogénéité a été constatée dans leurs comportements (faibles indices de Pareto, hauts coefficients de variation). Les faibles variabilités temporelle et spatiale mesurées pour ces usagers sont principalement dues à leur faible intensité d'usage. Il est plus difficile de tirer de nettes conclusions pour les combinaisons de cartes caractérisées par une plus grande diversité tarifaire puisque ces groupes sont composés d'usagers très hétéroclites (utilisant des types de produits très différents). Par conséquent, ils présentent des similarités avec de nombreuses autres combinaisons. D'autres analyses seraient nécessaires pour explorer ces groupes plus en profondeur. Néanmoins, d'après les premiers résultats de cet article, le comportement de ces usagers a tendance à être plus variable.

D'un point de vue analytique, les résultats de cet article ont prouvé l'existence de différents types de variations dans les comportements d'utilisation du transport en commun. Ils justifient la nécessité de prendre en compte la variabilité inter et intrapersonnelle des usagers dans les modèles de prévision de la demande. De plus, une relation a été révélée entre les titres de transport utilisés et la variabilité des comportements, validant ainsi l'hypothèse émise au début de cette étude. Dans un contexte de tarification intégrée, des analyses plus poussées permettraient aux opérateurs de déceler des opportunités de fidélisation de leurs clients. Ces analyses pourraient également inciter à la création de nouveaux produits personnalisés en fonction des types d'utilisateurs observés. Une interprétation plus fine des résultats serait nécessaire pour arriver à de telles conclusions. Toutefois, cette interprétation serait propre au cas particulier de Montréal et à sa politique de tarification du transport en commun. L'objectif principal de cet article étant de présenter et d'illustrer une méthode, nous n'irons pas plus loin dans l'analyse de cet exemple d'application tarifaire.

Au contraire, la méthodologie proposée est transférable d'une société de transport à l'autre. Des indicateurs simples et reproductibles ont été mis à disposition : ils pourront être réutilisés pour quantifier la variabilité d'utilisation du transport en commun de différents blocs de cartes, pour différentes années ou villes, dans un objectif de comparaison des comportements de mobilité. Ces outils d'analyse de la variabilité d'utilisation du transport en commun peuvent servir aux opérateurs et planificateurs des réseaux de transport. Les indicateurs proposés pourraient notamment être intégrés à des modèles pour améliorer la prévision de la demande, ou être exposés dans les rapports de suivi des grands paramètres du transport. De même, ils pourraient aider à comparer plusieurs scénarios dans des outils de simulation : l'effet de différentes interventions stratégiques sur les comportements individuels pourrait ainsi être évalué (aux niveaux intrapersonnel et interpersonnel). Par ailleurs, cet article est un exemple concret de traitement et de valorisation de données massives dans le contexte

du transport urbain. Il confirme l'utilité des données de cartes à puce pour suivre de manière longitudinale et désagrégée les comportements de déplacement. En outre, cet article illustre les problèmes méthodologiques auxquels les analystes sont confrontés avec les données du Big Data. Il propose ainsi une nouvelle approche (la taille d'effet) pour mieux interpréter les tests statistiques appliqués aux grandes bases de données.

De futurs travaux seront consacrés à explorer les outils développés dans cet article et à dépasser certaines limites de la méthode proposée. Le travail qui a été réalisé dans cet article pourrait être inversé : les indicateurs proposés seraient alors utilisés pour segmenter les usagers puis la typologie obtenue serait croisée avec la composition tarifaire des cartes. Les conclusions tirées permettraient de valider la régularité de certains types d'utilisateurs et de repérer les usagers les plus réguliers en vue de les retenir dans le système. Certains groupes d'usagers pourraient être analysés plus spécifiquement (par exemple, seulement ceux qui sont présents sur le réseau toute l'année). De même, plusieurs années pourraient être étudiées afin de vérifier la stabilité des comportements et la fidélité des usagers sur une plus longue durée. Des techniques de reconnaissance d'un même usager pour plusieurs cartes et des procédures d'estimation des lieux de débarquements pourraient également être appliquées afin d'enrichir les données utilisées dans cet article. En outre, les indicateurs de variabilité temporelle et spatiale proposés pourraient être normalisés de sorte à être moins dépendants de la fréquence d'utilisation. La sensibilité de leur définition pourrait également être testée en se basant sur d'autres échelles spatio-temporelles ou en examinant l'influence des valeurs extrêmes (qui impactent notamment les moyennes obtenues). De plus, leur intégration à des modèles de prévision de la demande prenant en compte la variabilité d'usage pourrait être envisagée, de même que leur transférabilité à d'autres modes de transport pourrait être vérifiée. Enfin, le critère de Cohen devrait être adapté à un contexte de transport.

Les auteurs souhaitent remercier Jean-Simon Bourdeau pour son aide dans le prétraitement des données. Ils sont également reconnaissants envers la Société de Transport de Montréal pour l'autorisation d'accès aux données de cartes à puce OPUS et envers la Chaire de recherche du Canada sur la mobilité des personnes pour son soutien financier.

Bibliographie

Adjengue, L. (2014). *Méthodes statistiques: concepts, applications et exercices* ; Luc Adjengue. Montréal: Presses internationales Polytechnique.

Agard, B., Morency, C. et Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, 39(3), p. 399-404. doi:10.3182/20060517-3-FR-2903.00211.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: American Psychological Association.

Bagchi, M. et White, P. R. (2004). What role for smart-card data from bus systems? *Proceedings of the Institution of Civil Engineers - Municipal Engineer*, 157(1), p. 39-46. doi:10.1680/muen.2004.157.1.39.

Barabási, A.-L., González, M. C. et Hidalgo, C. A. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196), p. 779-782. doi:10.1038/nature06958.

Briand, A. S., Côme, E., Trépanier, M. et Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, 79, p. 274-289. doi:10.1016/j.trc.2017.03.021.

Chira-Chavala, T. et Coifman, B. (1996). Effects of Smart Cards on Transit Operators. *Transportation Research Record: Journal of the Transportation Research Board*, 1521, p. 84-90. doi:10.3141/1521-12.

Cleophas, T. J. et Zwinderman, A. H. (2011). Non-Parametric Tests. *Statistical Analysis of Clinical Data on a Pocket Calculator: Statistics on a Pocket Calculator*. Dordrecht: Springer Netherlands, p. 9-13.

Coe, R. (2002). *It's the effect size, stupid: What effect size is and why it is important.* Communication présentée à Annual Conference of the British Educational Research Association, University of Exeter, England.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2e éd.). Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, 49(12), p. 997-1003.

Conover, W. J., Johnson, M. E. et Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23(4), p. 351–361.

Cooper, H. et Hedges, L. V. (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation.

Deschaintres, E. (2018). *Analyse de la variabilité individuelle d'utilisation du transport en commun à l'aide de données de cartes à puce*. Montréal : École Polytechnique de Montréal. Tiré de https://publications.polymtl.ca/3284/.

Eagle, N. et Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), p. 255-268. doi:10.1007/s00779-005-0046-3.

El Mahrsi, M. K., Côme, E., Oukhellou, L. et Verleysen, M. (2017). Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 18(3), p. 712-728. doi:10.1109/TITS.2016.2600515.

Fritz, C. O., Morris, P. E. et Richler, J. J. (2012). Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of experimental psychology: General*, 141(1), p. 2-18. doi:10.1037/a0024338.

Ghosh, A., Chattopadhyay, N. et Chakrabarti, B. K. (2014). Inequality in societies, academic institutions and science journals: Gini and k-indices. *Physica A: Statistical Mechanics and its Applications*, 410, p. 30-34.

Goulet-Langlois, G., Koutsopoulos, H. N., Zhao, Z. et Zhao, J. (2017). Measuring Regularity of Individual Travel Patterns. *IEEE Transactions on Intelligent Transportation Systems*.

Huang, J., Xu, L. et Ye, P. (2015). Exploring Transit Use Regularity Using Smart Card Data of Students. In Q. Peng, K. C. P. Wang, X. Liu et B. Chen (éd.), *ICTE 2015*. Dalian, China, p. 617-625.

Inoue, J.-I., Ghosh, A., Chatterjee, A. et Chakrabarti, B. K. (2015). Measuring social inequality with quantitative methodology: Analytical estimates and empirical data analysis by Gini and k indices. *Physica A: Statistical Mechanics and its Applications*, 429(C), p. 184-204.

Joh, C.-H., Arentze, T., Hofman, F. et Timmermans, H. (2002). Activity pattern similarity: a multidimensional sequence alignment method. *Transportation Research Part B*, 36(5), p. 385-403. doi:10.1016/S0191-2615(01)00009-1.

Joh, C.-H. et Timmermans, H. (2011). Applying Sequence Alignment Methods to Large Activity-Travel Data Sets Heuristic Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2231, p. 10-17. doi:10.3141/2231-02.

Jones, P. et Clarke, M. (1988). The significance and measurement of variability in travel behaviour. *Transportation*, 15(1-2). doi:10.1007/BF00167981

Khalilzadeh, J. et Tasci, A. D. A. (2017). Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management*, 62, p. 89-96. doi:10.1016/j.tourman.2017.03.026.

Kieu, L. M., Bhaskar, A. et Chung, E. (2014). *Transit passenger segmentation using travel regularity mined from Smart Card transactions data*. Communication présentée au 93rd Annual Meeting at the Transportation Research Board, Washington, D.C.

Kim, M. et Kotz, D. (2007). Periodic properties of user mobility and access-point popularity. *Personal and Ubiquitous Computing*, 11(6), p. 465-479. doi:10.1007/s00779-006-0093-4.

Kitamura, R., Yamamoto, T., Susilo, Y. O. et Axhausen, K. W. (2006). How routine is a routine? An analysis of the day-to-day variability in prism vertex location. *Transportation Research Part A*, 40(3), p. 259-279. doi:10.1016/j.tra.2005.07.002.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F. et Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36, p. 1-12. doi:10.1016/j.trc.2013.07.010.

Manley, E., Zhong, C. et Batty, M. (2016). Spatiotemporal variation in travel regularity through transit user profiling. *Transportation*. doi:10.1007/s11116-016-9747-x.

Moiseeva, A., Tinnmermans, H., Choi, J. et Joh, C. H. (2014). Sequence Alignment Analysis of Variability in Activity Travel Patterns Through 8 Weeks of Diary Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2412, p. 49-56. doi:10.3141/2412-06.

Morency, C., Trépanier, M. et Agard, B. (2006). *Analysing the Variability of Transit Users Behaviour with Smart Card Data*. Communication présentée à 2006 IEEE Intelligent Transportation Systems Conference, Toronto, Canada.

Morency, C., Trépanier, M. et Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, 14(3), p. 193-203. doi:10.1016/j.tranpol.2007.01.001.

Nishiuchi, H., King, J. et 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), p. 1-10. doi:10.1007/s13177-012-0051-7.

Pallant, J. (2007). *SPSS survival manual* (4th ed.). Berkshire, England: McGraw-Hill.

Pas, E. I. et Koppelman, F. S. (1987). An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation*, 14(1), p. 3.

Pelletier, M. P., Trépanier, M. et Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C*, 19(4), p. 557-568. doi:10.1016/j.trc.2010.12.003.

Raux, C., Ma, T.-Y. et Cornelis, E. (2012). *Variability and anchoring points in weekly activity-travel patterns*. Communication présentée au 91st Annual Meeting of the Transportation Research Board Washington DC.

Raux, C., Ma, T.-Y. et Cornelis, E. (2016). Variability in daily activity-travel patterns: the case of a one-week travel diary. *European Transport Research Review*, 8(4), p. 1-14. doi:10.1007/s12544-016-0213-9.

Roorda, M. J. et Ruiz, T. (2008). Long- and short-term dynamics in activity scheduling: A structural equations approach. *Transportation Research Part A*, 42(3), p. 545-562. doi:10.1016/j.tra.2008.01.002.

Schlich, R. et Axhausen, K. W. (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, 30(1), p. 13-36. doi:10.1023/A:1021230507071.

Spurr, T., Chu, A., Chapleau, R. et Piché, D. (2015). A Smart Card Transaction "Travel Diary" to Assess the Accuracy of the Montréal Household Travel Survey. *Transportation Research Procedia*, 11, p. 350-364. doi:10.1016/j.trpro.2015.12.030.

Trépanier, M. (2012). L'exploitation des données de cartes à puce à des fins de planification des transports collectifs urbains. *Recherche Transports Sécurité*, 28(2), p. 139-152. doi:10.1007/s13547-011-0019-z.

Van der Laan, M., Hsu, J.-P., Peace, K. E. et Rose, S. (2010). Statistics ready for a revolution: Next generation of statisticians must build tools for massive data sets. *AMSTAT news: the membership magazine of the American Statistical Association*, 399, p. 38-39.

White, P., Bagchi, M., Bataille, H. et East, S. M. (2010). *The role of smartcard data in public transport*. Communication présentée au 12th World Conference on Transport Research, Lisbon, Portugal.

Williams, M. J., Whitaker, R. M. et Allen, S. M. (2012). *Measuring Individual Regularity in Human Visiting Patterns*, p. 117-122. doi:10.1109/SocialCom-PASSAT.2012.93.

Wilson, W. C. (1998). Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning A*, 30(6), p. 1017-1038. doi:10.1068/a301017.

Xianyu, J., Rasouli, S. et Timmermans, H. (2017). Analysis of variability in multi-day GPS imputed activity-travel diaries using multi-dimensional sequence alignment and panel effects regression models. *Transportation*, 44(3), p. 533-553. doi:10.1007/s11116-015-9666-2.