Titre: Title:	Analyse de la variabilité individuelle d'utilisation du transport en commun à l'aide de données de cartes à puce
Auteur: Author:	Elodie Deschaintres
Date:	2018
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Deschaintres, E. (2018). Analyse de la variabilité individuelle d'utilisation du transport en commun à l'aide de données de cartes à puce [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/3284/

UNIVERSITÉ DE MONTRÉAL

ANALYSE DE LA VARIABILITÉ INDIVIDUELLE D'UTILISATION DU TRANSPORT EN COMMUN À L'AIDE DE DONNÉES DE CARTES À PUCE

ELODIE DESCHAINTRES

DÉPARTEMENT DES GÉNIES CIVIL, GÉOLOGIQUE ET DES MINES

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L’OBTENTION

DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(GÉNIE CIVIL)

AOÛT 2018

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

ANALYSE DE LA VARIABILITÉ INDIVIDUELLE D'UTILISATION DU TRANSPORT EN COMMUN À L'AIDE DE DONNÉES DE CARTES À PUCE

présenté par : DESCHAINTRES Elodie

en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

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

M. AGARD Bruno, Ph. D., président

Mme MORENCY Catherine, Ph. D., membre et directrice de recherche

M. TRÉPANIER Martin, Ph. D., membre et codirecteur de recherche

Mme MC DONOUGH Kim, M. Sc. A., membre

REMERCIEMENTS

Je tiens tout d'abord à remercier ma directrice Catherine MORENCY, dont les compétences, le dévouement et le dynamisme sont désormais légendaires. Toutes ses interventions ont été une réelle source d'admiration et d'inspiration pour moi. Je remercie également mon codirecteur Martin TRÉPANIER. Ses conseils avisés et sa grande expertise dans le domaine des cartes à puce ont été d'une aide inestimable dans l'avancement de mon projet. Travailler avec eux deux a été un immense privilège.

Je remercie ensuite l'École Polytechnique de Montréal et l'ESTP Paris, école québécoise et école française auxquelles je suis rattachée, de m'avoir offert l'opportunité d'effectuer ce double diplôme. Les échanges internationaux sont des expériences très lucratives qui devraient être développées et largement recommandées à de futurs ingénieurs.

Je suis également reconnaissante envers les organismes qui ont participé de loin comme de prêt à ma recherche : merci à la Chaire de recherche du Canada sur la mobilité des personnes d'avoir financé mon projet, merci à la Chaire Mobilité de l'avoir encadré et merci à la STM de m'avoir autorisé l'accès à leurs données de cartes à puce. Je tiens à remercier tout particulièrement Kim MC DONOUGH, Jean-François CANTIN et Catherine PLOUFFE pour leur suivi, leurs idées et leur disponibilité.

De plus, je remercie tous mes collègues du bureau B-330 (anciens et actuels), sans qui cette maîtrise n'aurait pas été la même. Merci notamment à Nicolas PELÉ de m'avoir initiée au logiciel R et merci à Gabriel LEFEBVRE-ROPARS d'avoir répondu à mes nombreuses questions (et d'être venu me chercher à la frontière américaine lorsque je me suis retrouvée coincée à la douane !). Un grand merci également aux associés de recherche, avec une mention particulière pour Jean-Simon BOURDEAU, qui m'a apporté une aide très précieuse dans le prétraitement des données.

D'un point de vue plus personnel, je remercie mes amis français et québécois de Montréal qui ont toujours été là pour moi depuis mon arrivée « en terre inconnue », merci en particulier à Stephen WOODALL-KALFAIAN pour ses corrections d'anglais et nos traditionnelles pauses café, sans oublier mes amis restés en France, que la distance n'aura pas réussi à éloigner : merci de penser encore à moi comme si rien n'avait changé.

Je tiens finalement à remercier mes parents, Michel et Christine DESCHAINTRES, pour leur soutien aussi bien moral que financier. Merci d'avoir accepté mon départ à l'étranger, même si je sais que cela n'a pas été facile tous les jours, et merci de m'avoir transmis votre courage et votre détermination, sans lesquels je ne serai pas arrivée jusque-là.

Pour clôturer cette longue liste de remerciements, merci infiniment à mon conjoint, Alexis CSEJTEI, d'avoir tout sacrifié pour me suivre dans ce périple transatlantique. Merci aussi de m'avoir soutenue, encouragée (et supportée) lors des grosses crises de stress et des longues nuits d'insomnie qui ont parfois ponctué ma maîtrise.

RÉSUMÉ

L'achalandage des réseaux de transport en commun varie en fonction de nombreux paramètres. Des facteurs exogènes tels que la météo sont souvent rapportés dans la littérature mais il existe aussi des facteurs individuels : en effet, chaque usager a une utilisation temporelle et spatiale du transport en commun qui lui est propre. D'une part, des différences sont visibles entre les individus. Cette variabilité interpersonnelle est particulièrement prononcée dans les réseaux qui desservent un grand nombre de personnes du fait de la grande hétérogénéité des comportements observés. D'autre part, des variations intrapersonnelles peuvent être décelées au sein du comportement de chaque individu, à plus ou moins court terme selon l'évolution de ses contraintes et de ses préférences. Pourtant, la plupart des modèles de planification supposent encore un usage moyen et constant, de même que les services proposés sont peu adaptés à ces fluctuations. Les variations observées rendent en effet difficiles la prévision de la demande et l'ajustement de l'offre à cette demande.

Toutefois, une meilleure compréhension des comportements individuels de mobilité pourrait contribuer non seulement à développer des modèles plus justes, mais aussi à envisager des micro-ajustements et une personnalisation des services. C'est pourquoi de nombreux auteurs s'intéressent aujourd'hui à l'étude de la variabilité de ces comportements. Pour cela, des données longitudinales et individualisées sont généralement nécessaires. Les données de cartes à puce sont particulièrement adaptées à cette fin puisque, au-delà de ses fonctions primaires de collecte des recettes et de prévention de la fraude, ce système de perception automatique des titres permet de collecter de grandes quantités de données riches en informations temporelles et spatiales.

Dans cette perspective, ce mémoire vise à analyser qualitativement et quantitativement la variabilité individuelle d'utilisation du transport en commun en exploitant un an de données de cartes à puce. Ces données proviennent de la Société de Transport de Montréal (STM), opérateur des réseaux de métro et de bus de Montréal. Tout d'abord, ce projet est mis en contexte par une brève présentation de la STM et de son système de perception-validation, basé sur la carte à puce OPUS. Une revue de littérature permet ensuite de mieux connaitre les caractéristiques générales de cette technologie et de présenter des exemples de travaux réalisables à partir de ces données, notamment pour décrire l'utilisation temporelle et spatiale du transport en commun. Différentes méthodes de mesure de la variabilité des comportements de mobilité sont également recensées

(calcul d'indicateurs, analyse de séquences, modélisation) et une section entière est dédiée à l'évaluation de la variabilité interpersonnelle par la segmentation des usagers du transport en commun. Ce mémoire se poursuit par la description de la base de données utilisée, dont la taille était un des plus grands défis de cette recherche : près de 430 millions de validations réalisées par environ 2 millions de cartes sont exploitées. Une méthode de prétraitement est ainsi mise en œuvre afin de convertir les validations en déplacements et de résumer la mobilité des usagers dans des vecteurs.

La méthodologie développée dans ce mémoire repose sur la combinaison de plusieurs outils d'exploration de données et est décomposée en plusieurs étapes. En premier lieu, quatre indicateurs sont construits pour mesurer différents types de variations individuelles dans la répartition des déplacements, la fréquence, ainsi que dans l'utilisation temporelle et spatiale du réseau. Ces indicateurs sont mis à l'essais sur plusieurs groupes de cartes afin de démontrer un lien entre la tarification et la variabilité d'utilisation du transport en commun. Ces indicateurs sont également vérifiés à l'aide de différentes statistiques. En particulier, une statistique non influencée par la grosseur des échantillons manipulés est introduite. Les résultats de cette application montrent des différences plus ou moins importantes entre les usagers en fonction des titres de transport qu'ils ont utilisés pendant l'année. La régularité des utilisateurs d'abonnements est notamment confirmée.

En outre, plusieurs processus de segmentation sont proposés afin d'examiner séparément la variabilité interpersonnelle et la variabilité intrapersonnelle des comportements selon des définitions préalablement fixées. Une typologie d'usagers est d'abord créée à partir de l'intensité et de la dispersion mensuelle des déplacements de chaque carte pendant l'année. Tous les passagers du réseau de la STM sont ainsi répartis en six groupes caractérisés par des intensités et des profils temporels d'utilisation différents. En plus des quatre indicateurs adoptés précédemment, plusieurs indices temporels et spatiaux (ou modaux) sont calculés afin d'aider l'interprétation de la typologie obtenue. À la suite de ce portrait global de la mobilité montréalaise, un bloc de cartes particulier est approfondi : celui des utilisateurs d'abonnements annuels avec une amplitude de 12 mois, eux aussi segmentés en six groupes. Cet échantillon d'environ 57000 cartes est manié dans tout le reste du mémoire car les analyses qui s'y trouvent requièrent une plus grande complexité de calcul.

Une typologie de semaines est produite en fonction de l'intensité et de la distribution quotidienne des déplacements de chaque carte-semaine. Cette typologie sert à déterminer la stabilité du

comportement hebdomadaire de chaque individu et permet ainsi de quantifier la variabilité intrapersonnelle moyenne à l'intérieur de chaque groupe d'usagers à l'aide d'indicateurs. Cette segmentation hebdomadaire est ensuite utilisée pour représenter chaque usager par une séquence de types de semaines. Une troisième typologie (une typologie de séquences) est alors conçue après avoir appliqué une distance originale et non euclidienne prenant en compte la chronologie et l'organisation des semaines de chaque usager.

Enfin, le croisement de tous les résultats obtenus pour les utilisateurs d'abonnements annuels (typologies et indicateurs) permet de conclure que les usagers les plus fréquents et les plus réguliers au niveau mensuel le sont également au niveau hebdomadaire. Ces usagers sont, pour la plupart, des travailleurs captifs du transport en commun pendant les jours ouvrables. Quelques groupes de comportements plus atypiques sont révélés, en particulier dus à l'émergence de la semaine de travail de quatre jours ou à une utilisation préférentielle du transport en commun pour des activités de fin de semaine. Quelques usagers occasionnels sont également repérés, mais leur minorité légitimise les intérêts de la STM à fidéliser ses clients avec un abonnement annuel.

En tant que travail pionnier sur la variabilité individuelle d'utilisation du transport en commun, ce mémoire se termine par une longue liste de limites à dépasser et de perspectives futures à explorer.

ABSTRACT

Transit ridership varies according to many parameters. Exogenous factors such as weather are often reported in the literature, but there are also individual factors, as every user has a specific temporal and spatial use of the transit system. On the one hand, differences can be found between individuals. This interpersonal variability is particularly pronounced in networks which serve a lot of people because of the great heterogeneity of observed behaviors. On the other hand, intrapersonal variations are noticeable within the behavior of each individual, more or less in a short-term scope, depending on the evolution of the individual's restrictions and preferences. Nevertheless, most planning models still assume an average and constant use, and the provided services are not really adapted to these fluctuations. Indeed, variations in public transit use make it difficult to forecast demand and adjust supply to this demand.

However, a better understanding of individual mobility behaviors may contribute not only to the development of more accurate models, but also to the enabling of micro-adjustments and customization of services. This is why many authors are now interested in studying the variability of these behaviors. To this end, longitudinal and individualized data are needed. Smart card data is particularly suitable for this purpose since, beyond its primary functions of revenue collection and fraud prevention, this automated fare collection system gathers large amounts of data, rich in temporal and spatial information.

As such, this thesis aims to qualitatively and quantitatively analyze the individual variability of public transit use by mining one year of smart card data. This data comes from the Société de Transport de Montréal (STM), operator of the Montreal subway and bus networks. First, this project is contextualized by a brief presentation of the STM and its perception-validation system, based on OPUS smart cards. A literature review then allows to better explain the typical characteristics of this technology and to present examples of work which can be done using this data, particularly to describe the temporal and spatial use of public transport. Various methods of measuring the variability of mobility behaviors are also identified (calculation of indicators, sequence analysis, modeling) and an entire section is dedicated to the evaluation of interpersonal variability by way of the segmentation of public transit users. This thesis continues with the description of the database used, the size of which was one of the biggest challenges of this research: almost 430 million validations made by about 2 million cards are considered. A

preprocessing method is thus implemented in order to convert validations into trips and to summarize users' mobility into vectors.

The methodology developed in this thesis relies on the combination of several data mining tools and is broken down into several stages. Firstly, four indicators are constructed to measure different types of individual variations in trip distribution, frequency, as well as in temporal and spatial use of the network. These indicators are tested on several groups of cards to demonstrate a relationship between fare and variability in transit use. These indicators are also verified using different statistics. A statistic which is not influenced by the large size of the samples handled is therefore introduced. The results of this application show more or less important differences between users according to the products they used during the year. The regularity of pass users is confirmed.

Moreover, several segmentation algorithms are suggested to separately study interpersonal variability and intrapersonal variability of behaviors according to previously formulated definitions. A typology of users is first created from the intensity and the monthly trip dispersion of each card during the year. All the passengers of the STM network are thus divided into six clusters characterized by different intensity levels and time profiles of use. In addition to the four previously adopted indicators, several temporal and spatial (or modal) indices are calculated to help interpret the typology obtained. Following this overview of Montrealers' mobility, a particular subset of cards is selected: annual pass users with an amplitude of 12 months, also segmented into six groups. This sample of about 57000 cards is handled throughout the rest of the thesis as the analyses therein require a greater computational complexity.

A typology of weeks is then produced according to the intensity and the daily distribution of the trips of each card-week. This typology is used to determine the stability of each individual's weekly behavior and thus makes it possible to quantify the average intrapersonal variability within each group of users using indicators. This weekly segmentation is then employed to represent each card user by a sequence of week types. A third typology (a typology of sequences) is then conceived after having applied an original and non-Euclidean distance taking into account the chronology and organization of the weeks of each user.

Finally, the comparison of all the results obtained for the annual pass users (typologies and indicators) allows to conclude that the most frequent and regular users at the monthly level are also the most frequent and regular at the weekly level. These passengers are mostly captive users of

public transit during working days. Some groups of more atypical behaviors are revealed, particularly due to the emergence of the four-day work week or because of a preferential use of public transit for weekend activities. Some occasional users are also identified, but their minority legitimizes the interests of the STM to retain customers with an annual subscription.

As a pioneering work on individual variability in transit use, this thesis is concluded with a long list of limitations to surpass and future perspectives to explore.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

AFC	*Automated fare collection*
AM	*Ante meridiem* (avant-midi)
AMT	Agence métropolitaine de transport
ANOVA	*Analysis of variance*
AOT	Autorité organisatrice de transport
ARTM	Autorité régionale de transport métropolitain
BPA	Boîte de perception dans l'autobus
BPSS	*Between-person sum of squares*
c.-à-d.	c'est-à-dire
CEB	Code d'emplacement billettique
CDD	*Cross-correlation distance*
CIT	Conseils intermunicipaux de transport
CMJ	Carte magnétique jetable
CMM	Communauté métropolitaine de Montréal
CO	Combinaison
COS	*Card operating system*
CPCT	Carte à puce commune de transport
CPO	Carte à puce occasionnelle
CRF	*Conditional random fields*
CRT	Conseil régional de transport
CTCUM	Commission de transport de la Communauté urbaine de Montréal
CTM	Commission de transport de Montréal
DAT	Distributrice automatique de titres

DBSCAN	*Density-based spatial clustering of applications with noise*
DTW	*Dynamic Time Warping*
EEPROM	*Electrically erasable programmable read-only memory*
FPT	*First Passage Time*
GPS	*Global Positioning System*
GTFS	*Global Transit Feed Specification*
HAC	*Hierarchical agglomerative clustering*
ISO	*International Organization for Standardization*
JO	Jours ouvrables
JNO	Jours non ouvrables
LCA	*Latent class analysis*
MRC	Municipalités régionales de comté
NYCT	*New York City Transit*
OD	Origine-Destination
OMIT	Organisme municipal et intermunicipal de transport
PCA	*Principal component analysis*
PM	*Post meridiem* (après-midi)
REM	Réseau express métropolitain
RFID	*Radio Fréquence IDentification*
RTL	Réseau de transport de Longueuil
RTM	Réseau de transport métropolitain (Exo)
SAM	*Sequential Alignment Method*
SEM	*Structural equation modeling*
SIM	*Subscriber Identity Module*

SQL	*Structured Query Language*
SCAFC	Smart Card Automated Fare Collection
STCUM	Société de transport de la Communauté urbaine de Montréal
STL	Société de transport de Laval
STM	Société de transport de Montréal
STO	Société de transport de l'Outaouais
SVM	*Support vector machine*
TC	Transport en commun
TTAPS	*Toronto Travel Activity Panel Survey*
TTS	*Total sum of squares*
WPSS	W*ithin-person sum of squares*

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

D'après Vuchic (2007), les origines du transport en commun en tant que service organisé remontent au XVIème siècle. Ce mode de transport apparaît d'abord sous la forme de bateaux et de véhicules à traction animale, principalement utilisés pour le déplacement inter et intra urbain de personnes aisées et de marchandises. Son développement et sa démocratisation se font surtout à partir du XIXème siècle pour permettre l'urbanisation et l'agrandissement des villes. Dès lors, il devient un moyen efficace pour transporter simultanément des volumes plus importants d'individus vers leur lieu de travail. Néanmoins, la vulnérabilité des chevaux et le mauvais état des routes entraînent peu à peu le déclin de la traction animale. Les omnibus et les tramways hippomobiles cèdent ainsi leur place aux tramways à vapeur puis aux tramways électriques. Après avoir connu une certaine période d'apogée, le tramway doit faire face à la compétition de l'automobile et sa conversion progressive vers le bus a lieu dès le début du XXème siècle. Parallèlement à ces évolutions technologiques, la gouvernance du transport collectif subit également des changements : les entreprises privées sont peu à peu étatisées, du moins en partie, et le transport collectif par autobus devient donc public. Cependant, l'apparition de problèmes liés à la congestion automobile conduit rapidement à un regain d'intérêt pour le rail, en particulier avec le développement du train et du métro. Par ailleurs, suite à l'avènement de ces modes plus lourds, l'automatisation des systèmes de perception des titres de transport (autrefois manuels) devient nécessaire.

L'histoire du transport en commun à Montréal s'inscrit dans cette longue lignée. Ainsi, plusieurs compagnies de transport collectif se sont succédé avant que la Société de transport de Montréal (STM) ne soit créée (Société de transport de Montréal, 2018a). Aujourd'hui, cette entreprise publique opère un large réseau de transport en commun à l'aide du système de validation et perception OPUS. Ce réseau couvre toute l'île de Montréal, soit une superficie totale d'environ 500 km^2, et dessert les plus de 4 millions d'habitants du Grand Montréal (17% d'entre eux utilisant le transport en commun au moins une fois par jour d'après l'enquête OD 2013). Cette grande envergure suscite de nombreuses problématiques de mobilité, notamment celle qui sera traitée dans ce mémoire.

1.1 Mise en contexte

1.1.1 La Société de transport de Montréal (STM)

Actuelle agence de Montréal, la STM est la dernière en date des nombreuses compagnies qui se sont succédé dans l'histoire du transport en commun à Montréal. En effet, plusieurs sociétés privées de tramway ont d'abord existé entre 1861 et 1951. Les premiers bus sont mis en service en 1919, puis des trolleybus sont ajoutés en 1937. En 1951, la gestion de l'ensemble des transports en commun de Montréal passe aux mains d'un organisme public, la Commission de transport de Montréal (CTM). C'est aussi la date à partir de laquelle les lignes de tramway commencent à être remplacées par des autobus. Le dernier tramway est retiré de Montréal en 1959. Les 20 premières stations du métro sont inaugurées en 1966 pour desservir l'exposition universelle de 1967. Cette inauguration annonce également la fin des trolleybus. En 1970, la CTM devient la Commission de transport de la Communauté urbaine de Montréal (CTCUM), qui est remplacée en 1985 par la Société de transport de la Communauté urbaine de Montréal (STCUM), avant de devenir à son tour la Société de transport de Montréal (STM) en 2002 (Société de transport de Montréal, 2018a). En 2017, la STM est la 15^{e} entreprise en importance au Québec, avec une valeur de remplacement des actifs estimée à 26 G$ (Société de transport de Montréal, 2017a).

Par ailleurs, la STM fait partie d'un cadre institutionnel plus large visant à organiser les transports collectifs du Grand Montréal. Le territoire métropolitain de Montréal est décomposé en 5 grandes régions : l'île de Montréal, Laval, Longueuil, et les couronnes nord et sud. La STM gère ainsi le réseau de transport en commun de l'île de Montréal et travaille en collaboration avec deux autres sociétés : la Société de transport de Laval (STL) et le Réseau de transport de Longueuil (RTL), qui administrent respectivement les réseaux de Laval et de Longueuil. De plus, le réseau de trains de banlieue et de bus dans les couronnes de Montréal, autrefois régi par l'Agence métropolitaine de transport (AMT) et onze sociétés municipales et intermunicipales de transport (9 CIT : conseils intermunicipaux de transport, 1 CRT : conseil régional de transport et 1 OMIT : organisme municipal et intermunicipal de transport), est désormais opéré par une seule entité : le Réseau de transport métropolitain (RTM), nouvellement appelé Exo. Le tout est chapeauté par l'Autorité régionale de transport métropolitain (ARTM). Cette organisation est le résultat d'une modification récente de la gouvernance du transport collectif dans la région métropolitaine de Montréal. Votée par les différentes municipalités de la Communauté métropolitaine de Montréal (CMM) dans un

but d'uniformité et d'interactions entre les différents organismes, la « loi modifiant principalement l'organisation et la gouvernance du transport collectif dans la région métropolitaine de Montréal » (Québec, 2016) a permis la décomposition de l'AMT en ARTM et RTM ainsi que la fusion des différentes sociétés municipales et intermunicipales avec le RTM, donnant naissance à 5 AOT (autorités organisatrices de transport) pour le Grand Montréal au lieu de 15 (Agence métropolitaine de transport, 2016). L'interdépendance et le rôle de chacune de ces entités sont donnés dans le schéma ci-dessous (d'après Exo (2018) et les sites officiels de chaque organisme). De surcroît, la Figure 1.2 montre le territoire de desserte de chacune des 5 entités et du Réseau express métropolitain (REM), un projet de train léger qui viendra compléter l'offre actuelle en 2021.

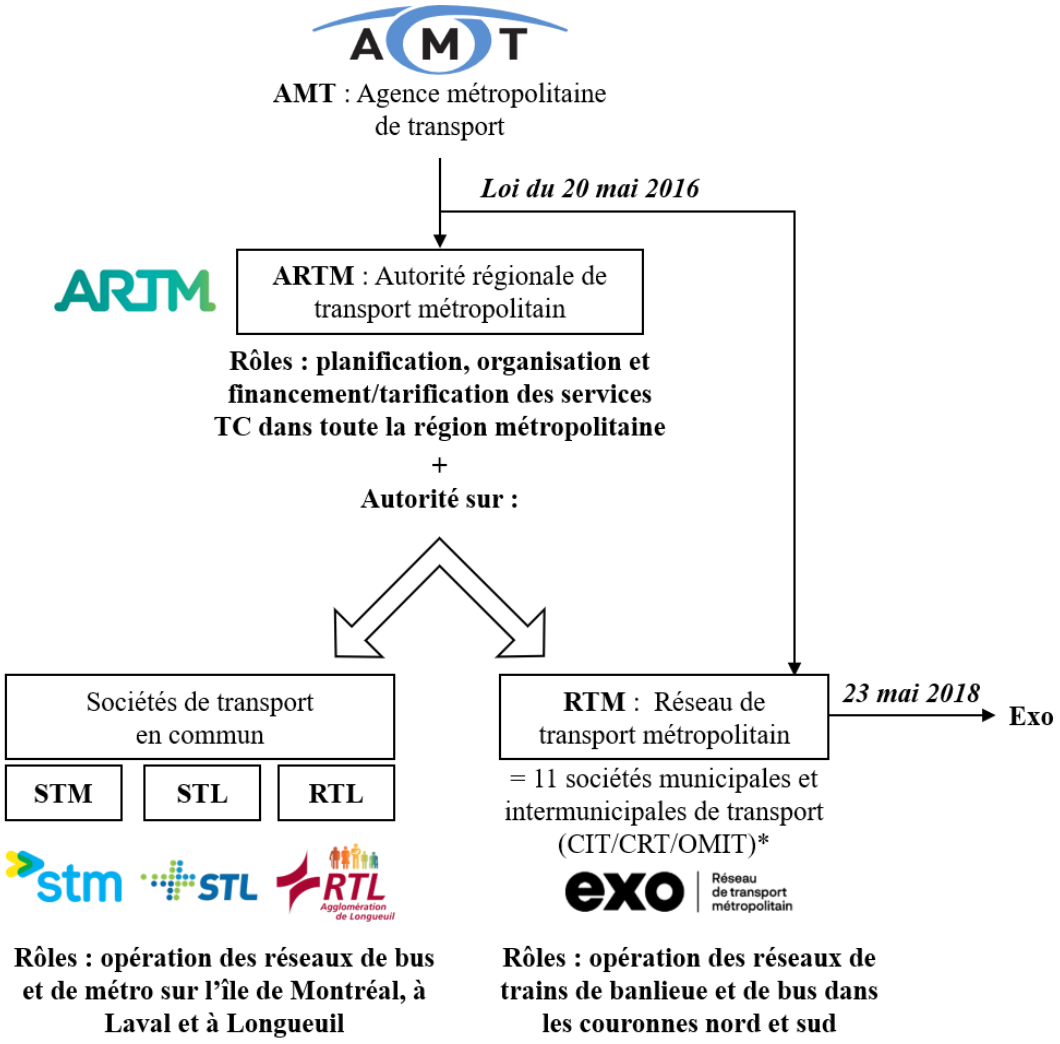


Figure 1.1 Autorités organisatrices de transport (AOT) du Grand Montréal

* un CIT regroupe plusieurs municipalités, un CRT réunit des municipalités régionales de comté (MRC) alors qu'un OMIT est géré par une seule municipalité (Meloche et al., 2012).

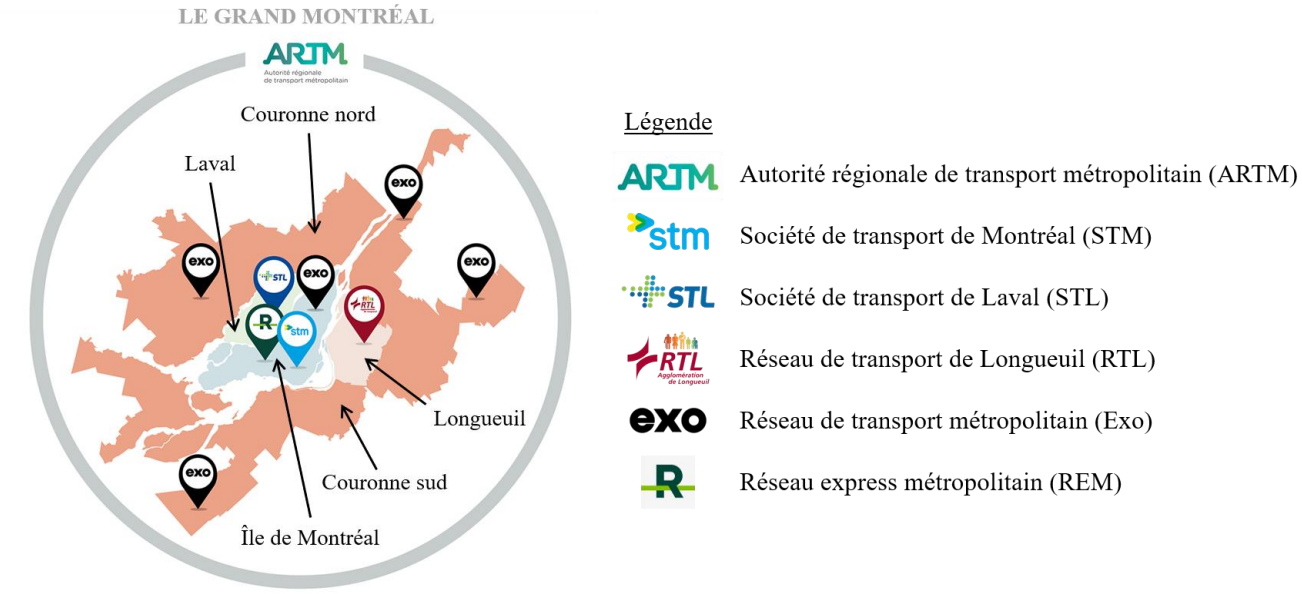


Figure 1.2 Territoire de desserte des 5 organismes de transport existants et du futur REM

(Carte initiale tirée de https://rtm.quebec/fr/actualites/nouvelles-evenements/nouvelles/exo-organismes-transport-qui-fait-quoi2)

De son côté, la STM « met en œuvre, exploite et maintient » (Société de transport de Montréal, 2017a) un réseau de métro, un réseau de bus ainsi qu'un service de transport adapté. Le réseau de métro est constitué de 4 lignes d'une longueur totale de 71 km et composées de 68 stations. Le réseau de bus est quant à lui organisé en 220 lignes régulières déployées en plusieurs types de service : le réseau local, le réseau 10 minutes max, le réseau de nuit, le réseau express, les Navettes Or réservées aux aînés et les lignes spéciales (dont fait notamment partie la ligne 747 qui se rend à l'aéroport international Pierre-Elliott-Trudeau). Un service de taxi collectif est également mis en œuvre dans les quartiers moins bien desservis. Enfin, le transport adapté est un service à la demande assuré par 86 minibus de la STM ainsi que par 13 fournisseurs de taxis. Des informations complémentaires peuvent être obtenues avec la Figure 1.3 ci-après extraite du Budget 2018 de la STM ou sur le site officiel de la STM (http://stm.info/fr). De plus, le plan des réseaux de métro et de bus est donné en ANNEXE A. Tous ces services ont généré un achalandage annuel de 429.5 millions déplacements en 2017, soit une moyenne de 1.2 million de déplacements par jour (Société de transport de Montréal, 2018e).

Réseau du métro[1]

Le métro comporte quatre lignes qui couvrent 71 kilomètres et desservent 68 stations. Le parc de matériel roulant compte 852 voitures, dont 195 MR-63, 423 MR-73 et 234 AZUR qui parcourront 88,1 millions de kilomètres commerciaux en 2018.

Réseau des bus[1]

La STM possède un parc de 1 827 bus, dont 1 570 réguliers (12 mètres) et 257 articulés (18 mètres), ainsi que 16 minibus pour les navettes OR et le service urbain. Son réseau couvre l'île de Montréal, soit un territoire de près de 500 kilomètres carrés. Il compte 220 lignes, dont 209 sont accessibles aux personnes à mobilité réduite et 23 sont dédiées au service de nuit. De plus, environ 375 kilomètres de voies comportant des mesures préférentielles pour bus, incluant des voies réservées, permettent des déplacements plus rapides. L'offre de service bus devrait atteindre 68,2 millions de kilomètres commerciaux en 2018.

Transport adapté[1]

La STM offre un service de transport adapté porte à porte pour les personnes ayant des limitations fonctionnelles avec son parc de 86 minibus. Ainsi, plus de 31 000 clients effectueront plus de 4,1 millions de déplacements en 2018 sur l'île de Montréal et en périphérie, à l'aide de minibus de la STM et des 13 fournisseurs de services de taxi réguliers et accessibles.

Note 1 : en date du 1er septembre 2017

Figure 1.3 Informations sur les services offerts par la STM (extrait du Budget 2018)

La politique actuelle de la STM repose sur le Plan stratégique organisationnel 2025 adopté en juin 2017. Ce plan est décomposé en 7 axes et 16 objectifs articulés autour des 4 grandes orientations stratégiques suivantes :

- Améliorer l'expérience client
- Adapter l'organisation à l'évolution de la gouvernance
- Maîtriser les finances
- Attirer, développer et mobiliser les talents

En outre, ce plan prévoit une hausse d'achalandage de 10 millions de déplacements par rapport à l'achalandage actuel, soit un total de 440 millions de déplacements en 2025 (Société de transport de Montréal, 2017b).

1.1.2 Le système Validation & Perception de la STM

Pour contrôler l'accès à ses réseaux de métro et de bus, la STM s'est dotée d'un système automatisé de perception des titres (*AFC : Automated fare collection* en anglais). Ainsi, l'usager doit faire

valider son titre de transport par un dispositif de perception tarifaire pour pouvoir entrer dans une station de métro ou embarquer dans un bus.

Le système utilisé à la STM est appelé le système OPUS. Implanté en 2008, il a fonctionné en parallèle avec l'ancien système pendant une année avant de le remplacer définitivement à partir de 2009. Aujourd'hui trois types de support coexistent dans ce système : la carte à puce OPUS (ou CPCT, carte à puce commune de transport), le carte à puce occasionnelle (CPO) et la carte magnétique jetable (CMJ). Leurs caractéristiques respectives sont données dans le tableau ci-dessous.

Tableau 1.1 Caractéristiques des trois types de supports utilisés à la STM

Type support	**Carte OPUS**	**CPO**	**CMJ**
Technologie	Puce (sans contact)	Puce (sans contact)	Bande magnétique (avec contact)
Matière	Plastique	Papier	Papier
Nombre max de contrats	4	1	1
Rechargeable	Oui	Non	Non
Utilisé en dehors de l'île	Oui	Non	Non
Photo			

La carte OPUS est une carte à puce traditionnelle sans contact. Cette technologie sera expliquée plus amplement dans la section 2.1 du Chapitre 2. Cette carte en plastique est rechargeable et peut contenir jusqu'à 4 contrats (ou titres de transport) différents. Toutefois, certains types de contrats ne peuvent pas cohabiter, en particulier lorsqu'il s'agit de contrats équivalents, car la machine automatique ne sait alors pas en sélectionner un préférentiellement à l'autre. De plus, la carte OPUS est commune aux trois sociétés de transport du Grand Montréal (STM, STL, RTL) et, bien que sa validation ne soit pas obligatoire à bord des trains, elle est également utilisée dans le réseau d'Exo.

Les deux autres supports (CPO et CMJ) disponibles sont des cartes en papier non rechargeables et propres à la STM. Leur principale différence réside dans leur mode de fonctionnement. Comme la carte OPUS, la CPO contient une puce alors que la CMJ est une carte à bande magnétique; elle nécessite donc un contact pour être lue par le système de perception. Ces deux supports ne peuvent

comporter qu'un seul type de contrat, mais la CPO autorise une plus grande diversité de titres que la CMJ.

Les différents titres de transport de la STM pouvant être chargés sur chacun de ces supports sont détaillés dans la grille tarifaire de l'ANNEXE B. Plusieurs titres existent selon le nombre de passages ou la durée d'abonnement souhaités, mais aussi selon le type de support utilisé. Certains titres sont également disponibles en différents tarifs. La STM propose deux tarifs : le tarif ordinaire et le tarif réduit, ce dernier se déclinant en deux catégories. La première de ces deux catégories de tarif réduit est destinée aux extrémités de la population : les séniors de 65 ans et plus, et les enfants de 6 à 17 ans (le transport en commun étant gratuit pour les enfants de moins de 6 ans). La deuxième catégorie est réservée aux étudiants entre 18 et 25 ans. Par ailleurs, la grille tarifaire donne la validité de chaque titre dans les deux autres sociétés de transport (STL et RTL).

D'une part, ces titres de transport peuvent être achetés à différents endroits :

- Aux distributrices automatiques de titres (DAT)
- Aux bornes de rechargement (pour recharger la carte OPUS uniquement)
- Auprès du changeur (contre argent comptant)
- Dans certains commerces (la dernière colonne de la grille tarifaire de l'ANNEXE B précise les titres pouvant être achetés chez ces détaillants)

D'autre part, ils doivent être validés à chaque entrée dans le réseau. Pour cela, l'usager doit passer son support sur (ou dans) un équipement spécialisé de validation et perception : un portillon ou un tourniquet pour le métro, une BPA (Boîte de Perception Autobus) pour le bus. Ces équipements, illustrés dans la Figure 1.4, vont lire les informations contenues dans le support puis coder le résultat de la validation sur l'équipement et sur le support. Un système de priorités est utilisé quand plusieurs produits tarifaires sont présents sur la même carte : à chaque produit est associé un niveau de priorité et le système choisit toujours le produit le plus avantageux pour l'usager.

Les données écrites sur l'équipement à chaque validation tarifaire sont ensuite recueillies à l'aide d'une sonde lectrice infrarouge. Cet échange entre les équipements et un système d'information centralisé, réalisé de manière asynchrone car aucun équipement ne communique en direct, est effectué de manière assez récurrente (en général quotidiennement), sans quoi des données seraient

perdues. En effet, les équipements de collecte ont un espace mémoire limité et ils sont programmés pour réécrire sur les premières données lorsque leur limite d'enregistrement est atteinte.

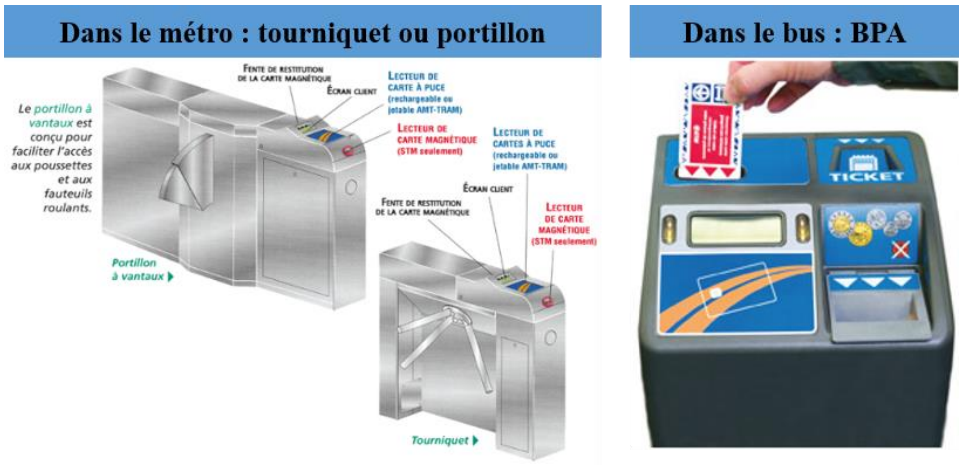


Figure 1.4 Équipements de validation et perception des titres (métro et bus)

Images tirées de : http://www.rtl-longueuil.qc.ca/fr-CA/tarifs/carte-opus/ et http://blog.fagstein.com/2008/08/06/stm-fare-card-problems/

Cet échange fait ainsi partie du système informationnel de la STM schématisé dans la Figure 1.5. La communication entre les équipements de perception (mais aussi de vente) et le système centralisé est en fait bidirectionnelle. Les informations fournies par les équipements sont des données d'activité (transaction monétaire ou transaction de déplacement) non modifiables. En sens inverse, le système centralisé renvoie des paramètres (données modifiables). Ces paramètres, tels que la grille tarifaire ou la durée de correspondance (fixée à 120 minutes à la STM), permettent à chaque AOT de personnaliser leur système.

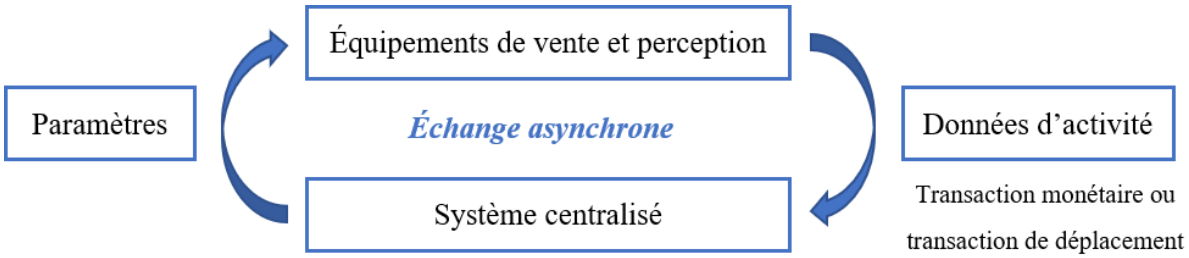


Figure 1.5 Schéma du système informationnel validation et perception OPUS

Les données collectées à chaque transaction de déplacement sont globalement composées d'un identifiant de validation, d'un numéro de support, du titre utilisé, d'informations temporelles et

spatiales sur la date, l'heure et le lieu d'embarquement de l'usager. Ces données seront explicitées dans la section 3.1 du Chapitre 3. Il est néanmoins à souligner que si les numéros de carte OPUS et CPO sont uniques, à l'inverse les numéros de CMJ ne le sont pas car ils sont réutilisés après un délai assez long.

Par ailleurs, quelques anomalies possibles dans ces données ont été répertoriées par la STM :

- Écriture d'un produit non valide sur la CMJ : d'après la grille tarifaire, la carte magnétique ne peut contenir que trois types de titres (1 passage, 2 passages, titre de groupe) ;
- Problème d'heures dans le tourniquet : suite à des problèmes techniques, l'horodatage des tourniquets peut être perturbé ;
- Ligne de bus moyennement fiable : en 2016, le numéro de ligne était rentré manuellement par le chauffeur, conduisant parfois à des oublis ou des erreurs.

De plus, le projet d'embarquement par toutes les portes initié en 2016 sur la ligne 121 (Société de transport de Montréal, 2016a), puis sur les lignes 139 et 439 (Société de transport de Montréal, 2016b), a entraîné une perte de données pour les validations de bus. En effet, cette mesure a permis aux détenteurs d'abonnements (annuels, mensuels, hebdomadaires) d'entrer par les portes à l'arrière sans avoir à valider leur titre à l'avant. Ces possibles sources d'erreurs ou de données manquantes n'ont pas été considérées et font donc partie des imperfections de ce projet ; celles-ci seront rappelées à la fin de ce mémoire.

1.1.3 La carte OPUS

Créée en 2008 avec le système OPUS qui lui a donné son nom, la carte OPUS est aujourd'hui le support le plus utilisé par les usagers de la STM. Quatre types de cartes se partagent le marché :

- Carte anonyme
- Carte personnalisée sans photo
- Carte personnalisée avec photo
- Carte d'accès

La carte anonyme est la carte de base possédée par un détenteur inconnu. Au contraire, la carte est dite personnalisée ou enregistrée lorsqu'elle est associée à un individu donné dont l'identité a été

répertoriée. Cette personnalisation, réalisée à l'initiative de l'usager, lui assure une garantie en cas de perte ou de vol de la carte : le cas échéant, l'usager pourra récupérer le solde de ses titres de transport sur une nouvelle carte. De plus, une photo doit obligatoirement être apposée sur la carte pour pouvoir bénéficier des titres à tarif réduit. Cette carte avec photo est automatiquement personnalisée. Enfin, la carte d'accès est réservée à du personnel qualifié qui travaille sur le réseau.

La carte personnalisée avec photo est non transférable (Règlement R-105 de la STM). En revanche, rien n'interdit explicitement de prêter une carte anonyme ou une carte personnalisée sans photo. Une même carte peut donc être utilisée pour réaliser les déplacements de différentes personnes. A l'inverse, les déplacements d'une même personne peuvent être collectés sur différentes cartes si un usager perd ou renouvelle sa carte au terme de sa période de validité : 4 ans en général, mais des durées spécifiques sont appliquées pour les cartes OPUS avec photo (Société de transport de Montréal, 2018b). Une nouvelle carte est alors produite, avec un numéro différent de la carte initiale, rendant impossible le suivi continu de l'individu. Ainsi, non seulement une carte peut correspondre à plusieurs personnes, mais plusieurs cartes peuvent également se rapporter à la même personne. Dans le système actuel de la STM il serait donc totalement faux de prétendre qu'une carte est équivalente à un usager.

En outre, le numéro de chaque carte est crypté de manière à être rendu anonyme. En effet, les données de la STM sont envoyées à un fournisseur extérieur chargé de les anonymiser. Les informations personnelles de chaque usager enregistré sont rattachées à un dossier client, mais aucune donnée n'est inscrite dans la carte sauf l'identité de l'usager (nom et prénom). De plus, ces informations personnelles ne sont pas consultables sans autorisation de la police, qui permet un accès limité pendant une durée de 48h.

Par ailleurs, chaque carte est caractérisée par un état et un statut. D'une part, son état est dichotomique : une carte est *valide* ou *invalide*. Les cartes invalides sont des cartes mises sur liste noire suite à un comportement frauduleux. Ces cartes sont repérées manuellement par la STM et sont ensuite refusées par tous les équipements du système. D'autre part, chaque carte possède un des statuts suivants : *initiale* (avant la vente), *détenue* (et fonctionnelle), *inutilisable* (mise sur liste noire), *retournée*, *reconstituée* (reconditionnée), *rebutée* (fin de vie).

1.2 Problématique et objectifs

Les agences de transport de grande taille comme la STM sont confrontées à de nombreuses problématiques liées à la grande hétérogénéité des individus transportés. En effet, l'achalandage des réseaux de transport en commun varie constamment en fonction de paramètres exogènes comme la météo, les événements, les congés ou les arrêts imprévus du service, qui influent sur les comportements individuels de mobilité, mais aussi en fonction de la variabilité (ou, par opposition, la régularité) intrinsèque de ces comportements. L'impact de cette deuxième cause de variabilité dans l'utilisation du transport en commun est souvent moins connu, car plus difficile à mesurer.

Pourtant, différents types de variabilités individuelles sont observés dans l'utilisation du système: des variations dans la fréquence d'utilisation ou le choix modal (métro versus autobus) des usagers, ainsi que des variations temporelles et spatiales en termes de jours d'utilisation, de plage horaire ou de lieu d'embarquements choisis par les usagers pour se déplacer. De plus, ces variations peuvent être appréciées selon deux angles différents. Il existe des variations entre les usagers (variabilité interpersonnelle) et des variations au sein du comportement d'un même usager (variabilité intrapersonnelle). La variabilité interpersonnelle est expliquée par la diversité des individus, qui ont chacun leurs propres habitudes de déplacements, alors que la variabilité intrapersonnelle est due à l'adaptation et à l'évolution de l'être humain, qui modifie son comportement de mobilité dans le temps en fonction de ses contraintes quotidiennes ou des évènements qui surviennent au cours de sa vie.

Or, ces variations individuelles rendent difficile l'ajustement de l'offre à la demande, ce qui peut induire des coûts d'opération supplémentaires et une affectation non optimale des véhicules sur le réseau. Pour permettre une meilleure adéquation offre-demande, une prévision plus précise de la demande est nécessaire, ce qui permettrait idéalement ensuite de faire des micro-ajustements du service voire de développer des systèmes d'information personnalisée en temps réel. Cependant, malgré l'existence connue de variations dans l'utilisation des réseaux, de nombreux modèles supposent encore un usage moyen et constant du transport en commun avec un taux de mobilité fixe sur toute l'année. Cette limitation est principalement due à un manque de données longitudinales et individualisées permettant de détecter des variations dans les comportements de mobilité.

En effet, les données des enquêtes Origine-Destination (enquêtes téléphoniques de mobilité qui recensent les déplacements d'environ 4 à 5% des Montréalais) ne suffisent pas. Celles-ci n'étant collectées que tous les cinq ans, elles ne permettent pas de suivre l'évolution de la demande de manière ininterrompue. De plus, elles ne rapportent les comportements de mobilité que d'une portion infime de la population. Les données de cartes à puce, recueillies à chaque validation tarifaire de l'usager, permettent cependant de pallier ce manque de continuité et de représentativité (dans le cas particulier du transport en commun).

D'une part, la disponibilité de données de carte à puce permet une analyse longitudinale du comportement des usagers du transport en commun. En effet, ces données fournissent des séries de validations tarifaires successives qui permettent de suivre chronologiquement les usagers. Selon le taux de pénétration de la carte à puce parmi les usagers et la durée de la période de collecte, de gros volumes de données peuvent être exploités. Ces données sont de nature temporelle, la date et l'heure de chaque validation étant enregistrées, mais aussi de nature spatiale puisque des informations sont disponibles sur le lieu de la validation. La précision des mesures temporelles est à la seconde près et les embarquements et/ou débarquements peuvent être géospatialisés avec les coordonnées de chaque arrêt ou station. La quantité et la qualité des données de cartes à puce permettent ainsi d'analyser temporellement et spatialement la variabilité d'utilisation du réseau de transport en commun à travers une longue période.

D'autre part, cette variabilité peut être caractérisée à la fois aux niveaux agrégé et désagrégé : les déplacements peuvent être suivis pour l'ensemble de la clientèle desservie, mais aussi à l'échelle individuelle de l'usager, car il est désormais possible de lier les données de ses déplacements à un numéro unique de carte. Si l'étude de la mobilité collective permet de déceler des phénomènes cycliques et saisonniers de la demande, l'étude des comportements individuels permet quant à elle de différencier différents types d'usagers, par exemple en distinguant les usagers réguliers des usagers occasionnels. Ainsi, au-delà de donner une simple tendance globale, une analyse individualisée permet d'expliquer les fluctuations de la demande à une granularité plus fine. De plus, elle permet de mieux connaître tous les usagers du réseau (et non seulement les plus dominants) afin de pouvoir ensuite mieux adapter le service aux besoins spécifiques de chacun.

Cette recherche s'inscrit donc dans un cadre général visant à mieux comprendre les comportements individuels de mobilité. Plus particulièrement, ce projet a pour objectif principal d'analyser et de

mesurer la variabilité d'utilisation du transport en commun à l'échelle individuelle (donc totalement désagrégée) à l'aide d'un an de données de cartes à puce couvrant les réseaux de bus et de métro de la STM. Des objectifs plus spécifiques ont été formulés comme suit :

- Sélectionner des techniques avancées permettant de traiter et de valoriser des données massives (données de cartes à puce)
- Qualifier et quantifier différents types de variations dans l'utilisation du transport en commun à l'aide d'outils de visualisation et d'indicateurs
- Développer des prototypes d'analyse dans lesquels seront distinguées la variabilité interpersonnelle et la variabilité intrapersonnelle
- Créer une typologie d'usagers et caractériser chaque groupe obtenu
- Examiner la relation entre l'usage du système de transport collectif et sa tarification

Par ailleurs, les grands défis méthodologiques, analytiques et opérationnels auxquels prend part ce projet sont les suivants :

- Évaluer le potentiel des données de cartes à puce à expliquer et prédire le comportement des usagers du transport en commun
- Contribuer à l'amélioration des modèles de prévision de la demande, à un meilleur ajustement de l'offre, et à la mise en place d'une tarification intégrée et personnalisée

1.3 Structure du mémoire

Ce mémoire est composé de huit chapitres organisés dans l'ordre chronologique des recherches réalisées. Tout d'abord, le présent chapitre pose quelques jalons et met en contexte le projet à l'étude. La Société de transport de Montréal (STM) et les spécificités de son système de validation et perception OPUS sont ainsi présentés, puis la problématique et les objectifs qui en découlent sont énoncés.

Le chapitre 2 propose une revue de littérature des différents thèmes qui seront abordés dans ce mémoire. Cette revue dresse un portrait détaillé de l'état de la recherche dans le domaine des cartes à puce, et en particulier de leur utilisation dans l'étude des comportements de mobilité. Différentes

méthodes pour analyser et mesurer la variabilité d'utilisation du transport en commun sont également recensées.

Le chapitre 3 décrit les données de cartes à puce exploitées dans cette recherche ainsi que la méthodologie générale du projet. Un schéma conducteur est présenté afin d'introduire les différentes étapes de la démarche proposée. Le premier travail de prétraitement des données est également expliqué dans ce chapitre avant de rentrer dans le cœur du sujet.

Le chapitre 4 définit quatre indicateurs pour quantifier la variabilité d'utilisation du transport en commun. Ces indicateurs sont ensuite appliqués dans une optique tarifaire : différents groupes de cartes sont construits en fonction des titres de transport utilisés pendant l'année, puis plusieurs types de variations sont évalués pour chacun de ces groupes. Le pouvoir explicatif des indicateurs proposés est également vérifié à l'aide de tests statistiques.

Les chapitres 5 et 6 fournissent des méthodes d'analyse pour étudier distinctement la variabilité interpersonnelle et la variabilité intrapersonnelle d'utilisation du transport en commun à l'aide d'outils de classification. Une typologie d'usagers et une typologie de semaines sont produites, puis des indicateurs sont calculés pour caractériser plus en détail les segmentations obtenues.

Le chapitre 7 tire profit de la typologie de semaines développée dans le chapitre 6 pour construire et analyser des séquences de types de semaines. Cette analyse séquentielle permet de prendre en compte l'ordre et l'organisation des comportements hebdomadaires de mobilité observés. Une méthode de segmentation basée sur une distance non euclidienne est également proposée pour créer une typologie de séquences. Cette dernière est alors mise en correspondance avec la typologie d'usagers produite dans le chapitre 5 et d'autres indicateurs de variabilité sont estimés à partir des séquences contenues dans chaque groupe.

Enfin, le chapitre 8 conclut ce mémoire en résumant les principaux résultats obtenus et en soulignant les contributions apportées par cette recherche. Les limites du projet sont également relevées et des perspectives futures sont formulées pour présager une suite possible à ces premiers travaux sur la variabilité d'utilisation du transport en commun.

CHAPITRE 2 REVUE DE LITTÉRATURE

Ce deuxième chapitre présente une revue de la documentation touchant au sujet de recherche traité dans ce mémoire. Cette revue vise à délimiter un cadre et à examiner les travaux précurseurs à la recherche réalisée. Outre sa vocation informative qui permettra une meilleure connaissance des questions abordées, cet état de l'art servira également à repérer des lacunes dans les travaux cités. Ces limitations permettront de fonder et de justifier l'approche méthodologique qui sera développée dans la suite de ce mémoire.

Le support sur lequel se base cette recherche, c'est-à-dire la carte à puce, est d'abord introduit. Il s'agit de situer cette technologie dans le temps, de comprendre son fonctionnement ainsi que d'en pointer les avantages et les inconvénients. Son utilité dans la recherche sur le transport en commun est en particulier examinée. En effet, les données produites par les cartes à puce ont déjà été exploitées dans de nombreux travaux, répertoriés ici en trois catégories. Plus spécifiquement, divers types d'analyses possibles de l'utilisation du transport en commun à partir des données de cartes à puce sont abordées dans la troisième section de ce chapitre. La revue se concentre ensuite sur la mesure de la variabilité des comportements de mobilité. Les méthodes et outils associés sont recensés et une section entière est dédiée aux techniques utilisées pour segmenter les usagers.

2.1 La carte à puce

La carte à puce est de plus en plus populaire dans les réseaux de transport en commun du monde entier; elle devient un outil indispensable et intégré à la structure même de leur fonctionnement. En effet, la carte à puce est le vecteur d'un système automatique de perception des titres, aujourd'hui nécessaire pour gérer les flux massifs de personnes déplacées en transport collectif. Néanmoins, ses fonctions ne se limitent pas à un simple acte de validation pour entrer dans les réseaux de transport ; ses applications sont en réalité beaucoup plus larges. Cette première section vise donc à présenter plus en détail cette carte « intelligente ». Après une brève mise en situation, les caractéristiques générales de la technologie sont renseignées, puis les avantages et les inconvénients de son utilisation en transport sont évalués selon différents points de vue.

2.1.1 Passé, présent et futur

D'après l'historique dressé par Shelfer et Procaccino (2002), le principe de la carte à puce n'est pas récent. Les deux inventeurs allemands, Jürgen Dethloff et Helmut Grötrupp, sont les premiers à breveter l'idée d'intégrer une puce électronique dans une carte en plastique en 1968. Cette invention est suivie par d'autres brevets déposés au Japon et en France dans les années 1970. La première puce est créée en 1977 par l'association de deux entités : l'entreprise américaine Motorola Semiconductor et la compagnie française Bull. Cependant, l'utilisation de la carte à puce ne devient importante qu'à partir des années 1990.

Aujourd'hui, la technologie de la carte à puce est adoptée dans de multiples domaines : entre autres, les banques (cartes de débit et de crédit), le gouvernement et la sécurité (contrôle d'identité, autorisation d'accès), la santé (carte d'assurance maladie), le commerce (programmes de fidélité), les télécommunications (cartes SIM), le transport (perception tarifaire), etc. (Secure Technology Alliance, 2018b; Smart Card Basics, 2018a). En transport, son utilisation est désormais assez répandue. En effet, d'un point de vue géographique, la technologie est implantée dans les réseaux du monde entier, particulièrement en Europe. Elle est également présente en Asie, sous la forme de cartes possédant différentes fonctions de paiement. En Amérique, plusieurs villes des États-Unis et du Canada sont dotées d'un système de carte à puce, et la technologie se développe en Amérique du Sud (Pelletier et al., 2011). En particulier, au Québec, un système de cartes à puce est exploité par les opérateurs de transport en commun dans les régions de Gatineau, Québec et Montréal.

Par ailleurs, la Figure 2.1 atteste que cette technologie continue de se déployer. Eurosmart, association internationale établie à Bruxelles depuis plus de 20 ans et impliquée dans la sécurité numérique, confirme la tendance de croissance globale de ce marché. En 2017, l'association a augmenté ses ventes de 3.6% par rapport à 2016, et elle prévoit expédier plus de 10 milliards d'éléments sécurisés (principalement des cartes à puce) en 2018. Cette hausse est essentiellement due à l'adoption des solutions sans contact : les cartes de paiement sans contact, les cartes de transport et les cartes d'identité électronique (eID) représentent plus de 2 milliards des unités vendues en 2017. De plus, une augmentation de 5% est attendue dans ce secteur en 2018 (Eurosmart, 2018).

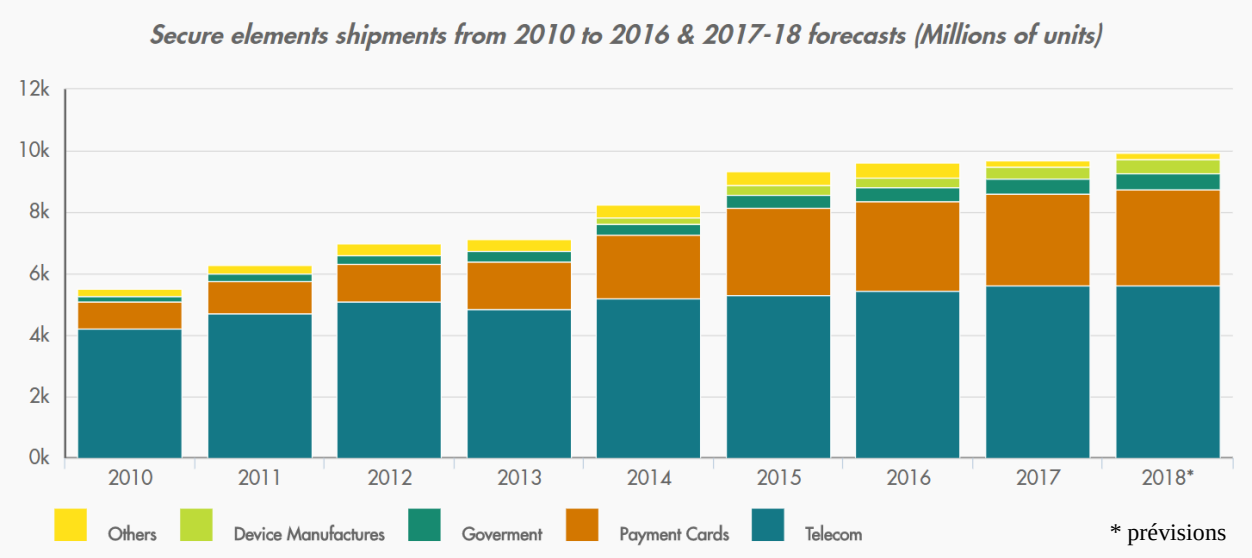


Figure 2.1 Expéditions réelles (2010 à 2017) et prévues (2018) d'éléments sécurisés par Eurosmart (http://www.eurosmart.com/facts-figures.html, page consultée le 17/06/2018)

Les sections qui suivent expliquent plus en détail les raisons de ce véritable succès, notamment dans le cas du transport en commun. De plus, un aperçu des différents types d'études réalisables à partir des données de cartes à puce est donné dans la section 2.2.

2.1.2 Caractéristiques de la technologie

Une carte à puce est, par définition, une carte généralement en plastique qui contient une puce à circuit intégré. C'est dans cette puce que réside « l'intelligence » de la carte : elle est à l'origine de toutes ses fonctions et c'est grâce à cette puce que la carte « communique » avec les appareils de lecture (Blythe, 2004). Différents types de cartes existent selon leur mode de fonctionnement. Les cartes les plus populaires, nommées sur la Figure 2.2, sont définies en deux niveaux (Smart Card Basics, 2018b):

1) Le mode de communication de la carte, c'est-à-dire la manière dont les données de la carte sont écrites et lues (avec contact, sans contact ou les deux modes combinés)

2) Le type de puce intégrée (carte mémoire ou microprocesseur) dont découlent les capacités de la carte en termes de stockage et de traitement des informations

La carte avec contact, construite dans le respect de la norme ISO 7816 (Blythe, 2004), nécessite un rapport physique pour être lue. Cette carte doit donc être insérée dans le lecteur afin que ce dernier

puisse se connecter à la puce par contact direct avec le conducteur métallique de la plaque. À l'inverse, la carte sans-contact doit seulement être portée à proximité du lecteur, à une distance de l'ordre de 10 cm s'il agit d'une carte de proximité [*« proximity card »*, norme ISO 14443] et jusqu'à 70 cm s'il s'agit d'une carte de voisinage [« *vicinity card* », norme ISO 15693] (Bagchi & White, 2004). Cette carte sans contact communique avec le lecteur par fréquences radio (RFID : Radio Fréquence IDentification) grâce à une antenne. De plus, les deux types d'interfaces (avec et sans contact) peuvent être disponibles sur une même carte : cette dernière est soit hybride soit à double interface. Une carte hybride est composée de deux puces distinctes. Les deux technologies avec contact et sans contact sont alors indépendantes et ne communiquent pas entre elles. En revanche, une carte double-interface ne comporte qu'une seule puce pouvant fonctionner avec les deux technologies (Secure Technology Alliance, 2018a).

Par ailleurs, chaque carte peut contenir une des deux catégories de puce suivantes : une puce mémoire ou un microprocesseur. La carte mémoire peut seulement stocker des informations et être lue par un appareil spécialisé, alors que la carte avec microprocesseur peut aussi traiter ces informations de manière dynamique et exécuter des programmes enregistrés dans la puce via un système d'exploitation de la carte (COS ou *card operating system*). Contrairement à la carte mémoire, la carte avec microprocesseur gère la mémoire de la carte dans une structure organisée : le processeur alloue la mémoire de la carte dans des sections indépendantes ou des fichiers affectés à une fonction particulière (Smart Card Basics, 2018b). L'association Secure Technology Alliance (anciennement appelée Smart Card Alliance) compare ainsi la carte mémoire à une petite disquette et la carte avec microprocesseur à un petit ordinateur. Du fait de leurs moins nombreuses fonctionnalités, les cartes mémoire sont moins dispendieuses, mais en contrepartie elles sont aussi moins sécuritaires (Secure Technology Alliance, 2018a). En effet, l'accès aux données enregistrées sur la carte est moins contrôlé que dans une carte avec microprocesseur.

Dans ces deux types de cartes, le stockage des données se fait généralement dans une mémoire EEPROM (*electrically erasable programmable read-only memory*, ou mémoire morte effaçable électriquement et programmable). La quantité de mémoire disponible dépend de la taille physique et du coût de la puce (Blythe, 2004). Même si des cartes avec plus de 100kB de mémoire existent, l'auteur spécifie qu'une mémoire comprise entre 2 et 4kB est suffisante pour enregistrer des informations sur l'équilibre financier et le contrat souscrit, ainsi qu'un historique d'une centaine de transactions.

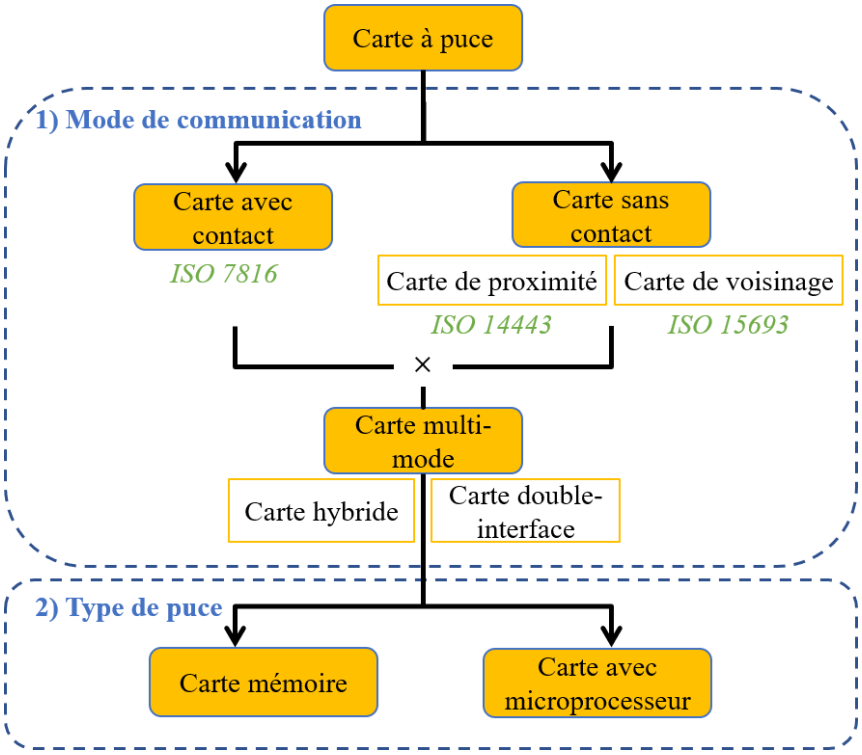


Figure 2.2 Différents types de cartes à puce selon deux niveaux de différenciation

Les cartes à puce utilisées en transport pour la validation tarifaire sont typiquement des cartes sans contact avec microprocesseur. En effet, la technologie sans contact est idéale pour permettre une entrée ou un paiement rapide, conduisant ainsi à une diminution des temps morts des bus. Selon le fabricant de cartes à puce Gemplus, le temps requis pour traiter une transaction est réduit d'un facteur 20 à 30 par la technologie sans contact (Shelfer & Procaccino, 2002). De plus, un microprocesseur est nécessaire pour rendre la carte plus sécuritaire et pour exécuter plusieurs applications. Entre autres, les applications de la carte à puce utilisée en transport doivent permettre trois fonctions : l'authentification pour identifier un numéro de carte, l'autorisation pour vérifier la validité de ce numéro dans une base de données, puis la transaction pour prélever un paiement ou un contrat sur la carte (Trépanier et al., 2004).

2.1.3 Avantages et inconvénients

L'utilisation de la carte à puce dans les réseaux de transport en commun présente de nombreux avantages, mais aussi plusieurs inconvénients. Ces atouts et écueils ont été recensés à partir de 15 références de la littérature scientifique et ont été résumés dans un tableau récapitulatif en ANNEXE

C selon les points de vue de trois protagonistes du domaine des transports : l'usager, l'opérateur et le planificateur/chercheur.

2.1.3.1 Du point de vue de l'usager

Du point de vue de l'usager, des gains de temps et de commodité sont évidents : la facilité d'utilisation de la carte et la rapidité de l'acte de validation rendent son attente pour embarquer moins longue et moins pénible (Bagchi & White, 2005; Chira-Chavala & Coifman, 1996; White et al., 2010). Sa perception du transport en commun est également améliorée par l'avènement de technologies plus modernes et sa satisfaction est augmentée grâce à un service plus performant (Ibrahim, 2003; McDonald, 2000). De plus, au lieu d'être sollicité pour répondre à une enquête, l'usager n'a plus qu'à taper son titre de transport sur un lecteur pour transmettre les informations concernant son déplacement (Spurr et al., 2015). Ce geste mécanique et automatique lui demande moins d'effort et de temps. En effet, White et al. (2010) évaluent à 1.5 seconde la durée d'une validation, contre 10 à 15 minutes de temps de réponse dans une enquête traditionnelle. Grâce aux données transactionnelles que l'utilisateur fournit souvent inconsciemment, il devient possible de mettre en place des services et des tarifs personnalisés selon ses habitudes de déplacement et sa régularité (Kieu et al., 2014).

Néanmoins, certaines personnes oublient parfois ou ne se sentent pas obligées de valider leur carte, notamment lorsqu'elles possèdent un forfait illimité (Bagchi & White, 2004) ; elles risquent alors une contravention. Un soin particulier doit aussi être apporté à la puce de la carte. La STM conseille de conserver sa carte dans un étui rigide pour mieux la protéger (Société de transport de Montréal, 2018d). Enfin, la confidentialité des données est un autre point crucial pour l'usager soucieux de la protection de sa vie privée. Cette préoccupation est inhérente à la nature même de la carte à puce, vulnérable aux usurpations d'identité et à l'utilisation abusive des données produites. En effet, les données de cartes à puce peuvent être piratées pour des fins malhonnêtes (Pelletier et al., 2011). Cependant, ces données peuvent aussi être utilisées par le gouvernement dans la lutte contre le terrorisme. Un compromis entre sécurité intérieure et vie privée doit donc être trouvé (Dempsey, 2008). Entre autres, des mesures de sécurité doivent être adoptées afin de renforcer la confiance des usagers dans le système. En transport, des procédures de séparation des données de déplacements et des données personnelles sont ainsi mises en place. Les numéros de carte sont

également recodés pour garder l'anonymat du propriétaire de chaque carte et rendre impossible le retraçage de ses mouvements.

2.1.3.2 Du point de vue de l'opérateur

Pour l'opérateur, le système de validation tarifaire par cartes à puce permet un contrôle de paiement plus fiable qu'une simple vérification par le chauffeur. Les comportements frauduleux sont ainsi freinés (White et al., 2010). En outre, l'automatisation de ce système rend possible la mise en place d'une grille tarifaire beaucoup plus complexe (Attoh-Okine & Shen, 1995; Deakin & Kim, 2001; McDonald, 2000; Pelletier et al., 2011; White et al., 2010). Elle facilite aussi l'intégration tarifaire entre les différents réseaux et/ou opérateurs partenaires, ce qui permet une meilleure gestion et allocation des revenus ainsi qu'une plus grande uniformité dans l'utilisation des réseaux (Attoh-Okine & Shen, 1995; Blythe, 2004; Deakin & Kim, 2001; Trépanier, 2012; White et al., 2010). À plus long terme, la technologie de la carte à puce permet une diminution des coûts d'opération (McDonald, 2000; Pelletier et al., 2011). En effet, le personnel assigné dans les stations de métro pour surveiller les portes d'entrée et de sortie peut être allégé. De même, l'écourtement des temps morts des bus conduit à des déplacements plus rapides et plus fiables, et donc à une réduction du nombre de véhicules/chauffeurs nécessaire pour opérer le service à fréquence constante (White et al., 2010). D'après le concept d'élasticité de la demande, la diminution du temps d'attente des usagers engendre également une hausse d'achalandage, d'où une augmentation des revenus d'exploitation (McDonald, 2000; White et al., 2010).

Cependant, l'implantation d'un système de cartes à puce représente un investissement majeur (Deakin & Kim, 2001; Pelletier et al., 2011; Trépanier, 2012; White et al., 2010). À titre d'exemple, d'après un communiqué datant du 21 avril 2008, le système OPUS de la région de Montréal a coûté 169 millions de dollars à la STM (Société de transport de Montréal, 2008). Le lancement de cette technologie nécessite également la formation d'un personnel qualifié pour gérer la complexité du système et des équipements (Pelletier et al., 2011; White et al., 2010). Cette complexité technique rend difficile la mise à jour du système qui, une fois mis en place, tend à rester figé pendant plusieurs années (Deakin & Kim, 2001). Par ailleurs, le succès de l'implantation d'un tel système dépend grandement de son acceptation sociale par les usagers (Deakin & Kim, 2001; Pelletier et al., 2011). Cette acceptation peut être accélérée par l'instauration de tarifs réduits. White et al. (2010) citent ainsi l'exemple de Londres où le prix du titre sur la carte est plafonné en dessous du

prix d'un ticket équivalent. Néanmoins, l'établissement de ce système implique une prise initiale de risques élevée sans aucune garantie de profitabilité (Deakin & Kim, 2001; McDonald, 2000).

2.1.3.3 Du point de vue du planificateur ou du chercheur

Au-delà de sa fonction principale de collecte des recettes et de prévention de la fraude, le système de cartes à puce fournit des données utiles pour les planificateurs et les chercheurs. De grosses quantités de données longitudinales sont collectées par ces systèmes (Bagchi & White, 2004; El Mahrsi et al., 2017; Trépanier, 2012; Utsunomiya et al., 2006). En effet, la longueur de la période de collecte de ces données 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). Le volume des données recueillies dépend néanmoins du taux de pénétration de la technologie parmi les usagers. De plus, la précision des données amassées est très grande au niveau temporel mais aussi au niveau spatial si les lieux de validation (embarquement et/ou débarquement) sont géolocalisés: les informations sont alors rapportées aux coordonnées et la seconde près (El Mahrsi et al., 2017; Trépanier, 2012; White et al., 2010). Ces données sont également individualisées puisqu'il est possible de lier les déplacements observés à un individu donné [ou à une carte donnée, car, comme expliqué précédemment, une même carte peut parfois être utilisée par plusieurs personnes] (Bagchi & White, 2004; El Mahrsi et al., 2017; Trépanier, 2012). Cette individualisation permet une étude totalement désagrégée des comportements de mobilité. De surcroît, les données sont enregistrées de manière passive. Elles ne sont donc pas biaisées par la faculté des usagers à rapporter leurs déplacements (Bagchi & White, 2004; Trépanier, 2012). Enfin, ce sont des données moins coûteuses en temps et en argent. Une fois que le système de paiement est bien installé, la collecte des informations de déplacements est peu onéreuse (Spurr et al., 2015; Trépanier, 2012).

Toutefois, certaines informations ne sont pas recueillies, en particulier le motif de déplacement, la perception et la satisfaction des usagers, l'origine et la destination finales du déplacement ainsi que les temps d'accès. Par souci de protection de la vie privée, les attributs socio-démographiques du propriétaire de la carte ne sont pas disponibles non plus. Souvent, aucune validation n'est requise à la sortie du métro ou du bus : ni le lieu ni l'heure de débarquement ne sont alors connus. Au demeurant, la carte à puce permet seulement de capturer la mobilité d'une portion de la population : elle ne prend pas en compte ni les utilisateurs des autres modes ni les usagers sans carte. Pour toutes ces raisons, les données de cartes à puce sont souvent dites partielles et non universelles (Bagchi

& White, 2005; El Mahrsi et al., 2017; Spurr et al., 2015; Trépanier, 2012; White et al., 2010). C'est pourquoi Bagchi et White (2004) ont insisté sur le fait que les données de carte à puce étaient des données complémentaires aux autres sources de données. En effet, elles ne peuvent pas les remplacer, car il manque parfois des informations indispensables à l'analyse souhaitée. Par ailleurs, initialement prévues pour l'opération des réseaux et le contrôle financier, les données de cartes à puce ne sont pas toujours adaptées à la planification et à la recherche. Des traitements importants sont souvent nécessaires avant leur exploitation. Étant donnés les gros volumes disponibles, ces traitements font appel à des programmes puissants avec une grande mémoire de stockage. Un échantillonnage est parfois inévitable (El Mahrsi et al., 2017; Trépanier, 2012). Enfin, la qualité et la continuité des données sont parfois altérées par des défaillances du système ou des erreurs humaines. Certaines transactions peuvent ainsi être incomplètes et inexploitables (Trépanier, 2012; White et al., 2010).

2.2 Exploitation des données de carte à puce en recherche

Malgré les quelques faiblesses évoquées dans la section précédente, les données de cartes à puce sont largement utilisées dans la recherche sur le transport en commun. Selon Pelletier et al. (2011), les travaux correspondants sont divisés en trois catégories :

- Des études au niveau stratégique (planification du réseau à long-terme, compréhension du comportement des usagers, enrichissement des données et prévision de la demande)
- Des études au niveau tactique (ajustement de l'offre à différentes échelles grâce à une meilleure estimation des profils de charge, estimation des correspondances, mise en évidence des besoins des usagers et développement des services)
- Des études au niveau opérationnel (calcul d'indicateurs mesurant la performance du réseau, notamment le respect des horaires planifiés, détection et correction des irrégularités et erreurs du système)

Plusieurs exemples sélectionnés dans la littérature scientifique sont donnés ci-après pour chacune de ces trois catégories. Nombreuses sont les études qui cherchent justement à dépasser les inconvénients mentionnés précédemment.

2.2.1 Études au niveau stratégique

Au niveau stratégique, la compréhension des comportements de mobilité est un des principaux thèmes investigués par les chercheurs. Leurs travaux portent généralement sur la caractérisation et la classification des usagers du transport en commun à l'aide de différents outils d'exploration de données (*data mining* en anglais). Agard et al. (2006) analysent ainsi quatre groupes d'utilisateurs définis selon leurs habitudes temporelles de déplacement. De plus, les études proposées sont souvent réalisées à différents niveaux d'agrégation pour pouvoir mettre en évidence des tendances collectives et/ou individuelles. Zhong et al. (2015) explorent par exemple la mobilité des usagers à trois échelles spatiales : à un niveau totalement désagrégé/individuel, à un niveau agrégé par station ou arrêt de bus et à un niveau encore plus agrégé pour l'ensemble du réseau. Dans ce dernier cas, la cité-État étudiée (Singapour) est vue comme un système de zones reliées par des flux de déplacements. White et al. (2010) donnent d'autres exemples d'applications à considérer pour mieux comprendre le comportement des usagers, notamment grâce à l'étude des chaînes de déplacements ou par l'estimation d'un taux de rotation.

Pour approfondir cette compréhension des comportements, notamment en les contextualisant, mais aussi pour combler leurs lacunes, les données de cartes à puce sont parfois enrichies avec d'autres sources de données. D'une part, elles peuvent être croisées avec des informations géographiques et opérationnelles. Tao et al. (2014) utilisent par exemple les données GTFS pour reproduire les mouvements correspondant aux chemins réellement empruntés par les usagers. La méthodologie développée par ces auteurs permet de reconstruire les trajectoires des usagers arrêt de bus par arrêt de bus puis de géospatialiser les dynamiques spatio-temporelles résultantes. D'autre part, les données de cartes à puce peuvent être complétées avec des données socio-économiques ou socio-démographiques. Utsunomiya et al. (2006) construisent ainsi des profils d'usagers par ligne et par station. Dans la même perspective, El Mahrsi et al. (2014) mettent en relation des clusters temporels avec des clusters socio-économiques pour examiner comment les caractéristiques socio-économiques des usagers influencent leurs habitudes temporelles. Par ailleurs, de nombreux auteurs s'accordent à dire que les données de cartes à puce et les données des Enquêtes Ménages Déplacements sont complémentaires (Bagchi & White, 2004). La comparaison de différents indicateurs entre ces deux sources de données révèlent notamment que les données de cartes à puce sont plus précises et plus justes, mais que les données des enquêtes sont plus complètes : contrairement aux données de cartes à puce, elles contiennent des renseignements sur les lieux

d'activités et les motifs de déplacement, ainsi que des informations socio-démographiques (Trépanier, Morency, & Blanchette, 2009). L'enrichissement des données de cartes à puce avec les données de ces enquêtes peut, entre autres, permettre de déterminer le motif le plus probable de chaque déplacement (Zhong et al., 2014) ou d'analyser les déplacements domicile-travail (Long & Thill, 2015).

Tous les travaux évoqués précédemment s'inscrivent dans une volonté stratégique de planification. En effet, une meilleure connaissance des comportements de mobilité est souhaitable pour mieux prédire la demande. Plus particulièrement, les données de cartes à puce sont parfois directement exploitées pour développer des modèles de prévision à plus ou moins court terme. Van Oort et al. (2015) ont ainsi créé un modèle élastique pour prévoir les effets des changements de service. Ce modèle repose sur une matrice Origine-Destination extrêmement détaillée obtenue grâce à des données de cartes à puce. À partir de l'historique des déplacements de chaque usager et de tendances collectives, certains auteurs réussissent à prédire des comportements individuels, notamment les jours d'activité (Foell et al., 2013) ou les durées de déplacement de chaque usager (Lathia et al., 2010). Au niveau spatial, une liste de stations d'intérêt peut également être estimée pour chaque usager (Foell et al., 2014; Lathia et al., 2010). Ces prévisions individuelles sont notamment convoitées par les chercheurs dans le but de développer des systèmes intelligents d'information personnalisée, adaptée aux besoins de chaque usager.

2.2.2 Études au niveau tactique

L'ajustement des services est le principal sujet des études réalisées au niveau tactique. Cet ajustement est nécessaire du fait des variations observées dans l'utilisation du transport en commun. En effet, en comparant le jour moyen de semaine obtenu avec l'enquête Origine-Destination de 2005 à chaque jour couvert par les données de cartes à puce sur la même période, Trépanier, Morency et Blanchette (2009) ont mis en évidence une grande variabilité dans l'achalandage quotidien. Actuellement, les horaires des services proposés sont les mêmes pour tous les jours de la semaine, mais l'idéal serait d'offrir un service différent et adapté à chaque jour. Outre cet ajustement temporel, un ajustement spatial au niveau de la ligne est envisageable, mais cela nécessite une estimation plus précise des profils de charge et des points de charge maximaux. Lorsque seules des informations sur l'embarquement sont disponibles dans les données, cela passe d'abord par la détermination de la destination de chaque déplacement.

Plusieurs travaux se sont penchés sur l'estimation de cette destination. En général, les algorithmes développés sont basés sur deux hypothèses principales : 1) Le lieu de destination du déplacement précédent est le même que le lieu d'origine du déplacement suivant, 2) Le dernier lieu de destination d'une journée est le même que le premier lieu d'origine de la journée suivante (Barry et al., 2002). De manière plus rigoureuse, le critère principal appliqué dans l'étude réalisée par Trépanier et al. (2007) pour déterminer le lieu de débarquement est la minimisation de la distance avec le point d'embarquement suivant. Munizaga et Palma (2012) proposent une méthode similaire adaptée à un réseau multimodal (bus et métro). Zhang et al. (2015) ont quant à eux adopté une approche probabiliste basée sur les champs aléatoires conditionnels (*conditional random fields* ou CRFs). Une fois l'origine et la destination de chaque déplacement connues, une matrice Origine-Destination détaillée peut être construite et les flux correspondants peuvent être affectés sur le réseau TC (Barry et al., 2002; Munizaga & Palma, 2012).

En plus de l'estimation des lieux de débarquement, la reconstruction de l'historique des déplacements individuels requiert également l'identification des correspondances. Hofmann et O'Mahony (2005) présentent ainsi un algorithme itératif de classification pour regrouper les embarquements en deux catégories : les déplacements avec correspondance versus les déplacements sans correspondance. Seaborn et al. (2009) considèrent quant à eux trois types de correspondance (bus-métro, métro-bus, bus-bus) et recommandent différents intervalles de temps seuils pour identifier deux transactions liées à un même déplacement. À une granularité plus fine, Nassir et al. (2015) énoncent une méthode pour distinguer les correspondances et les courtes activités à l'aide de plusieurs critères basés notamment sur le concept de "non optimalité". De même, Chu et Chapleau (2008) proposent un processus d'enrichissement permettant de reconstruire les itinéraires individuels à l'échelle de l'arrêt et d'analyser les activités de correspondance. L'analyse des transferts ainsi localisés permet aux opérateurs d'adapter le service et la géométrie du réseau aux pratiques des usagers (Pelletier et al., 2011).

Par ailleurs, le service doit être tactiquement ajusté aux besoins des utilisateurs. Pour mettre en évidence ces besoins, Trépanier et Morency (2010) ont développé des modèles permettant de caractériser la fidélité des usagers et de la mettre en relation avec différents facteurs tels que la « date de naissance » (première apparition sur le réseau), le type de titre et le lieu de domicile des usagers. Les opérateurs peuvent alors s'appuyer sur cette étude pour proposer des services et des tarifs spécifiques, mais aussi pour retenir les utilisateurs les plus fidèles. Blythe (2004) explique en

effet que les informations recueillies sur les comportements des usagers, leur profil et leurs préférences de paiement peuvent être exploitées pour mettre en place des plans de fidélité et de récompenses permettant d'augmenter la satisfaction de la clientèle visée. Cette meilleure connaissance des usagers desservis permet également d'évaluer l'impact d'une modification ou d'une perturbation du service (Trépanier & Morency, 2010).

2.2.3 Études au niveau opérationnel

Au niveau opérationnel, les données de cartes à puce permettent le développement d'indicateurs de performance. En effet, Bagchi et White (2005) ont écrit que les données de déplacement obtenues à partir d'un système de cartes à puce permettaient d'augmenter la qualité, mais aussi le nombre de statistiques disponibles. Trépanier et Vassiviere (2008) ont d'ailleurs créé un outil intranet pour fournir ces statistiques de manière quotidienne aux opérateurs du service. Reddy et al. (2009) reconnaissent dans leur papier que les données de cartes à puce permettent d'améliorer le calcul des deux indices nécessaires au calcul des subventions incitatives du New York City Transit (NYCT), à savoir le nombre de déplacements non liés et le nombre de passagers-kilomètres. Ils déclarent également que ces données sont moins coûteuses et moins problématiques que les données collectées jusqu'alors manuellement. Trépanier, Morency et Agard (2009) listent d'autres indicateurs de performance concernant le taux d'occupation des véhicules et le respect des horaires, ainsi que des mesures orientées offre (véhicules-kilomètres, véhicules-heures, vitesse commerciale, etc.) ou demande (passagers-kilomètres, passagers-heures, durée moyenne de déplacement, etc.)

De plus, la grande résolution des données de cartes à puce permet de détecter des irrégularités et des erreurs. Des défaillances dans les équipements, des fautes humaines ou des fraudes peuvent effectivement se traduire par des validations incomplètes ou erronées. C'est pourquoi Trépanier et al. (2007) conseillent de valider et de corriger (si possible) les données de cartes à puce avant de commencer l'estimation des destinations de chaque transaction. Les auteurs mentionnent les principales sources d'erreurs possibles et produisent une distribution de ces erreurs pour le mois de juillet 2003. Le plus souvent, les perturbations observées prennent la forme d'informations manquantes, c'est-à-dire de transactions incomplètes. Pour outrepasser ces problèmes et améliorer la continuité et la justesse des données de cartes à puce, Chapleau et Chu (2007) proposent une procédure permettant de repérer et d'imputer les erreurs et les incohérences dans les données. Dans

une autre optique, celle de déceler des comportements frauduleux, Zhao et al. (2014) construisent deux indicateurs (par passager), un indicateur d'anormalité temporelle et un indicateur d'anormalité spatiale, afin de détecter des écarts dans les déplacements.

2.3 Analyse descriptive de l'utilisation du transport en commun à partir des données de cartes à puce

Le sujet de recherche traité dans ce mémoire porte plus particulièrement sur l'étude de l'utilisation du transport en commun : où, quand et comment les usagers empruntent-ils ce mode de transport ? Ce thème s'inscrit à la fois dans les niveaux stratégiques et tactiques évoqués dans la section précédente. En effet, il s'agit ici de mieux comprendre les comportements de mobilité dans le but de planifier à long terme et de prédire la demande, mais aussi d'améliorer l'ajustement de l'offre à cette demande et de développer les services proposés en fonction des besoins identifiés.

La richesse des données de cartes à puce permet d'analyser l'utilisation du transport en commun dans des dimensions aussi bien temporelles que spatiales. En effet, la date, l'heure et le lieu sont des informations enregistrées à chaque validation tarifaire. De plus, les comportements de mobilité peuvent être examinés sur un plan individuel ou collectif selon si les validations sont considérées par carte ou agrégées à différentes échelles (par jour, par station, par type de titre, etc.).

Pour cela, différentes techniques d'exploration de données sont utilisées dans littérature scientifique. Anand et Büchner (1998) définissent cette exploration comme « la découverte d'informations non triviales, implicites, auparavant inconnues, potentiellement utiles et compréhensibles à partir de grands ensembles de données » [notre traduction]. De plus, Westphal et Blaxton (1998) catégorisent les différentes techniques disponibles en quatre fonctions : la classification, l'estimation, la segmentation et la description. Cette section est principalement consacrée à la fonction descriptive. En effet, les auteurs présentés ci-après se basent sur des données de cartes à puce pour décrire et visualiser l'utilisation du transport en commun au niveau temporel ou spatial, voire spatio-temporel lorsque les deux dimensions sont combinées. Ils recourent également parfois à des méthodes de segmentation pour identifier des comportements typiques.

2.3.1 Analyse temporelle

Tout d'abord, les informations temporelles contenues dans les données de cartes à puce peuvent être exploitées pour tracer des profils temporels de déplacements : la demande est agrégée par intervalle de temps puis un graphique est produit pour représenter le nombre d'embarquements en fonction du temps, pour tous les usagers ou par type de titre (White et al., 2010). À un niveau totalement désagrégé, Morency et al. (2006) réalisent ce même type de distribution pour deux cartes en particulier. Les profils décrivant l'achalandage sur une longue période d'étude permettent d'observer des phénomènes cycliques et saisonniers (Morency et al., 2007) alors que les profils dessinés sur quelques jours seulement mettent en évidence des différences de comportement entre la semaine et la fin de semaine. Des périodes de pointe AM et PM sont aussi généralement visibles en semaine (Huang et al., 2015; Liu, L. et al., 2009; Nishiuchi et al., 2013; Zhong et al., 2015). En s'appuyant sur les données de cartes à puce fournies par la Société de Transport de l'Outaouais (STO), Descoimps (2011) présente un éventail complet d'analyses temporelles possibles à partir de ces profils pour différentes échelles. De plus, des tests statistiques peuvent être appliqués pour comparer les profils d'achalandage obtenus entre eux et souligner les disparités temporelles observées. Zhong et al. (2015) utilisent ainsi une matrice de corrélation tandis que Nishiuchi et al. (2013) réalisent une analyse de variance (ANOVA). Dans les deux cas, la principale conclusion tirée est que les différences sont significatives entre les jours de semaine et les jours de fin de semaine. Elles ne le sont plus (ou moins) lorsque les jours de semaine ou les jours de fin de semaine sont comparés entre eux.

Les heures d'embarquement rapportées dans les données de cartes à puce sont également utilisées comme critère de classification dans de nombreuses études, notamment pour construire des profils temporels typiques. Ces profils ne sont alors pas des profils absolus totaux comme précédemment, mais des profils moyens (ou populaires) pour un ensemble d'individus ou un usager donné. En effet, chaque profil type est la moyenne (ou le plus populaire) des profils individuels observés appartenant au même groupe. Agard et al. (2006) résument ainsi les données transactionnelles dans des vecteurs cartes-semaines où chaque semaine est décomposée en 5 jours ouvrables et 4 grandes périodes journalières (pointe AM, midi, pointe PM, soir). Ces vecteurs, remplis de manière binaire en fonction des heures de validation de la carte, servent de données d'entrée à un algorithme de classification. Les quatre profils hebdomadaires types obtenus traduisent des horaires d'activité différents, le groupe des travailleurs étant par exemple caractérisé par des heures de pointe plus

achalandées. En appliquant une méthode similaire, Morency et al. (2007) segmentent les journées de chaque type de titre étudié (étudiants, aînés, adultes-régulier, adultes-express, adultes-interzone) en quatre groupes d'usagers-jours. Les centroïdes des groupes produits révèlent des heures habituelles de déplacements différentes d'un type d'usager à l'autre, mais aussi au sein de chaque type. De même, toujours sur une base quotidienne, Morency et al. (2006) construisent des groupes de jours à un niveau individuel. Ces groupes rassemblent des jours similaires en termes d'heures d'embarquement pour chacune des deux cartes étudiées. À une échelle encore plus désagrégée, Manley et al. (2016) identifient des groupes d'heures régulières de déplacement pour chaque usager.

2.3.2 Analyse spatiale (ou spatio-temporelle)

De plus, de nombreux efforts ont été faits pour ajouter une dimension spatiale à l'étude des comportements de mobilité. Différents outils d'analyse issus des systèmes d'information géographique peuvent être manipulés pour illustrer l'utilisation spatiale du réseau de transport en commun. Ces outils permettent de géospatialiser les validations de cartes à puce à différentes échelles : des flux entre les stations/arrêts visités ou des volumes sur les lignes empruntées peuvent ainsi être modélisés. Ces représentations géographiques permettent notamment de visualiser des tendances spatiales de mobilité et de les contextualiser. En outre, des analyses spatio-temporelles peuvent également être menées, car les deux dimensions sont souvent fortement liées.

En représentant les déplacements entrants et sortants de chaque station de métro, Manley et al. (2016) montrent que les stations origine et destination les plus régulières sont situées respectivement en banlieue et au centre de la ville. La dynamique pendulaire domicile – travail qui cause cette tendance explique également les fluctuations constatées par Liu, L. et al. (2009). Après avoir localisé et orienté les connections inter-stations dérivées d'une matrice Origine-Destination pour différents types de jour (semaine versus fin de semaine), les auteurs interprètent les mouvements observés avec le contexte économique et culturel de la ville. De même, Zhong et al. (2015) établissent des profils temporels de déplacement par station puis remarquent que les stations les plus corrélées entre elles sont caractérisées par les mêmes types d'utilisation du sol. Au niveau systémique, les auteurs utilisent une méthode de détection communautaire (*Community detection* et *PageRank*) pour regrouper les stations et arrêts du réseau qui interagissent entre eux. La

représentation géographique des groupes obtenus révèle que, malgré quelques échanges locaux, la structure globale de la ville reste la même, quel que soit le jour de la semaine.

Chu et Chapleau (2010) changent de paradigme en ne parlant plus de station ni d'arrêt, mais de point d'ancrage. Ces points d'ancrage, ou nœuds fréquemment visités, sont détectés pour chaque usager selon ses habitudes d'embarquement et associés à des lieux d'activité. Les déplacements sont ensuite analysés et visualisés entre ces points d'ancrage. Cette représentation permet de justifier les heures de départ en pointe du matin par la distance au lieu d'activité, les heures de départ les plus matinales correspondant à des distances plus grandes à parcourir. Tao et al. (2014) cherchent également à comprendre les dynamiques spatio-temporelles des usagers du transport en commun en visualisant les tendances des flux agrégés sur tout le réseau. Pour cela, ils utilisent une technique géovisuelle innovante qui crée des *flow-comaps* ou « cartes de flux ». Ces cartes permettent de représenter la variation des volumes d'achalandage sur chaque lien du réseau selon différentes périodes de la journée et par type de titre. Un coefficient de variation spatiale est également calculé pour mesurer la dispersion spatiale des déplacements. Par ailleurs, les différences observées entre deux cartes de flux sont quantifiées à l'aide de cartes pondérées (*weighted flow-comaps*). Ces dernières permettent de visualiser des changements dans les comportements par rapport à une référence.

2.4 Mesure de la variabilité individuelle des comportements de mobilité

En analysant l'utilisation du transport en commun, de nombreux auteurs ont également remarqué sa variabilité dans le temps et dans l'espace. Les fluctuations macroscopiques observées sont principalement dues à des variations individuelles (c.-à-d. relatives à l'individu), interpersonnelles ou intrapersonnelles, 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) soulignent des différences en termes de nombre d'embarquements par jour et d'heures de déplacements entre les deux cartes étudiées. Au contraire, les auteurs relèvent une grande régularité intrapersonnelle puisque la majorité des journées collectées pour chaque carte sont du même type. Une certaine régularité spatiale est aussi démontrée par Barabási et al. (2008) en examinant la trajectoire de 100,000 utilisateurs anonymes de téléphones cellulaires. Ils calculent en effet une forte probabilité

de retourner dans les mêmes endroits. Cependant, en construisant des profils spatio-temporels à une échelle individuelle à partir de données de cartes à puce, Nishiuchi et al. (2013) révèlent que les comportements de mobilité peuvent énormément varier selon le type de passagers étudiés.

Des observations similaires sont retrouvées dans de nombreux travaux d'analyse sur la mobilité, mais elles restent qualitatives. Ainsi, pour aller au-delà d'une simple description ou visualisation des comportements, certains auteurs ont développé des moyens pour quantifier leur variabilité. Ces moyens, centrés sur des caractéristiques individuelles de déplacement, sont, entre autres, des indicateurs, des comparaisons de séquences d'événements activités-déplacements et des modèles. La mise en œuvre de ces différents outils peut reposer sur des données de cartes à puce, mais aussi sur d'autres sources de données individualisées comme des données cellulaires ou des données d'enquêtes. De plus, ces outils peuvent s'appliquer quel que soit le mode de transport étudié. Les travaux présentés ci-après ne se cantonneront donc pas nécessairement à l'utilisation du transport en commun : les méthodes énoncées dans ces travaux mesurent d'une manière générale la variabilité des comportements individuels de mobilité.

2.4.1 Indicateurs de variabilité (ou régularité) individuelle

La mesure de la variabilité (ou, au contraire, de la régularité) des comportements individuels de mobilité fait l'objet de recherches depuis de nombreuses années. Jones et Clarke (1988) sont parmi les premiers à reconnaître la nécessité d'examiner les habitudes de déplacement des usagers sur plusieurs jours plutôt que sur un jour typique. De plus, ils soulignent à juste titre que plus la description des comportements est détaillée, plus leurs variations sont apparentes. Pour évaluer cette variabilité, ils proposent d'abord une méthode graphique, mais les analyses possibles avec cette représentation sont limitées à un petit nombre d'individus. Conscients de cela, ils présentent également une méthode numérique : ils créent un indice de similarité qui augmente quand la même activité est répétée dans le même intervalle de temps sur deux jours différents. Néanmoins, ils perdent ainsi la dimension horaire de la variabilité puisque l'indice ne donne qu'une seule valeur pour plusieurs journées entières. Ils développent finalement une méthode hybride, combinant des outils graphiques et numériques, mais elle reste difficile à automatiser.

D'autres indicateurs ont ensuite été suggérés dans la littérature scientifique, selon la définition donnée à la variabilité et les aspects étudiés. Ce sont principalement des indicateurs basés sur l'intensité d'utilisation du TC, la répétition des mêmes attributs de déplacements, la périodicité des

comportements, l'allocation du temps et les durées d'activité, mais aussi des indicateurs basés sur le calcul d'une variance et sur la diversité spatiale des lieux visités.

2.4.1.1 Indicateurs basés sur l'intensité d'utilisation du TC

Certains auteurs assimilent parfois la régularité à l'intensité d'utilisation du transport en commun. Avec cette simple (voire simpliste) définition, les usagers les plus réguliers sont donc aussi les plus fréquents et les plus actifs sur le réseau. Un indice populaire appartenant à cette première « famille » d'indicateurs est le taux de mobilité, c'est-à-dire le nombre de déplacements par personne par jour. Avec les données de cartes à puce, c'est généralement un nombre d'entrées ou d'embarquements par jour et par carte qui est calculé (White et al., 2010). Huang et al. (2015) l'estiment ainsi pour les étudiants de Chengdu, en Chine, et Morency et al. (2006) pour d'autres types de titres. Ces derniers analysent la variabilité du taux de mobilité entre différents types d'usagers, mais aussi pour une même carte. Entre autres statistiques, un coefficient de variation est estimé pour chacun des jours de la semaine. Par ailleurs, le taux d'activité, défini comme le rapport entre le nombre de jours durant lesquels l'usager s'est déplacé et le nombre total de jours observés (dernier jour observé – premier jour observé), est un autre indicateur mesuré et analysé dans plusieurs travaux (Huang et al., 2015; Morency et al., 2006; Nishiuchi et al., 2013). Cet indicateur n'évalue plus la fréquence, mais le niveau d'activité des usagers. Un usager moins présent sur le réseau est possiblement un usager qui n'utilise pas le transport en commun pour ses déplacements habituels, d'où la confusion entre les deux notions d'intensité et de régularité.

2.4.1.2 Indicateurs basés sur la répétition des mêmes attributs de déplacements

Plus couramment, les travaux sur la variabilité des comportements s'intéressent à la fréquence non pas de tous les déplacements de l'usager, mais seulement des déplacements qui sont similaires, faits aux mêmes endroits et aux mêmes heures. En effet, la stabilité des comportements de mobilité est souvent définie par la répétition de déplacements qui se ressemblent par leurs attributs. C'est pourquoi Morency et al. (2007) ou encore Nishiuchi et al. (2013) établissent des tendances quotidiennes types puis se basent sur le nombre ou la proportion de la tendance la plus fréquente pour déterminer la régularité des usagers.

Dans cette même perspective, Manley et al. (2016) s'appuient sur un seul attribut, à savoir l'heure de départ des déplacements, pour évaluer leur régularité. Ils appliquent un algorithme basé sur la

densité, le DBSCAN (*density-based spatial clustering of applications with noise*), appliqué à des données de cartes à puce, pour déterminer des groupes d'heures régulières pour chaque usager. Tous les déplacements appartenant à un de ces groupes sont alors définis comme des déplacements réguliers, les autres étant considérés comme occasionnels. De même, à partir d'un carnet hebdomadaire de déplacements, Raux et al. (2012) recherchent la régularité des comportements des usagers par l'identification de nœuds (*core stops*), définis par les auteurs comme des « déplacements, classés selon les caractéristiques de quatre attributs (activité, mode de transport, heure d'arrivée, lieu), survenant au moins trois jours différents dans la semaine » [notre traduction]. Dans ces deux précédentes études, après avoir déterminé des groupes de déplacements réguliers pour chaque usager, les auteurs analysent plusieurs indicateurs individuels : le nombre différent de groupes, la fréquence associée (c.-à-d. la taille de chaque groupe), ou encore la proportion des jours où les mêmes comportements sont observés.

De plus, Schlich et Axhausen (2003) passent en revue des indicateurs basés sur la cooccurrence et la répétition des mêmes attributs de déplacements, mesurées à l'aide d'une matrice de contingence. Ils citent ainsi l'indicateur de Hanson, indice qui mesure l'homologie des journées de chaque usager en prenant en compte deux attributs ainsi que le nombre de déplacements effectués par jour, et l'indicateur de Pas, basé sur le concept d'attributs "primaires et secondaires" associés à des poids différents. Cui, C. L. et al. (2014) proposent également un indicateur global construit à partir d'un modèle d'évaluation complète floue (*fuzzy comprehensive evaluation model*). Cet indicateur est fondé sur la stabilité de quatre attributs de déplacements : la fréquence hebdomadaire, l'heure de départ, les lignes empruntées et le tarif payé.

2.4.1.3 Indicateurs basés sur la périodicité des comportements

D'autres auteurs partent du principe que les comportements sont réguliers s'ils se répètent de manière cyclique. Ils développent ainsi des indicateurs qui mettent en évidence une certaine périodicité dans les comportements de mobilité. La transformation de Fourier est par exemple utilisée dans plusieurs travaux pour détecter des périodes typiques entre deux occurrences d'un même évènement (Eagle & Pentland, 2006; Kim & Kotz, 2007). Li et al. (2015) proposent quant à eux une mesure probabiliste de la périodicité à partir de données GPS et prouvent la robustesse de leur méthode même en cas d'observations imparfaites ou manquantes.

Dans certaines études, la régularité des comportements est mesurée directement par leur périodicité. Par exemple, Williams et al. (2012) construisent un indicateur de variabilité temporelle à partir des intervalles de temps identifiés entre deux visites consécutives au même endroit. Dans d'autres études, l'évaluation de cette régularité est basée indirectement sur un cycle prédéfini, en général quotidien (Elango et al., 2007) ou hebdomadaire (El Mahrsi et al., 2017). Un usager est ainsi dit régulier si sa mobilité se ressemble d'une journée ou d'une semaine à l'autre. Cependant, Goulet-Langlois et al. (2017) dénoncent cette définition de la régularité : les déplacements ne devraient pas avoir besoin de s'aligner avec une période particulière du calendrier pour être considérés comme réguliers. Ils donnent l'exemple d'une chaine de déplacements (docteur > pharmacie > retour à la maison) qui ne se répètent pas de manière cyclique. Par ailleurs, des cycles autres que quotidiens ou hebdomadaires peuvent être observés dans les comportements.

2.4.1.4 Indicateurs basés sur l'allocation du temps et les durées d'activité

De manière plus anecdotique, des indicateurs mesurant la régularité de l'allocation du temps et des durées d'activité sont parfois présentés. Dans leur revue, Schlich et Axhausen (2003) mentionnent notamment l'indicateur de similarité de Jones et Clarke, déjà évoqué précédemment dans cette section. Cet indicateur se base sur le temps alloué à différentes activités et non sur les déplacements eux-mêmes. La journée est divisée en plusieurs intervalles temporels et les activités choisies sur deux jours différents dans le même intervalle de temps sont comparées. Si ces activités sont les mêmes, l'indice de similarité augmente de 1 ; si l'activité d'une journée est retrouvée à l'intervalle de temps précédent ou suivant dans l'autre journée, alors l'indice augmente de 0.5 ; s'il n'y a pas de correspondance entre les activités réalisées, l'indice reste identique. L'indicateur final est alors défini comme la somme de toutes les comparaisons (une comparaison par intervalle de temps), divisée par sa valeur maximale (lorsque toutes les activités sont les mêmes pour tous les intervalles).

Liu, L. et al. (2009) analysent quant à eux la distribution du FPT (*First Passage Time*), égal à l'intervalle de temps entre deux apparitions consécutives au même endroit. Cet indicateur reflète une durée d'activité si les données exploitées sont disponibles au départ et à l'arrivée, ou bien un cycle si seule l'origine des déplacements est connue. Dans le premier cas, l'indicateur peut permettre de capturer une durée habituelle d'activité (par exemple, 8h pour un motif travail). Dans le deuxième cas, l'indicateur fournit seulement la période de temps écoulée entre deux passages au

même lieu d'embarquement (par exemple, 24h); on revient alors à la définition précédente de la régularité (assimilée à la périodicité des comportements selon un certain cycle).

2.4.1.5 Indicateurs basés sur le calcul d'une variance

Plusieurs autres travaux recensés dans la littérature calculent la variance de différents attributs de mobilité pour évaluer sa variabilité. Pas et Koppelman (1987) capturaient déjà la variabilité individuelle quotidienne de différents segments de la population en estimant une variance intrapersonnelle moyenne à partir du nombre de déplacements réalisés par jour par chaque individu.

Dans de but de déterminer à quel point les déplacements sont routiniers, Kitamura et al. (2006) utilisent des prismes spatio-temporels pour modéliser l'heure de départ du premier déplacement de la journée. Ils calculent également la variance de cette coordonnée temporelle et la dissocient en plusieurs types : ils différencient d'abord les variations systématiques dues aux contraintes de temps et les variations aléatoires, puis ils les redécomposent en une variabilité interpersonnelle et une variabilité intrapersonnelle, ceci pour distinguer la part de la variance expliquée par les différences entre les usagers et celle qui est due aux variations dans le comportement de chaque usager. De même, Chikaraishi et al. (2009) séparent la variance de l'heure du départ du premier déplacement en cinq parties : variation interpersonnelle, variation inter-ménage, variation spatiale, variation temporelle et variation intrapersonnelle.

D'autres attributs de mobilité peuvent être utilisés. Raux et al. (2016) s'appuient notamment sur la variance de trois indicateurs (la fréquence journalière des déplacements, l'utilisation individuelle du temps et la séquence d'activités quotidiennes), divisée en une variance interpersonnelle et une variance intrapersonnelle, pour étudier la variabilité journalière des comportements d'activité et de déplacement. L'analyse des séquences d'activités est réalisée grâce à la méthode d'alignement des séquences (SAM ou *Sequential Alignment Method*) expliquée plus tard dans la section 2.4.2.

2.4.1.6 Indicateurs basés sur la diversité spatiale des lieux visités

La variabilité spatiale d'un individu est généralement mesurée par la diversité des lieux qu'il visite. Ainsi, plus les déplacements d'un usager sont concentrés aux mêmes endroits, et donc moins les lieux où il passe sont diversifiés, plus cet usager est dit régulier au niveau spatial. Les données de cartes à puce sont particulièrement intéressantes pour réaliser cette mesure puisqu'elles permettent d'énumérer les différents arrêts de bus ou stations de métro empruntés pour embarquer voire

débarquer par chaque utilisateur. L'indicateur le plus courant est le nombre total d'arrêts différents utilisés pendant une certaine période de temps (Huang et al., 2015).

Morency et al. (2007) proposent d'autres indicateurs individuels : après avoir représenté graphiquement une structure cumulative de la première utilisation de tous les arrêts, les auteurs calculent un taux d'acquisition (nombre de nouveaux arrêts acquis par semaine) et s'intéressent à la fréquence d'utilisation des arrêts les plus fréquentés (proportion des montées réalisées à chacun des différents arrêts). Certains de ces indicateurs sont appliqués et modélisés par Liu, L. et al. (2009). De plus, un calcul d'entropie peut être effectué pour mesurer la diversité des lieux d'embarquement. Briand et al. (2017) appliquent notamment l'entropie de Shannon à la probabilité pour chaque carte à puce d'être validée dans une de ses stations les plus fréquentées.

Le principal défaut reconnu de ces indicateurs est lié à la non-prise en compte de la proximité entre les arrêts; chaque arrêt est considéré de manière indépendante, quelle que soit sa distance avec les autres arrêts utilisés par l'individu (Morency et al., 2007). Or, dans un réseau très dense, certains arrêts peuvent être équivalents pour l'usager, qui en choisira un à tour de rôle pour effectuer les mêmes déplacements. Néanmoins, des algorithmes tels que le DBSCAN peuvent être appliqués pour regrouper les arrêts similaires (Kieu et al., 2014; Ma et al., 2013).

2.4.2 Comparaison de séquences

La variabilité des comportements de mobilité peut aussi être mesurée par la comparaison de séquences individuelles d'événements activités-déplacements (*activity-travel events*). Ces séquences, construites pour chaque individu, sont généralement des suites ordonnées d'activités, de lieux ou d'autres attributs de déplacements comme le mode emprunté. Elles contiennent des informations structurelles et séquentielles, c'est-à-dire que la valeur de chaque élément, mais aussi la position de cette valeur sont disponibles. Autrement dit, elles permettent d'étudier la nature, mais aussi l'ordre et l'organisation des activités ou déplacements réalisés par l'usager.

L'analyse de telles séquences est souvent faite à l'aide de la méthode d'alignement des séquences (SAM ou *Sequential Alignment Method*). Cette méthode calcule la dissimilarité ou distance, appelée la distance Levenshtein, entre deux séquences de caractères en termes de nombre minimal d'opérations (délétion, insertion ou substitution) nécessaires pour égaliser les deux séquences. Wilson (1998) a été le premier à l'appliquer dans un contexte de mobilité. Dans l'article

correspondant, il décrit plusieurs méthodes d'alignement (globale, locale, par matrice de points, multiple) et plusieurs manières de construire la séquence d'activités d'une même personne : il suggère une séquence courte (liste des activités) ou une séquence longue (activités par intervalle de 5 minutes). À l'aide du logiciel CLUSTAL, il compare d'une part les séquences journalières d'activités d'une même personne pour étudier sa variabilité intrapersonnelle, d'autre part les séquences de différentes personnes entre elles pour examiner la variabilité interpersonnelle. Moiseeva et al. (2014) utilisent la même méthode d'alignement appliquée cette fois à deux types de séquences hebdomadaires : une séquence d'activités et une séquence de lieux. Ils estiment également les deux types de variabilité individuelle : la variabilité intrapersonnelle en comparant les huit semaines enquêtées d'un même usager, et la variabilité interpersonnelle en considérant toutes les semaines de tous les usagers. Ils analysent ensuite l'évolution de ces deux types de variations dans le temps pour des nouveaux-arrivants.

De plus, une méthode d'alignement multidimensionnelle a été développée par Joh et al. (2002) pour considérer les différents attributs des événements activités-déplacements. Les séquences comparées sont alors des séquences à plusieurs dimensions. Par exemple, Xianyu et al. (2017) mettent en pratique cette méthode pour mesurer le degré de dissimilarité intrapersonnelle entre les séquences journalières de chaque individu. Les auteurs choisissent d'étudier deux attributs de mobilité (motif d'activité et mode de déplacement), mais ils évoquent d'autres dimensions possibles, notamment spatiale (lieu d'activité) et temporelle (durées d'activité ou de déplacement discrétisées). Ils soulignent que cette dernière dimension n'est pas facile à intégrer, car elle dépend grandement des intervalles de temps choisis; en particulier, de longs intervalles vont entraîner une sous-évaluation des activités courtes. Toutefois, ces méthodes d'alignement nécessitent une grande puissance informatique (Wilson, 1998). Pour réduire le fardeau du temps de calcul, Joh et Timmermans (2011) proposent la recherche de sous-séquences partagées.

Par ailleurs, d'autres types de distances existent pour mesurer la dissimilarité entre deux séquences de caractères. Phithakkitnukoon et al. (2010) utilisent par exemple la distance de Hamming, égale au nombre de positions où les caractères sont différents, pour comparer des séquences d'activités parmi et entre plusieurs groupes de travailleurs.

Une autre façon d'inclure le caractère séquentiel des événements activités-déplacements est de calculer un taux d'entropie comme cela a été fait par Goulet-Langlois et al. (2017). En effet, ce

taux garde en mémoire les événements précédents et quantifie leurs dépendances. Pour argumenter leurs propos, les auteurs confrontent ensuite cet indice à une entropie classique qui mesure seulement la diversité et la répétitivité des événements. Leurs résultats confirment que la fréquence, mais aussi l'ordre des événements sont essentiels pour traiter la question de la variabilité des comportements. Dans un autre article (Goulet-Langlois et al., 2016), les mêmes auteurs proposent une méthode innovante pour déduire des comportements types à partir de séquences. À chaque usager ils associent une séquence continue d'aires d'activités sur quatre semaines, ensuite discrétisée et binarisée dans une matrice multidimensionnelle. À l'aide d'une analyse en composantes principales (PCA), ils réduisent les dimensions de cette matrice et projettent chaque séquence sur les composantes principales les plus importantes pour obtenir un plus petit nombre de variables par usager. Une simple distance euclidienne peut alors être calculée pour comparer les séquences initiales.

Outre des séquences de caractères, des séquences de valeurs quantitatives peuvent être intéressantes à analyser dans un contexte de mobilité. Lorsque la variable étudiée est continue, l'analyste doit choisir des intervalles de temps discrets pour rapporter les quantités mesurées dans une séquence (Wilson, 1998). Ces séquences discrètes sont aussi appelées séries temporelles, définies comme la suite de valeurs numériques obtenues par des mesures séquentielles au cours du temps. Esling et Agon (2012) fournissent une revue de littérature des différents types d'analyses qui peuvent être faites et des méthodes qui peuvent être appliquées sur ces séquences. En particulier, ils recensent différentes mesures de similarité. La déformation temporelle dynamique (DTW ou *Dynamic Time Warping*) est un algorithme populaire qui en fait partie. Contrairement aux méthodes d'alignement, cette technique ne ramène pas les séquences comparées à la même longueur avec des suppressions-insertions, car un seul élément d'une séquence peut être associé à plusieurs éléments d'une autre séquence et inversement. Ces principes d'expansion et de compression sont expliqués et discutés par Kruskal et Liberman (1983). He et al. (2018) illustrent cette technique dans un cas concret du transport en utilisant des données de cartes à puce. Les auteurs construisent un série temporelle binaire de validations pour chaque carte et calculent ensuite la distance entre les séquences produites à l'aide de deux métriques : la distance de corrélation croisée (CDD ou *cross-correlation distance*) et la distance de déformation temporelle dynamique (DTW ou *Dynamic Time Warping distance*).

2.4.3 Modélisation de la variabilité

Enfin, la variabilité individuelle des comportements de mobilité peut être évaluée à l'aide de modèles. Quelques auteurs seulement ont réussi à la modéliser directement. D'autres ont au moins tenté de la prendre en compte dans leurs modèles de prédiction de la demande, souvent basés activités. Dans ce deuxième cas, l'effet de la variabilité des comportements sur le paramètre modélisé peut être quantifié. En outre, de tels modèles constituent un premier pas vers des modèles de prédiction plus précis et adaptés à des circonstances particulières. Les travaux présentés ci-après sont donnés dans un ordre croissant d'intégration de la variabilité des comportements dans les modèles développés.

En premier lieu, certains auteurs reconnaissent la nécessité d'inclure cette variabilité à la modélisation des comportements et mettent en évidence son influence sur la mobilité. Par exemple, Habib et Miller (2008) ne modélisent pas directement les variations quotidiennes des comportements, mais ils les prennent en compte en construisant un modèle différent pour chaque jour de la semaine. Les auteurs utilisent des modèles aléatoires de maximisation de l'utilité (*random utility maximizing models*) pour prédire la fréquence de 15 motifs d'activité sur une semaine. Ils prouvent ainsi que la qualité d'ajustement des modèles calculés pour chaque jour est bien meilleure que celle du modèle agrégé à la semaine. De manière plus explicite, Bhat et al. (2005) développent un modèle à risque multivarié non paramétrique (*multivariate non-parametric hazard model*) pour modéliser dynamiquement la durée entre deux participations successives à la même activité. Leur formulation du modèle permet de rendre compte des variations entre les différents jours de la semaine et inclut l'existence d'une certaine hétérogénéité entre les usagers. De plus, les variations intrapersonnelles dues à des caractéristiques individuelles non observées sont considérées et modélisées par une distribution gamma.

D'autres auteurs sont parvenus à modéliser directement l'impact de la variabilité sur des paramètres de mobilité. Les modèles d'équations structurelles (SEM), qui permettent d'analyser des relations causales entre plusieurs types de variables (endogènes, exogènes, latentes) dans des structures complexes, sont particulièrement adaptés à cet effet. Appliqués à des données longitudinales issus d'enquêtes de panel, ils permettent de capturer des évolutions dans la mobilité. Ainsi, bénéficiant de l'enquête TTAPS (Toronto Travel Activity Panel Survey) réalisée à Toronto par vague sur trois années consécutives, Roorda et Ruiz (2008) cherchent en particulier à comprendre des dynamiques

à court terme (entre les jours de la semaine) et à long terme (d'année en année) dans les changements de comportement. À l'aide d'un modèle d'équations structurelles avec variables latentes, ils modélisent l'effet du jour (*same-day effects*) ainsi que les effets d'un jour sur le suivant (*next-day effects*) et d'une année sur l'autre (*year-to-year effects*) en termes de nombre de déplacements par mode et de durées d'activités par motif. Dans le contexte d'un pays en voie de développement, Dharmowijoyo et al. (2016) évaluent avec le même type de modèle les interactions entre différents paramètres individuels de mobilité (le nombre de déplacements par jour, le choix du mode de déplacement, le nombre de chaînes, la durée totale de déplacement et l'heure de départ) en considérant la variabilité des contraintes quotidiennes et d'autres variables explicatives (des caractéristiques individuelles socio-démographiques ainsi que des caractéristiques du ménage, de l'environnement habité et du réseau de transport à proximité). L'impact des variations journalières est modélisé à l'aide de coefficients associés à chacune des variables et l'hétérogénéité non observée entre les individus est prise en compte par un terme d'erreur.

Par ailleurs, la variabilité des comportements est parfois utilisée comme une variable indépendante du modèle. Zhu et al. (2017) prédisent notamment un statut individuel socio-démographique (6 catégories : employé plein-temps, employé temps partiel, étudiant, personne au foyer, retraité ou autre) à partir des variances de quatre attributs de déplacements (l'heure de départ, le lieu de destination, la durée de déplacement et la durée de conduite). Pour cela, ils appliquent à 18 mois de données GPS un modèle d'apprentissage supervisé basé sur les machines à vecteurs de support (SVM ou *support vector machine*), aussi appelé séparateurs à vaste marge.

Pour aller encore plus loin, certains auteurs modélisent la variabilité individuelle elle-même; elle devient la variable dépendante du modèle produit. Chikaraishi et al. (2009) modélisent notamment la variance de l'heure de départ du premier déplacement de la journée à l'aide de modèles multi-niveaux. Ces modèles sont en effet appropriés pour décomposer la variabilité totale en différents types, selon différentes sources et différents niveaux de variations. Les auteurs dissocient ainsi des effets au niveau microscopique (variation intrapersonnelle) et des effets au niveau macroscopique (variations entre les ménages, les zones et les jours de la semaine). Pour leur part, Xianyu et al. (2017) utilisent la distance Levenshtein obtenue à partir des séquences journalières de chaque individu comme variable dépendante de modèles de régression des effets de panel (*panel effects regression models*). Ils aspirent ainsi à évaluer l'effet des variables sociodémographiques et des jours de la semaine sur la variabilité intrapersonnelle.

2.5 Segmentation des usagers du transport en commun

La segmentation d'une population en plusieurs sous-ensembles d'individus peut également être un moyen de mesurer la variabilité interpersonnelle. En effet, elle permet de révéler des différences entre les individus et de les synthétiser en un nombre limité de groupes de comportements typiques. Dans le cas du transport en commun, la segmentation des usagers peut aider les opérateurs à mieux connaître leurs clients. Elle permet notamment d'identifier et de caractériser différents segments du marché, chaque segment étant associé à des besoins spécifiques. De nos jours, l'évolution de la société rend ce genre d'études particulièrement d'actualités. En effet, alors que le service est généralement adapté pour des travailleurs réguliers se déplaçant 5 jours par semaine en heures de pointe, d'autres types de comportements apparaissent. De plus en plus de personnes travaillent seulement 4 jours de la semaine ou à distance quelques jours par semaine. Par ailleurs, on constate un nombre croissant d'immigrants et de touristes, usagers associés à des comportements souvent atypiques et irréguliers (Ghaemi et al., 2017).

La segmentation (ou *clustering* en anglais) est une technique d'apprentissage non supervisé qui consiste à diviser des individus (ou des objets) en plusieurs groupes distincts. Les individus d'un même groupe partagent un comportement similaire entre eux, mais différent de celui des individus des autres groupes. Cette notion de similarité dépend de la métrique utilisée. De plus, différents algorithmes de segmentation peuvent être appliqués. Dans la littérature scientifique du transport, on retrouve notamment des algorithmes de partitionnement, basés sur l'optimisation d'un critère, comme celui des K-moyennes (Zhao et al., 2014), des algorithmes hiérarchiques basés sur une structure hiérarchique agglomérative ou divisive (Agard et al., 2006), des algorithmes basés sur des modèles statistiques comme les modèles de mélanges gaussiens (Briand et al., 2017), des algorithmes basés sur la densité comme le DBSCAN (Kieu et al., 2014), et des algorithmes plus récents comme les réseaux de neurones (Ma et al., 2013). La plupart de ces algorithmes sont expliqués et illustrés dans la revue de littérature proposée par Jain et al. (1999).

Cette section du mémoire ne cherche pas à reproduire une des multiples revues d'algorithmes déjà proposées dans la littérature ; elle s'intéresse surtout à l'implémentation de ces algorithmes. En effet, il existe différentes façons de décrire les comportements des utilisateurs pour pouvoir ensuite les segmenter. Les résultats obtenus dépendront bien sûr de cette description et du choix des paramètres qui seront étudiés. Même si des typologies d'usagers peuvent être faites pour tous les

modes de transport, cette section se concentre surtout sur des méthodes appliquées dans le cadre de l'utilisation du transport en commun. Beaucoup des travaux recensés ci-après s'appuient d'ailleurs sur des données de cartes à puce. Trois méthodes principales sont discernées : les usagers sont souvent segmentés en fonction des caractéristiques de leur mobilité, de leurs séquences de déplacements ou d'activités, ou directement à partir de la régularité de leurs déplacements.

2.5.1 En fonction des caractéristiques de leur mobilité

Premièrement, une typologie d'usagers peut être réalisée en se basant sur les caractéristiques de leur utilisation du transport en commun. La méthode la plus traditionnelle pour décrire cette utilisation consiste à construire un vecteur d'indicateurs pour chaque usager avant d'appliquer un algorithme de segmentation. Plusieurs exemples de cette méthode sont énoncés dans la revue de la thèse d'Ortega-Tong (2013). L'auteur lui-même définit, pour chacun des utilisateurs de la carte à puce Oyster de Londres, un vecteur de 20 variables rapportant la fréquence de leurs déplacements, leurs attributs temporels et spatiaux, la durée de leurs activités, leurs caractéristiques sociodémographiques et leurs choix modaux. D'autres auteurs comme De Oña et al. (2016) ou Machado et al. (2018) intègrent également des variables reflétant l'opinion des individus et leur perception de la qualité du service du transport en commun. Ces informations subjectives, collectées à partir d'une enquête de satisfaction, sont ajoutées à d'autres variables sur les habitudes de déplacement et les alternatives de transport, ainsi qu'à des informations socioéconomiques.

Ces indicateurs peuvent être calculés comme un total ou une moyenne pour toute la période d'étude (Ortega-Tong, 2013), mais aussi être agrégés à différents intervalles de temps. En effet, à l'aide de données longitudinales comme les données de cartes à puce, il est possible de représenter l'utilisation du transport en commun par des séries temporelles. La définition de telles séries a été donnée précédemment dans la section 2.4.2. Dans un contexte de transport, ces séries permettent de rapporter les caractéristiques de la mobilité de chaque utilisateur dans une séquence temporelle de valeurs. Par exemple, l'utilisation du transport en commun peut être décrite à différents intervalles de temps plus ou moins longs en discrétisant le nombre ou la présence de validations de cartes à puce dans des profils hebdomadaires (Agard et al., 2006; El Mahrsi et al., 2014) ou journaliers (Morency et al., 2007). Les vecteurs ainsi obtenus sont ensuite utilisés comme données d'entrée dans un algorithme de segmentation pour mettre en évidence différents types de comportements. Au lieu de considérer des périodes temporelles indépendantes, Zhao et al. (2014)

représentent les activités des usagers sur des périodes de trois heures superposées par intervalle d'une heure (exemple : 8h00 -10h59, 9h00-11h59, etc.), ceci pour pouvoir inclure toute la durée d'un même déplacement dans la même période de temps. Les auteurs segmentent ensuite les usagers au niveau temporel en s'appuyant sur deux indicateurs (le nombre de jours d'activité et le nombre d'heures d'activité), et au niveau spatial en regardant la fréquence de leurs déplacements sur chaque paire Origine-Destination. Le croisement de ces deux types de regroupements montre que les usagers réguliers au niveau temporel le sont aussi au niveau spatial.

Ces méthodes de description de la mobilité, basées sur l'agrégation d'attributs de déplacements dans un vecteur, sont généralement utilisées avec des métriques scalaires (exemple : la distance euclidienne). Ces métriques servent en effet à mesurer la similarité (ou la dissimilarité) entre les usagers afin de pouvoir ensuite appliquer un algorithme de segmentation. Cependant, de nombreux auteurs soulignent les limitations de ces approches sensibles aux unités utilisées et qui, dans le cas de séries temporelles, sont incapables de rendre compte de la progression temporelle des événements (He et al., 2018). C'est pourquoi certains auteurs préfèrent des méthodes non basées sur une distance, mais sur des modèles. Briand et al. (2017) utilisent notamment un modèle génératif à 2 niveaux fondé sur des mélanges de Gaussiennes pour représenter le temps de manière continue plutôt que de le décomposer en valeurs discrètes. De même, De Oña et al. (2016) ou Machado et al. (2018) choisissent d'appliquer une méthode d'analyse de classes latentes (LCA ou *Latent class analysis*) afin de ne pas avoir à normaliser leurs variables, action qui aurait impacté les résultats de leur segmentation. D'autres auteurs ont développé des méthodes innovantes pour calculer une distance qui permet de mieux capturer les similarités entre les usagers. Par exemple, Agard et al. (2013) présentent une métrique qui considère la position relative des éléments de chaque vecteur (dans leur cas, la position des « 1 » dans chaque vecteur binaire de 24h utilisé pour caractériser chaque usager). L'utilisation du transport en commun de chaque usager est alors résumée par trois coordonnées polaires. Ghaemi et al. (2017) adoptent une méthode similaire en calculant une distance à partir de la projection des comportements temporels sur un demi-cercle.

2.5.2 En fonction de leurs séquences de déplacements ou d'activités

Les segmentations traditionnelles, basées sur l'agrégation scalaire des caractéristiques de la mobilité imposée par la discrétisation du temps, sont largement critiquées dans la littérature. Elles sont souvent jugées non satisfaisantes, car elles expliquent seulement une faible part de la

variabilité observée à l'intérieur des groupes formés (Schlich, 2003). En effet, elles ignorent des informations essentielles qui concernent l'organisation et l'ordre dans lesquels les déplacements ou les activités se déroulent (Goulet-Langlois et al., 2016). Ces informations peuvent être retranscrites à l'aide d'une séquence d'événements, construite pour chaque usager à la place du vecteur d'attributs binaires ou scalaires utilisé dans les méthodes traditionnelles. De plus, ces séquences représentent plus justement les comportements humains qui, selon Hägerstrand (1970), peuvent être vus comme des séquences d'actions interdépendantes et non permutables dans le temps et dans l'espace. Elles sont donc plus appropriées lorsqu'on veut segmenter des usagers en fonction de leur comportement. Ainsi, les distances et méthodes discutées dans la section 2.4.2 pour mesurer la variabilité interpersonnelle à partir de séquences peuvent aussi être exploitées dans une perspective de segmentation.

À partir des résultats d'une enquête longitudinale recueillis sur un échantillon de 361 personnes, Schlich (2003) oppose une méthode de segmentation traditionnelle à une segmentation basée sur la comparaison de programmes quotidiens. Dans le deuxième cas, il emploie une méthode d'alignement multidimensionnelle avec des séquences de déplacements caractérisés par quatre attributs : le motif, le mode, la distance et l'heure de départ, puis il applique un algorithme hiérarchique de Ward de minimisation de la variance. Ses résultats montrent que, comparés à ceux de la segmentation traditionnelle, les groupes obtenus avec la méthode d'alignement des séquences contiennent des informations supplémentaires non corrélées avec les caractéristiques sociodémographiques des personnes. De même, Saneinejad et Roorda (2009) segmentent 282 individus en neuf groupes à partir de leur séquence d'activités hebdomadaires habituelles. Pour cela, ils utilisent le logiciel Clustal, qui exécute une méthode d'alignement multidimensionnelle pour calculer un score de similarité entre des séquences associant un motif et un lieu d'activités à chaque intervalle de 15 minutes des 5 jours de la semaine. La segmentation de ces séquences hebdomadaires est ensuite réalisée à l'aide d'un algorithme itératif de jointure par voisin (*iterative neighbour-joining algorithm*).

À partir de données plus massives comme les données cartes à puce, l'enjeu du temps de calcul limite les possibilités d'application de ces méthodes. Néanmoins, Goulet-Langlois et al. (2016) réussissent à segmenter 33 026 utilisateurs de cartes à puce grâce à leur procédure. Celle-ci consiste à représenter chaque passager par une séquence d'activités s'étendant sur 4 semaines avant d'appliquer une méthode de réduction de la dimension. Les auteurs obtiennent ainsi huit variables

par usager, sur lesquelles ils appliquent ensuite un algorithme des K-moyennes. De plus, Joh et Timmermans (2011) ont trouvé une approche heuristique pour appliquer les méthodes d'alignement à des données massives dans un contexte de segmentation : après avoir identifié des combinaisons courantes de sous-séquences, cette méthode recommande de choisir des groupes représentatifs dans un sous-ensemble de séquences tirées au hasard puis de procéder par adhésion additive, c'est-à-dire de déterminer le groupe des séquences restantes grâce à un arbre de décision.

2.5.3 En fonction de leur régularité de déplacement

Le plus souvent, une typologie d'usagers est d'abord créée puis la régularité des comportements est étudiée dans chaque groupe séparément du processus de classification. Cependant, des auteurs procèdent parfois en sens inverse, c'est-à-dire qu'ils segmentent les usagers en fonction de la régularité de leurs déplacements. Cette régularité doit donc être mesurée au préalable à l'échelle individuelle avant d'appliquer un algorithme de segmentation.

Pour cela, certains auteurs construisent des indicateurs pour quantifier la variabilité ou la régularité des comportements puis ils segmentent les usagers à partir de ces indicateurs. Ortega-Tong (2013) inclut notamment des indicateurs de variabilité spatiale comme le nombre de stations origines différentes pour plusieurs types de jours dans ses variables de segmentation. Ma et al. (2013) s'appuient également sur quatre indicateurs décrivant la régularité de chaque usager: le nombre de jours actifs (fréquence de l'usager), le nombre d'heures de départ similaires pour le premier embarquement de la journée, le nombre de séquences de lignes similaires et le nombre de séquences d'arrêts similaires. Ils utilisent ensuite l'algorithme des K-moyennes ++ et la théorie des ensembles approximatifs (*rough set theory*) pour segmenter les usagers en cinq niveaux de régularité, qualifiée de très élevée, élevée, moyenne, basse ou très basse.

De leur côté, Kieu et al. (2014) divisent les usagers du réseau de Brisbane, Australie, en quatre groupes : les réguliers au niveau spatial, les réguliers au niveau temporel, les réguliers aux niveaux spatial et temporel et les non réguliers. Les auteurs commencent par mesurer la régularité spatiale et temporelle de chaque passager en appliquant l'algorithme DBSCAN pour déterminer des ensembles de paires Origine-Destination régulières et d'heures habituelles pour chaque passager. Ils placent ensuite les usagers dans les quatre groupes susmentionnés à partir de règles définies a priori, en fonction de la proportion de leurs déplacements qui sont faits à des heures habituelles ou sur des paires Origine-Destination régulières.

CHAPITRE 3 DESCRIPTION DES DONNÉES ET MÉTHODOLOGIE

Le chapitre précédent a prouvé que la littérature sur les cartes à puce était très abondante et diversifiée. Les données produites sont en effet exploitées dans de nombreux travaux portant sur l'utilisation du transport en commun. De plus, la revue a montré que la variabilité de cette utilisation pouvait être quantifiée. En résumé, il existe de nombreuses méthodes possibles pour mesurer la variabilité individuelle des comportements de mobilité, allant de la définition d'indicateurs à l'application d'algorithmes de segmentation, en passant par l'analyse de séquences et la modélisation.

Cependant, la principale difficulté rencontrée par les chercheurs pour entreprendre des études comme celles mentionnées précédemment réside dans la collecte de données longitudinales et individuelles. Les enquêtes de mobilité de plusieurs jours ou les carnets de déplacements long-terme tels que Mobidrive (Kitamura et al., 2006; Schlich, 2003; Schlich & Axhausen, 2003) ont ces propriétés, mais le coût et le fardeau de telles collectes conduisent souvent à des échantillons de petite taille. Or, d'après les résultats de Schlich et Axhausen (2003), l'analyse de la variabilité des déplacements nécessite des données continues sur au moins 2 semaines. C'est pourquoi des données massives comme les données de cartes à puce, utilisées dans ce mémoire, ou alternativement les données GPS des cellulaires, sont très appropriées.

Ce manque de données explique aussi pourquoi les études décrites plus haut se concentrent principalement sur la variabilité des comportements à l'échelle quotidienne (ou parfois hebdomadaire). Pourtant, des variations pourraient être visibles à d'autres niveaux temporels, par exemple au niveau mensuel. Ainsi, l'originalité de ce projet de maîtrise réside dans la longueur de la période d'étude : alors que les travaux antérieurs se sont souvent limités à quelques semaines ou quelques mois de données tout au plus, ce projet bénéficie d'un échantillon complet de transactions permettant d'examiner l'utilisation du transport en commun sur une année entière.

Tous les travaux précédemment recensés dans la revue de littérature serviront de base à ce projet qui applique des outils d'exploitation de données similaires. Néanmoins, la nouveauté de la méthodologie développée tient dans la combinaison de plusieurs de ces outils pour mettre à disposition des prototypes d'analyse évolués permettant d'étudier séparément les variabilités interpersonnelle et intrapersonnelle des usagers du transport en commun.

Ce chapitre 3 vise donc à présenter les données utilisées et à décrire de manière générale la méthodologie proposée. La première étape de prétraitement des données sera également détaillée. Les autres étapes seront introduites, mais elles seront développées dans les chapitres suivants.

3.1 Description des données

Les données exploitées dans ce projet proviennent de la base de données transactionnelles de la Société de Transport de Montréal (STM). Elles ont été produites grâce au système OPUS décrit dans la section 1.1.2. L'échantillon utilisé couvre une période d'un an, du 1er janvier 2016 au 31 décembre 2016, et correspond à un total de plus de 481 millions de validations. Ces validations se répartissent comme suit entre les trois supports du système.

Tableau 3.1 Répartition des validations totales de 2016 parmi les trois supports du système OPUS

Support	**CPCT**	**CPO**	**CMJ**	**Total**
% des validations totales	89.1%	5.4%	5.6%	100%

La grande majorité des validations de 2016 (89% d'entre elles, soit près de 430 millions de validations) ont été réalisées à l'aide d'une carte OPUS (ou CPCT, carte à puce commune de transport). Un peu moins de 2 millions de cartes à puce OPUS différentes ont été employées pour atteindre ce chiffre. En plus d'être le support le plus populaire, la carte OPUS est aussi l'unique support que l'on peut suivre dans le temps puisque les autres supports ne sont pas rechargeables. Or, cette condition de longitudinalité est indispensable pour l'étude de la variabilité des comportements. Seul cet échantillon de presque 430 millions de transactions réalisées par environ 2 millions de cartes à puce en 2016 est donc utilisé dans ce projet.

Par ailleurs, aucun filtre sélectif n'a été appliqué en amont, ce qui signifie que toutes les cartes ayant été validées au moins une fois durant l'année sont incluses dans l'analyse. L'idée est d'englober ainsi tous les types de comportements, réguliers et irréguliers, pour donner un portait exhaustif de la mobilité des usagers de la STM. En effet, les usagers occasionnels sont souvent mis de côté dans la littérature (El Mahrsi et al., 2017; Goulet-Langlois et al., 2016) et leur comportement est donc très peu étudié. De même, aucun tri n'a été fait car les données fournies par la STM étaient déjà nettoyées : la base de données ne contenait initialement aucune transaction incomplète ou incohérente.

À chaque fois qu'un usager a validé sa carte sur un dispositif de perception tarifaire, les informations listées dans le Tableau 3.2 ont été recueillies. Chaque observation contient un identifiant de validation, un identifiant anonyme de carte, le code du produit utilisé, l'horodatage de la validation (date et heure) ainsi que des informations partielles sur l'endroit où la carte a été validée (un numéro de station pour le métro, un numéro de véhicule, de ligne et un identifiant de direction pour le bus). Dans la base de données fournie par la STM, les numéros de station et de véhicule sont référés par un code d'emplacement billettique (CEB). Contrairement aux enquêtes de mobilité, aucune information personnelle sur l'usager n'est rapportée. De plus, ces informations ne sont disponibles qu'à l'embarquement seulement. En effet, aucune validation n'étant requise à la sortie du réseau de Montréal, les données ne sont pas collectées au débarquement.

Tableau 3.2 Informations recueillies à chaque validation tarifaire par mode (à l'embarquement)

Information		Métro	Bus
ID validation		√	√
ID carte (anonymisé)		√	√
Code produit		√	√
Date, heure (AA/MM/JJ HH:MM:SS)		√	√
Station/arrêt	CEB : Code d'emplacement billettique	√	x
Véhicule		x	√
Ligne		x	√
Direction		x	?

Spatialement parlant, seule l'origine des validations faites dans le métro est donc connue. La destination ne l'est pas puisqu'aucune information n'est collectée au débarquement et ni l'origine ni la destination de la validation du bus ne sont disponibles puisque seul le numéro de la ligne empruntée est renseigné. La direction du bus est également recueillie, mais peu fiable en raison de possibles erreurs de manipulation chez les conducteurs. En revanche, la ligne et la direction de métro utilisées ne sont pas évidentes : la ligne peut être devinée pour certaines stations mais, si l'usager embarque à une des quatre stations de correspondance du réseau de la STM (Jean-Talon, Snowdon, Lionel-Groux ou Berri-Uqam), différentes lignes peuvent ensuite être empruntées. Des méthodes d'imputation existent pour combler certaines de ces données manquantes mais elles n'ont pas été appliquées dans le cadre de ce projet.

Les codes des produits utilisés viennent avec un dictionnaire (également fourni par la STM), composé de presque 1400 titres de transport différents et disponibles sur carte OPUS. Ces titres comprennent ceux de la grille tarifaire en ANNEXE B (titres vendus par la STM), mais aussi les titres mis en vente dans les autres AOT puisque la carte OPUS est commune dans tout le Grand Montréal; certains titres peuvent donc être achetés à Laval, à Longueuil ou dans les couronnes pour se déplacer jusqu'à l'île de Montréal, territoire géré par la STM. Toutes les validations réalisées sur le réseau de la STM sont considérées dans ce projet, quelle que soit la provenance du titre utilisé. Néanmoins, ce dictionnaire inclut également des titres spéciaux, qui ont été disponibles ponctuellement pour des événements spécifiques (congrès, marathons, festivals, etc.) ou dans le cadre d'éditions spéciales, de promotions, de problèmes de circulation ou de frais de privilèges. En réalité, seuls 124 titres différents ont été utilisés par les usagers de l'échantillon considéré sur le réseau de la STM en 2016 (soit une moyenne de 1.74 produit par usager). Pour simplifier leur analyse dans la suite de ce mémoire, ces titres ont été regroupés en plusieurs catégories. La classification appliquée est celle qui a été transmise par la STM, ceci afin de rendre les futures observations de ce travail cohérentes avec la politique tarifaire de la société. Ainsi, chaque titre est associé à un des 6 types de produits suivants :

- Billets unitaires (un seul passage);
- Carnets (2, 6 ou 10 passages);
- Abonnements journaliers ou hebdomadaires;
- Abonnements mensuels;
- Abonnements longue durée : annuels ou 4 mois;
- Titres spéciaux,

et à un des 5 types de tarifs suivants :

- Tarif ordinaire;
- Tarif réduit;
- Tarif étudiant;
- Gratuité;
- Tarif spécial.

La catégorie « tarif étudiant » ne contient pas le tarif ‘18-25 ans’ de la STM, celui-ci étant inclus dans les tarifs réduits et ne pouvant être dissocié du tarif ‘6-17 ans et 65 ans et plus’ dans le dictionnaire fourni. Cette catégorie est donc uniquement composée des tarifs étudiants appliqués dans les autres AOT (exemple : le forfait TRAM étudiant). Plus d’informations sur l’utilisation de ces différents titres seront apportées avec l’analyse tarifaire des données faite dans la section 4.2.

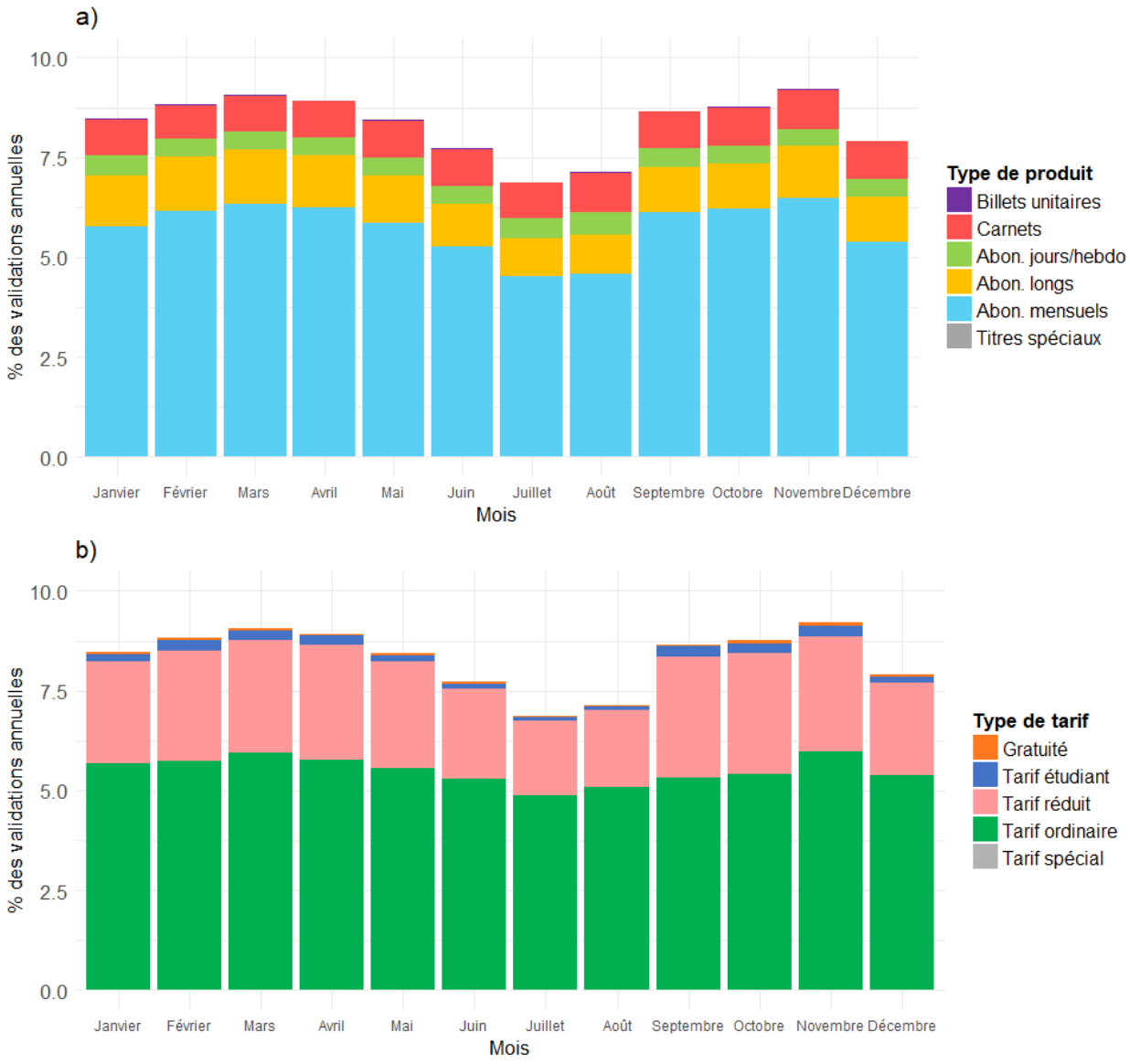


Figure 3.1 Distribution des validations totales de 2016 dans chaque mois de l’année a) par type de produit b) par type de tarif

Quelques statistiques descriptives sont présentées avec les Figure 3.1 et Figure 3.2 afin de donner un bref aperçu des données utilisées dans ce mémoire. Les graphiques produits représentent la

distribution des validations totales de 2016 agrégées par mois et par jour de la semaine. Ces validations sont également décomposées en fonction du type de produit ou du type de tarif utilisé. Les tendances collectives ainsi révélées font écho aux observations souvent faites dans la littérature. Au niveau annuel, l'achalandage du réseau diminue pendant les périodes estivales et hivernales, la proportion de validations la plus haute étant atteinte pour le mois de novembre. Au niveau hebdomadaire, le nombre de validations est plus faible en fin de semaine, notamment le samedi. On observe également des proportions plus réduites pour le lundi et le vendredi par rapport aux autres jours de la semaine, mais les jours fériés participent sûrement à ce phénomène. En termes de titres de transport utilisés, les abonnements mensuels et le tarif ordinaire sont les types de produit et de tarif qui ont généré le plus grand nombre de validations en 2016.

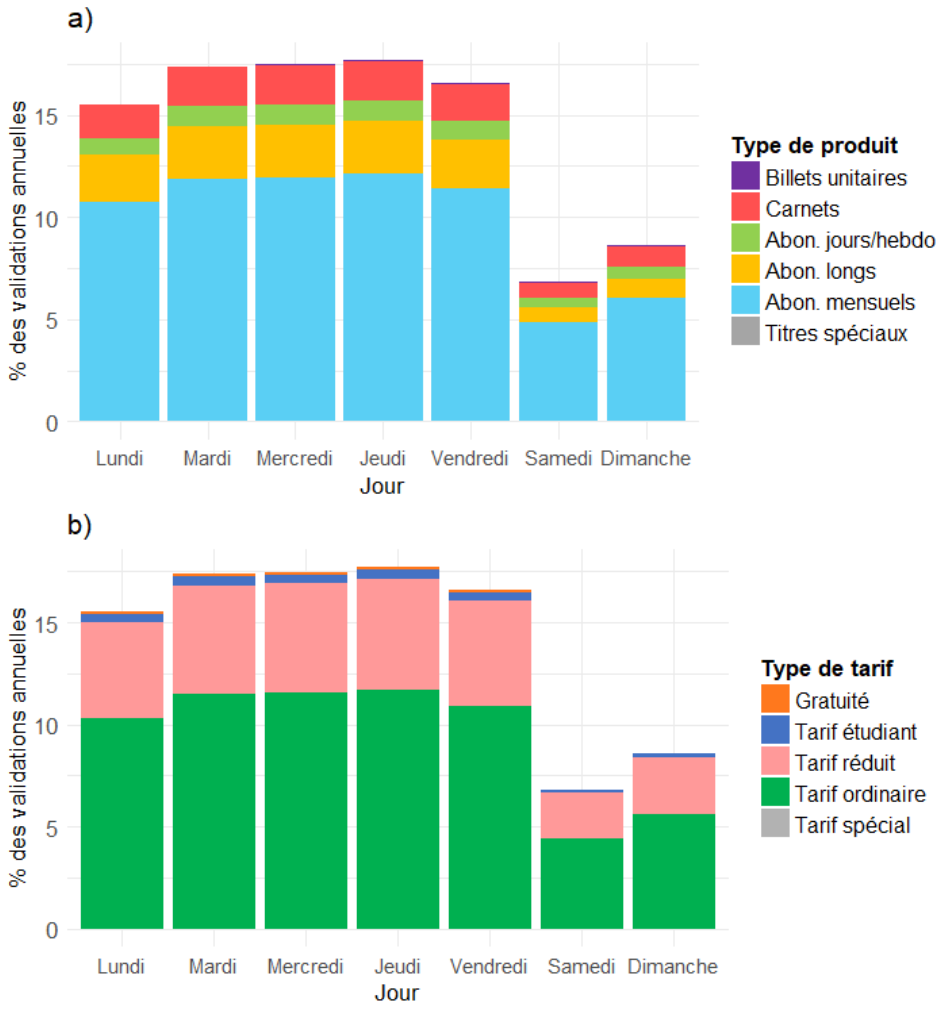


Figure 3.2 Distributions des validations totales de 2016 dans chaque jour de la semaine a) par type de produit b) par type de tarif

3.2 Méthodologie générale

Les fluctuations observées précédemment au niveau collectif sont induites par des variations individuelles : les usagers qui utilisent le transport en commun pour leurs déplacements domicile-travail ont tendance à moins se déplacer en été et pendant la fin de semaine car, pendant ces périodes-là, ils sont moins soumis à des contraintes de travail et donc moins captifs du réseau. Ces usagers, grands utilisateurs du transport en commun, sont les plus représentés dans les Figure 3.1 et Figure 3.2. Cependant, cette agrégation cache des singularités : d'autres types de comportements existent, mais ne sont pas visibles dans les profils collectifs car plus minoritaires. La courbe d'utilisation du transport en commun de ces usagers pourra suivre une autre évolution que celles observées précédemment, mais leur mobilité bien qu'atypique n'en sera pas moins régulière dans le temps (et/ou dans l'espace). L'objectif de ce mémoire est donc de mettre en évidence, mais aussi de quantifier toutes ces variations au niveau individuel.

Une définition de la variabilité individuelle (ou de la régularité individuelle, utilisée comme antonyme dans ce mémoire) doit tout d'abord être clarifiée et fixée pour toute la suite de cette recherche. Deux types de variabilité sont à dissocier. D'une part, la variabilité interpersonnelle est définie par les variations entre les individus. Cette première variabilité permet de distinguer différents types de mobilités (dont la traditionnelle dynamique domicile-travail). Ainsi, le comportement des usagers est dit variable au niveau interpersonnel dans le sens où leur utilisation du transport n'a pas les mêmes caractéristiques temporelles et spatiales d'un individu à l'autre. D'autre part, la variabilité intrapersonnelle est déterminée pour chaque individu par les variations de son comportement. Un usager est dit régulier au niveau intrapersonnel si son utilisation du transport en commun est uniforme dans le temps et dans l'espace. Un utilisateur régulier sera donc enclin à emprunter les mêmes stations et les mêmes lignes du réseau aux mêmes moments (mêmes jours, mêmes plages horaires), avec une fréquence et des durées d'activité relativement constantes. À l'inverse, un usager est dit variable ou irrégulier au niveau intrapersonnel si aucune tendance ou préférence ne peut être discernée dans son comportement, le rendant assez imprévisible.

Pour analyser et mesurer cette variabilité individuelle, une démarche en plusieurs étapes a été développée. Le cheminement et la structure de la méthodologie proposée ainsi que les dépendances entre les différences étapes qui la composent sont illustrés sur le schéma méthodologique de la Figure 3.3 suivante.

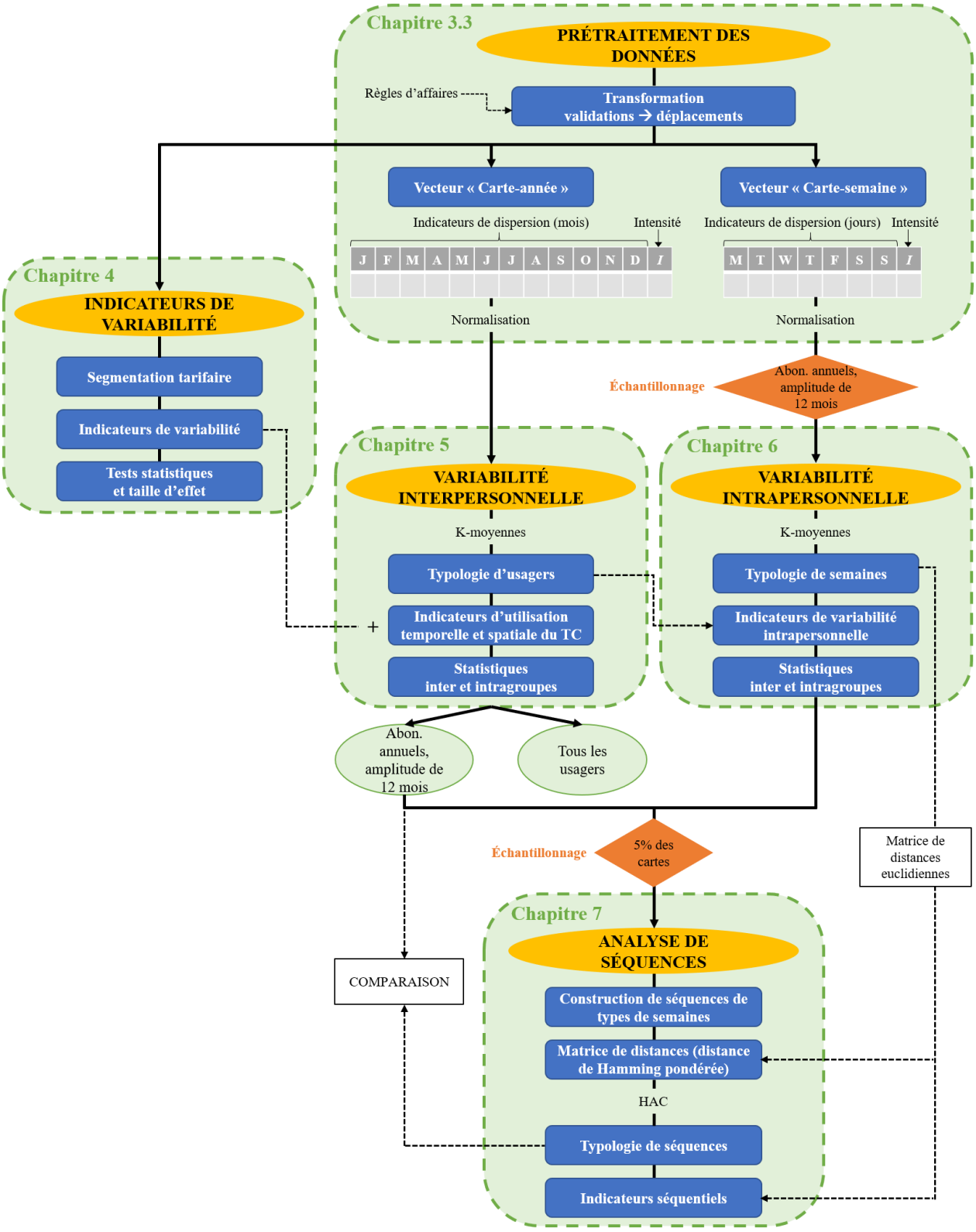


Figure 3.3 Schéma méthodologique général

Cette méthodologie est basée sur plusieurs techniques d'exploration de données qui, combinées, fournissent des prototypes complets d'analyse de la variabilité. Toutes les étapes de cette méthodologie sont introduites ci-après puis elles seront précisées dans la suite de ce mémoire.

La taille de la base de données de cartes à puce sur laquelle s'appuie cette recherche est un des principaux défis du projet. L'apprentissage de différents logiciels de gestion de données comme SQL et R est inéluctable afin de pouvoir manipuler des données aussi massives. De plus, une première étape de prétraitement est exécutée pour réduire la dimension des données à analyser. Les validations sont ainsi transformées en déplacements, puis la mobilité de chaque usager est résumée pour l'année au complet et pour chaque semaine de l'année. Deux types de vecteurs sont donc construits : des vecteurs « cartes-année » et des vecteurs « cartes-semaine ». Ces vecteurs servent ensuite à évaluer la variabilité d'utilisation du transport en commun sur deux niveaux temporels.

Tout d'abord, une analyse préliminaire des données est réalisée. En particulier, une segmentation tarifaire des cartes est établie pour montrer des différences de comportements selon les titres de transport utilisés. Plusieurs graphiques sont tracés pour mettre en évidence différents types de variations individuelles et des indicateurs, calculés pour chaque groupe de titres, sont définis pour quantifier les observations faites à partir de ces graphiques. Des tests statistiques sont également appliqués pour prouver le pouvoir explicatif des indicateurs proposés. Néanmoins, les limitations des tests classiques dues à la taille de l'échantillon étudié conduisent à l'utilisation d'une notion statistique particulière : la taille d'effet, plus connue sous le nom d'*effect size* en anglais.

La variabilité interpersonnelle est étudiée à partir des vecteurs « cartes-année ». Une typologie d'usagers (en réalité, de cartes) est créée en appliquant l'algorithme des K-moyennes. Cette typologie permet de déceler des différences entre les usagers en fonction de leur utilisation du transport en commun sur toute l'année 2016. Une première segmentation est faite pour tous les usagers confondus afin de permettre à la STM de mieux connaître l'ensemble de sa clientèle. Les grands types de comportements annuels observés sur le réseau de la STM sont ainsi exposés. Une deuxième typologie est ensuite effectuée sur les utilisateurs d'abonnements annuels seulement pour montrer qu'il existe des différences même parmi ce groupe d'usagers très fréquents et réguliers. Parmi ces usagers, seuls ceux avec une amplitude de 12 mois sont sélectionnés, c'est-à-dire ceux qui se sont déplacés au moins une fois en janvier et une fois en décembre, de manière à considérer uniquement les cartes présentes sur le réseau toute l'année. En outre, des indicateurs décrivant

l'utilisation individuelle moyenne du transport en commun sont calculés pour analyser les caractéristiques des groupes obtenus et introduire la dimension spatiale. Les indicateurs de variabilité précédemment évalués par titre de transport sont également estimés dans chaque groupe. Enfin, des tests sont appliqués pour confirmer les différences observées entre les groupes (statistiques intergroupes) et des coefficients de variations sont calculés pour mesurer la variabilité de chaque indicateur à l'intérieur de chaque groupe (statistiques intragroupes).

De même, la variabilité intrapersonnelle est analysée à partir des vecteurs « cartes-semaine ». Une typologie de semaines est produite et la régularité intrapersonnelle des usagers est mesurée par la répétition des mêmes types de semaines dans leur comportement au cours de l'année. Ainsi, un usager est considéré comme régulier au niveau intrapersonnel si sa mobilité se ressemble d'une semaine à l'autre. Cette typologie est faite seulement pour les utilisateurs d'abonnements annuels avec une amplitude de 12 mois, car la taille de la base de données « cartes-semaine » est bien plus élevée que celles des « cartes-année ». De plus, des indicateurs sont définis pour mesurer la variabilité intrapersonnelle moyenne à l'intérieur des groupes d'usagers obtenus précédemment et des statistiques sont calculées pour évaluer la variabilité inter et intragroupes de chaque indicateur.

Finalement, des séquences individuelles sont analysées dans le but de prendre en compte le caractère séquentiel et ordonné des déplacements des usagers. À partir de la typologie de semaines précédemment obtenue, une séquence de semaines types est construite pour chaque carte. Une matrice de dissimilarité entre toutes les séquences est ensuite calculée à l'aide d'une distance de Hamming pondérée (élaborée à partir d'une matrice de distances euclidiennes évaluées entre les types de semaines). Un échantillonnage est néanmoins nécessaire, car le calcul de cette matrice de distance est très exigeant en termes de mémoire et de temps. L'application d'un algorithme de segmentation hiérarchique agglomératif permet ensuite de créer une typologie de séquences, à comparer avec la typologie d'usagers produite précédemment. Encore une fois dans un souci de quantification, d'autres indicateurs sont présentés pour mesurer la variabilité d'utilisation du transport à commun à partir d'une séquence de comportements hebdomadaires.

3.3 Étape de prétraitement des données

La première étape de la méthodologie précédemment présentée consiste à prétraiter les données de cartes à puce avant de les exploiter dans de futures analyses. Cette première étape est composée de

deux sous-étapes : la transformation des validations en déplacements puis la construction de deux types de vecteurs pour résumer l'utilisation du transport en commun de chaque usager en 2016.

3.3.1 Transformation des validations en déplacements

Tout d'abord, les validations sont converties en déplacements. La Figure 3.4 rappelle la différence fondamentale entre ces deux concepts : un déplacement est un mouvement entre une activité d'origine et une activité de destination alors qu'une validation est faite à chaque nœud du déplacement impliquant un changement de mode ou de ligne. Seule la partie du déplacement effectué en transport en commun entre des arrêts ou stations est capturée par les données de cartes à puce : la vraie origine et la vraie destination du déplacement ne sont pas connues. À la STM, un usager doit revalider sa carte à chaque fois qu'il change de bus ou pour correspondre entre le bus et le métro. En revanche, il n'y a pas de validations métro-métro.

Cette conversion est faite afin d'éviter un biais spatial sur la fréquence d'utilisation (mesurée par le nombre de déplacements). En effet, l'utilisation des validations à la place des déplacements aurait pu conduire à une surestimation des longs parcours, c'est-à-dire qu'à nombre égal de déplacements, un usager qui fait des longs trajets, qui est donc susceptible d'avoir plus de correspondances et de valider plus souvent sa carte, aurait pu être considéré comme un utilisateur plus fréquent qu'un autre usager faisant des trajets plus courts (et donc moins de validations). Cette conversion des validations en déplacements permet aussi une réduction du volume des données puisque plusieurs validations peuvent avoir été effectuées dans un même déplacement.

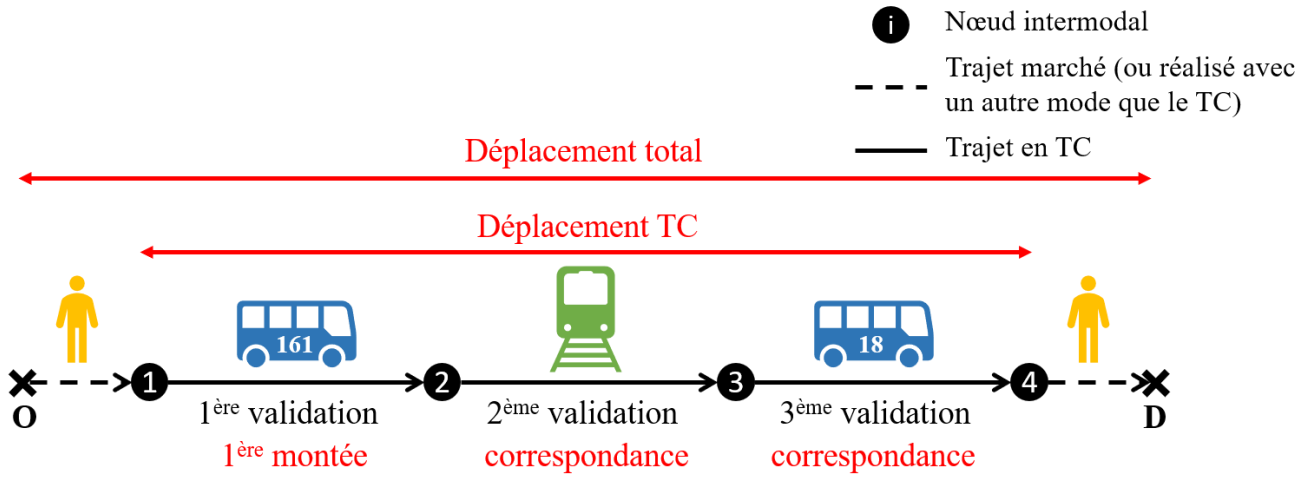


Figure 3.4 Schématisation d'un déplacement composé de trois validations

Pour réaliser cette transformation, des règles d'affaires basées sur la politique tarifaire de la STM ont été utilisées. Un usager du réseau de la STM ne peut pas reprendre la même ligne de bus et entrer deux fois dans le métro avec le même titre de transport. De plus, son titre est valide 120 minutes depuis sa première validation. Ainsi, un nouveau déplacement débute (première montée) lorsqu'une des trois conditions suivantes est rencontrée :

- Le temps depuis la première validation excède 120 minutes
- La ligne de bus courante est identique à une ligne précédente (sans distinction de la direction)
- La validation correspond à une deuxième entrée dans le métro

En appliquant ces règles aux validations successives de chaque utilisateur, il est possible d'associer à chaque validation un des deux types suivants: il s'agit d'une **première montée** lorsqu'elle satisfait les trois conditions précédentes, ou d'une **correspondance** sinon. Le type de chaque validation représentée dans la Figure 3.4 est donné en rouge. Seules les premières montées sont ensuite considérées pour compter le nombre de déplacements. Néanmoins, les déplacements ainsi calculés sont des déplacements faits sur le réseau de la STM. En effet, la possible utilisation d'autres réseaux avant l'arrivée de l'usager sur le réseau de la STM n'est pas prise en compte puisque les validations correspondantes ne sont pas rapportées dans la base de données de la STM.

Les résultats de cette conversion ont été vérifiés avec les calculs faits en interne par la STM. Une différence négligeable d'environ 0.02% a été obtenue dans le nombre total de déplacements de 2016, avec une déviation moyenne de 0.03 déplacement par usager. Cette différence est en particulier due aux bornes de l'échantillon étudié : les validations avant le 1er janvier et après le 31 décembre ne sont pas connues dans le cadre de ce projet, alors que la STM dispose de ces informations et peut donc faire des calculs plus justes. Toutefois, le nombre total de déplacements obtenu et toute autre information concernant les déplacements ne seront pas révélés dans ce mémoire pour des raisons de confidentialité. En effet, ils occupent une place importante dans l'estimation du financement de la STM chaque année. Or, l'achalandage officiel publié dans les rapports de la STM n'est pas calculé à partir des validations OPUS, mais à partir des ventes.

3.3.2 Création de deux types de vecteurs

Le nombre total de déplacements calculé, bien que réduit par rapport au nombre de validations, reste très élevé. Un moyen de décrire simplement et efficacement la mobilité des usagers a donc

dû être trouvé pour pouvoir manipuler une si grosse quantité de données à l'échelle individuelle. Pour cela, les déplacements de chaque usager sont ici résumés dans des vecteurs d'indicateurs. Inspirés du format utilisé par Morency et al. (2017) pour le vélopartage (Bixi), des vecteurs « cartes-année » sont construits pour rapporter l'utilisation mensuelle du système par chaque carte durant l'année et des vecteurs « cartes-semaine » sont constitués pour rendre compte de l'utilisation journalière de chaque carte pour chaque semaine de l'année. Ces deux types de vecteurs sont composés d'indicateurs d'intensité et de dispersion de l'utilisation du transport en commun. Le vecteur « carte-année » est un vecteur de 13 variables continues, incluant l'intensité mensuelle moyenne d'utilisation de la carte (le nombre moyen de déplacements par mois actif) et la distribution des déplacements réalisés dans chacun des 12 mois de l'année (le nombre de déplacements par mois, une variable par mois). Un mois est dit actif lorsque l'utilisateur de la carte a validé sa carte au moins une fois pendant ce mois. Cela signifie que seuls les mois avec plus de zéro déplacement sont considérés dans le calcul de l'intensité moyenne mensuelle de la carte, soient les mois durant lesquels l'utilisateur de la carte était présent sur le réseau. Par exemple, si un nouvel usager achète une carte en septembre, son intensité d'utilisation est calculée uniquement avec les mois venant après septembre (à condition qu'il se soit déplacé tous les mois). Cet indicateur d'intensité, le nombre de déplacements par mois actif, est choisi plutôt que le nombre moyen de déplacements par mois (actif ou pas actif) afin de ne pas répéter l'information contenue dans les indicateurs de dispersion. En effet, un vecteur rempli avec plusieurs zéros (par exemple tous les mois avant septembre) aurait conduit à un nombre moyen de déplacements par mois plus faible, ce qui n'est pas le cas pour le nombre moyen de déplacements par mois actif. De même, le vecteur « carte-semaine » contient le nombre moyen de déplacements par jour actif de la semaine et le nombre de déplacements effectués pendant chacun des 7 jours de la semaine, résultant en un vecteur de 8 variables continues.

Finalement, deux bases de données sont obtenues : une base « cartes-année » avec un nombre N de vecteurs correspondant au nombre de cartes dans l'échantillon considéré (environ 2 millions) et une base « cartes-semaine » constituée de N x 51 vecteurs. Chaque carte correspond à un vecteur « carte-année », mais à 51 vecteurs « cartes-semaine ». En effet, seules les semaines complètes de l'année 2016 sont considérées, c'est-à-dire les semaines du calendrier 2016 qui comprennent sept jours allant du lundi au dimanche. Cela correspond à 51 semaines, qui s'étalent du 4 janvier 2016 au 25 décembre 2016 (voir le calendrier en ANNEXE D). Cette condition de semaines complètes

est nécessaire pour la typologie de semaines qui sera expliquée plus tard dans le Chapitre 5. Les Tableau 3.3 a) et Tableau 3.4 a) montrent un extrait des deux bases de données ainsi produites.

Par la suite, les vecteurs sont normalisés afin d'attribuer aux variables des poids comparables. Plusieurs types de normalisation ont été testés, les distributions des variables ainsi normalisées ont été comparées, puis la méthode qui donnait les meilleurs résultats dans les segmentations des chapitres 5 et 6 a été conservée. Pour le vecteur « carte-année », la méthode de normalisation sélectionnée est la même que celle de Morency et al. (2017) : les nombres de déplacements par mois sont transformés en pourcentages des déplacements annuels et l'intensité mensuelle moyenne est normalisée en divisant le nombre de déplacements par mois actif par l'écart interquartile de sa distribution parmi les usagers multiplié par 1.5. Cette division par l'écart interquartile, aussi appliqué par Vogel et al. (2014), permet de recentrer les valeurs sur l'intervalle interquartile.

Tableau 3.3 Extrait de la base de données des cartes-année a) avant et b) après normalisation

a)

Indicateurs de dispersion — Indicateur d'intensité

ID Carte	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre	Moyenne mensuelle
i	51	41	47	46	36	39	44	54	33	34	46	31	41.83
ii	34	26	40	18	5	0	0	2	0	3	7	17	16.89
iii	0	2	0	0	0	0	2	0	0	2	3	0	2.25
iv	0	0	0	0	0	0	0	5	37	37	40	32	30.20
…	…	…	…	…	…	…	…	…	…	…	…	…	…

$$Mois_{norm} \leftarrow \frac{Mois}{\sum Année}$$

$$Intensité_{norm} \leftarrow \frac{Intensité}{1.5 * Écart\ interquartile}$$

b)

ID Carte	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre	Moyenne mensuelle
i	0.102	0.082	0.094	0.092	0.072	0.078	0.088	0.108	0.066	0.068	0.092	0.062	0.907
ii	0.224	0.171	0.263	0.118	0.033	0.000	0.000	0.013	0.000	0.020	0.046	0.112	0.366
iii	0.000	0.222	0.000	0.000	0.000	0.000	0.222	0.000	0.000	0.222	0.333	0.000	0.049
iv	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.245	0.245	0.265	0.212	0.655
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Pour le vecteur « carte-semaine », les nombres de déplacements par jour sont également convertis en pourcentages du nombre total de déplacements par semaine. De plus, l'intensité journalière moyenne est normalisée à l'aide d'une fonction logarithmique définie par l'équation 1. Comme la variable peut être nulle, on lui ajoute 1 puis on applique la fonction logarithme afin de réduire l'influence des valeurs aberrantes observées. La valeur obtenue est ensuite ramenée entre 0 et 1 à partir du minimum et du maximum comme dans l'équation 2. Les Tableau 3.3 a) et Tableau 3.4 a) présentent les résultats de ces normalisations.

$$Y_{i,norm} = [\log(Y_i + 1)]_{0\to1} \qquad \text{(Éq. 1)}$$

avec

$$[V_i]_{0\to1} = \frac{V_i - \min_i V_i}{\max_i V_i - \min_i V_i} \qquad \text{(Éq. 2)}$$

Tableau 3.4 Extrait de la base de données des cartes-semaine a) avant et b) après normalisation

a)

Indicateurs de dispersion — Indicateur d'intensité

ID Carte	Semaine	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche	Moyenne quotidienne
i	1	0	2	3	2	4	1	0	2.40
i	2	2	5	4	2	3	0	0	3.20
i	3	2	2	3	2	2	0	0	2.20
i	4	2	2	2	2	2	1	0	1.83
…	…	…	…	…	…	…	…	…	…

$Jour_{norm} \leftarrow \frac{Jour}{\sum Semaine}$

$Intensité_{norm} \leftarrow [\log(Intensité + 1)]_{0\to1}$

b)

ID Carte	Semaine	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche	Moyenne quotidienne
i	1	0.000	0.167	0.250	0.167	0.333	0.083	0.000	0.379
i	2	0.125	0.313	0.250	0.125	0.188	0.000	0.000	0.444
i	3	0.182	0.182	0.273	0.182	0.182	0.000	0.000	0.360
i	4	0.182	0.182	0.182	0.182	0.182	0.091	0.000	0.322
…	…	…	…	…	…	…	…	…	…

3.4 Introduction et justification des autres étapes méthodologiques

Les autres étapes méthodologiques constituent le cœur de ce travail de maîtrise sur la variabilité individuelle d'utilisation du transport en commun. Chacune de ces étapes correspond à un chapitre spécifique dans lequel la méthodologie utilisée sera précisée et les résultats de son application seront présentés. De plus, le caractère innovant et original de chacune de ces étapes peut être justifié en comparaison avec les travaux de la revue de littérature du chapitre précédent.

Dans les travaux recensés par le Chapitre 2, les indicateurs de variabilité calculés à partir de données de cartes à puce sont souvent présentés de manière anecdotique. Aucun auteur n'a formulé explicitement une série d'indicateurs simples et généralisables pour caractériser différents types de changements dans les comportements de mobilité. C'est là l'objectif du Chapitre 4: proposer des indicateurs reproductibles pour mesurer différentes sortes de variations dans l'utilisation du transport en commun, le tout grâce à des données de cartes à puce. Les indicateurs mis à disposition pourraient être utilisés pour comparer plusieurs années, villes ou groupes d'utilisateurs entre eux, du moment que les données nécessaires sont disponibles. Dans ce mémoire, ils sont appliqués pour évaluer des différences de comportement dans plusieurs groupes tarifaires (c.-à-d. des groupes d'utilisateurs construits en fonction des titres de transport utilisés durant l'année). Ils permettant ainsi d'analyser un lien entre la tarification et les comportements de mobilité.

Par ailleurs, les études précédemment décrites dans la revue de la littérature se sont principalement concentrées sur la variabilité journalière. Cela signifie que les auteurs supposent souvent que les comportements sont réguliers sur un cycle quotidien. Cependant, des variations pourraient être visibles à d'autres niveaux que le niveau quotidien, par exemple au niveau mensuel ou hebdomadaire. C'est pourquoi dans ce mémoire les comportements individuels sont étudiés à différentes échelles de l'année.

D'une part, la variabilité interpersonnelle est analysée au niveau mensuel dans le Chapitre 5. De cette façon, chaque utilisateur est caractérisé et dissocié des autres par son utilisation mensuelle du réseau de transport en commun sur l'année. Différents groupes d'utilisateurs sont donc formés en fonction de leur utilisation du transport collectif au cours des douze mois de 2016. La typologie résultante conduit à des grands types de comportements annuels spécifiés avec des profils mensuels.

D'autre part, la variabilité intrapersonnelle est examinée au niveau hebdomadaire dans le Chapitre 6. Toutes les semaines de chaque utilisateur sont considérées séparément et la régularité intrapersonnelle d'un utilisateur donné est définie comme la répétition des mêmes tendances hebdomadaires de mobilité dans son comportement au cours de l'année. Ainsi, une typologie des semaines est créée et la diversité des types de semaines observés dans le comportement individuel de chaque utilisateur est estimée.

De plus, de nombreux auteurs ont rappelé l'importance de prendre en compte la position et l'ordre dans lequel les événements de mobilité (déplacements ou activités) ont lieu. Des séquences individuelles sont donc analysées dans le Chapitre 7. Contrairement à ce qui est généralement vu dans la littérature scientifique, ces séquences ne sont pas des suites ordonnées d'activités ou de lieux d'embarquement mais des séquences de types de semaines issus d'une segmentation préalable (typologie de semaines du Chapitre 6). Une typologie de séquences est également produite afin de regrouper des individus avec des structures de comportements hebdomadaires similaires. Ainsi, les variabilités interpersonnelle et intrapersonnelle, distinguées dans les chapitres précédents, sont finalement croisées dans une même typologie à la fin de ce mémoire.

Par ailleurs, des indicateurs sont proposés dans chaque chapitre afin de ne pas perdre de vue l'aspect quantitatif désiré dans ce projet en plus du volet analytique. La dimension spatiale, souvent oubliée par les auteurs, est également apportée par ces indicateurs. La régularité spatiale des usagers est notamment révélée par l'utilisation des mêmes stations de métro et des mêmes lignes de bus durant l'échelle de temps étudiée.

CHAPITRE 4 MESURE DE LA VARIABILITÉ D'UTILISATION DU TRANSPORT EN COMMUN À L'AIDE D'INDICATEURS

Ce chapitre a pour objectif spécifique de mesurer la variabilité individuelle d'utilisation du transport en commun, définie par les variations observées à l'échelle individuelle dans l'utilisation du transport en commun. Initialement, ce chapitre devait présenter une analyse préliminaire des données permettant de mettre en évidence différents types de variations à l'aide de statistiques descriptives. Cependant, pour rendre cette analyse plus intéressante, des indicateurs ont ensuite été construits afin de quantifier les observations faites à partir des graphiques produits. Ainsi, à chaque type de variabilité observé a été associé un indicateur. Ces indicateurs calculés isolément ne permettent pas de conclure quoi que ce soit. Leur utilité, démontrée dans ce chapitre, réside dans la comparaison de plusieurs groupes d'utilisateurs.

4.1 Objectifs du présent chapitre

Le raisonnement général adopté dans ce chapitre est donné par le schéma de la Figure 4.1. L'idée directrice est d'évaluer la variabilité individuelle en fonction des titres de transport utilisés par le détenteur de chaque carte durant l'année, un titre étant considéré comme la combinaison d'un type de produit et d'un type de tarif. De cette manière, on suppose que l'utilisation de différents titres de transport mène à différents comportements de mobilité. Cette hypothèse, déjà validée par Habib et Hasnine (2017), sera également confirmée par les résultats. Par conséquent, ce chapitre vise à découvrir un certain pouvoir d'explication et de prédiction de la variabilité dans l'utilisation du transport en commun en se basant sur la composition tarifaire des cartes. D'ailleurs, le tarif est la seule information des données de cartes à puce pouvant être utilisée comme variable explicative dans un modèle de prévision de la demande (les autres informations sont des données d'usage).

Plusieurs types de variabilités peuvent être décelés avec les données de cartes à puce. Premièrement, la dispersion des déplacements parmi les utilisateurs de cartes montre des différences dans leur intensité d'utilisation annuelle: la majorité des déplacements sont souvent effectués par un petit nombre d'usagers. La variabilité de la fréquence d'utilisation peut également être déterminée en étudiant directement le nombre de déplacements réalisés quotidiennement, mensuellement ou annuellement par chaque utilisateur. De plus, l'instabilité de la plage horaire des embarquements de chaque usager, mais aussi la variance de son nombre de déplacements au cours

des mois peuvent faire ressortir des variations temporelles. La variabilité spatiale peut quant à elle être représentée par la diversité des lieux d'embarquement choisis.

Le but de ce chapitre est de proposer un indicateur pour quantifier chacun de ces types de variabilités. Pour cela, les indices de Pareto et de Gini, des mesures statistiques ainsi que l'entropie de Shannon sont utilisés. Par ailleurs, tous ces différents types de fluctuations peuvent être caractérisés selon deux points de vue : des variations peuvent être détectées entre les individus (variabilité interpersonnelle), mais également pour une même personne au fil du temps (variabilité intrapersonnelle). Certains indicateurs proposés sont donc décomposés en deux parties.

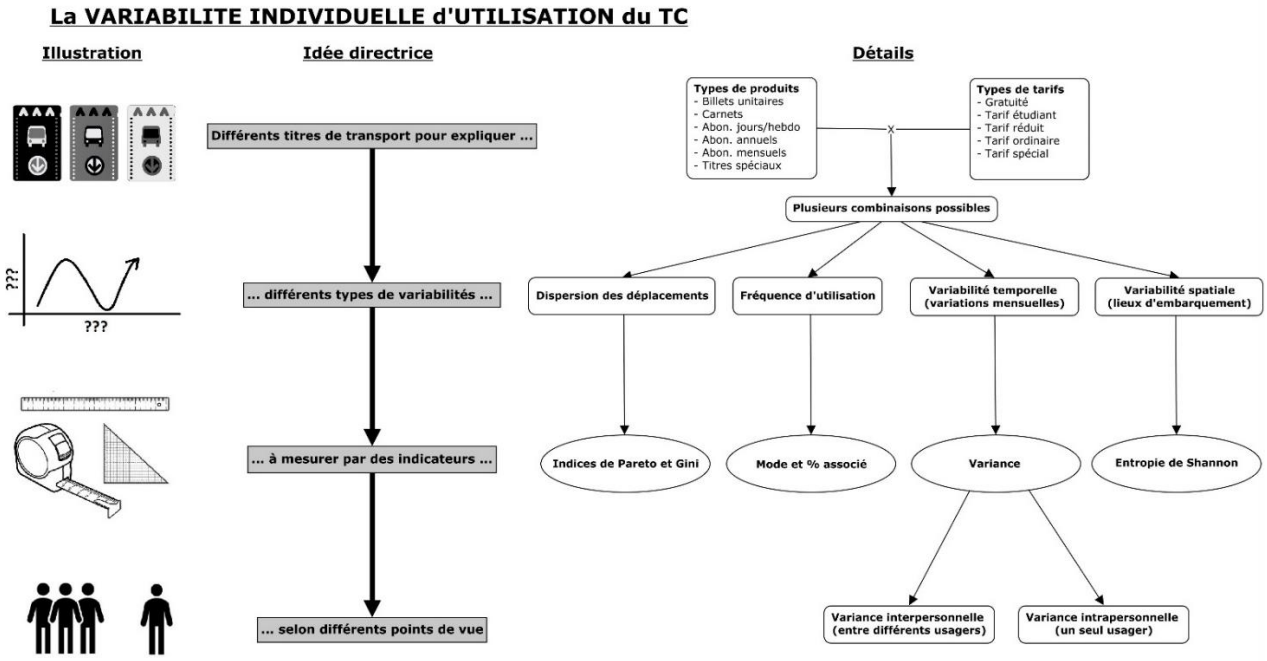


Figure 4.1 Schéma méthodologique du chapitre 4

Dans le cadre de cette démarche, les indicateurs présentés sont appliqués pour comparer plusieurs groupes de cartes définies en fonction des titres de transport validés pendant l'année. Les différences constatées entre les groupes sont ensuite testées statistiquement.

4.2 Segmentation tarifaire

L'objectif de cette deuxième section est de construire des blocs de cartes homogènes en taille en fonction de leur composition tarifaire. Avant d'aller plus loin, la terminologie utilisée est précisée grâce à l'équation suivante :

$$1\ titre = 1\ support\ \times 1\ produit\ \times 1\ tarif \qquad \text{(Éq. 3)}$$

Un titre de transport est défini comme la combinaison d'un support (carte OPUS, CPO ou CMJ), d'un produit (billet unitaire, carnet, abonnement mensuel, etc.) et d'un tarif (ordinaire, réduit, étudiant, etc.). Or, dans ce projet, seules les données de cartes à puce sont considérées, donc le support est toujours une carte OPUS. De plus, les produits et les tarifs ont été classés respectivement en 6 et 5 catégories dans la section 3.1. Un ensemble de 30 combinaisons de titres est donc disponible pour les analyses qui vont suivre.

Tout d'abord, les cartes sont distribuées dans le Tableau 4.1 en fonction du nombre total de types de produits et de tarifs différents utilisés au cours de l'année. Même si 6 catégories de produits et 5 catégories de tarifs ont été définies, le nombre maximum de types de produits et de tarifs utilisés en 2016 par les passagers de la STM est de 5 et 4 respectivement. Cependant, la grande majorité (63.2%) des détenteurs de cartes n'ont utilisé qu'un seul type de produit et un seul type de tarif durant l'année. En regardant les produits et les tarifs séparément, 92.5% des utilisateurs de cartes n'ont validé qu'un seul type de tarif et 63.9% n'ont validé qu'un seul type de produit sur toute l'année. Ces chiffres montrent que les usagers de la STM sont dans l'ensemble assez réguliers dans l'achat de leurs titres de transport.

Tableau 4.1 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%

Toutefois, les groupes obtenus dans le tableau précédent sont très hétérogènes. Pour répartir les cartes de manière plus équilibrée, les catégories auxquelles appartiennent les produits et les tarifs utilisés sont spécifiées en plus de leur nombre. La matrice de toutes les combinaisons possibles est disponible en ANNEXE E. (Remarque : le terme « combinaison » est employé plutôt que « séquence », car aucun ordre n'est supposé dans l'utilisation des différents titres de transport). En

particulier, les figures suivantes représentent la distribution des cartes ayant utilisé un seul type produit (Figure 4.2) et un seul type de tarif (Figure 4.3) durant toute l'année.

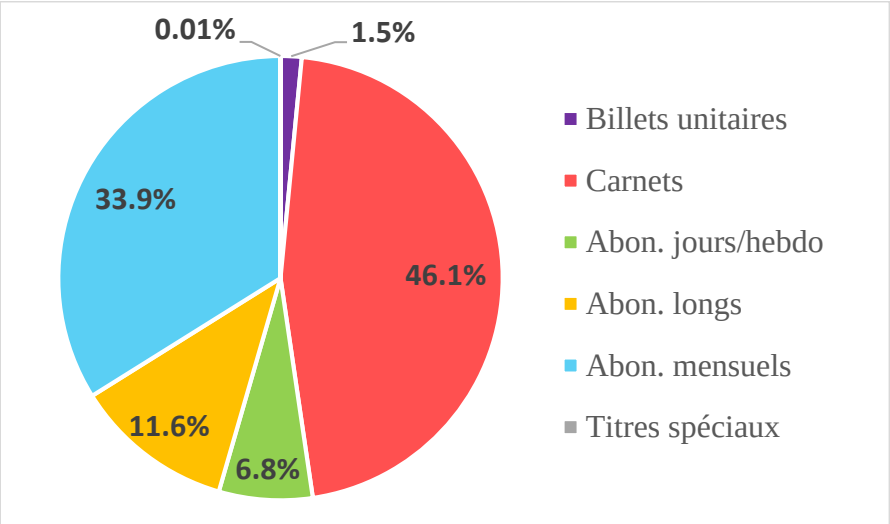


Figure 4.2 Distribution des cartes ayant utilisé un seul type de produit durant l'année (63.9% des cartes totales)

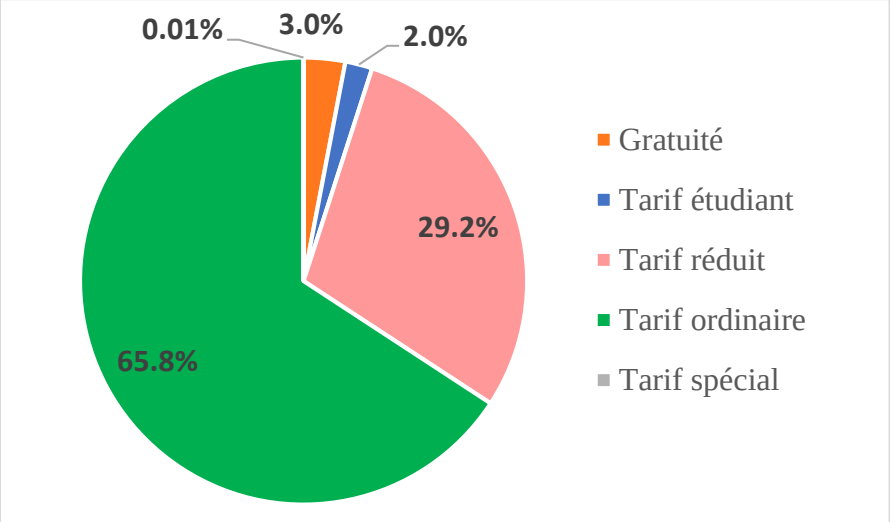


Figure 4.3 Distribution des cartes ayant utilisé un seul type de tarif durant l'année (92.5% des cartes totales)

Les groupes les plus volumineux du Tableau 4.1 comme celui des « 1 produit – 1 tarif » (en rouge) et celui des « deux produits – 1 tarif » (en orange) sont ainsi désagrégés en plusieurs sous-groupes, tandis que les autres groupes de cartes en jaune et en vert sont agrégés afin de rassembler les usagers ayant utilisé plus de 3 produits différents et ceux ayant utilisé plus de 2 tarifs différents durant

l'année. Le critère appliqué pour confectionner ces nouveaux groupes de cartes est de définir des ensembles réunissant plus de 5%, mais moins de 20% du nombre total de cartes. Les 10 combinaisons ainsi sélectionnées sont présentées dans le Tableau 4.2. Dans les graphiques de la section 4.3.2, chaque combinaison n°x sera notée COx.

Tableau 4.2 Les 10 combinaisons de cartes sélectionnées 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 carnets et les abonnements mensuels sont les types de produits les plus couramment utilisés sur le réseau de la STM, d'où leur présence dans toutes les combinaisons les plus populaires. Ces deux types de produits sont d'ailleurs employés séparément (combinaisons 1 à 4) ou ensemble (combinaison 7). Dans le deuxième cas, la Figure 4.4 atteste que les deux produits fonctionnent en complémentarité. Les carnets sont surtout utilisés en été puisque les usagers ont tendance à moins se déplacer durant cette période de vacances scolaires, alors que les abonnements mensuels sont achetés pour le reste de l'année. Au niveau des tarifs, le tarif ordinaire prédomine, suivi du tarif réduit. En effet, les combinaisons 1 et 2 regroupent les utilisateurs de carnets et d'abonnements mensuels à tarif ordinaire, puis les combinaisons 3 et 4 correspondent aux utilisateurs des mêmes types de produits avec un tarif réduit. En outre, la combinaison 5 est composée de cartes ayant utilisé des abonnements longs (annuels ou 4 mois) avec un tarif ordinaire. Or, d'après la grille tarifaire de la STM en ANNEXE B, l'abonnement de 4 mois n'est disponible qu'en tarif réduit. Les cartes de cette combinaison appartiennent donc uniquement à des utilisateurs d'abonnements annuels avec tarif ordinaire.

Les autres catégories de produits et de tarifs ont été moins utilisées par les passagers de la STM en 2016 et les combinaisons réalisées avec ces catégories conduisent à des groupes contenant moins de 5% du nombre total de cartes. Du fait du critère adopté, celles-ci ne peuvent pas être considérées comme des combinaisons indépendantes. Elles sont donc rassemblées dans les catégories « Autres » ou « Tous ».

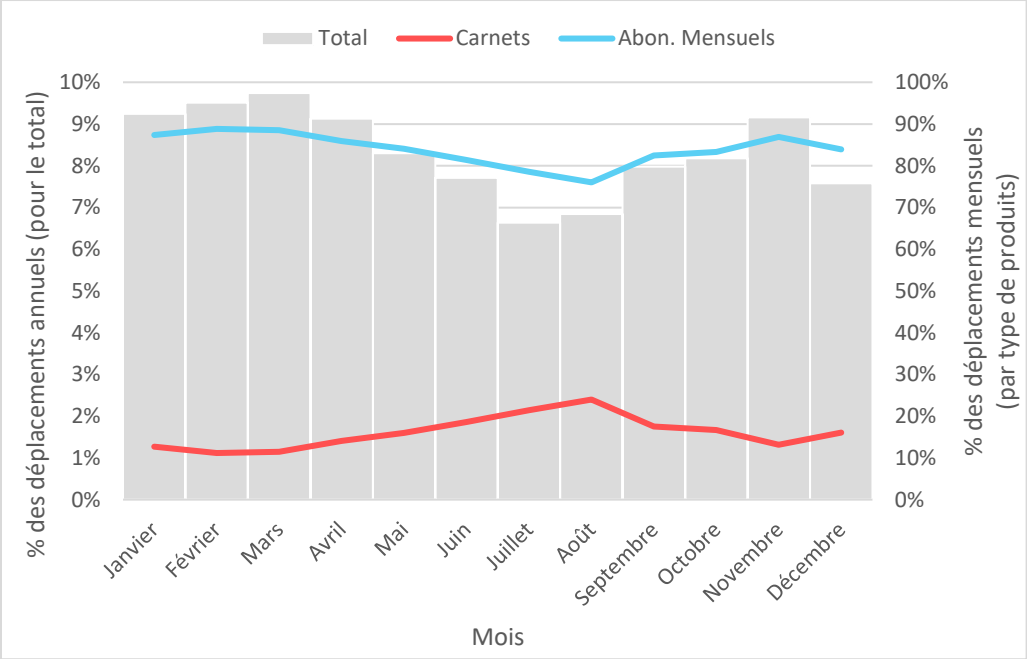


Figure 4.4 Distribution des déplacements des cartes appartenant à la combinaison 7 - Fluctuations de l'utilisation des deux types de produits

4.3 Indicateurs de variabilité

Un indicateur est associé à chacun des quatre types de variations mentionnés précédemment. Les quatre indicateurs correspondants sont définis ci-après puis ils sont calculés pour chacune des 10 combinaisons de cartes afin de mesurer la variabilité des comportements individuels en fonction des titres de transport utilisés.

4.3.1 Définition des indicateurs

4.3.1.1 Dispersion des déplacements parmi les usagers

Premièrement, la dispersion des déplacements parmi les usagers est quantifiée avec l'indice de Pareto, noté α (Pareto, 1896-97). Cet indice est généralement appliqué en économie pour rapporter

des inégalités dans la répartition des richesses d'une société : plus l'indice de Pareto est faible, plus la proportion des personnes à très hauts revenus est élevée et donc plus les richesses sont concentrées dans une petite fraction de la population. Cette mesure peut être transposée au transport en distribuant des nombres de déplacements (plutôt que des revenus) et en interprétant l'indice de Pareto comme suit : plus l'indice de Pareto est bas, plus la majorité des déplacements sont effectués par un petit nombre d'utilisateurs du réseau. Un indice de Pareto faible reflète donc une plus grande variabilité interpersonnelle de l'intensité annuelle d'utilisation. Inversement, un indice de Pareto élevé indique une distribution homogène des déplacements parmi les utilisateurs.

Quelques points théoriques doivent être énoncés pour expliquer les fondements de cet indicateur. Si le nombre annuel de déplacements réalisé avec une carte est considéré comme une variable aléatoire X avec une distribution de Pareto, la probabilité que X dépasse un certain nombre x est donnée par l'équation suivante :

$$P(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 1 & x < x_m \end{cases} \qquad \text{(Éq. 4)}$$

où x_m est la valeur minimale de X (ici, $x_m = 1$, car l'utilisateur de la carte doit avoir effectué au moins un déplacement pour faire partie de la base de données exploitée dans ce projet).

De cette définition, la fonction de répartition (ou fonction de distribution cumulative) suivante peut être déduite :

$$F_X(x) = P(X \leq x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 0 & x < x_m \end{cases} \qquad \text{(Éq. 5)}$$

Les autres éléments clé à introduire sont liés aux courbes de Lorenz qui, selon Eliazar (2016), sont « la base sous-jacente à la jauge sociogéométrique de la variabilité statistique de l'ensemble de données » [notre traduction]. Deux courbes doivent être définies : la courbe de Lorenz L et la courbe complémentaire de Lorenz $\bar{L}$. La première représente la distribution cumulative des valeurs de X (en ordonnée) en fonction de la distribution cumulée de la population (en abscisse), les valeurs de X étant triées en ordre croissant. En économie, chaque point (x, y) de cette courbe avec $y = L(x)$ est interprété comme « x% des personnes les plus pauvres possèdent y% des richesses ». La définition de la courbe complémentaire est similaire, mais les valeurs de X sont cette fois-ci rangées dans l'ordre décroissant. L'interprétation d'un point (x, y) avec $y = \bar{L}(x)$ devient : « x% des

personnes les plus riches possèdent y% des richesses ». Ces deux courbes sont dessinées sur la Figure 4.5 avec X le nombre annuel de déplacements par carte (à la place du revenu par personne). Lorsque les deux courbes sont sur la ligne $y = x$, il y a égalité parfaite entre tous les individus : c'est le cas quand par exemple, en économie, tout le monde gagne le même revenu ou, transposé au transport, quand tous les usagers font exactement le même nombre de déplacements par an.

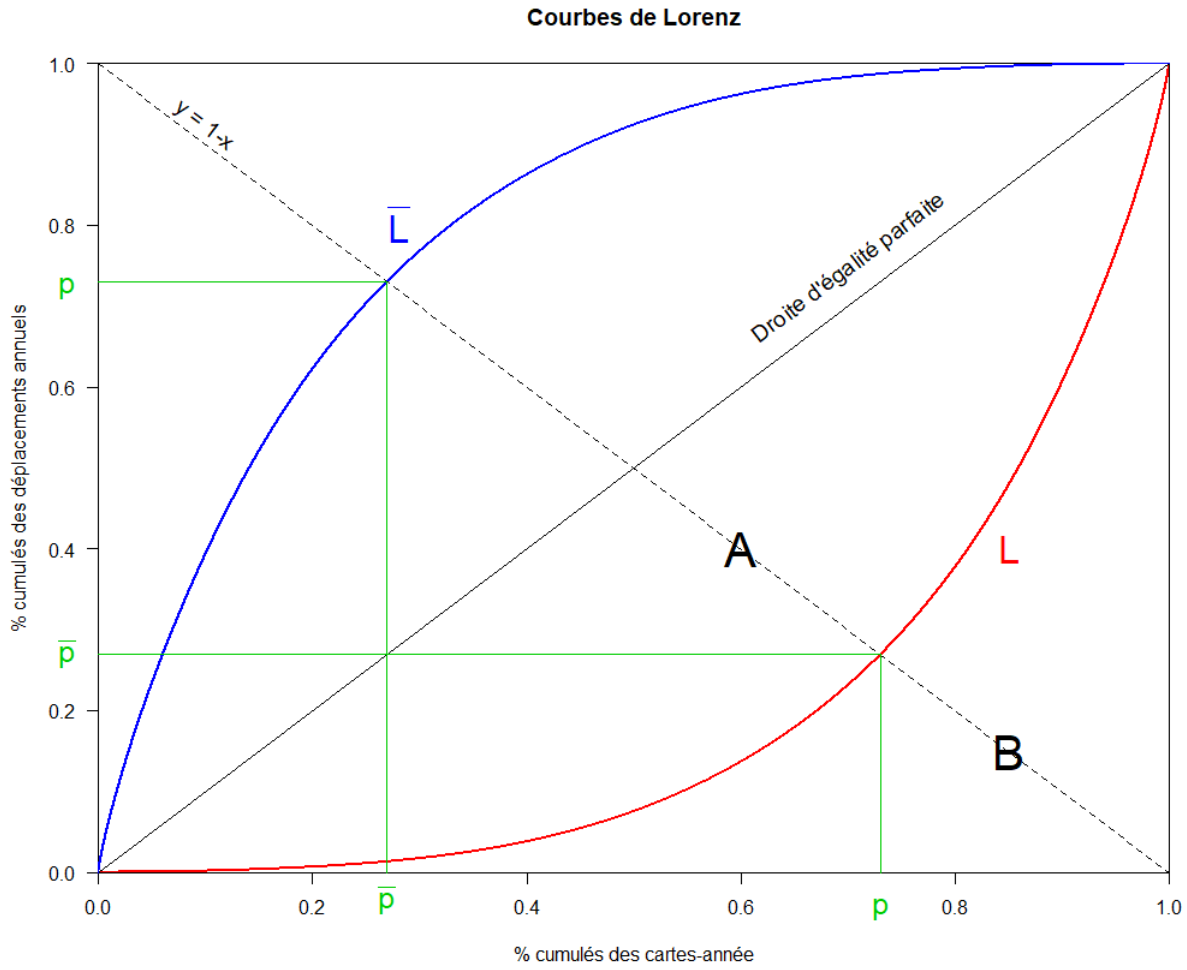


Figure 4.5 Courbes de Lorenz et proportions de Pareto appliquées au nombre annuel de déplacements par carte

Mathématiquement, pour une fonction de répartition $F(x)$ et sa bijection réciproque $x(F)$, la courbe de Lorenz est donnée par

$$L(F) = \frac{\int_0^F x(t)dt}{\int_0^1 x(t)dt} \qquad \text{(Éq. 6)}$$

Dans le cas particulier d'une distribution de Pareto,

$$x(F) = \frac{x_m}{(1-F)^{\frac{1}{\alpha}}} \qquad \text{(Éq. 7)}$$

et comme expliqué précédemment $x_m = 1$, d'où

$$L(F) = 1 - (1-F)^{1-\frac{1}{\alpha}} \qquad \text{(Éq. 8)}$$

À partir de cette courbe, les proportions de Pareto $(p, \bar{p})$ peuvent être déterminées. Elles sont également positionnées sur la Figure 4.5. 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 d'équation $y = 1 - x$. La proportion $\bar{p}$ est la valeur correspondante sur l'axe des y, telle 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 formules se traduisent par « p% des personnes les plus pauvres ont $\bar{p}$% des richesses » et « $\bar{p}$% des personnes les plus riches ont p% des richesses ». Transposé dans un contexte de transport, « p% des utilisateurs les moins fréquents ont effectué $\bar{p}$% des déplacements totaux de 2016 » et « $\bar{p}$% des utilisateurs les plus fréquents ont effectué p% des déplacements totaux de 2016 ».

Finalement, de $L(p) = 1 - p$ on tire l'expression suivante pour l'indice de Pareto :

$$\alpha = \frac{1}{1 - \frac{\log(p)}{\log(1-p)}} = \frac{1}{1 - \frac{\log(p)}{\log(\bar{p})}} = \log_{\frac{p}{\bar{p}}} \frac{1}{\bar{p}} \qquad \text{(Éq. 9)}$$

La fameuse « règle du 80-20 » $(p = 0.80, \bar{p} = 0.20)$, également connue sous le nom de principe de Pareto et selon laquelle 80% des effets proviennent de 20% des causes, conduit à $\alpha \approx 1.16$. Pour la distribution des déplacements produite sur la Figure 4.5, $p = 0.73$ et $\bar{p} = 0.27$ d'où $\alpha \approx 1.32 > 1.16$. La répartition étudiée ici dans un contexte de transport est donc plus homogène que celle énoncée par le principe de Pareto.

Le coefficient de Gini est une autre mesure de la dispersion statistique qui est souvent calculée en parallèle avec l'indice de Pareto (Ghosh et al., 2014; Inoue et al., 2015). Il peut également être évalué à partir des courbes de Lorenz par l'équation suivante :

$$G = \frac{A}{A+B} = 2A = 1 - 2B \qquad \text{(Éq. 10)}$$

où A et B sont les surfaces notées sur la Figure 4.5 (A est la surface comprise entre la droite d'égalité parfaite et la courbe de Lorenz L, tandis que B est la surface comprise entre la courbe de Lorenz L et les frontières du carré unitaire 1×1).

La valeur de cet indicateur varie entre 0 et 1: un coefficient de Gini de 0 indique une égalité parfaite entre les valeurs de la distribution ($A = 0$) alors qu'un coefficient de Gini de 1 exprime une inégalité totale ($B = 0$). De plus, l'indice de Pareto α et le coefficient de Gini G sont liés par l'équation suivante :

$$G = 1 - 2\int_0^1 L(F)dF = \frac{1}{2\alpha - 1} \qquad \text{(Éq. 11)}$$

Une augmentation de α entraîne donc une diminution de G.

4.3.1.2 Variabilité de la fréquence d'utilisation mensuelle

Pour quantifier la variabilité de la fréquence d'utilisation, les cartes sont d'abord distribuées en fonction de leur nombre moyen de déplacements par mois actif (voir l'exemple de la Figure 4.7 plus loin). Cet attribut, déjà défini dans la section 3.3.2, est calculé pour chaque carte comme le nombre moyen de déplacements par mois en considérant seulement les mois où il y a eu au moins un déplacement. De cette manière, seuls les mois pendant lesquels l'utilisateur de la carte était présent sur le réseau sont pris en compte. Les valeurs obtenues sont ensuite segmentées en 21 classes de même intervalle. Ces classes sont numérotées dans l'ordre croissant du nombre moyen de déplacements par mois actif, c'est-à-dire que la classe 1 regroupe les plus faibles fréquences d'utilisation mensuelle tandis que la classe 21 rassemble au contraire les plus grandes. Les extrémités de ces classes ne sont pas précisées par souci de confidentialité (ce sont des nombres de déplacements).

La proportion de cartes appartenant à chaque classe est déterminée. La classe modale, c'est-à-dire la classe la plus fréquente (avec la plus grande proportion de cartes), peut alors être désignée. 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 de cartes est élevé, plus leurs utilisateurs ont une fréquence d'utilisation mensuelle similaire. Cet indicateur mesure donc la variabilité interpersonnelle de la fréquence d'utilisation parmi les usagers du transport en commun.

4.3.1.3 Variabilité temporelle (variations du nombre de déplacements par mois)

La variabilité temporelle est traitée en examinant les variations du nombre mensuel de déplacements effectués avec chaque carte au cours des 12 mois de l'année. Ici on ne considère pas seulement les mois actifs, mais tous les mois de l'année puisque qu'on s'intéresse aux fluctuations du nombre de déplacements dans le temps (si un usager ne s'est pas déplacé en été, cela doit être pris en compte dans sa variabilité temporelle). De manière similaire à Raux et al. (2016) qui ont utilisé la variance du nombre de déplacements par jour pour mettre en évidence la variabilité temporelle au niveau hebdomadaire, la variabilité au niveau annuel est ici évaluée par la variance du nombre de déplacements par mois. La variance totale (TSS: *total sum of squares* ou somme totale des carrés) peut être décomposée en une variance intrapersonnelle (WPSS: *within-person sum of squares* ou somme des carrés par personne [par carte ici]) et une variance interpersonnelle (BPSS: *between-person sum of squares* ou somme des carrés entre personnes [entre cartes ici]).

$$TSS = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(n_{ij} - \bar{n}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(n_{ij} - \bar{n}_i\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{n}_i - \bar{n})^2 \qquad \text{(Éq. 12)}$$
$$= WPSS + BPSS$$

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 pour la carte i sur la période J, $\bar{n}$ le nombre moyen de déplacements par mois pour toutes les I cartes sur la période J. Les valeurs de $\bar{n}_i$ et $\bar{n}$ sont donc calculées par:

$$\bar{n}_i = \frac{1}{J}\sum_{j=1}^{J} n_{ij} \quad \text{et} \quad \bar{n} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \qquad \text{(Éq. 13)}$$

Dans les comparaisons qui suivront (comparaisons de plusieurs blocs de cartes en fonction de leur composition tarifaire), les proportions respectives des variabilités interpersonnelle ($\frac{BPSS}{TSS} \times 100$) et intrapersonnelle ($\frac{WPSS}{TSS} \times 100$) dans la variabilité totale seront déterminées pour chaque groupe afin de connaître la part de la variance qui expliquée par les différences entre les usagers et celle qui est due aux variations du comportement de chaque usager. Cependant, il est important de noter que, pour pouvoir comparer les variances totales de plusieurs groupes de cartes de tailles différentes, il

faudra diviser TSS par le nombre total de cartes I de chaque groupe de manière à rendre les valeurs obtenues comparables. Cette opération fournira une variance moyenne par utilisateur pour chaque groupe, comme Pas et Koppelman (1987) l'ont fait pour chaque segment de population étudié. En matière d'interprétation, plus cette variance moyenne est faible, plus les comportements sont réguliers au niveau temporel, c'est-à-dire plus les cartes du groupe considéré sont similaires entre elles en termes de mobilité mensuelle moyenne (au niveau interpersonnel), et plus le nombre de déplacements par mois de chaque carte est constant au cours de l'année (au niveau intrapersonnel).

4.3.1.4 Variabilité spatiale (diversité des lieux d'embarquements)

Enfin, la variabilité spatiale est évaluée par la variation des lieux d'embarquement de chaque utilisateur de carte au cours de l'année. Cette diversité d'utilisation spatiale est estimée en calculant un indicateur d'entropie individuel pour chaque carte. Plus précisément, comme dans les travaux de Briand et al. (2017), l'entropie de Shannon est utilisée:

$$H_i(X) = -\mathbb{E}\left[\log P\left(X = x_{ij}\right)\right] = -\sum_{j=1}^{n} P_{ij} \log P_{ij} \qquad \text{(Éq. 14)}$$

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

$$H_i^*(X) = \frac{H_i(X)}{\log(n)} \qquad \text{(Éq. 15)}$$

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 faites au lieu d'embarquement j. Il convient de souligner ici que les validations et non les déplacements sont exploitées, car un déplacement peut être composé de plusieurs validations effectuées avec différents modes ; plusieurs lieux d'embarquement sont donc rencontrés au cours d'un tel déplacement.

Néanmoins, cette entropie individuelle doit être calculée séparément pour le métro et pour le bus. Comme mentionné précédemment dans la présentation des données, les lieux d'embarquement sont connus au niveau de la station pour le métro, mais seule la ligne d'embarquement est disponible pour le bus. Ainsi, le lieu d'embarquement j dans le calcul de l'entropie du métro fait référence à une station de métro, avec un total de $n = 68$ stations dans le réseau de la STM. Pour l'entropie

du bus, le lieu d'embarquement j est une ligne de bus et $n = 233$ lignes différentes ont été empruntées par les usagers en 2016 (incluant des lignes non régulières).

D'après la définition ci-dessus, l'entropie normalisée H_i^* est proche de 0 quand beaucoup de probabilités P_{ij} sont proches de 0 ou 1, ce qui signifie que les lieux d'embarquement de l'utilisateur ne sont pas très diversifiés (c.-à-d. il 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 une avec les validations de bus) mesurent la variabilité spatiale intrapersonnelle de l'utilisateur sur l'année. Plus ces entropies sont faibles, plus l'utilisateur de la carte est régulier au niveau spatial. La variabilité interpersonnelle peut également être analysée en comparant les entropies individuelles de différentes cartes entre elles.

Pour comparer les différents groupes de cartes définis précédemment, une entropie spatiale moyenne et un coefficient de variation seront calculés à partir des cartes appartenant à chaque groupe. Des diagrammes violon seront également tracés afin de visualiser la courbe de densité de probabilité des valeurs d'entropie calculées dans chaque groupe en sus de la boîte à moustaches correspondante (voir la Figure 4.10 plus loin).

4.3.2 Application des indicateurs

Dans cette section, les différents types de variabilités évoqués précédemment sont visualisés à l'aide d'outils graphiques et les indicateurs correspondants sont calculés pour chacune des 10 combinaisons de cartes constituées dans la section 4.2. Les résultats de l'application de ces indicateurs sont donnés dans le Tableau 4.3.

Dans la première partie de ce tableau, les indicateurs proposés mesurent l'homogénéité de la répartition des déplacements parmi les utilisateurs de cartes à puce de chaque groupe. L'indice de Pareto α le plus élevé et donc le coefficient Gini le plus faible G sont obtenus pour la combinaison 7 (carnets et abonnements mensuels utilisés conjointement avec un tarif ordinaire); c'est le groupe le plus uniforme en termes de nombre annuel de déplacements par carte. De même, les indices de Pareto des combinaisons 9 (utilisation de plus de 3 types de produits différents), 5 (abonnements annuels) et 10 (utilisation de plus de 2 types de tarifs différents) sont assez élevés. Une plus grande diversité dans l'achat des titres de transport suggère donc une plus faible variabilité interpersonnelle. Les utilisateurs d'abonnements annuels sont également assez similaires entre eux

dans leur intensité d'utilisation annuelle. Au contraire, les combinaisons 1 et 3 (carnets) ont des indices de Pareto faibles et des coefficients de Gini élevés : certains usagers de ces groupes sont donc beaucoup plus assidus que d'autres. Cette conclusion semble être indépendante du tarif (ordinaire versus réduit), ce qui est confirmé par la Figure 4.6, les deux courbes de Lorenz CO1 et CO3 étant très proches l'une de l'autre. L'indice de Pareto le plus bas et le coefficient de Gini le plus élevé sont atteints pour la combinaison 6 (cartes avec un seul type de produit et de tarif différents de ceux des combinaisons les plus populaires). La courbe de Lorenz correspondante est en effet celle qui est la plus à gauche sur la Figure 4.6. Cela signifie que la répartition des déplacements parmi les cartes de ce groupe est très inégale et que la majorité des déplacements sont effectués par un petit groupe d'utilisateurs. Cependant, cette tendance peut être expliquée par le fait que les cartes regroupées dans cette combinaison sont assez diverses. En effet, ce groupe contient des utilisateurs d'abonnements longs, annuels ou 4 mois (tarif réduit), mais aussi des utilisateurs de billets unitaires, usagers qui ont probablement des comportements assez différents.

Tableau 4.3 Calcul des indicateurs de variabilité dans chacune des 10 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* (10^{-3})	0.38	9.07	0.30	7.19	6.13	2.43	5.37	5.49	6.56	5.59	5.14
	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%

*TSS est divisé par le nombre de cartes dans chaque combinaison

**Ces combinaisons ont été définies dans le Tableau 4.2

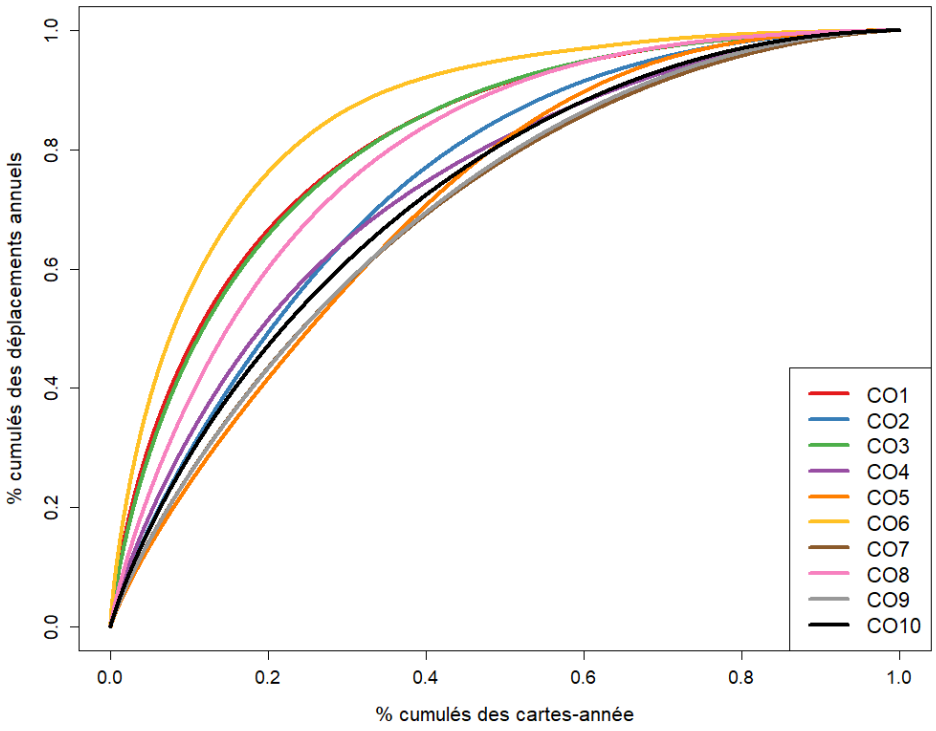


Figure 4.6 Courbes de Lorenz (complémentaires) par combinaison de cartes

La variabilité de la fréquence d'utilisation mensuelle est évaluée avec les indicateurs de la deuxième partie du Tableau 4.3 et illustrée par la Figure 4.7. La classe modale la plus élevée (classe n°8) se trouve dans les combinaisons 2 (abonnements mensuels avec tarif ordinaire) et 5 (abonnements annuels avec tarif ordinaire), mais avec une plus grande proportion de cartes dans le second cas. Ces deux types de produits sont donc utilisés par des usagers fréquents et cette forte utilisation du transport en commun est encore plus courante chez les abonnements annuels. Les utilisateurs d'abonnements mensuels à tarif réduit (combinaison 4) effectuent également un nombre moyen de déplacements par mois actif élevé, mais légèrement inférieur à celui obtenu pour le tarif ordinaire. À l'inverse, les combinaisons 1 et 3 (utilisateurs de carnets) regroupent des usagers peu fréquents, plus de 60% des cartes étant utilisées pour faire un nombre moyen de déplacements par mois actif compris dans l'intervalle de la classe la plus faible (classe n°1). Ces pourcentages élevés montrent que la faible intensité d'utilisation du transport collectif observée dans ces deux groupes est commune à la majorité des cartes. La classe modale des combinaisons 6 et 8 est également la classe n°1, probablement en raison du grand nombre d'utilisateurs d'abonnements quotidiens, de carnets et de billets unitaires dans ces deux groupes. Comme attendu, la classe modale de la combinaison 7 (carnets et abonnements mensuels utilisés ensemble, tarif ordinaire) est intermédiaire entre celles

des utilisateurs mono-produit des combinaisons 1 (carnets, tarif ordinaire) et 2 (abonnements mensuels, tarif ordinaire). De plus, l'utilisation de plusieurs types de produits ou de tarifs (combinaisons 7, 9, 10) conduit à des fréquences d'utilisation et des pourcentages mensuels comparables. La similarité des distributions CO7, CO9 et CO10 sur la Figure 4.7 peut d'ailleurs être soulignée.

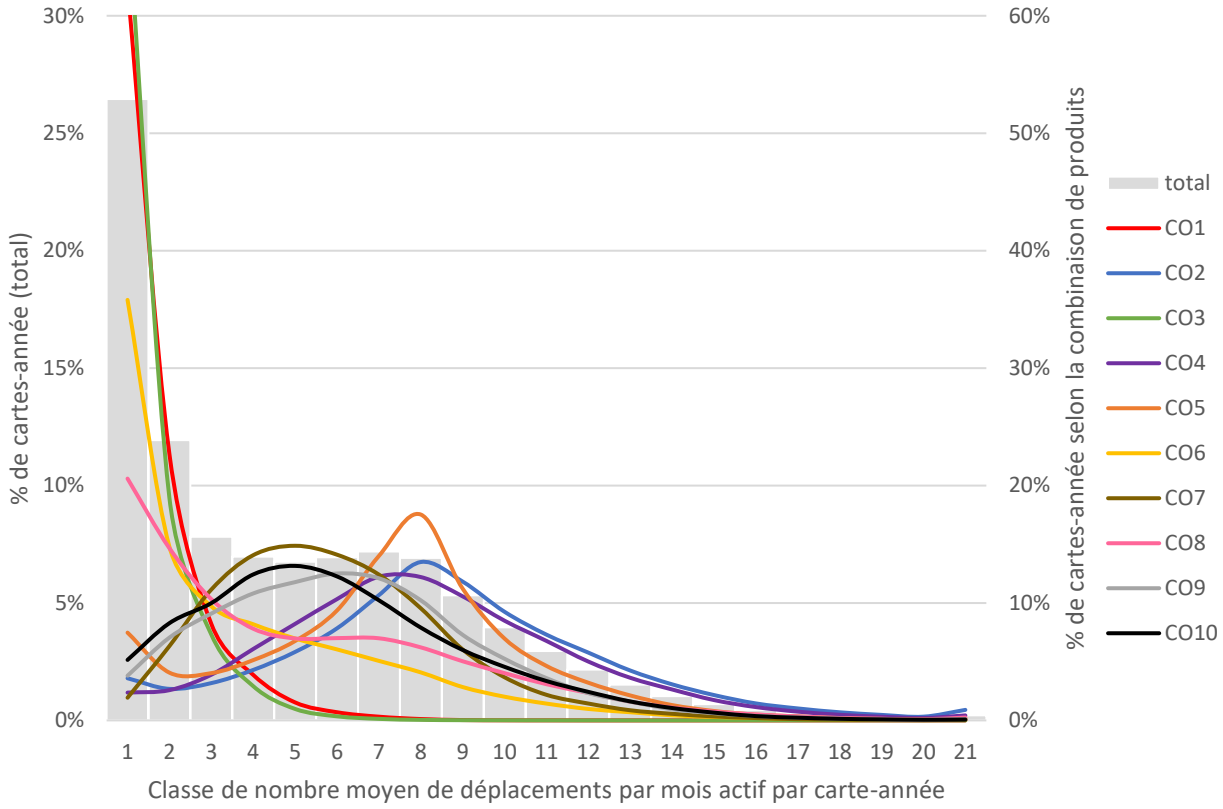


Figure 4.7 Distribution fréquentielle des cartes en fonction du nombre moyen de déplacements par mois actif par carte regroupé en 21 classes (pour le total et par combinaison)

La variabilité temporelle du nombre de déplacements par mois est visible sur la Figure 4.8, en particulier pour les combinaisons 4 et 10. Cette figure montre la variabilité d'un utilisateur moyen de chaque groupe au cours de l'année 2016. Les variances rapportées dans le Tableau 4.3 permettent quant à elles de quantifier cette variabilité en sommant les variations observées dans le comportement de chaque individu puis en calculant une variabilité individuelle moyenne pour chaque combinaison de cartes. D'après les variances totales moyennes TSS obtenues, les utilisateurs les plus variables au niveau temporel sont aussi les plus fréquents. En effet, il existe une forte relation positive entre la classe modale et la valeur de TSS. Le coefficient de corrélation

poly sérial ou *polyserial correlation coefficient* calculé entre la variable ordinale et la variable numérique est égal à 0.61. Cette tendance 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 moins mobiles. Cependant, même si les utilisateurs d'abonnement annuel (combinaison 5) sont plus fréquents que les utilisateurs d'abonnements mensuels (combinaisons 2 et 4), ils sont aussi plus réguliers au niveau temporel puisque leur variance moyenne de 6.13 est inférieure à 9.07 et 7.19 respectivement.

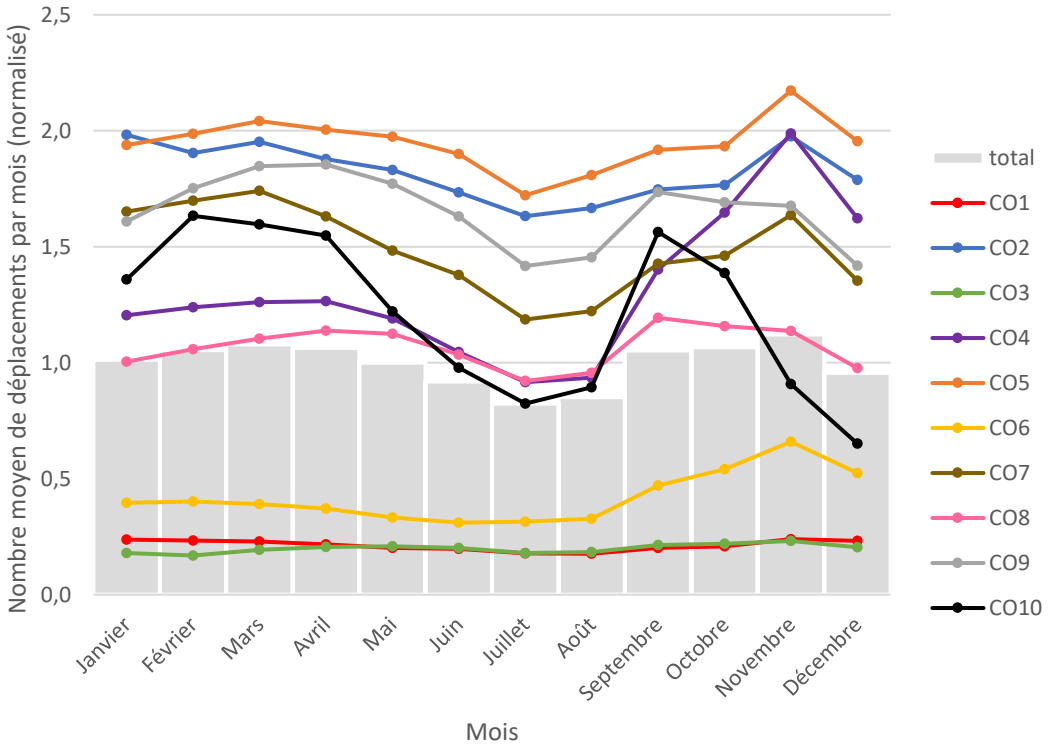


Figure 4.8 Nombre moyen de déplacements par mois par carte pour chaque mois de l'année 2016 (pour le total et par combinaison), normalisé avec le nombre moyen de déplacements par mois tous mois confondus

En outre, la variance interpersonnelle (BPSS) représente 64% de la variance totale mesurée dans cette 5ème combinaison. Ce chiffre élevé révèle que les utilisateurs d'abonnements annuels peuvent différer les uns des autres, mais que le comportement d'un même utilisateur tend à rester assez stable au cours de l'année. L'inverse est vrai pour les utilisateurs d'abonnement mensuel à tarif réduit (combinaison 4) et pour les utilisateurs de plus d'un type de produit ou de tarif (combinaisons 7, 9, 10). Dans ces groupes, la variabilité intrapersonnelle prédomine, indiquant que les

comportements de mobilité individuels changent durant l'année, mais de façon similaire pour tous les utilisateurs. Par exemple, pour la combinaison 7, la Figure 4.4 montrait une diminution significative du nombre de déplacements en été, corrélée à une augmentation de l'utilisation des produits de type carnets. Toutefois, cette variation intrapersonnelle est analogue pour la plupart des utilisateurs du groupe, d'où une plus faible variance interpersonnelle. Dans les autres combinaisons de cartes, les proportions des variances inter et intrapersonnelles sont assez comparables, même si on observe généralement une légère supériorité de la variabilité interpersonnelle (BPSS) par rapport à la variabilité intrapersonnelle (WPSS).

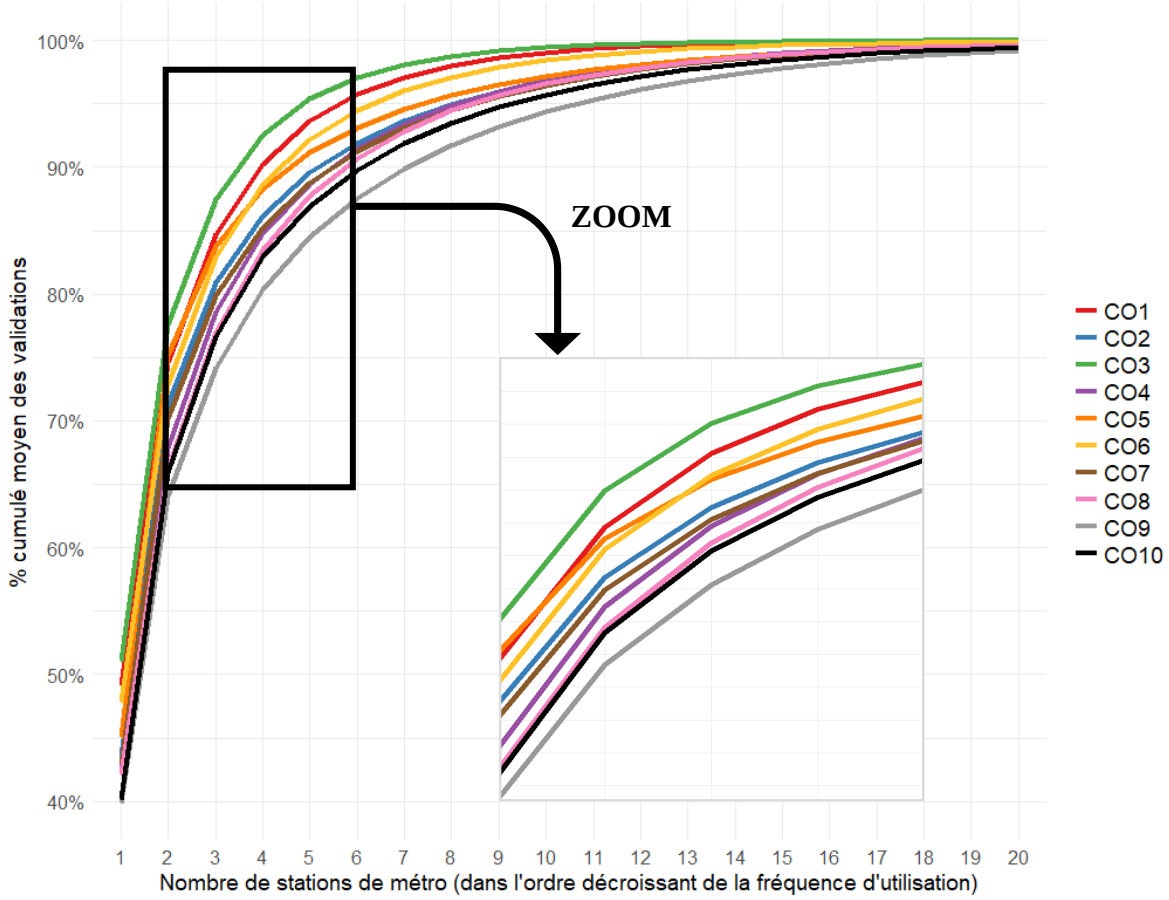


Figure 4.9 Distribution cumulée moyenne des validations dans l'ordre décroissant de la fréquence d'utilisation des stations de métro (par combinaison de cartes)

Enfin, la variabilité spatiale est illustrée (pour le métro uniquement) par la Figure 4.9, sur laquelle est représentée la distribution cumulée moyenne des validations dans l'ordre décroissant de la fréquence d'utilisation des stations de métro visitées par chaque utilisateur de carte. Un point (x, y) de cette distribution se traduit donc par : « les usagers du métro ont fait en moyenne y% de leurs

validations dans leurs x premières stations les plus utilisées pendant l'année ». Ainsi, plus la courbe est à gauche du graphique, plus les validations de l'utilisateur sont concentrées dans un petit nombre de stations de métro et donc plus cet utilisateur est dit régulier au niveau spatial. D'après cette figure, les utilisateurs de carnets (combinaisons 1 et 3) sont les utilisateurs les plus réguliers alors que les utilisateurs avec plus de 3 types de produits ou plus de 2 types de tarifs (combinaisons 9 et 10) sont les plus variables.

Ces observations sont confirmées par l'indicateur d'entropie rapporté dans le Tableau 4.3. Cette entropie est calculée comme la moyenne des entropies individuelles obtenues pour chaque usager du groupe ayant utilisé au moins une fois le métro pendant l'année (c'est-à-dire qu'on ne considère pas les non-utilisateurs du métro pour lesquels l'entropie évaluée serait nulle). Les valeurs les plus basses sont effectivement obtenues pour les combinaisons 1 et 3 alors que les plus élevées sont atteintes pour les combinaisons 9 et 10. D'après la définition donnée précédemment, une entropie plus faible indique un ensemble moins diversifié de stations de métro et donc une plus grande régularité spatiale. La régularité surprenante des utilisateurs de carnets peut être due à leur faible activité: ces usagers ont utilisé moins de stations durant l'année (4.8 stations en moyenne, contre respectivement 12.7 et 11.8 stations pour les utilisateurs d'abonnements annuels et mensuels), conduisant à de nombreuses probabilités égales à 0 et donc à une entropie qui tend également vers 0. De plus, les coefficients de variation calculés pour ces combinaisons sont élevés (environ 50%) et les graphes violon de la Figure 4.10 témoignent également d'une grande hétérogénéité parmi les utilisateurs de ces cartes. Au contraire, pour les combinaisons avec les plus hautes entropies (combinaisons 9 et 10), les valeurs des entropies individuelles sont regroupées autour de la moyenne et les coefficients de variation sont assez faibles (inférieurs à 30%). Ainsi, les usagers de ces deux groupes utilisent assurément plus de stations différentes que dans les autres combinaisons. Par ailleurs, non seulement les utilisateurs d'abonnements annuels (combinaison 5) sont plus réguliers au niveau temporel que les utilisateurs d'abonnements mensuels (combinaisons 2 et 4), mais ils sont aussi plus réguliers au niveau spatial puisque leur entropie moyenne est plus faible (0.34 contre 0.36 et 0.37). Les mêmes conclusions ou presque peuvent être tirées des entropies estimées avec les validations de bus sur les lignes. Toutefois, les valeurs moyennes sont plus basses en raison du plus grand nombre de lignes d'embarquement possibles.

En conclusion, l'application des quatre indicateurs confirme qu'il existe une relation entre les comportements de mobilité et les titres de transport utilisés. En effet, ces indicateurs décèlent

différents niveaux de régularité selon la composition tarifaire des cartes étudiées. Néanmoins, on peut remarquer dans les analyses précédentes que les indicateurs proposés ont tendance à moins bien fonctionner pour les groupes d'usagers les moins observés (par exemple les utilisateurs de carnets), leur faible intensité d'utilisation du transport en commun se répercutant sur l'estimation de leur variabilité spatiale et temporelle. Cette limitation sera discutée plus tard dans le mémoire et des pistes de solution seront avancées. En revanche, lorsqu'ils sont employés pour comparer des groupes d'usagers avec des fréquences d'utilisation similaires, ces indicateurs révèlent notamment que les utilisateurs d'abonnements annuels ont tendance à être plus réguliers au niveau spatial et temporel que les utilisateurs d'abonnements mensuels.

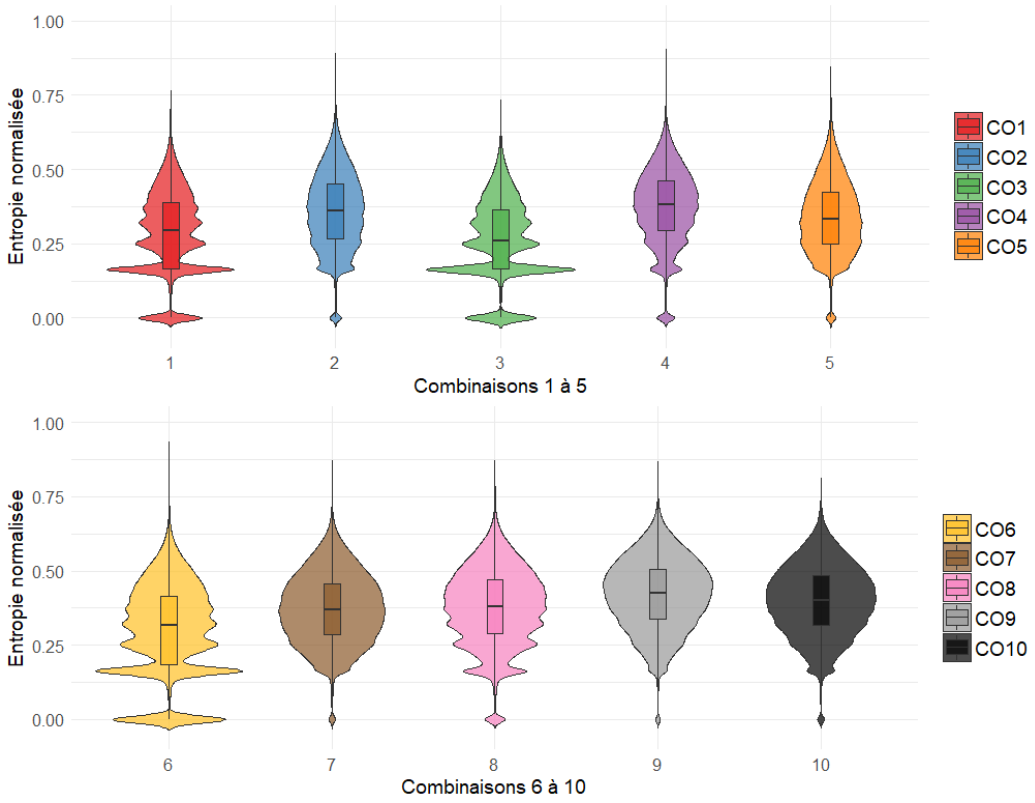


Figure 4.10 Diagramme violon des entropies individuelles pour les validations de métro (par combinaison de cartes)

4.4 Tests statistiques et taille d'effet

Dans la section précédente, des différences de variabilité dans l'utilisation du transport en commun ont été mises en évidence en fonction des titres de transport validés. Il s'agit maintenant d'évaluer ces différences à l'aide de tests statistiques et de la notion de taille d'effet.

4.4.1 Tests statistiques utilisés et résultats

Des tests statistiques bilatéraux sont tout d'abord utilisés afin de vérifier que les différences observées entre les 10 combinaisons de cartes sont significatives pour chaque indicateur. Contrairement aux tests unilatéraux qui examinent une seule direction (de supériorité ou d'infériorité), les tests bilatéraux regardent les deux directions et permettent donc de tester une égalité. Des tests non paramétriques sont choisis, car ceux-ci ne supposent pas une distribution spécifique des données et ils sont moins affectés par les valeurs aberrantes que les tests paramétriques (Cleophas & Zwinderman, 2011). Ils sont donc particulièrement adaptés à la situation puisque les variables étudiées ici ne suivent pas une loi gaussienne et, comme les données n'ont pas été préalablement filtrées, il y a des valeurs aberrantes dans l'échantillon traité. Ces tests sont appliqués sur toutes les paires de combinaisons de cartes possibles pour comparer deux à deux les 10 groupes. Le Tableau 4.4 détaille les tests utilisés pour chaque indicateur et précise la statistique associée à chaque test.

Tableau 4.4 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 différences observées avec le premier indicateur sont examinées avec le test de Kolmogorov-Smirnov. Ce test permet de comparer par paire les distributions cumulées de déplacements représentées sur la Figure 4.6 (qui sont en fait ici des courbes complémentaires de Lorenz) en calculant la distance maximale entre ces courbes. L'hypothèse nulle est que les deux groupes testés suivent la même distribution. Le test du χ^2 (test d'homogénéité des proportions) est appliqué pour comparer deux à deux les distributions de cartes de la Figure 4.7 produites à partir d'une variable catégorique, soit une classe de nombre de déplacements par mois actif (deuxième indicateur). Ce test contrôle si les proportions de cartes dans les mêmes classes sont égales ou, en d'autres termes, si les deux groupes comparés sont homogènes par rapport aux classes. De plus, le test

d'homogénéité des variances de Fligner-Killeen est réalisé pour tester l'égalité des variances totales du nombre de déplacements par mois calculées avec le troisième indicateur. Plus précisément, la procédure de Fligner-Killeen modifiée par Conover et al. (1981) avec une approximation asymptotique de la distribution exacte est utilisée, ce qui permet d'obtenir une statistique Z. Enfin, le test de Wilcoxon-Mann-Whitney est effectué pour comparer les entropies moyennes du quatrième indicateur. Du fait de la grande taille des données, la statistique U suit approximativement une distribution normale et une statistique Z asymptotique peut également être estimée (Adjengue, 2014). De manière plus générale, ce test évalue si les deux distributions comparées ont les mêmes tendances centrales (même centre, même moyenne, même médiane).

Les valeurs-p (*p-values*) obtenues à partir de ces tests statistiques sont données dans le tableau en ANNEXE F. Chaque ligne de ce tableau est un test 'COi v COj' (pour « COi versus COj »), qui correspond au test de la différence entre les cartes des combinaisons i et j. Une valeur-p petite (en général, le niveau de significativité est fixé à 0.05) indique que l'hypothèse nulle peut être rejetée. D'après ces résultats, presque toutes les différences sont significatives ($p < 0.05$), à l'exception des comparaisons 'CO2 v CO10' et 'CO7 v CO8' pour l'indicateur d'entropie calculé avec les validations de bus. Bien que cela puisse paraître une bonne nouvelle à première vue, ces conclusions ne permettent pas d'apprécier dans quelle mesure les différences entre les combinaisons de cartes sont importantes : la divergence de comportement entre un utilisateur d'abonnement annuel et un utilisateur d'abonnement mensuel est-elle la même que celle entre un utilisateur d'abonnement et un utilisateur de carnets ? Les résultats des tests statistiques ne permettent pas de répondre à cette question puisqu'il est impossible de nuancer les différences rapportées (toutes les valeurs-p ou presque sont nulles).

De plus, la taille de l'échantillon est probablement la cause de ces résultats « trop bons ». Ce problème a notamment été souligné par Van der Laan et al. (2010) : « Nous savons que pour des tailles d'échantillon suffisamment grandes, toute étude - y compris celles dans lesquelles l'hypothèse nulle de non effet est vraie - va déclarer un effet statistiquement significatif. » [notre traduction]. Quand on dispose de données massives comme les données de cartes à puce, le test réussit toujours à détecter des différences mêmes minimes et les résultats sont donc constamment significatifs. Pour résoudre ce problème de taille d'échantillon trop grande (ou trop petite dans d'autres cas), des auteurs se sont intéressés à la taille d'effet, notion qui permet de quantifier la taille d'un effet indépendamment de la taille de l'échantillon.

4.4.2 Définition de la taille d'effet

La taille d'effet est une mesure qui permet de quantifier l'ampleur ou l'importance d'un effet, tel que la relation entre une variable indépendante et une variable dépendante ou la différence entre deux populations. Elle est à distinguer de la significativité statistique, qui elle calcule 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:

$$Significativité\ statistique = taille\ d'effet \times\ taille\ d'échantillon \quad \text{(Éq. 16)}$$

Cette relation implique qu'une faible valeur-p peut refléter un effet important, mais aussi un effet moindre dans un grand échantillon (Cooper et al., 2009). Cela signifie également que tout test est statistiquement significatif si la taille de l'échantillon est suffisamment grande (Coe, 2002; Cohen, 1988; Fritz et al., 2012). Au contraire, d'après cette équation, la taille d'effet est une mesure indépendante de la taille de l'échantillon.

Depuis les travaux de Cohen (Cohen, 1969, 1988, 1994), la notion de taille d'effet a été principalement utilisée dans les sciences comportementales, en particulier en médecine et en psychologie, par exemple pour mesurer l'efficacité d'un traitement sur des patients. Une taille d'effet élevée indique que la différence avant versus après traitement est importante. Coe (2002) mentionne ainsi un changement de paradigme: cette mesure permet de passer d'un « *does it work ?* » (est-ce que ça marche) à un « *how well does it work ?* » (à quel point ça marche?). Dans un contexte plus général et statistique, Cohen (1988) définit la taille de l'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.

Par conséquent, la taille d'effet permet une meilleure connaissance et une meilleure interprétation d'un effet dit significatif. C'est pourquoi son calcul a été officiellement encouragé par l'American Psychological Association (American Psychological Association, 1994). Dans ses directives de publication, l'association conseille d'inclure des calculs de taille d'effet aux analyses pour estimer l'ampleur et l'importance de tous les résultats obtenus. Cependant, la taille d'effet reste peu utilisée dans le domaine scientifique, car les chercheurs ont moins d'expérience avec cette notion et ils ne savent souvent pas comment la mesurer ou l'interpréter (Fritz et al., 2012). Pourtant, cette mesure

peut devenir particulièrement intéressante dans l'ère du Big Data. Les bases de données étudiées étant de plus en plus grosses, les tests statistiques vont finir par toujours être significatifs et les chercheurs risquent de surévaluer les effets observés s'ils n'examinent pas les tailles d'effet correspondantes.

4.4.3 Indices de taille d'effet utilisés

Dans ce chapitre, la taille d'effet est explorée pour quantifier l'ampleur des différences observées entre les 10 classes tarifaires construites (par paires) et ce, pour chacun des quatre indicateurs proposés. Ces différences ont été trouvées significatives par les tests statistiques appliqués précédemment, mais la taille d'effet va permettre d'évaluer dans quelle mesure les hypothèses nulles sont fausses, c'est-à-dire à quel point les groupes comparés sont différents, indépendamment de la taille de l'échantillon. Ainsi, la question traitée avant était « est-ce qu'il existe des différences? », alors qu'on tente maintenant de répondre à la question : « quelle est l'importance de ces différences? ». Plus largement encore, la taille d'effet permet ici de quantifier l'influence de l'utilisation d'un certain produit sur les comportements individuels.

Plusieurs indices (paramétriques et non paramétriques) existent pour mesurer cette taille d'effet. Ces indices augmentent avec le degré de discordance entre les deux distributions, variances ou ensembles de proportions comparés. Ils sont basés sur les statistiques des tests appliqués dans la section précédente, mais leur formulation les rend indépendants de la taille de l'échantillon (cette taille est d'ailleurs généralement au dénominateur). Comme les indices paramétriques sont très sensibles à la non-conformité des hypothèses de normalité (Coe, 2002), les équivalents non paramétriques sont utilisés ici.

Pour le premier indicateur, la distance du test de Kolmogorov-Smirnov définie par l'équation suivante est rapportée comme indice de taille d'effet :

$$d = \max_{x} |F_A(x) - F_B(x)| \qquad \text{(Éq. 17)}$$

où A et B sont deux échantillons comparés, leur fonction de répartition étant notée $F(x)$. Cette statistique n'est pas à proprement parler un indice de taille d'effet, mais elle permet de mieux évaluer l'étendue de la différence entre les deux courbes que la valeur-p. La littérature consultée ne rapporte pas de meilleur indice pour le test de Kolmogorov-Smirnov.

Pour la variabilité de la fréquence d'utilisation (deuxième indicateur), la taille d'effet est calculée avec l'indice V de Cramér, défini par l'équation ci-dessous. Selon Cohen (1988), c'est l'indice de taille d'effet le plus général et le plus facile à interpréter dans le cas des tableaux de contingence de fréquences ou de proportions.

$$V = \sqrt{\frac{\chi^2}{N.t}} \qquad \text{(Éq. 18)}$$

où χ^2 est la statistique du test du khi-deux, $N = n_A + n_B$ la taille totale de l'échantillon 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 la situation étudiée, cet indice est égal au coefficient Φ et à l'indice w de Cohen, car il y a $r = 2$ groupes testés à chaque fois et $c = 11$ classes de nombre de déplacements par mois actif avec au moins 5 observations par classe, donc $t = \min(1,10) = 1$. Voir les références (Cohen, 1988) et (Fritz et al., 2012) pour plus de détails.

$$V = \sqrt{\frac{\chi^2}{N}} = \Phi = w \qquad \text{(Éq. 19)}$$

Pour les deux indicateurs restants (mesurant les variabilités temporelle et spatiale), le coefficient de corrélation r énoncé par Cooper et Hedges (1994) et utilisé par Pallant (2007) est estimé à partir de la statistique Z comme suit :

$$r = \sqrt{\frac{Z^2}{N}} = \frac{|Z|}{\sqrt{N}} \qquad \text{(Éq. 20)}$$

avec $N = n_A + n_B$ la taille totale de l'échantillon. (Note : tous les indicateurs de taille d'effet présentés précédemment sont définis avec un signe positif, car l'ordre attribué aux deux conditions testées n'a pas d'importance ici.)

Pour améliorer sa compréhension de la taille de l'effet, le lecteur peut également penser en termes de Φ^2 et r^2. La valeur au carré de ces indices est interprétée comme la proportion de la variance totale de l'indicateur mesurée dans la population combinée des groupes A et B expliquée par l'appartenance à un des deux groupes. En d'autres termes, appartenir à une combinaison de cartes plutôt qu'à une autre, c.-à-d. utiliser un produit plutôt qu'un autre, représente $\Phi^2\%$ ou $r^2\%$ de la variance totale de l'indicateur dans les deux groupes combinés. Ce point de vue justifie bien le fait

que la taille d'effet permet de mesurer la relation entre les produits utilisés et la variabilité des comportements de mobilité.

Par ailleurs, il est pertinent de souligner ici que les tailles d'effet exprimées ci-dessus sont en fait des tailles d'effet *estimées*: elles décrivent l'échantillon, mais elles estiment toute la population. Comme c'est le cas pour beaucoup d'autres statistiques descriptives (par exemple, les moyennes sont en général déclarées avec des écarts-types ou des coefficients de variation), une statistique de variabilité est associée à la mesure de la taille d'effet. La statistique recommandée est un intervalle de confiance (Fritz et al., 2012). Cependant, la variabilité des résultats diminue et la précision de la taille d'effet augmente avec la taille de l'échantillon. Son calcul ne présente donc aucun intérêt pour l'étude réalisée dans ce chapitre (des intervalles de confiance très étroits sont obtenus du fait des gros échantillons testés).

4.4.4 Application et interprétation des indices de taille d'effet

Les résultats de l'application des indices de taille d'effet énoncés dans la section précédente sont donnés dans le Tableau 4.5. Des couleurs ont été ajoutées pour faciliter leur interprétation. Des règles de couleurs indépendantes sont appliquées au premier indicateur (mesure de la dispersion des déplacements) puisque la statistique d n'est pas un vrai indice de taille d'effet. Le gradient de couleurs rouge-blanc-bleu est ainsi administré dans l'ordre décroissant des valeurs. Pour les autres indicateurs, leur interprétation peut être faite avec le critère de Cohen défini dans le Tableau 4.6. Dans son livre, Cohen (1988) propose les adjectifs qualitatifs « petit », « moyen » et « grand » auxquels il associe une valeur représentative pour désigner une taille d'effet peu, moyennement ou très importante. Cependant, cette convention est arbitraire et les valeurs assignées à chaque adjectif ont été déterminées sur la base de résultats expérimentaux dans la recherche comportementale et biologique. Cohen avertit le lecteur que cette échelle de trois points peut différer selon le domaine de recherche et qu'elle doit donc être adaptée au contexte, par exemple en prenant des résultats préalables dans le domaine comme base de comparaison. Quand aucune référence antérieure n'est disponible, ce qui est le cas en transport, cette convention est quand même recommandée. Ce critère a donc été utilisé pour organiser les couleurs des quatre dernières colonnes du Tableau 4.5. Néanmoins, pour s'affranchir du critère de Cohen, et comme plusieurs valeurs de taille d'effet ont été calculées, elles peuvent aussi être comparées entre elles pour déterminer quelles différences sont plus importantes par rapport à d'autres (comparaisons relatives).

Tableau 4.5 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
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 4.6 Critère de Cohen

Valeur	Taille d'effet
0.1	petite
0.3	moyenne
0.5	grande

Dans l'ensemble, les nombreuses cases vertes du Tableau 4.5 (signalant des petites tailles d'effet) révèlent que beaucoup des différences trouvées significatives dans le tableau des valeurs-p en ANNEXE F ne sont en fait pas si importantes. Ce résultat confirme que la significativité précédemment obtenue était principalement due à la grosse taille des échantillons comparés. Pour commencer, ni les combinaisons 1 et 3 ('CO1 v CO3', c.-à-d. carnets à tarif ordinaire versus carnets à tarif réduit) ni les combinaisons 2 et 4 ('CO2 v CO4', c.-à-d. abonnements mensuels à tarif ordinaire versus abonnements mensuels à tarif réduit) ne sont vraiment différentes. En effet, pour ces deux paires de combinaisons, la distance de Kolmogorov-Smirnov et les indices de taille d'effet sont faibles. On en conclut donc que le type de tarif utilisé (ordinaire versus réduit) a peu d'effet sur la variabilité de l'utilisation du transport en commun. De plus, les tests 'CO2 v CO5' et 'CO4 v CO5' montrent que les indicateurs de variabilité calculés pour la combinaison 5 (utilisateurs d'abonnements annuels) ne sont pas très éloignés de ceux des combinaisons 2 et 4 (utilisateurs d'abonnements mensuels): d'après ces résultats, ces usagers ont donc des niveaux de régularité semblables. Les différences entre la combinaison 6 (utilisateurs d'un seul type de produit et de tarif autres que ceux des combinaisons les plus populaires) et les combinaisons 1 et 3 (utilisateurs de carnets) ne sont pas signifiantes non plus, sauf pour l'indicateur de fréquence d'utilisation. Ce constat peut être dû à la grande proportion d'utilisateurs d'abonnements quotidiens et de billets unitaires dans la combinaison 6, ce qui les rend similaires aux combinaisons 1 et 3 (utilisateurs de carnets). En outre, la ressemblance entre les combinaisons 7, 8, 9 et 10 (utilisation de plusieurs types de produits ou de tarifs) est attestée avec les faibles valeurs de tailles d'effet rapportées dans la dernière partie du tableau. Inversement, les grandes distances et tailles d'effet mesurées entre les utilisateurs d'abonnements mensuels ou annuels (combinaisons 2, 4, 5) et les utilisateurs de carnets (combinaisons 1 et 3), notamment pour le second indicateur (fréquence d'utilisation), soulignent de grandes différences dans le comportement de ces usagers. Ces résultats sont plutôt intuitifs, mais la taille d'effet, contrairement aux valeurs-p, a permis de les quantifier de manière plus précise.

CHAPITRE 5 ANALYSE DE LA VARIABILITÉ INTERPERSONNELLE: CRÉATION D'UNE TYPOLOGIE D'USAGERS

L'objectif de ce cinquième chapitre est d'analyser plus particulièrement la variabilité interpersonnelle de l'utilisation du transport en commun. Pour ce faire, une typologie d'usagers (en réalité, une typologie de cartes) est créée afin de séparer les utilisateurs de cartes à puce en plus petits groupes. Par définition, les individus d'un même groupe présentent des comportements similaires, mais les individus de deux groupes distincts ont des comportements plus éloignés. Cette typologie permet ainsi de mettre en évidence des différences entre les usagers, chaque groupe étant caractérisé par ses propres habitudes de mobilité. Une segmentation globale de tous les usagers de la STM est d'abord réalisée, puis un bloc de cartes particulier est étudié : celui des utilisateurs d'abonnements annuels avec une amplitude de 12 mois.

5.1 Méthode de segmentation utilisée

Dans les deux cas, la même méthode de segmentation est utilisée. Cette segmentation est basée sur l'observation de comportements individuels d'utilisation à travers une longue période d'un an. À cet effet, la mobilité de chaque utilisateur de carte à puce a été précédemment résumée par mois dans un vecteur carte-année (voir la section 3.3.2 pour rappel). La méthode employée consiste donc à segmenter ces vecteurs cartes-années (déjà normalisés) afin d'obtenir une typologie des propriétaires de ces cartes en fonction des caractéristiques de leur utilisation individuelle mensuelle du transport en commun pendant l'année 2016.

Pour cela, l'algorithme de segmentation appliqué ici est celui des K-moyennes. Cet algorithme a été choisi car il est particulièrement bien adapté aux gros ensembles de données (He et al., 2018). Son fonctionnement est expliqué en détail par Steinley (2006) ou James et al. (2013). En quelques mots, l'approche par K-moyennes permet de partitionner des vecteurs en K groupes distincts en minimisant la variabilité à l'intérieur de chaque groupe (*within-cluster variation* ou *within-cluster sum of squares*). Cette variabilité est exprimée par la somme des distances au carré entre les observations à l'intérieur de chaque groupe et le centre de ce groupe. Lorsque la métrique euclidienne est utilisée, le problème d'optimisation posé est énoncé comme suit :

Problème d'optimisation des K-moyennes

Soient N observations prenant la forme de vecteurs à P dimensions regroupés dans une matrice $X_{N\times P} = \{x_{ij}\}_{N\times P}$ *. Si on cherche à segmenter ces observations en K groupes notés* $C_1, \dots, C_K$ *de tailles* $n_1, \dots, n_K$ *telles que* $n_1 + \dots + n_K = N$*, la fonction à minimiser est la suivante :*

$$\min_{C_1,\dots,C_K} \sum_{k=1}^{K} \sum_{i\in C_k} \sum_{j=1}^{P} \left(x_{ij} - \bar{x}_{kj}\right)^2 \qquad \text{(Éq. 21)}$$

avec $\bar{x}_{kj}$ *la moyenne de la dimension j dans le groupe* C_k *telle que :*

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i\in C_k} x_{ij} \qquad \text{(Éq. 22)}$$

où $i \in C_k$ *désigne la* $i^{ème}$ *observation dans le* $k^{ème}$ *groupe.*

Dans le cas étudié, N est égal au nombre de vecteurs cartes-année dans l'échantillon analysé et $P = 13$ dimensions par vecteur. Pour résoudre ce problème d'optimisation sans avoir à calculer les K^N partitions possibles, l'algorithme suivant (appelée algorithme de Lloyd) est appliqué :

Algorithme des K-moyennes

1. <u>Initialisation</u> *Partitionner aléatoirement les observations en K groupes et déterminer le centre de chaque groupe. Ce centre est calculé comme le barycentre du groupe (moyenne de tous les vecteurs appartenant au groupe).*
2. <u>Itération</u> *Répéter les 2 étapes suivantes jusqu'à convergence de l'algorithme (c.-à-d. jusqu'à ce que les groupes obtenus soient stables) :*

   (a) <u>Affectation</u>

   Réassigner chaque observation au groupe dont le centre est le plus proche (où la notion de proximité dépend de la métrique utilisée).

   (b) <u>Actualisation</u>

   Calculer le nouveau centre de chacun des groupes obtenus à l'étape précédente.

Cet algorithme est mis en pratique avec le logiciel R (fonction *kmeans*) et la métrique choisie est la distance euclidienne. Pour éviter d'obtenir un optimum local au lieu d'un optimum global,

important problème à considérer dans l'application de l'algorithme des K-moyennes (Steinley, 2006), dix initialisations aléatoires différentes sont réalisées.

De plus, pour déterminer la valeur optimale du paramètre K (c.-à-d. le nombre de groupes dans la partition), deux critères sont utilisés. Premièrement, la proportion de la variance expliquée par la partition est tracée en fonction du nombre de groupes : la valeur de K est alors déterminée par le « critère du coude » (*elbow criterion*), c'est-à-dire au point où un angle se forme dans le graphique. Cette règle empirique permet de choisir un nombre de groupes K tel que l'addition d'un nouveau groupe $K + 1$ n'apporte pas suffisamment d'informations supplémentaires pour être considéré. En effet, les premiers groupes identifiés fournissent beaucoup d'informations, mais ce gain ralentit à partir d'un certain nombre de groupes, d'où l'apparition d'un angle dans le graphique (Madhulatha, 2012). Le deuxième critère utilisé est basé sur la méthode en deux étapes appliquée par Morency et al. (2017). Cette méthode consiste à produire 30 groupes avec l'algorithme des K-moyennes pour réduire la dimension des données, puis à dessiner un dendrogramme à partir des 30 centres obtenus en utilisant un algorithme hiérarchique agglomératif (voir section 7.1.3). La valeur de K est ensuite fixée en coupant horizontalement le dendrogramme produit à la hauteur souhaitée.

Par ailleurs, plusieurs typologies sont possibles selon les filtres appliqués en amont du processus de segmentation. Différentes combinaisons envisageables avec les données de cartes à puce sont schématisées dans la Figure 5.1. Les données de cartes à puce contenant trois types d'informations (des informations temporelles, spatiales/modales et tarifaires), trois niveaux de distinction sont représentés (une couleur par niveau). Ainsi, des typologies différentes peuvent être réalisées selon si on sélectionne des déplacements faits sur certains types de jours (ouvrables versus non ouvrables). De même, une sélection peut être faite sur les modes empruntés, en prenant les déplacements effectués avec un seul mode (métro uniquement, bus uniquement) ou deux modes (métro + bus combinés). À une granularité encore plus fine, des typologies spécifiques peuvent être faites pour certains types de stations de métro ou lignes de bus. Enfin, une typologie peut être créée pour chaque combinaison de produits utilisés et cette typologie peut également être spécifiée en fonction de l'amplitude d'utilisation des produits (où l'amplitude est définie comme la durée entre le dernier mois observé et le premier mois observé du produit pendant l'année, sans impliquer toutefois que le produit a été utilisé durant tous les mois intermédiaires). Pour les exemples proposés sur le schéma de la Figure 5.1, cela revient à 10290 segmentations potentielles.

Parmi cet immense champ des possibles, seuls deux « chemins » sont empruntés dans ce mémoire : ils sont représentés en rouge et en violet sur le schéma. Le premier correspond à la section 5.2 pour laquelle aucun filtre n'est appliqué, c'est-à-dire que tous les déplacements pour tous les types de titres utilisés sur n'importe quelle durée sont considérés. Le deuxième cas d'étude est traité dans la section 5.3. Dans cette section, un filtre est appliqué sur les informations tarifaires afin de sélectionner les utilisateurs d'abonnements longs (annuels ou 4 mois) avec une amplitude de 12 mois. Ce groupe de cartes est en réalité composé majoritairement d'utilisateurs d'abonnements annuels (on comptabilise moins de 1% d'abonnements de 4 mois à tarif réduit), c'est pourquoi on parlera uniquement d'« utilisateurs d'abonnements annuels avec une amplitude de 12 mois » dans la suite de ce chapitre.

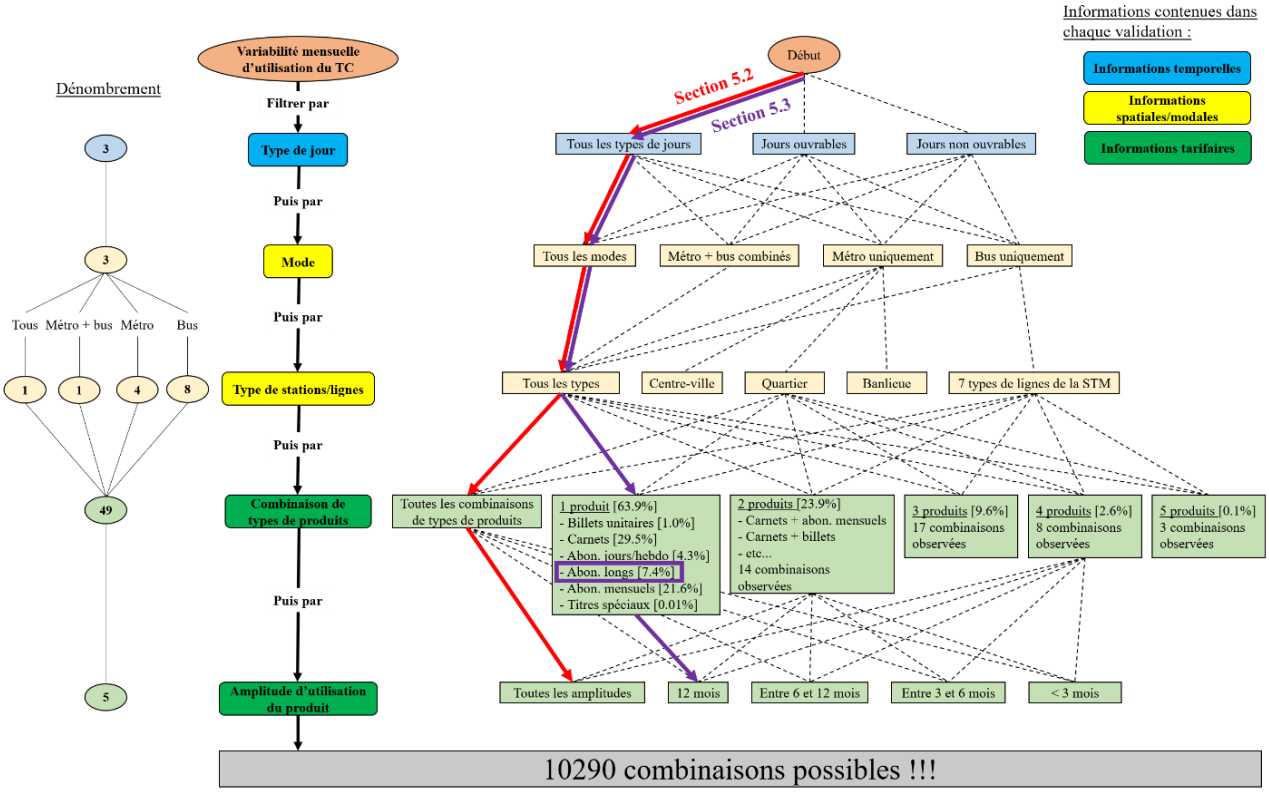


Figure 5.1 Exemples de segmentations possibles avec les données de cartes à puce

5.2 Cas de tous les usagers

Le processus de segmentation expliqué précédemment est d'abord appliqué sur toute la base « cartes-année » au complet, soit sur environ 2 millions de vecteurs. La typologie ainsi développée

comprend tous les usagers d'une carte OPUS s'étant déplacés sur le réseau de la STM au moins une fois en 2016. On s'attend donc à observer deux extrêmes : des usagers très réguliers présents toute l'année sur le réseau et au contraire des usagers très occasionnels apparaissant sur des intervalles de temps plus ou moins courts.

Il s'agit ici de donner un portrait global de la mobilité des utilisateurs du transport en commun afin de permettre une meilleure connaissance de la clientèle de la STM. Cet exercice est particulièrement intéressant pour les grandes villes comme Montréal : les nombreuses opportunités offertes par le réseau de transport de telles villes entraînent généralement une très grande hétérogénéité dans les comportements observés (Goulet-Langlois et al., 2016).

5.2.1 Typologie obtenue

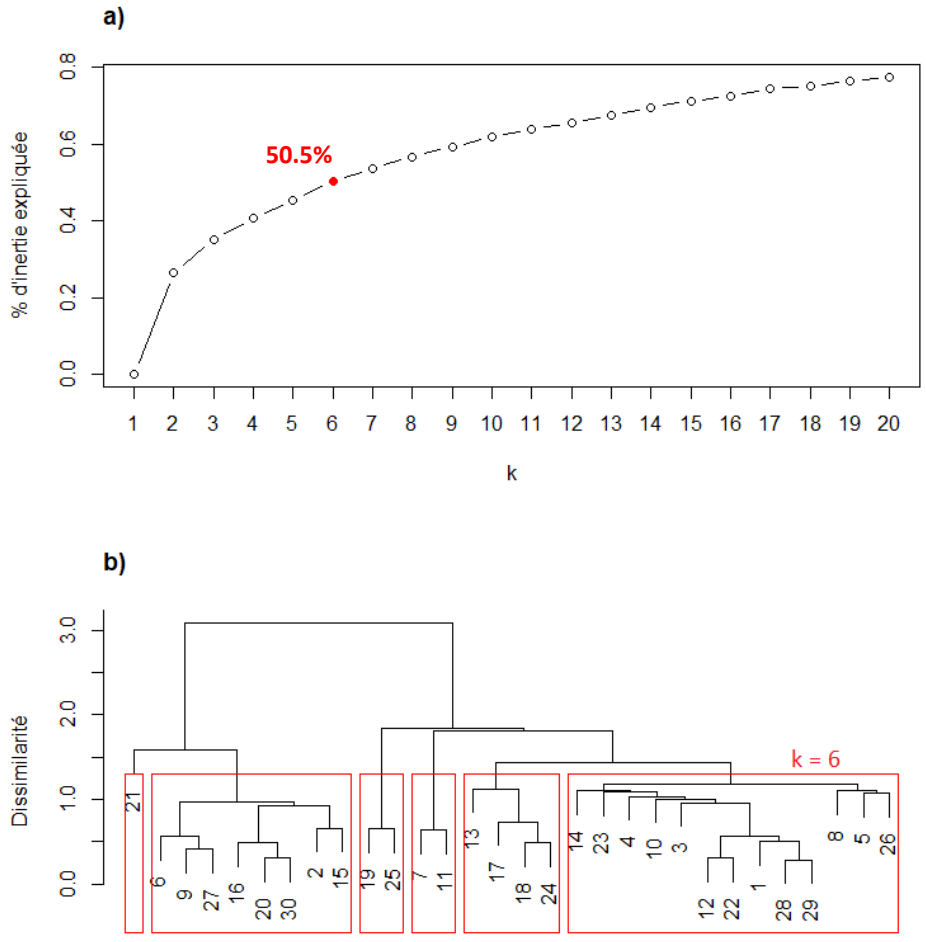


Figure 5.2 Choix du nombre de groupes K avec a) le pourcentage d'inertie expliquée b) un dendrogramme réalisé sur 30 groupes initiaux – Cas de tous les usagers

Tout d'abord, le nombre de groupes K dans la typologie est déterminé à l'aide des deux critères précédemment énoncés. Les résultats de l'application de ces deux critères sont représentés sur la Figure 5.2. La Figure 5.2 a) montre qu'à partir de $K = 3$ groupes, l'ajout d'un segment supplémentaire n'augmente pas significativement la part d'inertie expliquée par la partition : c'est donc le nombre minimum de groupes à sélectionner. Cependant, ce nombre est trop faible pour pouvoir prétendre que la typologie obtenue sera exhaustive. Le dendrogramme de la Figure 5.2 b) permet néanmoins de compléter la recherche du nombre optimal de segments à considérer. En partant du haut du graphique, les 30 centres obtenus en amont avec l'algorithme des K-moyennes se séparent d'abord en deux branches en fonction de leur intensité mensuelle d'utilisation: les cartes fortement utilisées (intensité mensuelle moyenne normalisée = 1.12) sont rassemblées à gauche du dendrogramme et les cartes faiblement utilisées (intensité mensuelle moyenne normalisée = 0.26) sur la droite. Le paramètre $K = 6$ est choisi afin de diviser le groupe à faible usage en plusieurs sous-groupes. Ce chiffre procure une bonne granularité pour l'interprétation des résultats (le nombre de groupes analysés ni n'est trop faible ni trop élevé) et il permet également d'obtenir des groupes de tailles relativement homogènes.

Tableau 5.1 Taille et centre de chaque groupe (distribution moyenne des déplacements annuels par mois et intensité mensuelle moyenne normalisée) – Cas de tous les usagers

Groupe		**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**Total**
Taille (% cartes-année)		12.0%	26.6%	12.4%	5.4%	5.1%	38.6%	100%
Distribution des déplacements annuels par mois	**Janvier**	6.9%	9.6%	0.2%	0.3%	80.5%	5.7%	9.7%
	Février	6.6%	11.4%	0.1%	0.3%	11.0%	9.1%	7.9%
	Mars	6.7%	10.9%	0.1%	0.3%	2.7%	9.4%	7.5%
	Avril	6.6%	10.1%	0.2%	0.3%	1.4%	9.2%	7.1%
	Mai	6.4%	8.8%	0.2%	0.3%	0.9%	9.0%	6.7%
	Juin	6.1%	7.6%	0.3%	0.4%	0.7%	9.4%	6.4%
	Juillet	6.2%	6.6%	0.3%	0.4%	0.6%	9.6%	6.3%
	Août	6.5%	7.0%	0.6%	0.5%	0.5%	9.4%	6.4%
	Septembre	8.4%	9.6%	3.8%	0.9%	0.4%	10.0%	8.0%
	Octobre	10.7%	8.2%	14.1%	2.4%	0.4%	10.3%	9.3%
	Novembre	14.7%	5.8%	55.0%	6.8%	0.4%	4.7%	12.3%
	Décembre	14.0%	4.5%	25.0%	87.1%	0.5%	4.2%	12.3%
Moyenne mensuelle normalisée		1.31	0.73	0.42	0.26	0.26	0.15	0.49

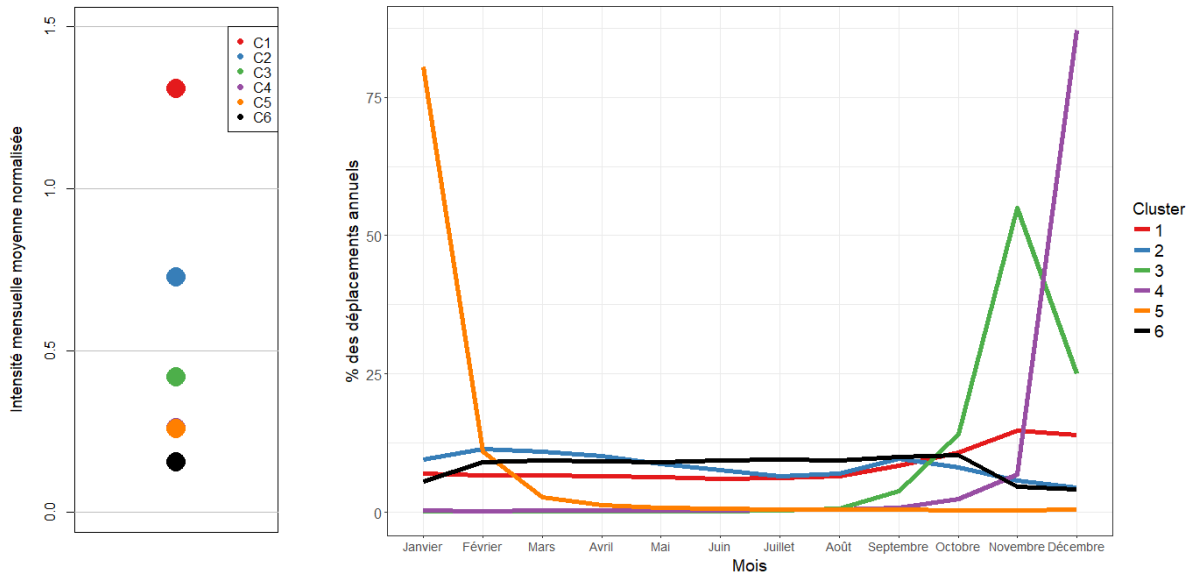


Figure 5.3 Représentation des 6 centres : intensité mensuelle normalisée (à gauche) et distribution des déplacements annuels par mois (à droite) – Cas de tous les usagers

La taille (en pourcentage de cartes-années) et les centres des six groupes ensuite produits avec l'algorithme des K-moyennes sont donnés dans le Tableau 5.1 et illustrés sur la Figure 5.3. Chaque centre correspond à la moyenne des vecteurs appartenant au groupe. Étant donnée la composition des vecteurs cartes-année manipulés (voir section 3.3.2), ce calcul fournit une distribution moyenne des déplacements annuels par mois ainsi qu'une intensité mensuelle moyenne normalisée (c.-à-d. un nombre moyen normalisé de déplacements par mois actif). Les six groupes sont classés en ordre décroissant de cette intensité d'utilisation. Le $i^{\text{ème}}$ groupe ou *cluster* est noté C*i*, et une couleur est attribuée à chaque groupe.

D'après les résultats obtenus, les groupes plus fréquents (C1 et C2) sont aussi les plus réguliers. En effet, sur la Figure 5.3 les déplacements réalisés avec les cartes appartenant à ces groupes sont équitablement distribués au cours des mois, avec une faible augmentation (pour le groupe C1) ou diminution (pour le groupe C2) de l'utilisation du transport en commun en fin d'année. Ces deux groupes probablement captifs du transport en commun se différencient surtout par leur intensité mensuelle moyenne d'utilisation : les usagers du groupe C1 font en moyenne plus de déplacements par mois actif que ceux du groupe C2. De plus, une légère diminution des déplacements est visible en été (juin-juillet-août) pour le groupe C2, ce qui n'est pas le cas pour C1. Les cartes des groupes C3 et C4 sont essentiellement utilisées en fin d'année, d'août à septembre pour les cartes du groupe C2, avec un pic de déplacements en novembre, et principalement en décembre pour les cartes du

groupe C4. Les usagers du groupe C3 sont possiblement de nouveaux clients, par exemple des immigrés arrivés durant l'été, ou des usagers ayant renouvelé leur carte, notamment des étudiants. En effet, les cartes à tarif réduit étudiant (18-25 ans) doivent être renouvelées tous les ans d'après la politique tarifaire de la STM. De plus, le mois de septembre est généralement un mois de promotions, encourageant de nombreux usagers ordinaires à changer eux aussi leur carte. Au contraire, le groupe C5 représente des départs (ou du moins des arrêts dans l'utilisation de la carte après mars-avril). Ce groupe est l'antipode du groupe C4, la majorité des déplacements effectués par ce groupe étant concentrée dans le mois de janvier. Les tailles (environ 5% des cartes-année) et les fréquences d'utilisation similaires de ces deux groupes signalent peut-être une relation entre leurs membres respectifs. Les usagers du groupe C5 pourraient être des personnes, par exemple des cyclistes, qui n'utilisent pas leur carte durant l'été et qui en rachètent une autre en septembre pour reprendre le transport en commun pendant l'hiver : ces usagers réapparaîtraient alors dans le cluster C4 avec une nouvelle carte. Cette possible jonction entre les deux groupes rappelle la principale limite de ce projet : des cartes et non des usagers sont étudiées, ce qui implique que si un usager change de carte au cours de l'année, il peut se retrouver dans deux groupes différents. Enfin, le groupe C6 est le groupe à plus faible usage du transport en commun, mais aussi le groupe le plus volumineux (38.6% des cartes-année). Son profil moyen mensuel est similaire à celui des groupes C1 et C2 de février à octobre mais, à l'inverse des groupes C4 et C5, une diminution de l'utilisation du transport en commun est observée en début et en fin d'année. Les usagers appartenant à ce groupe sont donc des utilisateurs occasionnels de leur carte, mais ils sont présents sur le réseau presque toute l'année : ce sont des usagers « occasionnels-réguliers ». Ainsi, les comportements observés dans ce groupe C6 peuvent être associés à des usagers qui utilisent de temps à autre le transport collectif pour un motif loisir. Par ailleurs, en regardant de plus près les cartes appartenant à ce groupe, on constate que les comportements touristiques (cartes présentes sur le réseau seulement en été) sont également inclus dans ce groupe.

Certaines des hypothèses d'interprétation faites précédemment peuvent être supportées par la composition tarifaire des cartes appartenant à chaque groupe. Les tableaux ci-dessous présentent deux distributions de déplacements (une distribution horizontale et une distribution verticale) en fonction des titres de transport validés pour les réaliser et de l'appartenance des cartes utilisées à chacun des six groupes. Le Tableau 5.2 distribue ainsi les déplacements de chaque groupe par type de produits et de tarifs utilisés alors que, inversement, le Tableau 5.3 distribue les déplacements

par type de produits et de tarifs dans chacun des six groupes. Le nombre moyen de titres de transport différents utilisés par carte durant l'année est également calculé dans chaque groupe. Pour cela, les titres réels sont considérés (et non les 30 combinaisons type de produit × type de tarif).

Tableau 5.2 Distribution des déplacements de chaque groupe par type de produits et de tarifs – Cas de tous les usagers

	Groupe	C1	C2	C3	C4	C5	C6	Tous
	Nb moyen de titres	1.67	2.09	1.55	1.34	1.41	1.68	1.74
TYPES DE PRODUITS	**Billets unitaires**	0.1%	0.2%	0.4%	1.0%	0.9%	1.1%	0.3%
	Carnets	0.6%	6.8%	16.6%	30.6%	25.4%	61.3%	11.7%
	Abon. jours/hebdo	4.9%	4.9%	5.1%	14.5%	11.7%	6.4%	5.2%
	Abon. longs	14.1%	19.4%	9.3%	11.7%	8.7%	6.9%	15.5%
	Abon. mensuels	80.2%	68.7%	68.7%	42.1%	53.2%	24.3%	67.3%
	Titres spéciaux	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
TYPES DE TARIFS	**Gratuité**	0.7%	0.7%	1.5%	4.9%	0.6%	1.7%	0.9%
	Tarif étudiant	1.3%	3.3%	10.1%	1.6%	2.8%	2.8%	2.9%
	Tarif réduit	30.6%	28.5%	47.4%	21.8%	20.2%	30.9%	30.3%
	Tarif ordinaire	67.5%	67.5%	41.0%	71.7%	76.5%	64.7%	66.0%
	Tarif spécial	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Tableau 5.3 Distribution des déplacements par type de produits et de tarifs dans chaque groupe – Cas de tous les usagers

	Groupe	C1	C2	C3	C4	C5	C6	Total
	Nb moyen de titres	1.67	2.09	1.55	1.34	1.41	1.68	1.74
TYPES DE PRODUITS	**Billets unitaires**	12.5%	32.6%	7.3%	2.3%	2.5%	42.8%	100.0%
	Carnets	1.8%	27.9%	6.8%	1.7%	1.7%	60.0%	100.0%
	Abon. jours/hebdo	32.6%	45.0%	4.7%	1.8%	1.8%	14.0%	100.0%
	Abon. longs	31.5%	59.6%	2.9%	0.5%	0.4%	5.1%	100.0%
	Abon. mensuels	41.2%	48.7%	4.9%	0.4%	0.6%	4.1%	100.0%
	Titres spéciaux	25.0%	33.0%	1.0%	0.3%	24.8%	15.8%	100.0%
TYPES DE TARIFS	**Gratuité**	27.1%	37.7%	8.6%	3.8%	0.5%	22.3%	100.0%
	Tarif étudiant	15.3%	55.7%	16.9%	0.4%	0.8%	10.9%	100.0%
	Tarif réduit	34.9%	44.9%	7.5%	0.5%	0.5%	11.6%	100.0%
	Tarif ordinaire	35.4%	48.8%	3.0%	0.7%	0.9%	11.2%	100.0%
	Tarif spécial	42.3%	48.9%	2.4%	0.2%	0.0%	6.2%	100.0%

La majorité des déplacements réalisés par les trois premiers groupes à plus fort taux d'utilisation du transport en commun (C1, C2, C3) sont faits avec des abonnements mensuels ou longs. Dans les autres groupes (C4, C5, C6), les proportions de déplacements effectués avec des produits de type carnets sont plus élevées, notamment pour le groupe C6, composé d'usagers occasionnels. Cette tendance est encore plus frappante si on regarde la distribution inverse (distribution des déplacements par type de produits dans chacun des six groupes) : 42.8% des déplacements avec billets unitaires et 60% des déplacements avec carnets sont réalisés par les membres du groupe C6. Au contraire, les déplacements faits avec des abonnements (quotidiens, hebdomadaires, mensuels, annuels) sont principalement réunis dans les deux premiers groupes, C1et C2. Au niveau des tarifs, on constate une grande prédominance du tarif ordinaire pour tous les groupes. Néanmoins, les plus hautes proportions de déplacements réalisés avec des tarifs étudiants ou réduits sont atteintes dans le groupe C3, confirmant ainsi que les cartes de ce groupe appartiennent en grande partie à des étudiants (d'où leur apparition en septembre). Par ailleurs, le nombre moyen de titres de transport différents utilisés par carte est plus élevé pour les groupes les plus fréquents (C1, C2), mais aussi pour le groupe le plus occasionnel (C6), témoignant ainsi d'une plus grande diversité dans les achats tarifaires de ces groupes. Dans le cas du groupe C2, cette tendance peut être expliquée par l'utilisation combinée d'abonnements mensuels et de carnets en été : en effet, 27.9% des déplacements avec carnets ont été réalisés par ce groupe.

5.2.2 Indicateurs d'utilisation et de variabilité

Pour aider à l'interprétation de la typologie obtenue et pour mieux connaître les caractéristiques de mobilité de chaque groupe, différents indicateurs sont proposés dans cette sous-section. Tout d'abord, des indicateurs temporels et spatiaux (ou modaux) décrivant l'utilisation du transport en commun sont examinés. Ces indicateurs sont des distributions de cartes ou de déplacements, ainsi que des moyennes de variables individuelles calculées à l'intérieur de chaque groupe. Ensuite, les indicateurs de variabilité du Chapitre 4 sont appliqués à la typologie de cette section.

5.2.2.1 Indicateurs temporels d'utilisation du TC

Différents indicateurs peuvent être évalués pour décrire l'utilisation du transport en commun dans le temps. Ces indicateurs peuvent notamment servir à quantifier l'intensité d'usage, la longueur de la période d'activité observée pour chaque carte, ou encore la concentration des déplacements sur

certains types de jours (ouvrables versus non ouvrables). Ainsi, le Tableau 5.4 rassemble les indicateurs temporels suivants, définis dans leur ordre d'apparition :

- La **répartition des déplacements** totaux de 2016 dans chacun des six groupes de cartes (en pourcentage de déplacements).
- Le **nombre moyen de déplacements par an** par carte, soit le nombre total moyen de déplacements réalisés par carte en 2016. Du fait de sa confidentialité, cet indicateur est normalisé en divisant la moyenne obtenue dans chaque groupe par la moyenne calculée avec l'ensemble des cartes (moyenne générale). Un nombre supérieur à 1 indique donc un nombre moyen de déplacements supérieur à la moyenne générale alors qu'un nombre inférieur à 1 signale un nombre moyen de déplacements inférieur à la moyenne générale.
- Le **nombre moyen de déplacements par mois actif** par carte. C'est l'indicateur d'intensité qui a été utilisé dans les vecteurs « cartes-année » : il s'agit donc d'une donnée entrante dans le processus de segmentation. Cet indicateur est également confidentiel et est donc normalisé comme expliqué précédemment.
- Le **nombre moyen de mois observés** par carte (ou nombre moyen de mois d'activité par carte), c'est-à-dire le nombre moyen de mois en 2016 où la carte a été validée au moins une fois. Ces mois peuvent être discontinus.
- La **proportion de cartes actives chaque mois** de l'année par rapport au nombre total de cartes dans le groupe considéré. Une carte est dite active si elle a été utilisée au moins une fois dans le mois.
- La **proportion de cartes actives au moins 10 mois** de l'année par rapport au nombre total de cartes dans le groupe considéré.
- L'**amplitude moyenne d'utilisation (en mois)**, soit le nombre moyen de mois écoulés entre la première apparition et la dernière apparition de chaque carte dans le réseau de la STM (c'est-à-dire entre le premier et le dernier déplacement de l'année). Ce sont des mois continus, mais l'usager ne s'est pas nécessairement déplacé tous les mois durant cette période. L'équation appliquée est :

$$Amplitude = dernier\ mois\ observé - premier\ mois\ observé + 1 \quad \text{(Éq. 23)}$$

- L'**amplitude moyenne d'utilisation (en jours)**, soit le nombre de jours écoulés entre la première apparition et la dernière apparition de l'usager. Même chose que pour l'amplitude comptée en mois.
- La **proportion de cartes avec une amplitude de 12 mois** par rapport au nombre total de cartes dans le groupe considéré.
- Le **ratio d'activité moyen**, soit le ratio moyen entre le nombre de mois observés et l'amplitude d'utilisation (exprimée en mois). Cet indicateur représente l'activité moyenne des usagers de chaque groupe : si le ratio est égal à 1, les usagers sont en moyenne actifs tous les mois entre leur première apparition et leur dernière apparition sur le réseau.
- La **proportion moyenne des déplacements faits en jours ouvrables** (JO). De cet indicateur on peut également calculer la proportion moyenne des déplacements faits en jours non ouvrables par : $P_{JNO}^{déplacements} = 1 - P_{JO}^{déplacements}$. Les jours non ouvrables correspondent aux fins de semaine et aux jours fériés de 2016.

De plus, lorsque cela est possible, c'est-à-dire lorsque l'indicateur présenté est une moyenne de facteurs individuels, un coefficient de variation est calculé pour évaluer la variabilité de l'indicateur à l'intérieur de chaque groupe. Pour chaque variable étudiée, le coefficient de variation mesure la dispersion des valeurs autour de la moyenne obtenue dans le groupe. Il est défini comme le rapport entre l'écart-type σ et la moyenne μ:

$$CV = \frac{\sigma}{\mu} \cdot 100\% \qquad \text{(Éq. 24)}$$

Les résultats du calcul de ces indicateurs sont donnés dans le Tableau 5.4 ci-après. Les groupes C1 et C2 pris ensemble (38.6% des cartes-année) représentent à eux deux plus de 80% des déplacements réalisés en 2016. Le groupe C6 équivaut exactement à la même proportion de cartes-année (38.6%) que ces deux groupes réunis, mais il correspond seulement à 11.4% des déplacements totaux de 2016. Cette différence d'intensité d'utilisation se traduit également dans les nombres moyens de déplacements par an et par mois actif, très supérieurs à 1 pour les groupes C1 et C2, et au contraire très inférieurs à 1 pour le groupe C6. Les trois autres groupes (C4, C5, C6) se partagent les 6.3% de déplacements restants. Ces groupes étant observés sur un faible nombre de mois (environ 2 ou 3 mois), leur nombre moyen de déplacements par an est également

bas. Néanmoins, en plus d'avoir des tailles similaires, les groupes C4 et C5 présentent des fréquences d'utilisation mensuelles et annuelles analogues.

Tableau 5.4 Indicateurs temporels d'utilisation calculés dans chacun des six groupes d'usagers

INDICATEUR	C1	C2	C3	C4	C5	C6	Total
Répartition des cartes-année	12.0%	26.6%	12.4%	5.4%	5.1%	38.6%	100.0%
Répartition des déplacements	34.6%	47.8%	4.8%	0.7%	0.8%	11.4%	100.0%
Nb moyen de déplacements par an (normalisé) *Coefficient de variation*	2.89 *(61.7%)*	1.79 *(50.5%)*	0.39 *(76.1%)*	0.12 *(106.3%)*	0.16 *(107.7%)*	0.30 *(117.0%)*	1.00 *(125.2%)*
Nb moyen de déplacements par mois actif (normalisé) *Coefficient de variation*	2.68 *(23.0%)*	1.49 *(22.5%)*	0.86 *(66.4%)*	0.54 *(101.4%)*	0.53 *(97.4%)*	0.32 *(76.4%)*	1.00 *(87.9%)*
Nb moyen de mois observés par carte *Coefficient de variation*	7.0 *(54.3%)*	7.9 *(43.1%)*	2.9 *(50.8%)*	1.6 *(62.2%)*	2.0 *(69.9%)*	5.3 *(67.7%)*	5.5 *(68.5%)*
Proportion de cartes actives chaque mois	23.4%	23.3%	0.2%	0.0%	0.0%	7.1%	11.7%
Proportion de cartes actives au moins 10 mois	36.4%	39.9%	0.7%	0.1%	0.3%	16.9%	21.6%
Amplitude moyenne d'utilisation (en mois) *Coefficient de variation*	7.2 *(54.1%)*	8.3 *(41.6%)*	3.1 *(66.5%)*	2.2 *(106.8%)*	2.9 *(102.5%)*	6.5 *(62.4%)*	6.2 *(65.3%)*
Amplitude moyenne d'utilisation (en jours) *Coefficient de variation*	210 *(57.2%)*	239 *(45.6%)*	74 *(86.8%)*	45 *(157.9%)*	65 *(138.3%)*	174 *(73.3%)*	171 *(74.9%)*
Proportion de cartes avec une amplitude de 12 mois	26.3%	29.9%	1.9%	1.8%	3.3%	17.4%	18.3%
Ratio moyen nb de mois actifs/amplitude *Coefficient de variation*	0.99 *(6.7%)*	0.95 *(11.5%)*	0.97 *(12.5%)*	0.93 *(20.6%)*	0.89 *(25.5%)*	0.85 *(23.6%)*	0.92 *(18.5%)*
Proportion moyenne JO *Coefficient de variation*	78.1% *(9.7%)*	86.7% *(12.3%)*	85.9% *(21.6%)*	77.9% *(34.3%)*	77.8% *(37.3%)*	79.9% *(29.9%)*	82.0% *(24.2%)*

La distribution mensuelle des déplacements parmi les six groupes sur la Figure 5.4 confirme que les proportions de déplacements les plus élevées sont réalisées chaque mois par les cartes des groupes C2 en première position, puis par celles du groupe C1 en deuxième position, notamment parce qu'elles sont moins nombreuses. La part du groupe C1 reste relativement constante au cours de l'année alors que celle de C2 diminue pendant les périodes estivale et hivernale. Les proportions

des déplacements mensuels réalisés par les cartes du groupe C6 sont faibles, mais assez stables, nonobstant une légère diminution pendant les premiers et les derniers mois de l'année. En revanche, les déplacements faits par les trois groupes restants (C2, C3, C4) sont concentrés sur une courte période de temps, au début ou à la fin de l'année.

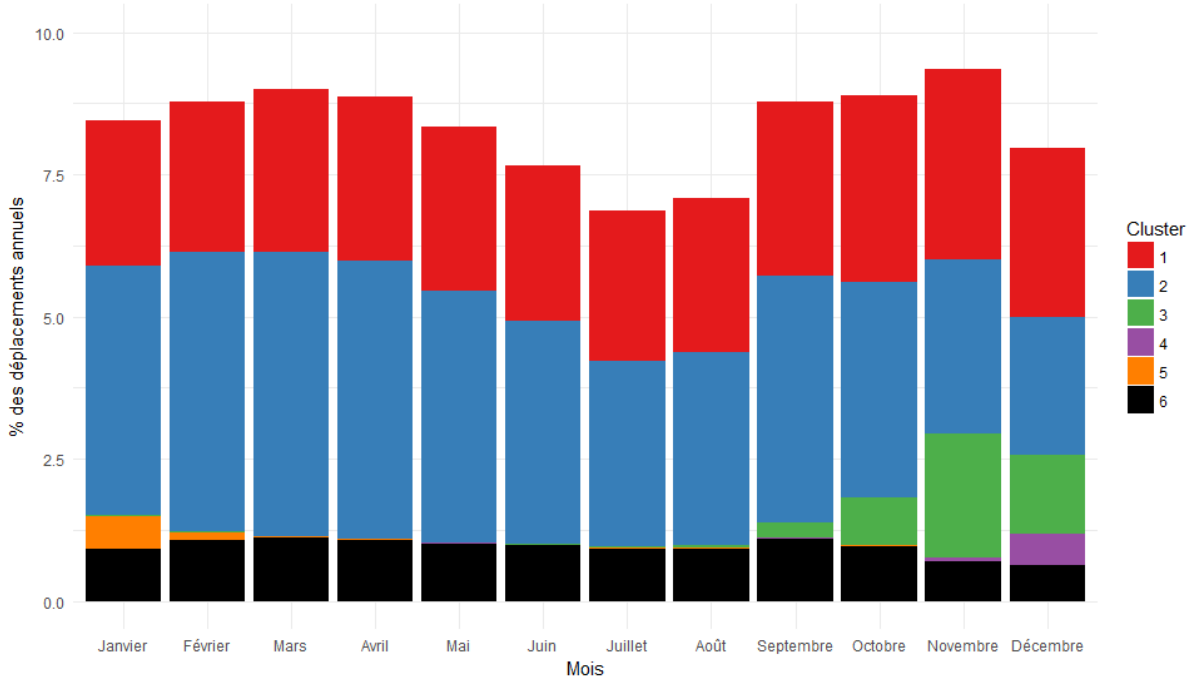


Figure 5.4 Distribution des déplacements de 2016 par mois et par groupe de cartes

De plus, les groupes C1 et C2 rassemblent des cartes avec des durées totales d'activité assez longues, de 7 à 8 mois en moyenne, 23% d'entre elles étant observées tous les mois de l'année. Au contraire, quasiment la totalité des cartes appartenant aux groupes C3, C4 et C5 sont actives moins de 10 mois et une petite partie seulement des cartes du groupe C6 sont validées chaque mois de l'année. Malgré sa courte période d'observation, le groupe C3 présente un des ratios d'activité les plus élevés, ce qui signifie que la majorité des cartes de ce groupe ont été activées au moins une fois tous les mois durant leur amplitude d'utilisation. Les groupes C1 et C2 possèdent également des moyennes assez hautes, alors que le groupe C6 correspond au plus faible ratio d'activité : les cartes de ce groupe ne sont donc utilisées que partiellement durant leurs 6.5 mois d'amplitude moyenne. La Figure 5.5 montre la distribution des cartes-année dans neuf classes de ratios d'activités. Le total des cartes est ramené à 100% dans chaque classe afin de pouvoir visualiser l'affiliation des cartes aux six groupes (sans cela, la majorité des cartes se trouvant dans la classe]0.9;1], la composition des autres classes étaient peu visibles). On constate que les ratios d'activité

les plus faibles sont principalement observés sur des cartes appartenant aux groupes C4 et C5, alors que la majorité des ratios les plus élevés sont produits par les cartes des groupes C2 et C6 (groupes également les plus volumineux, donc plus de ratios différents sont observés).

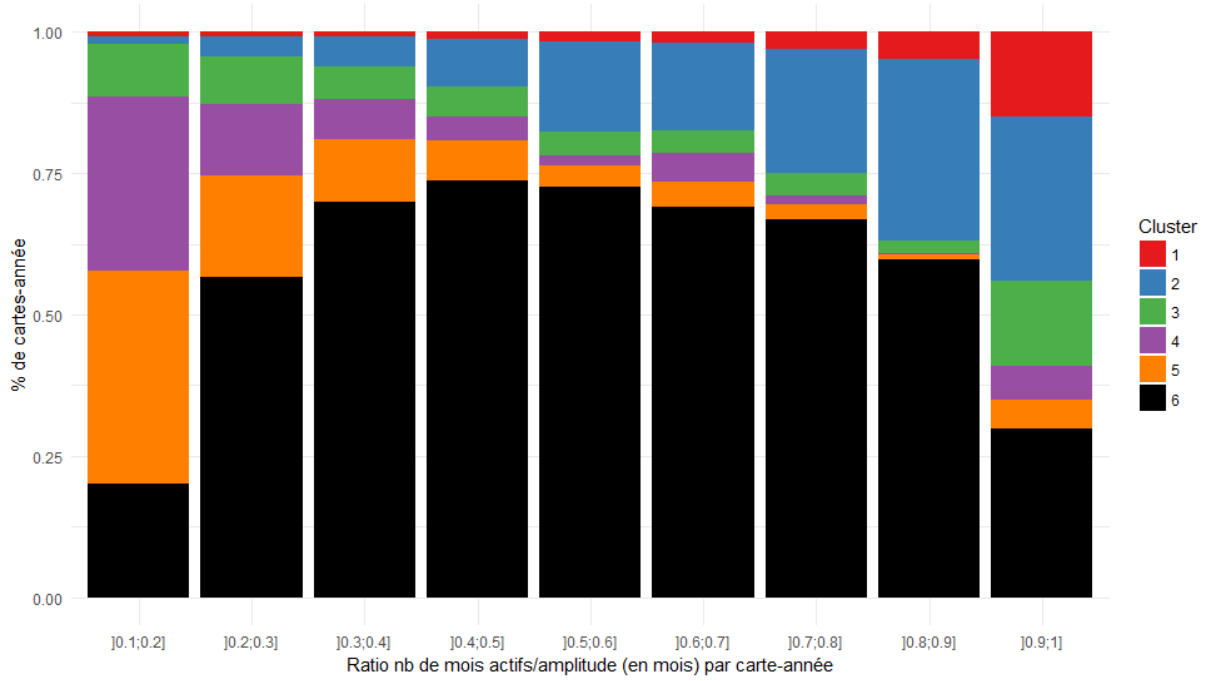


Figure 5.5 Distribution des cartes-année en fonction de leur ratio d'activité et du groupe auquel elles appartiennent

Par ailleurs, le Tableau 5.4 rapporte également la proportion moyenne des déplacements faits en jours ouvrables. Cette proportion est dans l'ensemble très élevée pour tous les groupes, traduisant ainsi que la majorité des déplacements sont effectués pendant des jours travaillés. Cependant, on note que cette concentration des déplacements dans un seul type de jour est moins importante pour le groupe C1 que pour le groupe C2 : les usagers du groupe C1 utilisent le transport en commun pour faire en moyenne presque 22% de leurs déplacements en jours non ouvrables (c.-à-d. jours de fin de semaine et jours fériés), et donc possiblement pour des motifs autres que le travail (même s'il y a aussi des personnes qui travaillent pendant ces jours). Cette diversification de leur usage du transport en commun justifierait leur plus grande intensité d'utilisation comparée à celle du groupe C2, les usagers appartenant à ce dernier groupe empruntant peu le transport en commun pendant les jours non ouvrables et comptabilisant donc moins de déplacements. Cela expliquerait aussi pourquoi le nombre de déplacements du groupe C1 ne diminue pas pendant la période estivale; les usagers de ce groupe continuent probablement d'utiliser le transport en commun pendant les

périodes de vacances scolaires. La proportion des déplacements faits en jours ouvrables est également moins élevée pour les groupes C4 et C5, certainement parce que les utilisateurs des cartes de ces groupes se déplacent pendant la période des fêtes de fin d'année.

5.2.2.2 Indicateurs spatiaux (ou modaux) d'utilisation du TC

Dans ce mémoire, l'utilisation spatiale du transport en commun est liée aux choix modaux des usagers, car les stations de métro et les lignes de bus empruntées durant l'année sont les seules informations contenues dans les données de cartes à puce de la STM pouvant être spécialisées. C'est pourquoi ces deux dimensions de l'utilisation du transport en commun (espace et modes utilisés) sont présentées dans la même famille d'indicateurs ici. Plusieurs des indicateurs définis ci-après sont basés sur les validations et non sur les déplacements, car un déplacement composé de plusieurs validations peut être réalisé avec plusieurs modes. Au contraire, une validation est effectuée à l'embarquement d'un seul mode (bus ou métro). Afin de pouvoir faire un lien entre ces deux concepts, le nombre moyen de validations par déplacement est calculé. Ce chiffre peut également être considéré comme un indicateur proxy pour estimer la longueur des déplacements (celle-ci ne pouvant pas être mesurée réellement car la destination des déplacements n'est pas connue). En effet, des déplacements plus longs auront tendance à générer plus de correspondances et donc plus de validations, mais ce n'est pas toujours le cas. Quelques-uns des indicateurs proposés ensuite servent à distribuer les validations ou les déplacements en fonction des modes empruntés. Une validation peut être réalisée à l'entrée d'une station de métro ou à l'embarquement d'une ligne de bus (donc une validation est associée soit au métro soit au bus), tandis qu'il existe trois types de déplacements : des déplacements faits avec les modes métro et bus combinés, des déplacements faits avec le métro uniquement et des déplacements faits en bus uniquement. Les autres indicateurs calculés sont orientés de façon à mesurer la variabilité spatiale, définie comme l'utilisation répétée des mêmes stations de métro et des mêmes lignes de bus durant l'année. Ainsi, plus un usager recourt souvent aux mêmes modes et aux mêmes lieux d'embarquement, plus il est dit régulier au niveau spatial.

Voici la liste des indicateurs répertoriés dans le Tableau 5.5 :

- Le **nombre moyen de validations par déplacement**, soit le quotient moyen entre le nombre total de validations et le nombre total de déplacements par carte. Un déplacement

est composé de plusieurs validations s'il implique des correspondances métro-bus ou bus-bus (les correspondances métro-métro n'étant pas observables avec les données exploitées).

- La **proportion moyenne des validations de chaque carte faites dans le métro**. De cet indicateur on peut également calculer la proportion moyenne des validations de chaque carte faites dans le bus par : $P_{bus}^{validations} = 1 - P_{métro}^{validations}$.
- La **proportion moyenne de déplacements métro-bus combinés**, c'est-à-dire la proportion moyenne des déplacements réalisés avec les deux modes.
- La **proportion moyenne de déplacements réalisés en métro uniquement**, c'est-à-dire que le métro est le seul mode utilisé pour faire ces déplacements. On peut également calculer la proportion moyenne des déplacements réalisée en bus uniquement par : $P_{bus}^{déplacements} = 1 - P_{métro}^{déplacements} - P_{métro+bus}^{déplacements}$.
- Le **nombre moyen de stations de métro différentes** visitées par l'usager pendant l'année. Chaque station doit avoir été utilisée au moins une fois pendant l'année pour être comptée.
- Le **nombre moyen de lignes de bus différentes** empruntées par l'usager pendant l'année. Chaque ligne doit avoir été utilisée au moins une fois pendant l'année pour être comptée.
- La **proportion moyenne des validations de métro réalisées dans les deux stations les plus empruntées** par l'usager pendant l'année.
- La **proportion moyenne des validations de bus réalisées dans les deux lignes les plus empruntées** par l'usager pendant l'année.
- Le **nombre moyen de stations différentes utilisées pour faire 80% des validations** métro de l'usager pendant l'année.
- Le **nombre moyen de lignes différentes utilisées pour faire 80% des validations** bus de l'usager pendant l'année.
- L'**espace d'action moyen par carte** [en km^2] (pour le métro uniquement), c'est-à-dire l'aire moyenne de l'enveloppe convexe de toutes les stations de métro visitées par un usager pendant l'année. Cette enveloppe est l'ensemble convexe le plus petit qui englobe toutes les stations de métro utilisées, comme sur le schéma ci-dessous. Elle peut également être assimilée à un élastique autour des stations.

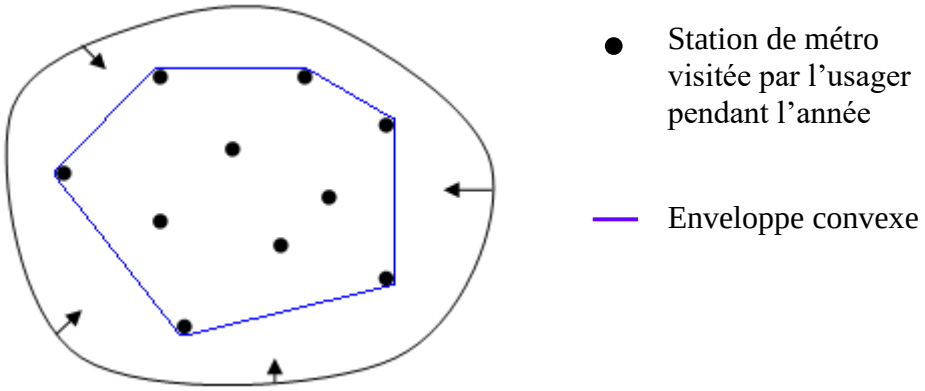


Figure 5.6 Schéma de l'enveloppe convexe englobant toutes les stations de métro visitées par un usager durant l'année (image tirée de https://fr.wikipedia.org/wiki/Enveloppe_convexe)

D'après les résultats du Tableau 5.5, le nombre moyen de validations par déplacement est plus élevé dans les groupes C1 et C2 : les usagers de ces groupes tendraient donc à faire des déplacements plus longs, ou au moins des déplacements plus multimodaux (une validation étant réalisée à chaque changement de mode entre le métro et le bus). Cette dernière observation est confirmée par la répartition de leurs validations et de leurs déplacements dans les modes. Ces distributions sont particulièrement équitables dans le cas du groupe C1 : 51.5% et 48.5% des validations sont faites respectivement dans le métro et dans le bus; 32.6%, 39.0% et 28.4% des déplacements sont réalisés respectivement avec le métro et le bus combinés, avec le métro uniquement et avec le bus uniquement. Dans les autres groupes, le métro est le mode privilégié : environ la moitié des déplacements sont faits en métro uniquement et autour de 60% des validations sont effectuées dans le métro, même si ces proportions sont légèrement plus faibles pour le groupe C2. En termes de nombre moyen de stations de métro et de lignes de bus utilisées par carte, le groupe C1 est largement devant les autres. Les usagers du groupe C2 empruntent également des stations et des lignes assez diversifiés. Toutefois, ces nombres moyens diminuent pour les autres groupes (dans l'ordre, C3, C6, C5 puis C4). L'utilisation de stations de métro variées se répercute sur l'espace d'action moyen. On observe une quasi-relation de proportionnalité (positive) entre le nombre moyen de stations utilisées et l'aire convexe moyenne englobant toutes ces stations. Cette tendance implique que plus le nombre de stations visitées est important, plus la couverture individuelle moyenne du réseau de métro augmente. Les stations utilisées par les usagers sont donc assez bien réparties sur le réseau et non concentrées en un seul endroit.

Tableau 5.5 Indicateurs spatiaux d'utilisation calculés dans chacun des six groupes d'usagers

INDICATEUR	C1	C2	C3	C4	C5	C6	Total
Nb moyen de validations par déplacement	1.53	1.42	1.35	1.34	1.38	1.33	1.38
Coefficient de variation	*(24.2%)*	*(27.3%)*	*(28.7%)*	*(30.3%)*	*(31.5%)*	*(28.9%)*	*(28.5%)*
Proportion moyenne validations métro	51.5%	57.6%	60.5%	64.4%	60.5%	62.9%	59.8%
Coefficient de variation	*(53.1%)*	*(53.9%)*	*(55.8%)*	*(51.2%)*	*(56.0%)*	*(52.7%)*	*(54.0%)*
Proportion moyenne dépl. métro-bus combinés	32.6%	27.8%	23.1%	23.1%	25.1%	21.4%	24.9%
Coefficient de variation	*(73.8%)*	*(99.5%)*	*(121.5%)*	*(126.7%)*	*(121.7%)*	*(125.9%)*	*(110.5%)*
Proportion moyenne dépl. métro uniquement	39.0%	46.7%	51.3%	55.2%	50.3%	54.6%	50.0%
Coefficient de variation	*(85.6%)*	*(80.6%)*	*(77.4%)*	*(72.7%)*	*(81.0%)*	*(72.2%)*	*(77.5%)*
Nb moyen de stations métro différentes par usager	20.0	14.2	6.7	4.4	4.6	6.6	10.0
Coefficient de variation	*(45.7%)*	*(53.0%)*	*(70.6%)*	*(78.3%)*	*(84.1%)*	*(78.2%)*	*(80.5%)*
Nb moyen de lignes bus différentes par usager	19.0	11.4	4.9	2.9	3.2	4.2	7.9
Coefficient de variation	*(58.1%)*	*(72.1%)*	*(97.6%)*	*(119.9%)*	*(114.8%)*	*(112.5%)*	*(107.9%)*
% moy validations dans les 2 stations les + empruntées	60.9%	69.3%	70.9%	71.8%	71.3%	67.3%	67.9%
Coefficient de variation	*(25.8%)*	*(27.8%)*	*(34.0%)*	*(37.4%)*	*(39.9%)*	*(37.2%)*	*(33.9%)*
% moy validations dans les 2 lignes les + empruntées	64.4%	68.0%	63.3%	57.7%	61.2%	60.5%	63.2%
Coefficient de variation	*(28.5%)*	*(37.0%)*	*(56.5%)*	*(71.2%)*	*(65.0%)*	*(62.7%)*	*(52.6%)*
Nb moyen de stations utilisées pour faire 80% des validations	5.0	4.0	3.1	2.9	2.8	3.5	3.7
Coefficient de variation	*(49.7%)*	*(58.5%)*	*(59.1%)*	*(61.8%)*	*(65.8%)*	*(61.4%)*	*(61.1%)*
Nb moyen de lignes utilisées pour faire 80% des validations	4.5	3.3	2.2	1.7	1.8	2.1	2.7
Coefficient de variation	*(60.2%)*	*(71.1%)*	*(85.8%)*	*(100.3%)*	*(94.5%)*	*(92.7%)*	*(85.9%)*
Espace d'action moyen par usager [km2]	62.6	45.3	20.6	11.6	12.7	20.0	31.1
Coefficient de variation	*(50.7%)*	*(66.9%)*	*(110.0%)*	*(150.0%)*	*(149.9%)*	*(118.2%)*	*(99.1%)*

Par ailleurs, les usagers sont dans l'ensemble assez réguliers dans le choix de leurs lieux d'embarquement : la majorité de leurs validations sont concentrées dans les deux stations ou lignes les plus empruntées. Ces proportions sont un peu plus basses pour le groupe C1. De même, les nombres de stations ou de lignes utilisées pour faire 80% des validations de chaque usager sont en moyenne plus élevés pour le groupe C1, puis le groupe C2. D'après la définition adoptée, ces deux

groupes (C1 et C2) seraient donc plus irréguliers au niveau spatial que les autres groupes. Cependant, cette plus grande diversification de leurs lieux d'embarquement peut être liée à leur plus forte intensité utilisation : comme les usagers de ces groupes se déplacent plus, leur acquisition du réseau s'en voit également élargie. Néanmoins, même si les groupes C3 et C6 ont des fréquences d'utilisation et des périodes d'activité assez différentes, leur variabilité spatiale est similaire. D'ailleurs, on remarque que les groupes C3 et C6, ainsi que les groupes C4 et C5 ont des usages spatial et modal assez équivalents : tous les indicateurs calculés dans le Tableau 5.5 se ressemblent pour ces deux paires de groupes.

De plus, la Figure 5.7 permet de visualiser les six groupes obtenus dans l'espace. Cette figure donne un groupe dominant dans chacune des 68 stations du réseau de métro de la STM. Pour déterminer ce groupe dominant, la première station de métro la plus utilisée par chaque carte est d'abord identifiée (c'est-à-dire la station dans laquelle il y a eu le plus de validations). Seules les cartes ayant utilisé le métro au moins une fois dans l'année sont considérées, soit 96% des cartes de l'échantillon exploité. Ensuite, chaque carte est assignée à sa station la plus utilisée ; toutes les cartes sont ainsi distribuées dans les 68 stations de métro. Pour chaque station il est alors possible de connaître la répartition des cartes par groupe, et le groupe dominant dans la station est celui pour lequel la proportion de cartes est la plus élevée. Ainsi, la Figure 5.7 montre que les groupes C4 et C5 sont majoritaires dans les stations du centre-ville de Montréal. Le groupe C5 est également le groupe dominant de la station Mont-Royal, située au cœur de l'arrondissement du Plateau-Mont-Royal. Cet arrondissement est justement connu pour son très grand nombre de piétons et de cyclistes, potentiels utilisateurs du transport en commun pendant la période hivernale, et donc potentiels membres des groupes C4 et C5. À l'inverse, les cartes du groupe C1 sont plutôt localisées à l'extérieur du centre-ville, ce qui pourrait motiver des déplacements plus longs. Aucune tendance spatiale n'est discernable pour les groupes C2 et C6. En revanche, on constate que le groupe C3 est dominant dans des stations avoisinant des écoles ou des universités. En particulier, les stations Côte-des-Neiges, Université de Montréal et Edouard-Montpetit sont proches des écoles Polytechnique, HEC et UdeM (Université de Montréal). De même, l'université de Concordia et les collèges Montmorency, Dawson et LaSalle se trouvent à proximité de stations colorées en vert sur la carte (couleur du groupe C3). Cette contiguïté pourrait être fortement liée au fait que les membres du groupes C3 sont en bonne partie des étudiants. Les observations faites sur la carte de la Figure 5.7 coïncident donc avec les analyses faites précédemment. Toutefois, des analyses

supplémentaires seraient nécessaires pour savoir si la station la plus utilisée de chaque carte représente plutôt un lieu de domicile ou un lieu de travail.

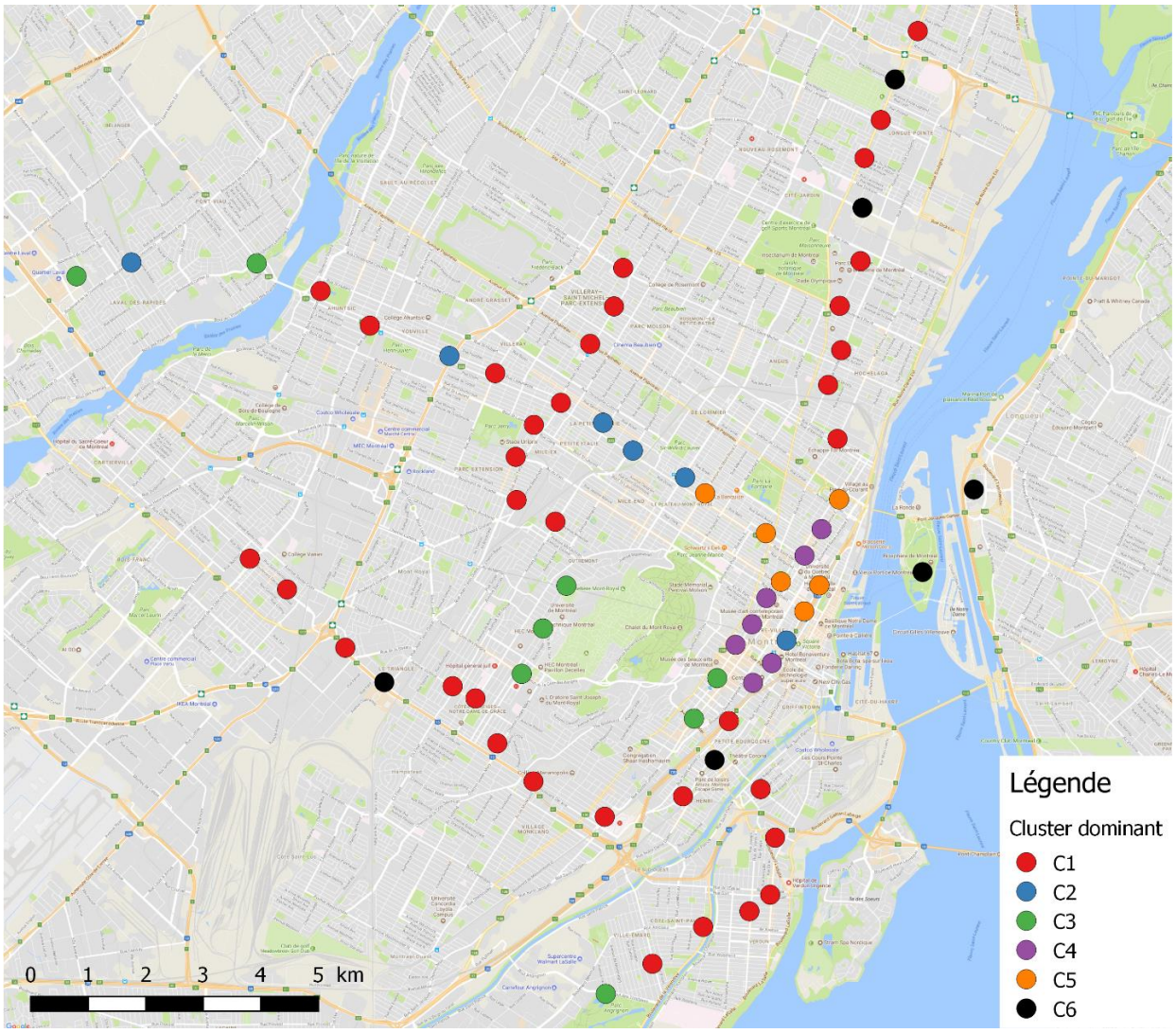


Figure 5.7 Groupe dominant dans chaque station de métro du réseau de la STM (à partir de la station la plus utilisée de chaque carte)

5.2.2.3 Indicateurs de variabilité (chapitre 4)

Pour finir, les indicateurs définis dans la section 4.3.1 pour mesurer différents types de variabilités dans l'utilisation du transport en commun sont calculés ici dans chacun des six groupes de la typologie obtenue. Les résultats de leur application sont présentés dans le Tableau 5.6.

Les indices de Pareto les plus élevés (et donc les coefficients de Gini les plus faibles) sont atteints dans les groupes C1 et C2. La distribution de tous les déplacements de 2016 parmi les cartes est

donc plus homogène à l'intérieur de ces deux groupes. En revanche, une plus grande hétérogénéité est décelée entre les usagers des autres groupes : environ 30% des cartes les plus fréquentes représentent 70% des déplacements. Cette plus grande variabilité interpersonnelle fait écho aux coefficients de variation élevés précédemment obtenus pour la variable « Nb moyen de déplacements par an » dans le Tableau 5.4. Néanmoins, les indices de Pareto évalués dans ces groupes restent supérieurs à celui calculé avec l'ensemble des cartes ($\alpha = 1.32$) : les usagers rassemblés dans chaque groupe ont donc des intensités d'utilisation plus similaires entre elles que dans le total des cartes, ceci étant voulu par la segmentation réalisée.

De plus, la fréquence d'utilisation mensuelle du transport en commun est majoritairement très élevée pour les cartes du groupe C1, dépassant largement celle du groupe C2. Le pourcentage de cartes associé à ce fort taux d'utilisation est également plus important dans C1 que dans C2, signalant une moins grande variabilité interpersonnelle entre les cartes du groupe C1. La classe de nombre de déplacements par mois actif la plus fréquente obtenue pour tous les autres groupes est la classe minimale (classe n°1). Cependant, la proportion de cartes correspondant à cette classe modale est plus basse pour C3 que pour les autres groupes : 78.5% des usagers du groupe C3 sont plus fréquents que ceux rassemblés dans la classe modale minimale. Au contraire, pour les groupes C4, C5, et C6, le faible taux d'utilisation mesuré est commun à environ 40% à 50% des cartes.

La variabilité temporelle moyenne estimée à l'aide de la variance du nombre de déplacements par mois est ici aussi fortement liée à la fréquence d'utilisation : plus les usagers empruntent le transport en commun, plus des variations ont des chances d'être observées dans leur comportement au fil du temps. Cette variabilité est notamment très élevée pour le groupe C1 par rapport aux autres groupes. Par ailleurs, dans tous les groupes, la part de la variance temporelle interpersonnelle est inférieure à celle de la variance intrapersonnelle. Ce résultat est la conséquence de l'application de l'algorithme des K-moyennes, dont l'objectif était des regrouper des usagers similaires entre eux (d'où la diminution des variations entre les usagers).

Pour la variabilité spatiale, les mêmes tendances sont observées qu'avec les indicateurs spatiaux calculés précédemment. Les groupes C1 et C2 ont une plus grande variabilité spatiale (entropies moyennes plus élevées) du fait de leur plus grande acquisition du réseau. Néanmoins, l'entropie moyenne évaluée avec les cartes du groupe C6 est également assez haute, indiquant une grande diversité de lieux d'embarquement malgré leur faible utilisation du transport en commun.

Tableau 5.6 Application des indicateurs de variabilité du chapitre 4 dans chacun des six groupes d'usagers

Type de variabilité	Indicateur	GROUPE / CLUSTER						
		C1	C2	C3	C4	C5	C6	Total
Dispersion des déplacements	p	62.3%	60.4%	65.0%	70.2%	70.6%	72.1%	73.0%
	$\bar{p}$	37.7%	39.6%	35.0%	29.8%	29.4%	27.9%	27.0%
	α	1.94	2.20	1.70	1.41	1.40	1.34	1.32
	G	0.34	0.29	0.43	0.54	0.55	0.58	0.61
Fréquence d'utilisation	Classe modale	11	8	1	1	1	1	1
	% de cartes	24.0%	21.1%	21.5%	44.0%	41.8%	50.1%	26.5%
Variabilité temporelle	TSS* (10^{-3})	14.07	4.95	1.62	0.48	0.62	0.56	5.14
	BPSS	41.2%	30.4%	9.8%	6.5%	8.3%	38.7%	55.7%
	WPSS	58.8%	69.6%	90.2%	93.5%	91.7%	61.3%	44.3%
Variabilité spatiale MÉTRO	Entropie moyenne	0.45	0.38	0.31	0.27	0.27	0.32	0.35
	CV	23.6%	32.3%	41.1%	51.5%	52.9%	44.4%	40.7%
Variabilité spatiale BUS	Entropie moyenne	0.31	0.25	0.19	0.16	0.16	0.19	0.22
	CV	34.2%	46.8%	60.9%	75.1%	74.5%	66.4%	57.6%

* TSS est divisé par le nombre de cartes dans chaque groupe

5.2.3 Vérifications statistiques

Des tests statistiques classiques ont été appliqués afin de vérifier que les différences observées entre les six groupes (par paires) étaient significatives pour tous les indicateurs proposés. Cependant, comme dans le Chapitre 4, la taille importante des échantillons manipulés a conduit à des résultats utopiques et sans intérêt, avec des valeurs-p toutes nulles. La notion de taille d'effet peut néanmoins être utilisée pour quantifier l'importance des différences mises en évidence. Pour les indicateurs de variabilité, les mêmes indices que ceux présentés précédemment sont appliqués pour mesurer cette taille d'effet. Pour les autres indicateurs d'utilisation temporelle et spatiale du transport en commun (uniquement ceux calculés comme une moyenne de variables individuelles), une statistique Z est déterminée avec le test asymptotique de Wilcoxon Mann-Whitney puis le coefficient de corrélation r défini dans la section 4.4.3 est calculé. Les résultats de l'application de ces indices de taille d'effet sont donnés dans le Tableau 5.7 pour les indicateurs d'utilisation, et dans le Tableau 5.8 pour les indicateurs de variabilité. Les couleurs attribuées aux cases de chaque tableau correspondent au critère de Cohen (voir Tableau 4.6 pour rappel).

Tableau 5.7 Mesure de la taille d'effet pour les indicateurs temporels et spatiaux – Cas de tous les usagers

	INDICATEURS TEMPORELS							INDICATEURS SPATIAUX										
Test	Nb moyen de déplacements par an	Nb moyen de déplacements par mois actif	Nb moyen de mois observés par carte	Amplitude moyenne d'utilisation (en mois)	Amplitude moyenne d'utilisation (en jours)	Ratio moyen nb de mois actifs/amplitude	Proportion moyenne JO	Nb moyen de validations par déplacement	Proportion moyenne validations métro	Proportion moyenne dépl. métro-bus combinés	Proportion moyenne dépl. métro uniquement	Nb moyen de stations métro différentes par usager	Nb moyen de lignes bus différentes par usager	% moy validations dans les 2 stations les + empruntées	% moy validations dans les 2 lignes les + empruntées	Nb moyen de stations utilisées pour faire 80% des validations	Nb moyen de lignes utilisées pour faire 80% des validations	Espace d'action moyen par usager
C1 v C2	0.28	0.79	0.10	0.13	0.08	0.17	0.40	0.16	0.10	0.13	0.09	0.30	0.34	0.22	0.13	0.24	0.24	0.25
C1 v C3	0.82	0.86	0.56	0.53	0.59	0.03	0.44	0.29	0.17	0.27	0.15	0.71	0.69	0.30	0.13	0.43	0.48	0.63
C1 v C4	0.80	0.79	0.70	0.63	0.67	0.14	0.17	0.29	0.22	0.27	0.18	0.73	0.72	0.28	0.05	0.44	0.53	0.68
C1 v C5	0.79	0.79	0.64	0.53	0.58	0.24	0.21	0.24	0.15	0.23	0.11	0.72	0.70	0.28	0.10	0.45	0.51	0.66
C1 v C6	0.70	0.74	0.21	0.09	0.15	0.34	0.20	0.28	0.18	0.26	0.16	0.60	0.61	0.16	0.07	0.28	0.43	0.54
C2 v C3	0.71	0.50	0.64	0.62	0.63	0.14	0.10	0.13	0.06	0.14	0.06	0.49	0.41	0.08	0.02	0.16	0.24	0.41
C2 v C4	0.63	0.52	0.60	0.55	0.56	0.03	0.02	0.14	0.11	0.14	0.09	0.51	0.44	0.09	0.01	0.18	0.28	0.45
C2 v C5	0.62	0.53	0.56	0.49	0.49	0.05	0.01	0.10	0.05	0.11	0.03	0.49	0.41	0.10	0.02	0.20	0.26	0.43
C2 v C6	0.77	0.85	0.35	0.23	0.27	0.28	0.02	0.17	0.11	0.18	0.10	0.52	0.48	0.01	0.01	0.09	0.28	0.45
C3 v C4	0.44	0.27	0.49	0.41	0.39	0.10	0.09	0.05	0.06	0.04	0.05	0.24	0.21	0.04	0.02	0.06	0.12	0.22
C3 v C5	0.38	0.28	0.36	0.25	0.25	0.20	0.06	0.00	0.01	0.01	0.01	0.23	0.17	0.05	0.01	0.09	0.09	0.20
C3 v C6	0.17	0.41	0.26	0.32	0.29	0.32	0.09	0.03	0.03	0.03	0.03	0.03	0.08	0.07	0.02	0.07	0.03	0.03
C4 v C5	0.07	0.00	0.14	0.14	0.10	0.10	0.02	0.05	0.06	0.03	0.06	0.00	0.05	0.00	0.03	0.03	0.03	0.01
C4 v C6	0.17	0.09	0.37	0.35	0.31	0.18	0.00	0.01	0.02	0.01	0.01	0.14	0.09	0.08	0.00	0.10	0.06	0.13
C5 v C6	0.12	0.09	0.32	0.28	0.26	0.11	0.02	0.02	0.02	0.01	0.03	0.13	0.06	0.08	0.02	0.11	0.04	0.12

D'après les valeurs surlignées en rouge dans la première partie du Tableau 5.7, les groupes C1 et C2 sont très distincts des autres groupes pour beaucoup d'indicateurs. Au niveau temporel, les indicateurs mesurant la fréquence d'utilisation et la période d'activité des cartes (nombre de mois observés et amplitude) affichent les différences les plus importantes. On remarque également la particularité déjà soulignée des usagers du groupe C1 en ce qui a trait à la répartition de leurs

déplacements dans les différents types de jours (jours ouvrables versus non ouvrables). Au niveau spatial, les plus grandes différences sont rapportées pour le nombre moyen de stations ou de lignes empruntées par chaque usager durant l'année, le nombre moyen de stations ou de lignes utilisées pour faire 80% des validations métro ou bus, et pour l'espace moyen d'action. Ces trois indicateurs sont en effet liés. Les groupes C1 et C2 présentent en revanche des similitudes. La seule grosse différence entre les deux groupes est obtenue pour le nombre moyen de déplacements par mois actif (taille d'effet égale à 0.79). Cette divergence s'explique par le fait que la variable d'intensité mensuelle a été prise en compte dans le processus de segmentation pour départager les usagers.

Les différences observées entre les autres groupes sont modérées, notamment au niveau temporel, ou faibles. En outre, les usagers des groupes C4 et C5 sont extrêmement semblables : presque tous les indices de taille d'effet associés au test 'C4 v C5' sont inférieurs à 0.1. En effet, il a déjà été constaté précédemment que tous les indicateurs calculés pour ces deux groupes se ressemblaient énormément. Cette importante similarité entre les deux groupes va dans le sens de l'hypothèse émise au départ selon laquelle les mêmes usagers appartiennent peut-être à ces deux groupes avec une carte différente. De même, les groupes C3 et C6 sont moyennement différents pour les indicateurs temporels, mais leur utilisation spatiale du transport en commun est très similaire. D'ailleurs, au niveau spatial, tous les groupes après C3 sont plus ou moins équivalents.

Certains indicateurs fonctionnent moins bien pour dissocier les groupes. Au niveau temporel, les ratios moyens d'activité sont assez comparables d'un groupe à l'autre, le groupe C6 étant l'unique bloc de cartes se différenciant légèrement des autres. Au niveau spatial, les distributions moyennes des déplacements ou des validations par mode, ainsi que les pourcentages de validations faites dans les deux stations ou lignes les plus empruntées sont analogues pour tous les groupes. Seule l'appartenance au groupe C1 semble parfois démontrer un effet moyen sur ces variables.

Par ailleurs, la taille d'effet est mesurée pour les indicateurs de variabilité dans le Tableau 5.8. Comme pour les indicateurs d'utilisation, les groupes C1 et C2 se démarquent des autres groupes : toutes les différences observées entre ces deux groupes et les autres sont très ou moyennement importantes. Néanmoins, la ressemblance de ces deux groupes est moins probante que précédemment : les usagers de C1 et C2 auraient donc des niveaux de régularité assez éloignés. Les indicateurs de variabilité calculés dans les autres groupes sont assez semblables, sauf pour la distribution de leurs fréquences d'utilisation. Les tailles d'effet les plus faibles sont encore une fois

relevées pour la paire ‘C4 v C5’, quel que soit l’indicateur mesuré. De même, les groupes C3 et C6 présentent des niveaux de variabilité temporelle et spatiale assez similaires.

Tableau 5.8 Mesure de la taille d'effet pour les indicateurs de variabilité – Cas de tous les usagers

Test	Dispersion des déplacements	Fréquence d’utilisation	Variabilité temporelle	Variabilité spatiale – métro	Variabilité spatiale – bus
C1 v C2	0.05	0.92	0.44	0.27	0.24
C1 v C3	0.13	0.98	0.67	0.52	0.47
C1 v C4	0.16	0.97	0.66	0.56	0.48
C1 v C5	0.19	0.98	0.64	0.54	0.48
C1 v C6	0.20	1.00	0.73	0.40	0.44
C2 v C3	0.15	0.63	0.45	0.23	0.21
C2 v C4	0.21	0.83	0.47	0.28	0.24
C2 v C5	0.23	0.83	0.45	0.27	0.24
C2 v C6	0.25	0.98	0.65	0.19	0.26
C3 v C4	0.09	0.26	0.14	0.14	0.12
C3 v C5	0.10	0.26	0.11	0.14	0.12
C3 v C6	0.13	0.61	0.06	0.04	0.03
C4 v C5	0.04	0.03	0.04	0.01	0.01
C4 v C6	0.05	0.37	0.16	0.12	0.07
C5 v C6	0.03	0.36	0.14	0.12	0.06

5.3 Cas des utilisateurs d’abonnements annuels

Dans cette section, un groupe de cartes spécifique est étudié : les utilisateurs d’abonnements longs (soit environ 12% des usagers ayant utilisé un seul type de produits pendant l’année) avec une amplitude de 12 mois (soit 39% des utilisateurs d’abonnements longs) ont été sélectionnés, c’est-à-dire les usagers qui ont utilisé uniquement un abonnement annuel ou un abonnement de 4 mois (tous tarifs confondus) durant l’année 2016, et qui ont validé leur carte au moins une fois en janvier et au moins une fois en décembre. Cette amplitude de 12 mois n’implique pas que les usagers se sont déplacés tous les mois de l’année. Néanmoins, le Tableau 5.9 ci-dessous montre que c’est quand même le cas pour la majorité des usagers de l’échantillon étudié : presque 90% des cartes ont été observées au moins une fois par mois durant les 12 mois de l’année. Par ailleurs, pour pouvoir garder le même échantillon dans les chapitres suivants, les usagers qui n’ont pas fait au moins un déplacement entre le 4 janvier et le 25 décembre 2016 ont été exclus. En effet, les analyses réalisées au niveau hebdomadaire dans les Chapitre 6 et 7 requièrent de ne considérer que les 51 semaines complètes de l’année. Cette opération a conduit à la suppression (très négligeable) d’un

seul utilisateur atypique qui s'était déplacé avant le 4 janvier et après le 25 décembre, mais pas entre temps. Finalement, l'échantillon analysé dans cette section et dans les chapitres qui vont suivre est composé de 56988 cartes, soit environ 3% de la base de données initiale. Cet échantillon contenant moins de 1% d'utilisateurs d'abonnements de 4 mois, on parlera uniquement d'utilisateurs d'abonnements annuels dans la suite.

Tableau 5.9 Distribution des cartes des utilisateurs d'abonnements annuels avec une amplitude de 12 mois en fonction du nombre de mois actifs par carte

Nombre de mois actifs	1	2	3	4	5	6	7	8	9	10	11	12
% de cartes	0.0%	0.1%	0.1%	0.2%	0.3%	0.4%	0.6%	0.9%	1.1%	1.9%	4.6%	89.8%

Cet échantillonnage a été effectué dans le but de réduire la dimension des données étudiées, permettant ainsi de développer des méthodes plus complexes et plus calculatoires, mais aussi pour pouvoir approfondir l'analyse d'un groupe particulier. Les utilisateurs d'abonnements annuels avec une amplitude de 12 mois ont été choisis car leur carte a été présente sur le réseau toute l'année (c.-à-d. elle n'a pas été renouvelée ou perdue au cours de l'année, mais elle peut encore avoir été utilisée par plusieurs personnes). Ces utilisateurs sont donc observés sur une longue période et ont un certain niveau d'activité d'après le Tableau 5.9. D'après les résultats du Chapitre 4, ces usagers sont aussi très réguliers. Néanmoins, l'ANNEXE G prouve l'existence de variations dans leur nombre de déplacements par mois en jours ouvrables. De même, la suite de cette section va montrer que des différences interpersonnelles peuvent également être décelées entre ces usagers.

5.3.1 Typologie obtenue

La méthode de segmentation détaillée dans la section 5.1 est ici appliquée à la base « cartes-année » échantillonnée, composée de presque 57000 vecteurs. Les deux critères utilisés pour choisir un nombre de groupes K approprié sont reproduits sur la Figure 5.8. Pour les mêmes raisons que celles énoncées précédemment, $K = 6$ groupes sont sélectionnés. Ainsi, deux petits groupes à très forte intensité d'utilisation (à gauche du dendrogramme) et quatre autres groupes constitués d'usagers moins fréquents (à droite) sont formés. D'après la distribution des cartes-années donnée dans la première ligne du Tableau 5.10, le groupe le plus volumineux est celui qui correspond au quatrième niveau d'intensité d'utilisation (groupe C4). Les usagers des groupes C1 et C2 à très fort taux

d’utilisation se font plus rares tandis que les groupes restants (C3, C5, C6) ont des tailles relativement homogènes comprises entre 15 et 20% du total des cartes.

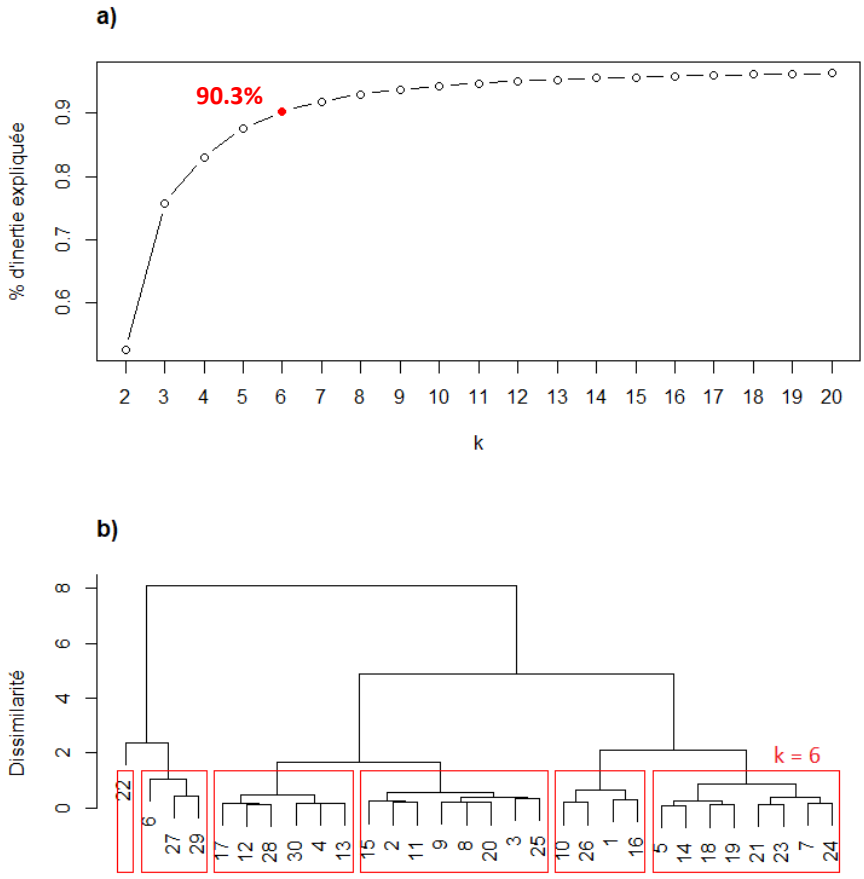


Figure 5.8 Choix du nombre de groupes K avec a) le pourcentage d'inertie expliquée b) un dendrogramme réalisé sur 30 groupes initiaux – Cas des abonnements annuels

Les six centres obtenus sont explicités dans le Tableau 5.10 et illustrés sur la Figure 5.9. Sur cette figure, la ligne grise en pointillés représente une utilisation uniforme du transport en commun, c’est-à-dire lorsque le même nombre de déplacements est effectué tous les mois (d’où une proportion constante de 8.33% des déplacements annuels réalisés chaque mois). On constate que les profils mensuels moyens de déplacements de tous les groupes ne sont pas tant éloignés que ça de cette distribution uniforme : les variations dans les proportions de déplacements par mois sont beaucoup moins grandes que celles qui avaient été observées avec la typologie précédente. En effet, entre 6.5% et 11.5% des déplacements annuels sont faits chaque mois. Les six groupes d’usagers produits se différencient surtout par leur intensité d’utilisation. Le graphique à gauche de la Figure 5.9 montre ainsi une belle distribution de l’indicateur d’intensité. Avec le graphique de droite

(distribution des déplacements par mois), on s'aperçoit qu'une intensité plus faible est liée à une période estivale plus longue durant laquelle le nombre de déplacements diminue.

Tableau 5.10 Taille et centre de chaque groupe (distribution moyenne des déplacements annuels par mois et intensité mensuelle moyenne normalisée) – Cas des abonnements annuels

Groupe		C1	C2	C3	C4	C5	C6	Total
Taille (% cartes-année)		1.4%	8.1%	19.3%	38.8%	18.6%	13.8%	100.0%
Distribution des déplacements annuels par mois	Janvier	8.3%	8.2%	8.3%	8.4%	9.3%	11.4%	9.0%
	Février	8.2%	8.2%	8.5%	8.8%	9.5%	9.5%	8.9%
	Mars	8.6%	8.7%	8.8%	8.8%	9.3%	8.8%	8.9%
	Avril	8.5%	8.6%	8.6%	8.8%	9.0%	8.2%	8.7%
	Mai	8.7%	8.6%	8.6%	8.7%	8.5%	7.4%	8.4%
	Juin	8.5%	8.2%	8.3%	8.3%	8.0%	7.4%	8.1%
	Juillet	8.3%	8.2%	7.8%	7.2%	6.8%	6.7%	7.3%
	Août	8.4%	8.3%	8.2%	7.7%	7.0%	6.5%	7.5%
	Septembre	8.2%	8.3%	8.3%	8.5%	8.0%	7.2%	8.1%
	Octobre	8.2%	8.3%	8.3%	8.4%	8.1%	7.8%	8.2%
	Novembre	8.3%	8.6%	8.8%	9.1%	9.1%	8.8%	9.0%
	Décembre	7.8%	7.8%	7.6%	7.3%	7.5%	10.4%	7.9%
Moyenne mensuelle normalisée		3.66	2.43	1.82	1.40	0.93	0.31	1.36

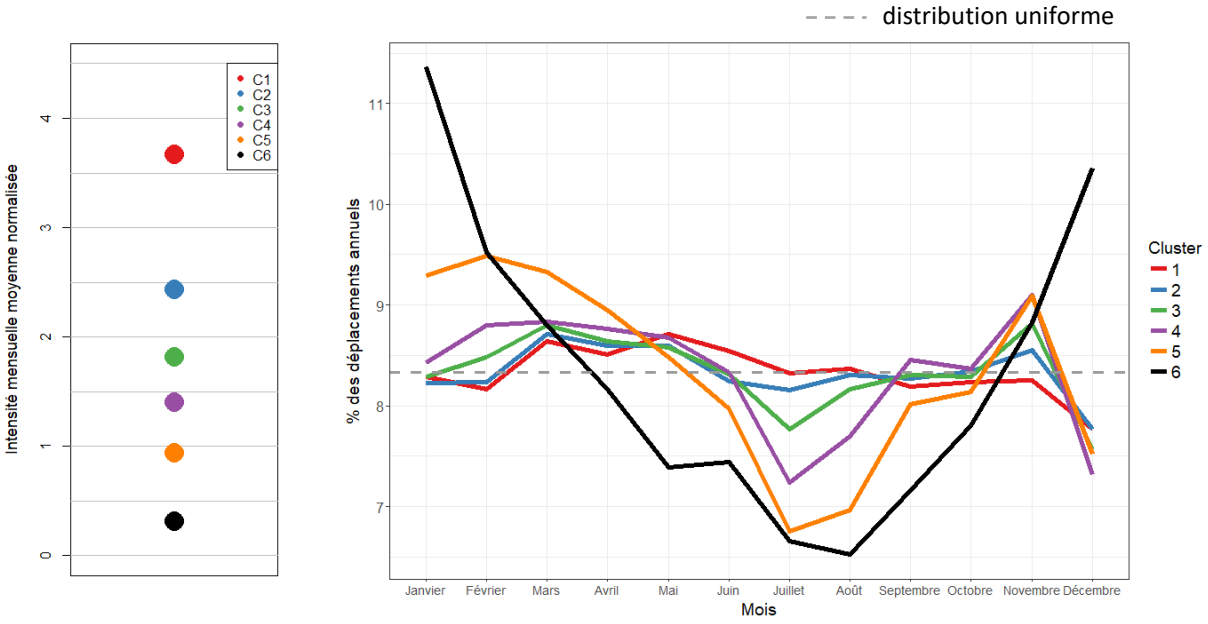


Figure 5.9 Représentation des 6 centres : intensité mensuelle normalisée (à gauche) et distribution des déplacements annuels par mois (à droite) – Cas des abonnements annuels

Dans les groupes C5 et C6, cette tendance est plus apparente et conduit à des proportions de déplacements plus élevées aux deux extrémités de l'année (car la somme des 12 points doit être égale à 1). De plus, un pic de déplacements en novembre et une chute en décembre sont visibles pour presque tous les groupes. Seuls les deux groupes opposés C1 et C6 présentent des profils mensuels plus atypiques. Les usagers du groupe C1, en plus d'être très fréquents, sont caractérisés par une répartition presque uniforme de leurs déplacements, avec des proportions légèrement plus élevées dans la première partie que dans la deuxième partie de l'année et une diminution plus abrupte en décembre. Les usagers du groupe C6 voient quant à eux leur nombre de déplacements diminuer pendant la période estivale, mais pas pendant la période hivernale.

De plus, les Tableau 5.11 et Tableau 5.12 permettent d'analyser la composition tarifaire des cartes de chaque groupe. Les usagers de l'échantillon considéré n'ont validé qu'un seul titre durant toute l'année : leur abonnement annuel, avec un certain type de tarif. La distribution des tarifs utilisés est donc précisée dans ces tableaux.

Tableau 5.11 Distribution des déplacements de chaque groupe par type de tarifs – Cas des abonnements annuels

	Groupe	C1	C2	C3	C4	C5	C6	Tous
	Nb moyen de titres	1.0	1.0	1.0	1.0	1.0	1.0	1.0
TYPES DE TARIFS	**Gratuité**	25.3%	5.0%	4.4%	4.1%	11.2%	31.6%	6.8%
	Tarif étudiant	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Tarif réduit	3.7%	2.1%	1.3%	1.0%	2.2%	1.8%	1.5%
	Tarif ordinaire	71.1%	92.9%	94.3%	94.9%	86.6%	66.6%	91.7%
	Tarif spécial	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Tableau 5.12 Distribution des déplacements par type de tarifs dans chaque groupe – Cas des abonnements annuels

	Groupe	C1	C2	C3	C4	C5	C6	Total
	Nb moyen de titres	1.0	1.0	1.0	1.0	1.0	1.0	1.0
TYPES DE TARIFS	**Gratuité**	13.8%	10.8%	16.8%	24.0%	20.9%	13.8%	100.0%
	Tarif étudiant	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
	Tarif réduit	8.9%	20.4%	21.6%	26.8%	18.7%	3.5%	100.0%
	Tarif ordinaire	2.9%	14.8%	26.6%	41.5%	12.0%	2.2%	100.0%
	Tarif spécial	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

La majorité des déplacements ont été effectués avec des abonnements à tarifs ordinaires ou gratuits; la part des tarifs réduits est moindre et celle des autres tarifs est nulle. Dans chaque groupe les usagers ont principalement utilisé des tarifs ordinaires, mais on relève des proportions de déplacements faits avec des abonnements gratuits plus importantes pour les groupes C1 et C6. Ces abonnements gratuits sont notamment des abonnements offerts par les entreprises à leurs employés (voire à leurs anciens employés désormais retraités), ceux du groupe C1 s'en servant beaucoup plus que ceux du groupe C6. Dans le sens inverse (Tableau 5.12), les déplacements par type de tarif sont répartis proportionnellement à la taille des groupes. Ainsi, la majorité des déplacements faits avec un tarif ordinaire ou gratuit ont été réalisés par les usagers du groupe C4 (groupe le plus volumineux). Cependant, seuls 2.2% des déplacements effectués avec un tarif ordinaire ont été réalisés par les usagers du groupe C6 du fait leur plus faible intensité d'utilisation.

5.3.2 Quelques indicateurs d'utilisation

Pour aider à l'interprétation de la typologie obtenue, des indicateurs caractérisant l'utilisation temporelle et spatiale du transport en commun sont évalués dans chaque groupe. Quelques-uns seulement des indicateurs proposés précédemment ont été sélectionnés afin d'abréger ce chapitre, d'autant plus que certains indicateurs se sont révélés peu intéressants ou redondants dans l'étude de la typologie précédente. De même, les indicateurs de variabilité du Chapitre 4 ne seront pas calculés ici, car l'analyse de la variabilité des utilisateurs d'abonnements annuels sera largement traitée avec d'autres méthodes dans la suite de ce mémoire.

Les résultats de l'application des indicateurs choisis sont rapportés dans le Tableau 5.13. La taille (en proportion de cartes-année) et la distribution des déplacements de 2016 parmi les six groupes sont données dans la première partie de ce tableau, puis des indicateurs temporels et spatiaux sont séparés en deux autres parties. Les deux groupes les plus fréquents et les plus constants dans le temps (du moins au niveau mensuel), C1 et C2, correspondent à peine à 10% des cartes, mais ils ont réalisé ensemble près de 19% des déplacements totaux. Les groupes C4, puis C3, sont les plus volumineux et comptabilisent à eux deux plus de 65% des déplacements totaux. Comme les membres de ces groupes sont les plus représentés, leur intensité d'utilisation moyenne est égale ou très proche de la moyenne générale calculée avec toutes les cartes. Néanmoins, ce nombre de déplacements par mois actif (normalisé) varie nettement entre le groupe C1 et le groupe C6, signifiant différents niveaux d'utilisation. En outre, on remarque que le ratio d'activité est très élevé

et similaire entre les quatre voire cinq premiers groupes (avec des coefficients de variations très faibles à l'intérieur de chaque groupe), mais ce ratio est un peu plus bas pour le groupe C6. Les utilisateurs des cartes de ce dernier groupe ont tendance à être actifs moins souvent durant leur amplitude de 12 mois, notamment en milieu d'année.

Tableau 5.13 Quelques indicateurs d'utilisation calculés dans chacun des six groupes d'utilisateurs d'abonnements annuels

	INDICATEUR	C1	C2	C3	C4	C5	C6	Total
DISTRI-BUTIONS	**Répartition des cartes-année**	1.4%	8.1%	19.3%	38.8%	18.6%	13.8%	100%
	Répartition des déplacements	3.7%	14.6%	25.9%	40.1%	12.7%	3.0%	100%
INDICATEURS TEMPORELS	**Nb moyen de déplacements par mois actif (normalisé)**	2.7	1.8	1.3	1.0	0.7	0.2	1.0
	Coefficient de variation	*(18.4%)*	*(9.8%)*	*(7.9%)*	*(8.3%)*	*(16.8%)*	*(53.4%)*	*(47.7%)*
	Ratio moyen nb de mois actifs/amplitude	0.998	0.999	0.999	0.997	0.985	0.873	0.978
	Coefficient de variation	*(2.7%)*	*(1.5%)*	*(1.5%)*	*(2.4%)*	*(5.9%)*	*(20.5%)*	*(8.6%)*
	Proportion moyenne JO	75.2%	78.5%	86.6%	94.5%	92.3%	86.6%	89.9%
	Coefficient de variation	*(11.2%)*	*(9.6%)*	*(9.7%)*	*(8.0%)*	*(11.4%)*	*(18.6%)*	*(12.4%)*
INDICATEURS SPATIAUX	**Proportion moyenne dépl. métro-bus combinés**	26.2%	31.4%	29.5%	20.7%	16.5%	11.3%	21.2%
	Coefficient de variation	*(86.2%)*	*(80.3%)*	*(96.8%)*	*(139.8%)*	*(148.1%)*	*(154.6%)*	*(127.4%)*
	Proportion moyenne dépl. métro uniquement	57.6%	50.1%	56.0%	68.6%	65.3%	71.5%	64.3%
	Coefficient de variation	*(60.7%)*	*(68.5%)*	*(64.1%)*	*(53.4%)*	*(56.8%)*	*(45.2%)*	*(56.6%)*
	Nb moyen de stations métro différentes par usager	32.2	26.6	21.3	14.2	13.5	10.7	16.2
	Coefficient de variation	*(39.2%)*	*(31.3%)*	*(34.2%)*	*(47.7%)*	*(50.7%)*	*(51.1%)*	*(52.8%)*
	Nb moyen de lignes bus différentes par usager	25.4	22.2	15.2	8.0	7.7	4.8	10.3
	Coefficient de variation	*(68.4%)*	*(51.1%)*	*(60.0%)*	*(93.4%)*	*(90.8%)*	*(99.8%)*	*(92.9%)*

La proportion de déplacements effectués en jours ouvrables est très importante dans les groupes C4 et C5. Une bonne partie des usagers de ces deux groupes utilisent donc essentiellement le transport en commun pendant les jours où ils travaillent. Cette proportion est plus faible pour les cartes des groupes C1 et C2 qui empruntent aussi le transport en commun pendant les jours non ouvrables, d'où peut-être leur plus fort taux d'utilisation du transport en commun. Au niveau spatial (ou modal), la majorité des déplacements se font uniquement en métro, notamment pour le groupe C6. La répartition des déplacements du groupe C2 est la moins « mono-mode ». De plus, le nombre

moyen de stations de métro et de lignes de bus empruntées au cours de l'année diminue avec la fréquence d'utilisation, de C1 à C6. Les lieux d'embarquement de C1 et C2 sont particulièrement diversifiés: les usagers ont utilisé au moins une fois presque la moitié des stations de métro du réseau (composé de 68 stations au total). Ainsi, comme déjà remarqué précédemment, un nombre plus important de déplacements semble encourager un taux d'acquisition plus élevé du réseau.

5.3.3 Vérifications statistiques

Enfin, des statistiques sont calculées pour quantifier les différences intergroupes observées précédemment avec les indicateurs. Comme les groupes manipulés dans cette section sont un peu moins gros, certaines différences testées se révèlent non significatives (valeur-p > 0.05) : celles-ci correspondent aux cases rouges et orange du Tableau 5.14. En croisant ces résultats avec ceux des indices de taille d'effet répertoriés dans le Tableau 5.15, on constate que les différences trouvées non significatives sont effectivement des différences peu importantes (taille d'effet < 0.1). Les valeurs-p supérieures au seuil de significativité sont situées aux mêmes endroits que les valeurs de taille d'effet les plus faibles. De même, de nombreuses différences significatives (cases vertes dans le Tableau 5.14) correspondent à des tailles d'effet très importantes (taille d'effet > 0.5).

Tableau 5.14 Résultats des tests statistiques appliqués sur les indicateurs temporels et spatiaux – Cas des abonnements annuels

Test	Nb moyen de dépl.par mois actif	Ratio d'activité	% moyenne JO	% moyenne dépl. métro-bus combinés	% moyenne dépl. métro uniquement	Nb moyen de stations métro par usager	Nb moyen de lignes de bus par usager
C1 v C2	0.000	0.019	0.000	0.000	0.000	0.000	0.000
C1 v C3	0.000	0.107	0.000	0.019	0.047	0.000	0.000
C1 v C4	0.000	0.382	0.000	0.000	0.000	0.000	0.000
C1 v C5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C1 v C6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C2 v C3	0.000	0.144	0.000	0.000	0.000	0.000	0.000
C2 v C4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C2 v C5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C2 v C6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C3 v C4	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C3 v C5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C3 v C6	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C4 v C5	0.000	0.000	0.000	0.000	0.000	0.000	0.143
C4 v C6	0.000	0.000	0.000	0.000	0.906	0.000	0.000
C5 v C6	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Tableau 5.15 Mesure de la taille d'effet pour les indicateurs temporels et spatiaux – Cas des abonnements annuels

Test	Nb moyen de dépl.par mois actif	Ratio d'activité	% moyenne JO	% moyenne dépl. métro-bus combinés	% moyenne dépl. métro uniquement	Nb moyen de stations métro par usager	Nb moyen de lignes de bus par usager
C1 v C2	0.61	0.03	0.16	0.08	0.09	0.16	0.06
C1 v C3	0.43	0.01	0.29	0.02	0.02	0.24	0.16
C1 v C4	0.31	0.01	0.27	0.07	0.06	0.26	0.19
C1 v C5	0.44	0.07	0.34	0.13	0.06	0.36	0.27
C1 v C6	0.50	0.27	0.28	0.22	0.12	0.45	0.36
C2 v C3	0.79	0.01	0.43	0.07	0.08	0.30	0.29
C2 v C4	0.66	0.04	0.55	0.24	0.23	0.50	0.47
C2 v C5	0.80	0.16	0.58	0.35	0.22	0.62	0.58
C2 v C6	0.84	0.51	0.41	0.46	0.32	0.75	0.72
C3 v C4	0.82	0.04	0.50	0.23	0.21	0.45	0.39
C3 v C5	0.87	0.19	0.42	0.30	0.17	0.51	0.43
C3 v C6	0.85	0.60	0.17	0.38	0.24	0.66	0.60
C4 v C5	0.81	0.17	0.06	0.03	0.05	0.05	0.01
C4 v C6	0.76	0.61	0.20	0.10	0.00	0.24	0.19
C5 v C6	0.86	0.47	0.16	0.08	0.06	0.21	0.21

Grâce à la réduction de la dimension des données, une bonne partie des résultats des tests statistiques coïncident ici avec ceux de la taille d'effet. Cependant, quelques différences trouvées significatives ne sont en réalité pas très importantes d'après la taille d'effet, en particulier pour les indicateurs de répartition des déplacements par mode (indicateurs pour lesquels les coefficients de variation intragroupes calculés dans le Tableau 5.13 étaient d'ailleurs très élevés).

Ainsi, les différences de fréquence d'utilisation (nombre moyen de déplacements par mois actif) sont moyennement ou très importantes pour toutes les paires de groupes comparés. Ces divergences sont dues à la segmentation elle-même, basée entre autres sur cet indicateur d'intensité. Au contraire, le ratio d'activité est similaire pour tous les groupes, sauf C6 ; les usagers de ce dernier groupe sont en effet moins actifs. Les disparités dans les proportions de déplacements faits en jours ouvrables, plus basses dans les groupes C1, C2 voire C3 et C6, sont également notables. La répartition des déplacements par mode est assez semblable dans tous les groupes, mais C4, C5 et C6 se distinguent modérément ou fortement des autres groupes par leur utilisation plus importante du métro. De même, le nombre moyen de stations de métro ou de lignes de bus est distinct entre C1, C2, C3 et les autres groupes. Cette différence est moins nette pour le groupe C1, sans doute à cause des plus grands coefficients de variations mesurés entre ses membres.

CHAPITRE 6 ANALYSE DE LA VARIABILITÉ INTRAPERSONNELLE: CRÉATION D'UNE TYPOLOGIE DE SEMAINES

Ce sixième chapitre porte sur la variabilité intrapersonnelle des utilisateurs du transport en commun. L'échelle de temps étudié est la semaine : il s'agit donc d'analyser les variations du comportement hebdomadaire de chaque usager au cours de l'année 2016. Une typologie de semaines est d'abord construite puis la régularité intrapersonnelle de chaque usager est évaluée en mesurant l'appartenance de ses semaines de déplacements aux mêmes groupes types. Ainsi, un usager est dit régulier au niveau intrapersonnel si sa mobilité se ressemble sur toutes les semaines de l'année. Seul le cas des abonnements annuels avec une amplitude de 12 mois est examiné ici, car la granularité hebdomadaire des analyses effectuées nécessite une réduction de la dimension des données. L'organisation de ce chapitre est symétrique à celle de la section 5.3 du chapitre précédent : la typologie de semaines obtenue est d'abord présentée, puis des indicateurs de variabilité intrapersonnelle sont proposés et vérifiés statistiquement.

6.1 Typologie de semaines

6.1.1 Méthode de segmentation utilisée

La méthode de segmentation appliquée pour créer une typologie de semaines est la même méthode que celle qui a été décrite dans la section 5.1 pour la typologie d'usagers. L'approche par K-moyennes est donc mise en œuvre avec l'algorithme de Lloyd, la métrique adoptée étant la distance euclidienne. Comme précédemment, dix initialisations aléatoires sont réalisées et les deux mêmes critères (inertie expliquée et dendrogramme sur 30 groupes préalablement calculés) sont utilisés pour le choix du nombre de groupes K.

La seule différence est que cette méthode n'est ici pas appliquée sur une base de cartes-année, mais sur une base de cartes-semaine, composée de 56,988×51 = 2,906,388 vecteurs à huit dimensions dans le cas de l'échantillon analysé (abonnements annuels avec une amplitude de 12 mois). Chaque ligne (ou vecteur, ou carte-semaine) de cette base correspond à une semaine de déplacements d'un usager, décrite au niveau journalier par les proportions de déplacements faits dans chacun des sept jours de la semaine et par un indicateur d'intensité moyenne quotidienne (le nombre moyen de déplacements par jour actif pour la semaine considérée). Seules les 51 semaines complètes de

l'année 2016 sont prises en compte puisque, pour pouvoir composer des vecteurs de même dimension (contrainte exigée par l'algorithme de segmentation utilisé), il est nécessaire d'avoir les nombres de déplacements de chaque carte-semaine du lundi au dimanche. Chaque semaine (ou vecteur) étant traitée séparément dans le processus de segmentation, les 51 semaines d'un même usager peuvent se retrouver dans différents groupes. L'objectif sera justement de mesurer la régularité de l'appartenance de ces 51 semaines aux mêmes groupes.

6.1.2 Résultats

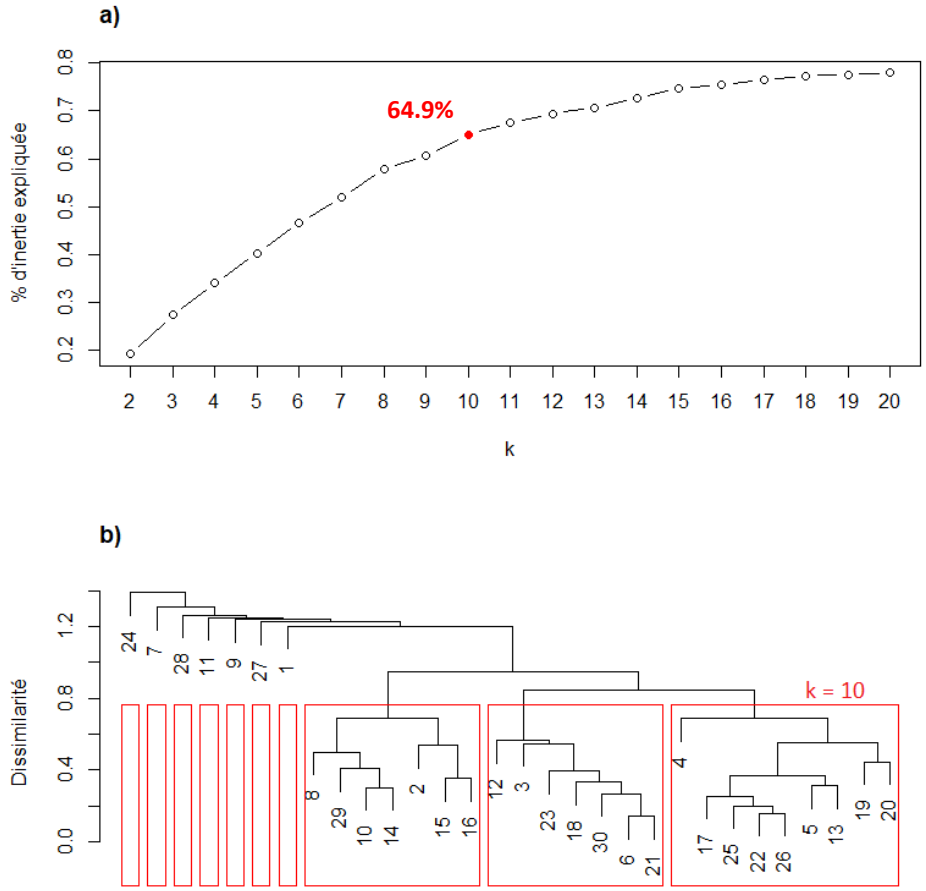


Figure 6.1 Choix du nombre de groupes K avec a) le pourcentage d'inertie expliquée b) un dendrogramme réalisé sur 30 groupes initiaux – Typologie de semaines

Les résultats des deux critères utilisés pour déterminer le nombre de groupes à constituer sont rapportés sur la Figure 6.1. Ces résultats conduisent à fixer $K = 10$ types de semaines. En effet, le coude du graphique de l'inertie expliquée est peu dessiné, mais il commence à se former à partir de $K = 8$. Un regain d'informations est néanmoins produit (la pente réaugmente) pour $K = 10$

avant de ralentir. De plus, le dendrogramme créée d'abord sept groupes individuels très distincts (sur la gauche du graphe) avant de rassembler tous les groupes restants dans un seul et même ensemble. Le choix de $K = 10$ permet de diviser en trois cet ensemble pour mieux l'analyser.

Dans la suite, on note Wi le $i^{\text{ème}}$ groupe de la typologie obtenue (W pour *week*), avec i allant de 1 à 10. La taille (en pourcentage de cartes-semaine) et les centres des 10 groupes produits, calculés comme la moyenne des vecteurs à l'intérieur de chaque groupe, sont donnés dans le Tableau 6.1. Ces centres sont également illustrés sur la Figure 6.2 avec l'intervalle interquartile (intervalle entre le premier et le troisième quartile) de chaque variable. De plus, la valeur médiane (deuxième quartile) de chaque variable est rapportée dans le Tableau 6.2.

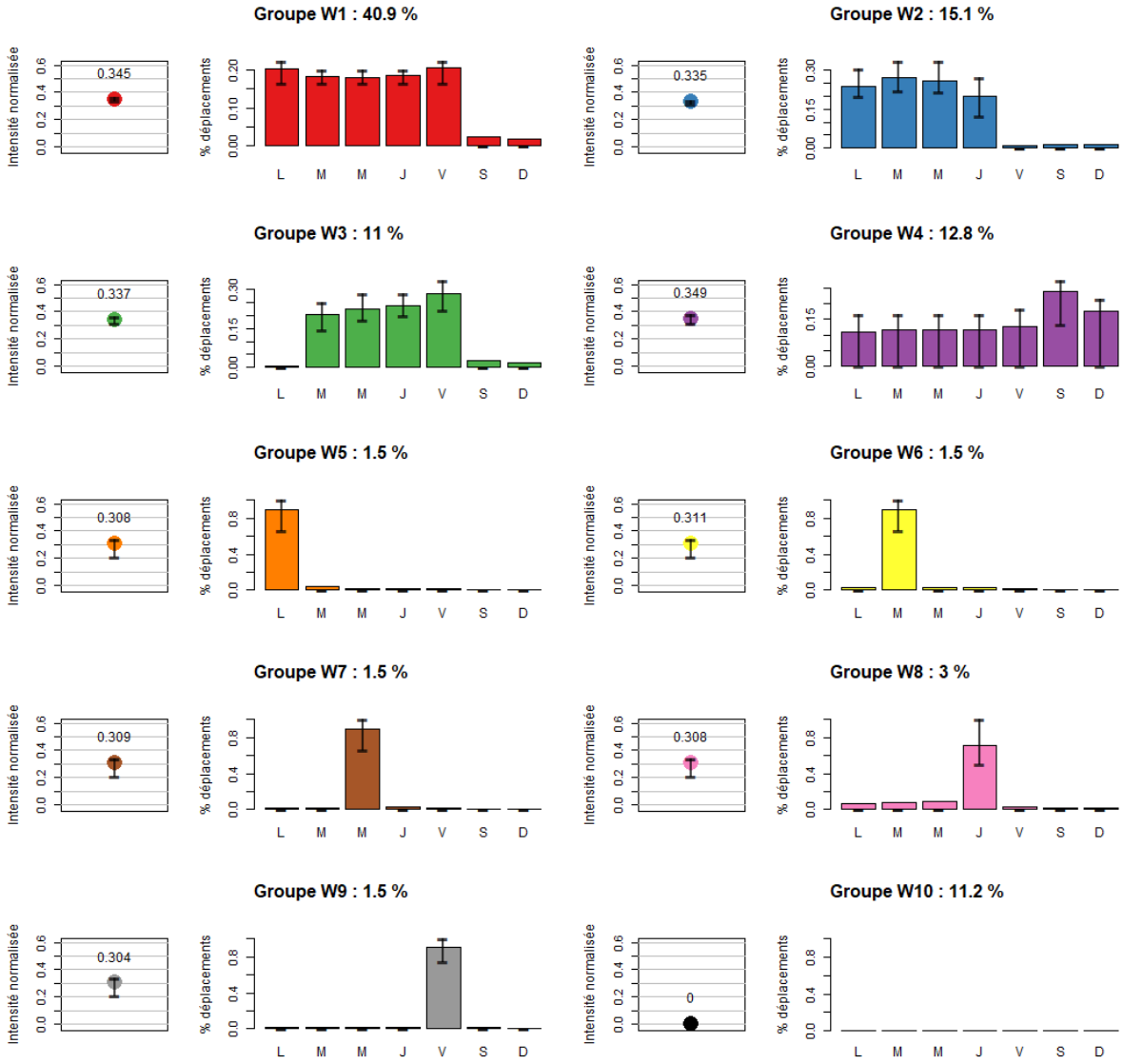


Figure 6.2 Représentation des 10 centres de la typologie de semaines

Le premier groupe W1 est de loin le plus populaire : il rassemble à lui seul plus de 40% de toutes les cartes-semaine considérées. Ce groupe représente la semaine typique de travail, avec des déplacements qui se font surtout pendant la semaine et très peu en fin de semaine. D'après les valeurs des quartiles, la majorité des déplacements des cartes-semaine appartenant à ce groupe sont répartis uniformément entre les 5 premiers jours de la semaine : les proportions médianes sont d'ailleurs toutes égales à 20% du lundi au vendredi, et nulles pour les deux jours restants en fin de semaine. En moyenne, en revanche, les proportions de déplacements sont plus élevées aux deux extrémités de la semaine de travail (le lundi et le vendredi), impliquant que certains usagers se déplacent beaucoup plus pendant ces deux jours. Au contraire, très peu de déplacements sont effectués le vendredi dans le groupe W2, ou le lundi dans le groupe W3. Ces semaines types peuvent correspondre à des semaines avec un jour férié ou bien à des comportements hebdomadaires caractéristiques de personnes qui ne travaillent que 4 jours ouvrables par semaine.

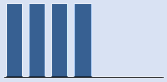
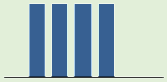
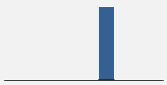
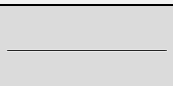
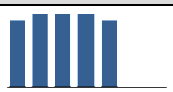
Tableau 6.1 Taille (en % de cartes-semaine) et centre de chaque groupe (distribution moyenne des déplacements par jour et intensité quotidienne moyenne normalisée)

Groupe	Taille	L	M	M	J	V	S	D	Intensité
W1	40.9%	20.3%	18.3%	18.1%	18.6%	20.5%	2.3%	1.8%	0.345
W2	15.1%	23.8%	27.0%	25.7%	20.0%	0.8%	1.3%	1.4%	0.335
W3	11.0%	0.6%	20.3%	22.5%	23.9%	28.4%	2.5%	1.7%	0.337
W4	12.8%	11.0%	11.7%	11.6%	11.6%	12.5%	23.8%	17.7%	0.349
W5	1.5%	89.7%	3.5%	2.0%	1.9%	1.6%	0.7%	0.6%	0.308
W6	1.5%	3.1%	89.1%	2.2%	2.4%	1.9%	0.8%	0.6%	0.311
W7	1.5%	2.1%	2.1%	89.5%	2.9%	2.1%	0.8%	0.6%	0.309
W8	3.0%	6.1%	8.0%	9.1%	71.4%	2.2%	1.8%	1.4%	0.308
W9	1.5%	1.7%	1.3%	1.4%	2.1%	91.3%	1.3%	0.8%	0.304
W10	11.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.000
Total	100.0%	15.0%	17.0%	16.9%	17.0%	14.7%	4.6%	3.5%	0.301

Dans le groupe W4, les déplacements sont plutôt accumulés en fin de semaine, supposant une utilisation du transport en commun pour des motifs autres que le travail. Néanmoins, l'intervalle interquartile révèle une grande variation du nombre de déplacements le dimanche. De même, les valeurs médianes du Tableau 6.2 montrent qu'en réalité la plupart des déplacements des cartes-semaine de ce groupe sont réalisés le samedi. Quelques déplacements sont également faits en semaine, leur répartition étant croissante du lundi au vendredi. Alors que les groupes W2, W3 et

W4 sont assez homogènes en taille et réunissent 10% à 15% des cartes-semaine, les groupes suivants, de W5 à W9, sont plus minoritaires. Dans ces groupes plus atypiques, les déplacements sont concentrés sur un seul jour de la semaine, du lundi au vendredi respectivement. Les valeurs médianes des proportions de déplacements réalisés sur ces jours sont même égales à 100% des déplacements. Enfin, aucun déplacement n'est enregistré pendant la semaine type W10. Ce groupe contient tous les vecteurs nuls (0,0,0,0,0,0,0,0,0,0) de la base initiale, soit 12% des cartes-année. Ce faible pourcentage confirme que les utilisateurs d'abonnements annuels avec une amplitude de 12 mois sont très actifs durant l'année.

Tableau 6.2 Valeur médiane de chaque variable utilisée dans le processus de segmentation pour les dix groupes de semaines

Groupe	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche	Graphique	Intensité
W1	20.0%	20.0%	20.0%	20.0%	20.0%	0.0%	0.0%		0.340
W2	25.0%	25.0%	25.0%	25.0%	0.0%	0.0%	0.0%		0.340
W3	0.0%	25.0%	25.0%	25.0%	25.0%	0.0%	0.0%		0.340
W4	12.5%	12.5%	12.5%	12.9%	13.3%	20.0%	14.3%		0.340
W5	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%		0.340
W6	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%		0.340
W7	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%		0.340
W8	0.0%	0.0%	0.0%	60.0%	0.0%	0.0%	0.0%		0.340
W9	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%		0.340
W10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%		0.000
Total	16.7%	18.2%	18.2%	18.2%	16.7%	0.0%	0.0%		0.340

En termes de fréquence d'utilisation quotidienne, l'indicateur d'intensité moyenne normalisée (le nombre moyen de déplacements par jour actif) est assez similaire d'un groupe à l'autre. Cette

variable a donc peu été prise en compte dans la segmentation, peut-être du fait de sa normalisation logarithmique (choisie pour atténuer les nombreuses valeurs aberrantes). Les semaines types W1 et W4 présentent les intensités moyennes les plus hautes malgré leurs utilisations complètement différentes du transport en commun (semaine versus fin de semaine). Toutefois, les valeurs médianes de cet indicateur d'intensité quotidienne sont égales pour tous les groupes, sauf pour le groupe W10 caractérisé par une intensité nulle.

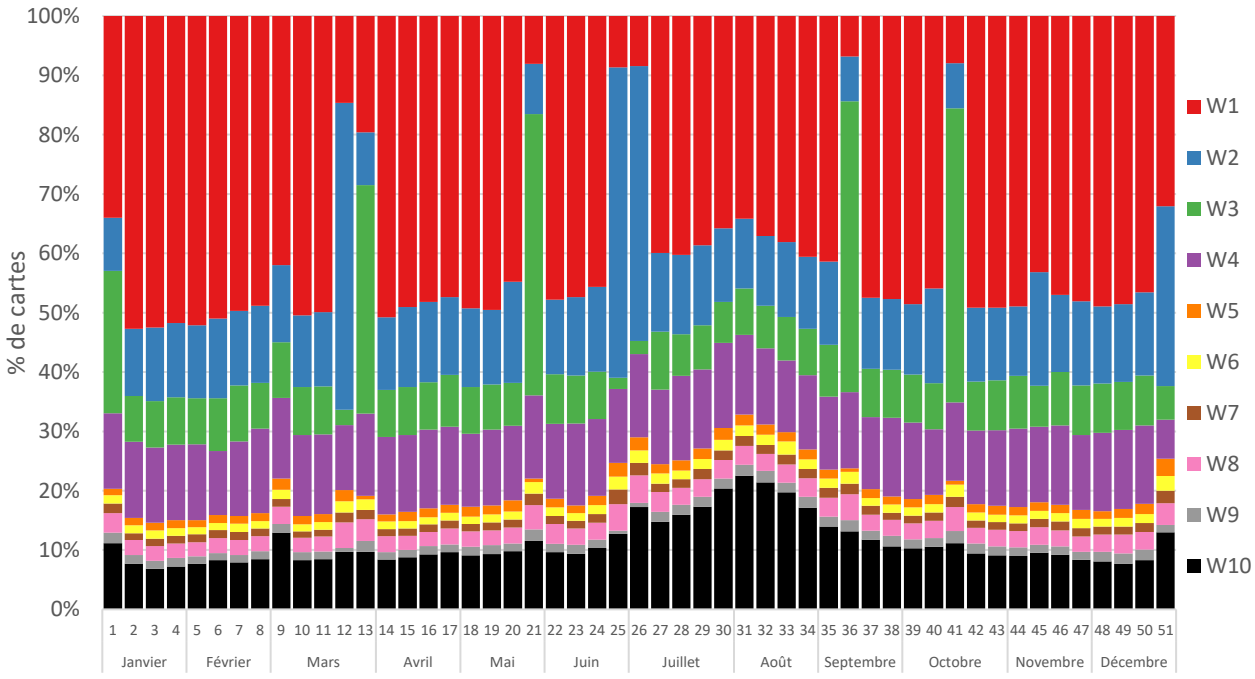


Figure 6.3 Distribution des 51 semaines de l'année par carte dans les 10 groupes

De plus, la distribution de toutes les cartes-semaines dans les 10 groupes est représentée sur la Figure 6.3 pour chacune des 51 semaines de l'année étudiées. L'axe des y est en fait un pourcentage de cartes (et non de cartes-semaine), car chaque carte est associée à une seule semaine n°1, une seule semaine n°2, …, et une seule semaine n°51. Sur ce graphique on constate tout d'abord la prédominance du groupe W1 dans une grande partie des semaines de l'année. À l'inverse, les groupes W5 à W9, dans lesquels les déplacements sont concentrés sur un seul jour ouvrable de la semaine, occupent une petite part de la distribution : en moyenne, pour chacune des 51 semaines considérées, un peu moins de 9% des cartes appartiennent à ces cinq groupes. Cette part reste néanmoins relativement constante au cours de l'année, tout comme celle du groupe W4. En revanche, la proportion des cartes appartenant au groupe W1 diminue et celle de W10 augmente pendant la période estivale. En effet, il a été vu précédemment que les utilisateurs des cartes

échantillonnées ont tendance à moins se déplacer durant l'été, conduisant à plus de semaines sans déplacements (toutes réunies dans W10). On remarque également une légère augmentation du nombre de cartes dans le groupe W10 durant les semaines 9 et 41, semaines qui correspondent à des périodes de congés scolaires (relâches des sessions d'hiver et d'automne), ainsi que durant la période des fêtes de fin d'année (semaines 1 et 51), et durant des semaines avec un jour férié (semaines 21 et 26). Le groupe W2, dans lequel sont rassemblées les semaines avec peu de déplacements le lundi, contient quant à lui plus de cartes pendant les semaines 12, 25 et 26. En croisant ces numéros de semaines avec le calendrier en ANNEXE D, on constate qu'ils coïncident avec des jours fériés survenus un vendredi (Vendredi Saint, Fête nationale du Québec et Fête du Canada). De même, la proportion des semaines appartenant au groupe W3 augmente pendant les semaines 13, 21, 36 et 41, car celles-ci comportent des lundis fériés (Lundi de Pâques, Journée nationale des patriotes, Fête du travail et Action de Grâce).

Ainsi, la répartition des semaines dans les 10 groupes change au cours de l'année. Il existe donc des variations intrapersonnelles dans le comportement hebdomadaire de certaines cartes. Dans le cas contraire, c'est-à-dire si toutes les 51 semaines de chaque carte avaient appartenu au même groupe, la distribution de la Figure 6.3 aurait été uniforme (mêmes proportions de cartes dans chaque groupe pour toutes les semaines de l'année). Toutefois, il est nécessaire d'évaluer la ressemblance entre les semaines d'un même usager qui appartiennent à plusieurs groupes : si un usager se déplace avec un comportement de type W1 pendant une semaine puis avec un comportement de type W2 pendant une autre semaine, son comportement est-il vraiment différent d'une semaine à l'autre ? Est-il du moins plus différent que s'il s'était déplacé avec un comportement de type W1 puis un comportement de type W5 ? En d'autres termes, à quel point le comportement hebdomadaire de cet usager varie en fonction de la similitude des groupes auxquels appartiennent ses semaines ? Pour pouvoir répondre à ces questions et donner un poids différent à chaque paire de semaine comparées, une matrice de dissimilarités est calculée entre les 10 types de semaines obtenus précédemment. Le but des travaux qui suivront étant d'insister sur l'utilisation de ces résultats plus que sur ces résultats eux-mêmes, la métrique la plus simple est choisie, c'est-à-dire la distance euclidienne, malgré ses nombreuses limitations qui seront discutées plus tard dans ce mémoire (Chapitre 8). La dissimilitude entre deux types de semaine W_i et W_j est alors estimée par la distance mesurée entre les centres de ces deux groupes, chaque centre étant un

vecteur à 8 dimensions. Cette distance, définie par l'équation suivante, est basée sur la somme des différences au carré entre toutes les dimensions :

$$d_E(W_i, W_j) = \sqrt{\sum_{p=1}^{8} (w_{i,p} - w_{j,p})^2} \qquad \text{(Éq. 25)}$$

où $w_{i,p}$ est la dimension p du vecteur représentant le centre du $i^{\text{ème}}$ groupe. Toutes les distances ainsi calculées pour chaque paire de groupes sont répertoriées dans la matrice du Tableau 6.3. Cette dernière sera utilisée dans la suite de ce chapitre, mais aussi dans le Chapitre 7. Le gradient de couleurs vert-jaune-rouge est attribué dans l'ordre croissant des valeurs obtenues. Cette matrice est symétrique puisque la distance entre W_i et W_j est la même que celle entre W_j et W_i. De plus, la diagonale est nulle car la distance d'un groupe avec lui-même est égale à zéro.

Tableau 6.3 Matrice des distances euclidiennes entre les 10 types de semaines

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
W1	0.000	0.232	0.224	0.316	0.770	0.786	0.793	0.593	0.787	0.430
W2	0.232	0.000	0.371	0.397	0.761	0.718	0.739	0.600	1.013	0.486
W3	0.224	0.371	0.000	0.376	0.992	0.795	0.772	0.574	0.724	0.480
W4	0.316	0.397	0.376	0.000	0.860	0.846	0.852	0.668	0.860	0.396
W5	0.770	0.761	0.992	0.860	0.000	1.218	1.238	1.091	1.256	0.898
W6	0.786	0.718	0.795	0.846	1.218	0.000	1.233	1.067	1.253	0.892
W7	0.793	0.739	0.772	0.852	1.238	1.233	0.000	1.059	1.255	0.896
W8	0.593	0.600	0.574	0.668	1.091	1.067	1.059	0.000	1.134	0.727
W9	0.787	1.013	0.724	0.860	1.256	1.253	1.255	1.134	0.000	0.914
W10	0.430	0.486	0.480	0.396	0.898	0.892	0.896	0.727	0.914	0.000

Les distances les plus grandes (en rouge) sont mesurées parmi les groupes W5, W6, W7, W8 et W9. Ces profils hebdomadaires sont effectivement très différents les uns des autres puisque les déplacements sont concentrés sur des jours uniques et distincts. La comparaison par paires de ces groupes entraîne donc la mise au carré de deux proportions de déplacements élevées, résultant en une distance euclidienne plus importante. Ces semaines types sont également assez distantes des autres groupes. La dissimilarité évaluée entre W3 et W5 est en particulier assez haute puisque ces deux groupes sont totalement opposés : dans le premier groupe, il n'y a pas de déplacements le lundi et dans l'autre il n'y a des déplacements que le lundi. De même pour la distance entre W2 et W9 avec l'absence ou la présence de déplacements le vendredi. Les distances obtenues entre les groupes W5 à W9 d'une part et W10 (aucun déplacement) ou W4 (déplacements en fin de semaine)

d’autre part sont aussi notables. Le fait de se déplacer ou non et le type de jour privilégié (ouvrable versus non ouvrable) sont ainsi pris en compte dans l’évaluation de la dissimilitude de ces groupes.

Les distances les plus faibles sont associées aux comparaisons des groupes W1, W2, W3 et W4 entre eux, groupes pour lesquels les déplacements sont mieux répartis dans la semaine. Les distances calculées entre W4 (déplacements en fin de semaine) et les autres semaines types (déplacements en jours ouvrables) tendent néanmoins à être plus élevées. De plus, on remarque que les distances entre W1, W2, W3 ou W4 (déplacements répartis sur la semaine ou en fin de semaine) d’une part et W10 (pas de déplacements) d’autre part, sont inférieures aux distances mesurées entre W1, W2, W3 ou W4 d’une part, et W5, W6, W7, W8 ou W9 (déplacement sur un seul jour) d’autre part. Ainsi, la différence entre une répartition assez régulière des déplacements versus une concentration des déplacements sur un seul jour est supposée plus importante que la différence observée entre la non-mobilité et la mobilité.

6.2 Indicateurs de variabilité intrapersonnelle

Dans la section précédente, chacune des 51 semaines de chaque utilisateur a été associée à un groupe (ou type) de semaines. La variabilité intrapersonnelle de l’utilisation du transport en commun peut donc maintenant être mesurée grâce à des indicateurs basés sur la répartition des semaines de chaque usager dans les 10 groupes formés. Une fois définis, ces indicateurs seront appliqués pour comparer les six groupes d’usagers construits dans le chapitre précédent (section 5.3 Cas des utilisateurs d’abonnements annuels), puis les différences mises en évidence entre ces groupes seront analysées et contrôlées.

6.2.1 Définition des indicateurs

La régularité intrapersonnelle d'un utilisateur donné est ici définie par la répétition des mêmes types de semaines dans son comportement au cours de l’année, c’est-à-dire par l’appartenance récurrente de ses semaines aux mêmes groupes. Plusieurs indicateurs basés sur cette définition sont proposés dans cette section. Ces indicateurs présentent une certaine analogie avec les indicateurs spatiaux (ou modaux) énoncés précédemment dans la section 5.2.2.2. En effet, comme la régularité spatiale était évaluée par la concentration des validations dans les mêmes stations de métro ou lignes de bus, la régularité intrapersonnelle est ici mesurée par la concentration des semaines dans les mêmes groupes ou types de semaines. Ainsi, plus le nombre de groupes auxquels appartiennent les

semaines de l'utilisateur est faible, moins son comportement hebdomadaire est diversifié et plus il est dit régulier au niveau intrapersonnel. Les indicateurs présentés ci-dessous seront calculés comme une moyenne d'indicateurs individuels à l'intérieur de chacun des six groupes de cartes (C1 à C6) produits précédemment. Ils permettront ainsi de quantifier la variabilité intrapersonnelle moyenne des membres de ces groupes basée sur un cycle hebdomadaire. Ces indicateurs sont :

- Le **nombre moyen de groupes** auxquels appartiennent les 51 semaines de chaque usager.
- Le **nombre moyen de semaines sans déplacements**, c'est-à-dire le nombre moyen de semaines qui appartiennent au groupe W10. Cet indicateur mesure l'immobilité moyenne des usagers du groupe.
- La **proportion moyenne des semaines qui appartiennent au groupe le plus populaire** de chaque usager. Cet indicateur mesure la concentration des semaines des usagers dans leur groupe de semaines le plus fréquent.
- La **proportion moyenne des semaines qui appartiennent aux deux groupes les plus populaires** de chaque usager. Cet indicateur mesure la concentration des semaines des usagers dans leurs deux groupes de semaines les plus fréquents.
- Le **nombre moyen de groupes nécessaires pour contenir 80% des semaines** de chaque usager.
- L'**entropie moyenne** appliquée aux proportions de semaines de chaque usager dans les 10 groupes. On utilise ici la même entropie normalisée que celle définie dans la section 4.3.1.4 (entropie de Shannon), avec n le nombre de groupes différents possibles ($n = 10$ ici) et P_{ij} la proportion des semaines de la carte i qui appartiennent au groupe de semaines W_j. Cet indicateur mesure la diversité des comportements hebdomadaires individuels.
- La **distance moyenne** (ou variabilité pondérée moyenne) entre les groupes de toutes les semaines d'un même usager. Cet indicateur est basé sur les distances euclidiennes du Tableau 6.3. Il est calculé comme la moyenne de toutes les comparaisons (uniques) entre toutes les semaines d'un même usager, chaque semaine étant comparée à toutes les autres semaines de l'usager. Le résultat de chaque comparaison est la distance euclidienne mesurée entre les groupes auxquelles appartiennent les deux semaines comparées. Ainsi, pour un usager u donné, cet indicateur est défini par l'équation suivante :

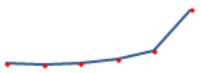
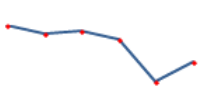
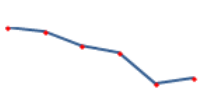
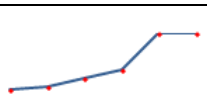
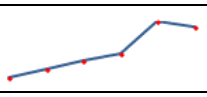
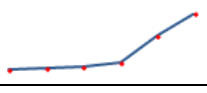
$$D^{u}_{moy} = \frac{1}{n_c}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_E(W_{(u,i)}, W_{(u,j)}) \qquad \text{(Éq. 26)}$$

avec N = 51 semaines, $n_c = \frac{N(N-1)}{2}$ le nombre de comparaisons uniques possibles entre toutes les semaines de l'usager, $d_E(.,.)$ la distance euclidienne et $W_{(u,k)}$ le groupe de semaines auquel appartient la $k^{\text{ème}}$ semaine de l'usager u. Cet indicateur mesure l'instabilité ou la variabilité moyenne de l'appartenance des semaines d'un usager donné aux mêmes groupes. Il prend en compte la quantité de semaines dans chaque groupe, mais aussi la ressemblance de ces groupes puisqu'un poids différent est attribué à chaque paire de groupes de semaines comparées.

6.2.2 Application des indicateurs

Les résultats du calcul des indicateurs définis à la section précédente dans chacun des six groupes d'utilisateurs d'abonnements annuels sont donnés dans le Tableau 6.4 ci-dessous. Les coefficients de variation de chaque indicateur à l'intérieur de chaque groupe sont également rapportés et la variabilité de chaque indicateur entre les groupes est illustrée par une courbe d'évolution.

Tableau 6.4 Indicateurs de variabilité intrapersonnelle calculés dans chacun des six groupes d'utilisateurs d'abonnements annuels

INDICATEUR	C1	C2	C3	C4	C5	C6	Évolution	Total
Nb moy de groupes	4.34	4.65	5.46	5.80	7.68	8.00		6.28
Coeff de variations	*(35.7%)*	*(25.7%)*	*(22.0%)*	*(25.0%)*	*(21.7%)*	*(21.3%)*		*(29.5%)*
Nb moy de semaines sans dépl.	1.30	0.90	1.37	2.89	5.98	22.63		5.72
Coeff de variations	*(204.4%)*	*(192.9%)*	*(128.6%)*	*(80.6%)*	*(83.6%)*	*(53.5%)*		*(152.9%)*
%moy semaines dans 1er groupe	61.7%	58.7%	59.7%	56.6%	41.6%	48.7%		53.6%
Coeff de variations	*(24.8%)*	*(21.7%)*	*(22.0%)*	*(22.4%)*	*(34.6%)*	*(41.5%)*		*(29.6%)*
%moy semaines dans 2 groupes	87.7%	85.5%	79.5%	76.5%	62.6%	64.7%		73.8%
Coeff de variations	*(12.5%)*	*(10.1%)*	*(10.6%)*	*(11.7%)*	*(21.2%)*	*(25.8%)*		*(18.3%)*
Nb moyen groupes 80% des semaines	2.10	2.17	2.44	2.67	3.81	3.82		2.95
Coeff de variations	*(32.0%)*	*(25.7%)*	*(27.7%)*	*(28.0%)*	*(31.3%)*	*(41.3%)*		*(39.0%)*
Entropie moy	0.40	0.44	0.49	0.52	0.69	0.66		0.56
Coeff de variations	*(35.5%)*	*(24.3%)*	*(22.0%)*	*(21.4%)*	*(21.1%)*	*(30.9%)*		*(28.5%)*
Distance moy	0.18	0.19	0.19	0.22	0.36	0.48		0.27
Coeff de variations	*(39.0%)*	*(27.5%)*	*(31.4%)*	*(32.9%)*	*(32.0%)*	*(31.0%)*		*(51.1%)*

Premièrement, on observe que le nombre moyen de groupes de semaines différents par usager augmente de C1 à C6. Les utilisateurs les plus fréquents, qui sont aussi les plus réguliers au niveau mensuel d'après l'uniformité de la distribution de leurs déplacements sur la Figure 5.9 , sont donc également les plus réguliers au niveau hebdomadaire puisque toutes leurs semaines appartiennent principalement aux mêmes types. Au contraire, les deux derniers groupes C5 et C6, composés d'utilisateurs plus occasionnels, ont des comportements hebdomadaires plus variés (en moyenne 8 types différents). De même, le nombre moyen de semaines sans déplacements, qui appartiennent donc au groupe W10, est plus élevé pour les derniers groupes, indiquant un plus grand taux d'immobilité parmi leurs membres. Ces résultats font écho à ce qui a été vu précédemment : les usagers de ces groupes ont tendance à moins se déplacer, surtout en milieu d'année. Cela est d'ailleurs particulièrement le cas pour le groupe C6 : en moyenne, 44% des 51 semaines des usagers de ce groupe sont des semaines sans déplacements. Malgré sa plus forte intensité d'utilisation du transport en commun, le nombre moyen de semaines non mobiles du groupe C1 est plus important que celui du groupe C2. Cependant, les coefficients de variation très élevés mesurés dans ces deux groupes suggèrent qu'il y a beaucoup de variabilité entre leurs membres.

En outre, les premiers groupes obtiennent les proportions de semaines appartenant au premier ou aux deux premiers types de semaines les plus fréquents les plus hautes. Les semaines des usagers de ces groupes sont donc plus concentrées dans certains types de semaine. La tendance de la courbe s'inverse pour C6, car le type de semaines le plus fréquent des usagers contenus dans ce groupe est généralement le type W10, et presque la moitié des semaines des usagers sont effectivement regroupées dans cet ensemble, car sans déplacements. De plus, la plupart (80%) des semaines des usagers sont réunies dans un nombre de groupes croissant de C1 à C6. Les usagers des premiers groupes ont en moyenne deux types de comportements hebdomadaires qui se répètent souvent au cours de l'année. La même tendance est révélée par l'indice d'entropie, plus bas pour ces premiers groupes, confirmant que leurs comportements hebdomadaires sont moins diversifiés. Leur utilisation hebdomadaire du transport en commun est donc plus stable au cours de l'année et, par conséquent, ils sont plus réguliers au niveau intrapersonnel d'après la définition adoptée.

Par ailleurs, la distance moyenne entre les groupes de toutes les semaines des usagers est plus petite pour les premiers groupes d'usagers que pour les derniers : les premiers groupes ont donc des comportements hebdomadaires qui se ressemblent plus au cours de l'année que les derniers groupes. La courbe de cet indicateur montre que les valeurs obtenues sont assez semblables entre

les quatre premiers groupes (C1 à C4) puis augmentent beaucoup pour les groupes C5 et C6. Cette croissance abrupte pour les deux derniers groupes est certainement due à leur plus grand nombre de semaines appartenant au type W10. En effet, les distances euclidiennes données dans le Tableau 6.3 entre W10 et les autres types de semaines sont généralement élevées. De plus, la moins grande diversité des comportements hebdomadaires observés dans les premiers groupes a entrainé une plus grande quantité de distances nulles, faisant baisser leur moyenne. Néanmoins, la similarité observée entre les quatre premiers groupes révèle que, malgré leur plus grande variété de comportements hebdomadaires, les groupes C2, C3 et C4 sont presque autant réguliers que le groupe C1, car les différents types de semaines auxquels leurs semaines appartiennent sont peu éloignés (distances euclidiennes faibles).

6.2.3 Vérifications statistiques

Encore une fois, les différences observées précédemment entre les six groupes d'usagers sont vérifiées pour tous les indicateurs. Les tests statistiques appliqués (tests asymptotiques de Wilcoxon Mann-Whitney) n'ont retourné que des valeurs-p inférieures à 5%, excepté pour la différence 'C1 v C2' concernant l'indicateur du nombre moyen de semaines sans déplacements (valeur-p = 0.33). Cependant, les résultats de la taille d'effet (coefficient de corrélation r) donnés dans le Tableau 6.5 ci-dessous permettent de tirer plus de conclusions. Les couleurs des cases correspondent au critère de Cohen, défini dans le Tableau 4.6.

Les faibles et moyennes tailles d'effet colorées respectivement en vert et en jaune indiquent que les différences observées entre les premiers groupes (C1 à C4) sont généralement moins importantes que les autres. Ces groupes d'usagers ont notamment des concentrations similaires de semaines dans leur groupe de semaines le plus populaire. De plus, parmi ces groupes, les deux premiers C1 et C2 sont particulièrement proches : tous leurs indicateurs sont équivalents (tailles d'effet faibles voire très faibles). De même, les deux derniers groupes C5 et C6 ont une variabilité intrapersonnelle moyenne assez comparable, sauf pour deux indicateurs : le nombre moyen de semaines sans déplacements et la distance moyenne. Ces deux groupes ont en effet des taux d'immobilité assez différents, les usagers du groupe C6 ayant beaucoup plus de semaines appartenant au type W10 que les usagers du groupe C5. D'ailleurs, cet indicateur d'immobilité (nombre moyen de semaines sans déplacements) est celui qui dissocie le plus tous les groupes d'usagers (tailles d'effet plus grandes), car il est lié à leur intensité d'utilisation et à leur niveau

d'activité, pris en compte initialement dans le processus de segmentation. Les différences les plus importantes sont relevées pour les tests comparant les premiers groupes (C1, C2, C3 ou C4) avec les deux derniers groupes (C5 ou C6). Cette divergence est un peu moins nette pour le groupe C1, car, d'après les résultats du Tableau 6.4, ce groupe présente des coefficients de variation plus élevés pour la majorité des indicateurs proposés. Certains des usagers appartenant à ce groupe sont donc possiblement peu éloignés de ceux contenus dans les deux derniers groupes. Entre autres, il existe une grande variabilité du nombre moyen de semaines non-mobiles entre les membres de ce groupe (CV = 204.4%), d'où les tailles d'effet plus faibles entre C1 et C5 ou C6.

Tableau 6.5 Mesure de la taille d'effet pour les indicateurs de variabilité intrapersonnelle

Test	Nb moy de groupes	Nb moy sem sans dépl.	% semaines 1er groupe	% semaines 2 groupes	Nb moy groupes 80%	Entropie moyenne	Distance moyenne
C1 v C2	0.10	0.01	0.07	0.13	0.05	0.12	0.07
C1 v C3	0.21	0.08	0.02	0.22	0.13	0.16	0.06
C1 v C4	0.17	0.17	0.05	0.19	0.14	0.16	0.09
C1 v C5	0.37	0.33	0.29	0.37	0.35	0.36	0.36
C1 v C6	0.43	0.48	0.20	0.37	0.32	0.33	0.45
C2 v C3	0.30	0.19	0.06	0.32	0.21	0.19	0.04
C2 v C4	0.30	0.42	0.04	0.36	0.27	0.26	0.14
C2 v C5	0.67	0.66	0.52	0.67	0.64	0.66	0.68
C2 v C6	0.72	0.83	0.29	0.58	0.53	0.53	0.75
C3 v C4	0.10	0.40	0.11	0.15	0.14	0.14	0.13
C3 v C5	0.61	0.66	0.57	0.61	0.60	0.64	0.70
C3 v C6	0.65	0.85	0.31	0.45	0.47	0.47	0.76
C4 v C5	0.49	0.38	0.47	0.50	0.48	0.52	0.59
C4 v C6	0.51	0.73	0.22	0.33	0.36	0.35	0.65
C5 v C6	0.11	0.71	0.16	0.06	0.02	0.03	0.42

CHAPITRE 7 ANALYSE DE SÉQUENCES DE TYPES DE SEMAINES

Dans le chapitre précédent, des indicateurs ont permis de mesurer la répartition globale des 51 semaines de chaque usager dans différents types de comportements hebdomadaires. Cependant, le caractère séquentiel et ordonné de ces comportements qui s'enchaînent depuis la semaine n°1 jusqu'à la semaine n°51 n'a pas été considéré. C'est là l'objectif de ce septième chapitre : une séquence de types de semaines est d'abord construite pour chaque usager, puis une typologie de séquences est créée. La comparaison de cette typologie à celle qui a été obtenue précédemment dans le Chapitre 5 à partir d'indicateurs d'utilisation du transport en commun permet alors de découvrir quelles sont les informations apportées par la prise en compte de l'organisation des semaines entre elles. Deux nouveaux indicateurs basés sur ces séquences sont également présentés.

Par ailleurs, ce chapitre constitue l'aboutissement même des deux chapitres précédents. Il donne lieu au croisement des deux types de variabilités : une typologie de séquences individuelles est réalisée pour montrer des différences entre les usagers (variabilité interpersonnelle) en fonction de l'appartenance de leurs semaines successives aux groupes formés dans la typologie de semaines (variabilité intrapersonnelle). Il s'agit donc de développer une typologie basée sur une autre typologie, et de faire ressortir la variabilité interpersonnelle à partir de la variabilité intrapersonnelle. Comme dans les chapitres précédents, le cas des abonnements annuels avec amplitude de 12 mois est étudié. Néanmoins, un sous-échantillon de 5% des usagers a dû être sélectionné, car les calculs requis par la méthodologie proposée sont assez lourds.

7.1 Méthodologie

La méthodologie proposée dans ce chapitre se décompose en trois sous-étapes : une séquence de types de semaines est d'abord construite pour chaque usager, puis une matrice de distances entre tous les usagers est calculée pour pouvoir finalement appliquer un algorithme hiérarchique et obtenir une typologie de séquences.

7.1.1 Construction des séquences

Pour chaque utilisateur d'abonnements annuels, un nouveau vecteur « carte-année » est constitué comme une séquence (par définition, une suite ordonnée) de groupes hebdomadaires. Ainsi, chacune des 51 semaines de l'utilisateur est associée à un des 10 types de semaines précédemment

produits dans le Chapitre 6, à savoir le type dans lequel elle a été classée par le processus de segmentation. Ces types sont placés dans le vecteur en respectant l'ordre des semaines de l'année. À titre d'exemple, le Tableau 7.1 ci-dessous présente la séquence (fictive) d'une carte-année. La première semaine de l'utilisateur de cette carte appartient au groupe W2, la deuxième semaine au groupe W3 et ainsi de suite pour toutes les 51 semaines complètes de l'année, le groupe de la $i^{\text{ème}}$ semaine étant positionné dans la $i^{\text{ème}}$ dimension du vecteur.

Tableau 7.1 Exemple (fictif) d'une séquence de types de semaines pour une carte-année

Semaine	1	2	3	4	5	6	7	8	9	10	…	50	51
Type	W2	W3	W5	W5	W5	W5	W9	W9	W1	W3		W9	W2

Le vecteur ainsi formé correspond en fait à une chaîne de caractères. Cette chaîne a une longueur (constante) de 51 caractères, mais elle est composée uniquement de 10 caractères différents, soient les 10 types de semaines W1 à W10. Au total, 10^{51} combinaisons sont donc possibles. Dans l'échantillon de 56,988 cartes considéré (abonnements annuels avec une amplitude de 12 mois), 56,972 séquences différentes ont été observées : seules 8 séquences sont répétées parmi 24 usagers. La chaine la plus fréquente, partagée par 10 usagers, est celle qui n'est composée que de semaines W4 (concentration des déplacements en fin de semaine). Par conséquent, il y a très peu de chaînes identiques, témoignant d'une grande diversité entre les usagers dans la composition de leur séquence de comportements hebdomadaires. Toutefois, cette grande variété s'explique notamment par la longue taille de la chaîne examinée; il aurait été intéressant ici de rechercher plutôt des sous-séquences (ou séquences ouvertes) communes, comme l'ont fait Hay et al. (2004) avec des séquences de pages web visitées.

7.1.2 Calcul d'une matrice de distances non euclidiennes

La première étape de la méthodologie proposée a abouti à la création d'une table de 56,988 séquences (une séquence par carte-année). Une matrice de distances est maintenant calculée entre toutes les paires de séquences afin de quantifier leur dissimilarité. Pour cela, plusieurs types de distances applicables à des chaînes des caractères ont été essayées, notamment la distance Levenshtein, caractéristique de la méthode d'alignement des séquences (introduite dans la revue de littérature). Cependant, les résultats obtenus avec cette distance n'étaient pas aussi intéressants que ceux qui vont être présentés avec la distance choisie.

La distance utilisée ici est une distance de Hamming modifiée, généralement appelée « distance de Hamming pondérée ». Pour rappel, la distance de Hamming traditionnelle compte le nombre de caractères différents entre deux séquences de même longueur : une unité est ajoutée à chaque fois que deux caractères situés à la même position dans les séquences sont différents. Cependant, au lieu d'appliquer un poids de substitution égal à 1 à chaque fois qu'il y a une différence entre les deux séquences, un poids correspondant à la distance euclidienne entre les deux types de semaines comparés est appliqué ici. Cette distance euclidienne est évaluée entre les centres des deux groupes et est donc récupérée dans le Tableau 6.3 établi précédemment. Ainsi, pour calculer la distance entre deux séquences, chaque semaine de la première séquence est comparée à la même semaine dans l'autre séquence (comparaisons par paires de semaines à la même position), puis les distances euclidiennes mesurées entre les groupes auxquels ces semaines appartiennent sont sommées sur les 51 semaines de l'année. L'équation 27 ci-dessous donne l'expression mathématique de cette distance (ou dissimilarité) entre les séquences de deux cartes i et j.

$$d(i,j) = \sum_{k=1}^{N} d_E(W_{(i,k)}, W_{(j,k)}) \qquad \text{(Éq. 27)}$$

avec N = 51 semaines, $d_E(.,.)$ la distance euclidienne et $W_{(i,k)}$ le groupe de semaines auquel appartient la $k^{\text{ème}}$ semaine de la carte i. Un exemple est illustré dans le Tableau 7.2 avec deux séquences de cinq semaines chacune. Pour obtenir la distance (ou dissimilarité) entre ces deux cartes, les semaines de leur séquence sont comparées deux à deux de la semaine n°1 jusqu'à la semaine n°5 et les distances euclidiennes mesurées entre les cinq paires de groupes auxquels ces semaines appartiennent sont additionnées. La distance résultante pour cet exemple est donnée par l'équation 28.

Tableau 7.2 Exemple de deux séquences sur une période de 5 semaines

	1	2	3	4	5
Carte 1	W2	W3	W5	W5	W5
Carte 2	W2	W9	W9	W2	W1

$d_E(W_{(1,3)}, W_{(2,3)}) = d_E(W_5, W_9)$

$$\begin{aligned} d(\text{carte 1}, \text{carte 2}) &= d_E(W_2, W_2) + d_E(W_3, W_9) + d_E(W_5, W_9) + d_E(W_5, W_2) + d_E(W_5, W_1) \\ &= 0.000 + 0.724 + 1.256 + 0.761 + 0.770 \\ &= 3.511 \end{aligned} \qquad \text{(Éq. 28)}$$

D'après Deza et Deza (2013), cette distance de Hamming pondérée est en fait plus qu'une distance : il s'agit d'une métrique qui a donc les propriétés de non-négativité, de symétrie, et qui vérifie l'identité des indiscernables et l'inégalité triangulaire. D'après sa formulation, cette distance prend en compte la position des semaines puisque chaque semaine est comparée à la semaine avec le même numéro et un poids différent est attribué en fonction des groupes auxquelles elles appartiennent. Ce poids est une distance euclidienne (prise dans le Tableau 6.3) qui traduit la ressemblance entre les deux groupes.

À la fin de cette deuxième étape, une matrice de distances évaluées entre toutes les paires de séquences (ou de cartes) est fournie. Dans la méthodologie proposée, c'est cette deuxième étape qui prend beaucoup de temps. Un échantillonnage des données a donc été réalisé : 5% des utilisateurs d'abonnements annuels (soit 2850 cartes) ont été sélectionnés aléatoirement et une matrice de distances de dimensions 2850x2850 a été calculée entre toutes ces cartes. Même après automatisation et optimisation de l'algorithme utilisé dans R (bien sûr, uniquement des fonctions de type *apply* et non des boucles ont été exploitées), ce calcul prend une trentaine de minutes.

7.1.3 Segmentation hiérarchique

Finalement, un algorithme de segmentation hiérarchique est appliqué sur cette matrice de distances pour créer une typologie de séquences (et donc une typologie d'usagers, car chaque séquence correspond à un usager, ou plutôt une carte). Plus précisément, un algorithme hiérarchique agglomératif est utilisé ici. Son fonctionnement est expliqué en détail par Jain et al. (1999) ou encore par James et al. (2013). L'objectif de la segmentation hiérarchique est de produire un dendrogramme, c'est-à-dire un arbre qui représente le regroupement imbriqué des observations en fonction du niveau de dissimilarité des groupes formés. L'algorithme agglomératif construit cet arbre itérativement en allant des feuilles au tronc : il commence donc par le bas du dendrogramme, chaque observation étant alors considérée comme son propre groupe (singleton), puis il remonte dans le dendrogramme en fusionnant les deux groupes les plus similaires à chaque nouvelle itération, jusqu'à ce que toutes les observations appartiennent à un seul et même groupe tout en haut du dendrogramme.

Différents types d'algorithmes existent selon la définition choisie pour mesurer la dissimilarité entre deux groupes. En effet, la distance entre les paires d'observations individuelles est connue, mais il faut étendre ce calcul à deux groupes composés de plusieurs observations pour pouvoir

ensuite remonter dans le dendrogramme. La méthode appliquée ici est la méthode de Ward (1963), basée sur la minimisation de la variance intragroupes : les deux groupes rassemblés à chaque étape sont alors ceux qui font le moins augmenter cette variance totale. D'autres algorithmes ont été testés (méthodes du saut minimum ou *single linkage*, du saut maximum ou *complete linkage*, du lien moyen ou *centroïd linkage*), mais les résultats obtenus n'étaient pas aussi bons et pas aussi faciles à interpréter. La méthode de Ward est mise en œuvre récursivement par l'algorithme de Lance-Williams. Plus particulièrement, le critère de regroupement de Ward (Murtagh & Legendre, 2014) est employé ici afin que les distances soient mises au carré à chaque étape. Il a été vérifié que cette méthode pouvait être généralisée à d'autres distances que la distance euclidienne : en effet, les coefficients de la formule de Lance-Williams utilisés pour mettre à jour les distances intergroupes à chaque itération restent les mêmes quelle que soit la dissimilarité d adoptée (Batagelj, 1988; Strauss & von Maltitz, 2017). L'application de cet algorithme est faite dans R avec la fonction *hclust* et la méthode *ward.D2*, la matrice des distances individuelles calculée précédemment étant donnée en argument.

Une fois le dendrogramme dessiné, le nombre de groupes est choisi en fonction du niveau de dissimilarité désiré (donné par l'axe vertical du graphique produit). La typologie de séquences finalement obtenue correspond à une typologie de 5% des cartes du groupe initialement étudié. Pour vérifier la stabilité des résultats exposés dans la section suivante, une première typologie a été créée à partir d'un premier échantillon, puis d'autres tirages aléatoires de 5% ont été réalisés: les mêmes conclusions pouvaient être tirées des différentes typologies produites, démontrant ainsi la fiabilité de la typologie présentée ci-après.

7.2 Typologie de séquences

Les résultats de la segmentation des utilisateurs d'abonnements annuels en fonction de leurs séquences de comportements hebdomadaires sont présentés dans cette section. Des indicateurs de variabilité basés sur ces séquences et prenant donc en compte l'ordre des types de semaines observés dans les comportements individuels sont également calculés à l'intérieur de chaque groupe. Finalement, la typologie obtenue est comparée à celle du Chapitre 5 (section 5.3) afin de croiser tous les résultats des précédents chapitres et d'en tirer des conclusions globales.

7.2.1 Résultats

Le dendrogramme produit par l'algorithme hiérarchique agglomératif appliqué est dessiné sur la Figure 7.1. La coupure de ce dendrogramme au niveau de dissimilarité choisi conduit à la formation de 7 groupes, chaque groupe étant encadré en rouge sur la figure. On constate qu'un des groupes (le deuxième en partant de la droite) est beaucoup plus gros que les autres. Toutefois, même si on diminue le niveau de dissimilarité en descendant la ligne de la coupure sur l'axe vertical, d'autres groupes vont se segmenter avant celui-ci. Les séquences appartenant à ce groupe sont donc très similaires et il est difficile de les départager sans devoir créer un nombre très élevé de segments. Pour permettre une interprétation plus facile des résultats, il a été décidé de conserver cette granularité de 7 groupes puis d'analyser ensuite le groupe majoritaire de manière indépendante.

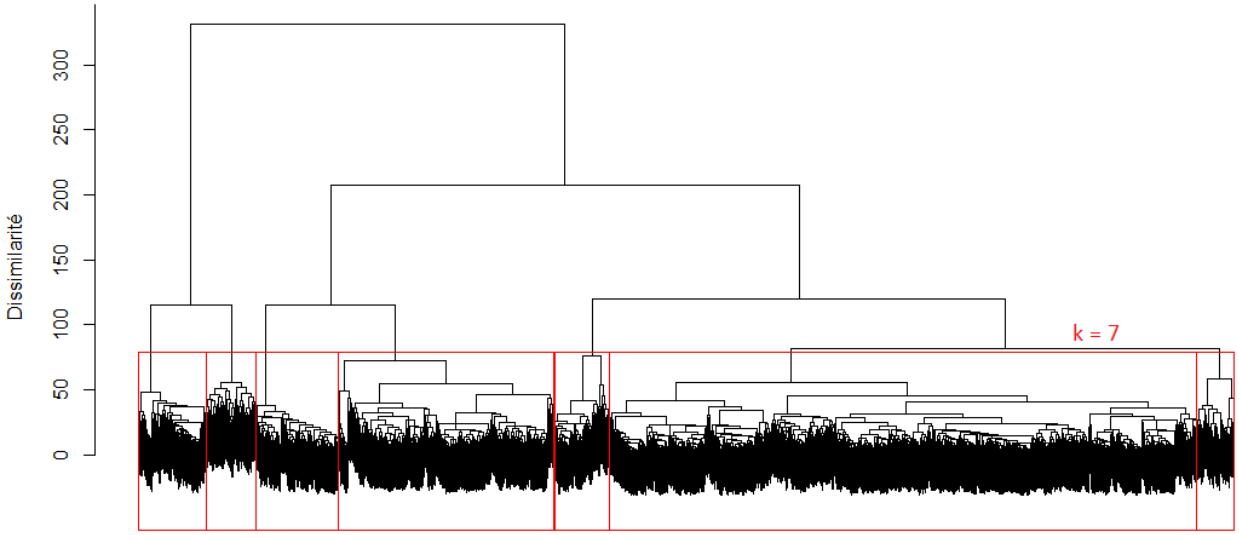


Figure 7.1 Dendrogramme - Segmentation de 5% des utilisateurs d'abonnements annuels en fonction de leur séquence de types de semaines

Tout d'abord, les résultats de la segmentation des séquences en 7 groupes sont présentés par la Figure 7.2. Ces groupes ont été rangés dans l'ordre décroissant de leur taille, donnée en pourcentage de séquences (ou de cartes-année) au-dessus de chaque graphique. La notation Sx est utilisée pour désigner le $x^{\text{ème}}$ groupe de la typologie obtenue. Toutes les séquences observées dans chaque groupe sont affichées dans une sorte de carte de chaleur : les 51 semaines de l'année sont énumérées en X et les cartes des usagers sont listées le long de l'axe Y. Ainsi, chaque ligne du graphique correspond à l'utilisateur d'une carte pour lequel on a représenté la séquence de ses 51 semaines,

chaque semaine étant colorée en fonction du type de semaines auquel elle appartient. Les couleurs attribuées à chaque type de semaines sont les mêmes que celles qui avaient été choisies précédemment dans le Chapitre 6. De plus, sur la droite de ce graphique est fournie la distribution de toutes les semaines des usagers du groupe dans les 10 types de semaines.

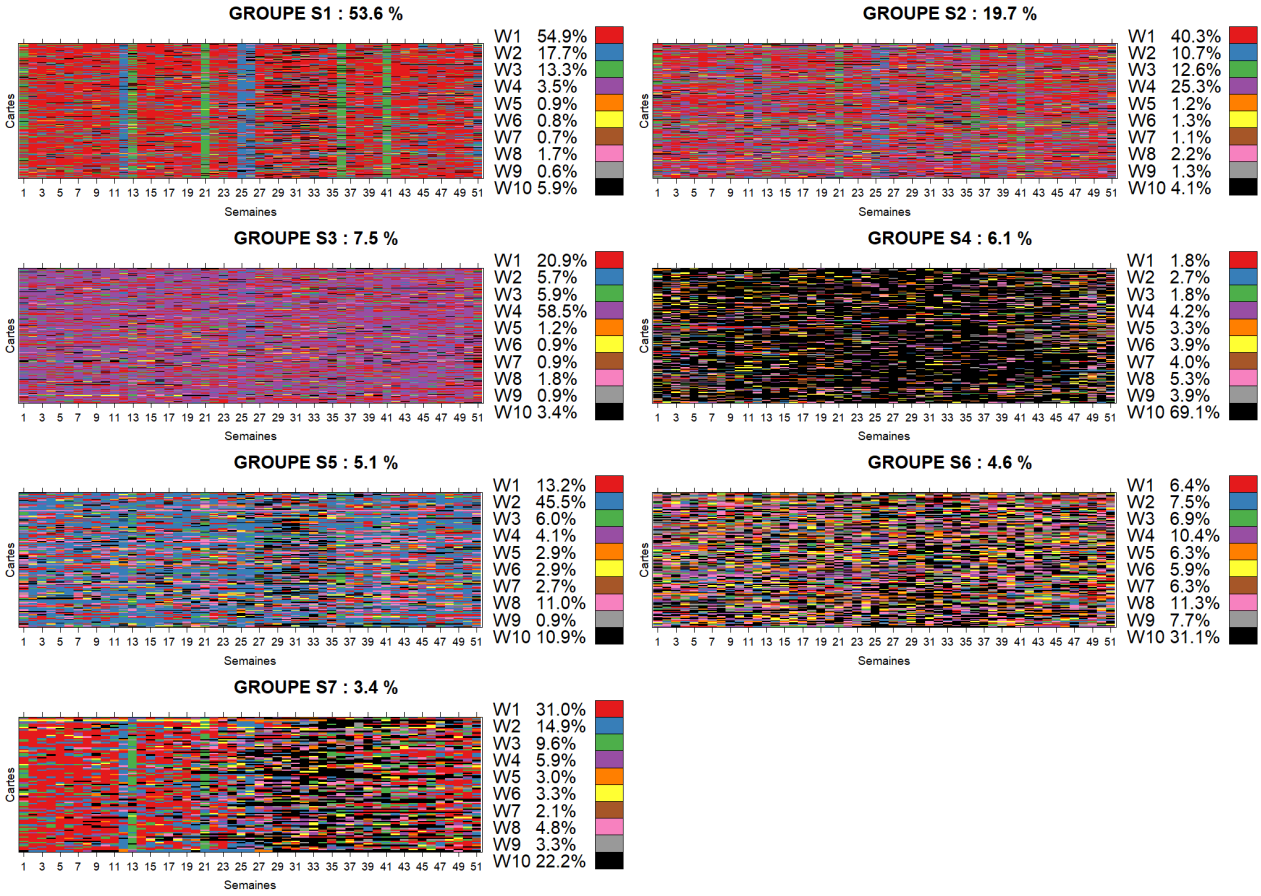


Figure 7.2 Séquences et distribution des types de semaines dans chacun des 7 groupes d'utilisateurs d'abonnement annuels

Le groupe S1 correspond au groupe volumineux remarqué précédemment dans le dendrogramme : il réunit à lui seul plus de la moitié des séquences segmentées. D'après la carte de chaleur et la distribution des semaines de ce groupe, la semaine type de travail W1 (en rouge) prédomine dans le comportement hebdomadaire de tous les usagers du groupe. Différents microphénomènes sont également capturés, notamment les jours fériés, repérés aux mêmes numéros de semaines que ceux évoqués précédemment avec la Figure 6.3. Ces numéros sont plus visibles sur la Figure 7.3 qui présente uniquement le groupe S1. En mettant en parallèle ce graphique avec le calendrier en ANNEXE D, on situe les vendredis fériés par l'accumulation des semaines du type W2 (en bleu)

dans les semaines 12, 25 et 26, ainsi que les lundis fériés par la concentration des semaines du type W3 (en vert) dans les semaines 13, 21, 36 et 41. De plus, l'impact des vacances scolaires sur les comportements individuels de ce groupe est apparent en été, car on observe plus de semaines dans le type W10 (en noir) entre les semaines 27 et 34 (correspondant aux mois de juillet et août). Ainsi, les usagers contenus dans ce groupe sont majoritairement des travailleurs soumis à des contraintes de type travail-congé et qui se déplacent uniquement pendant les jours ouvrables.

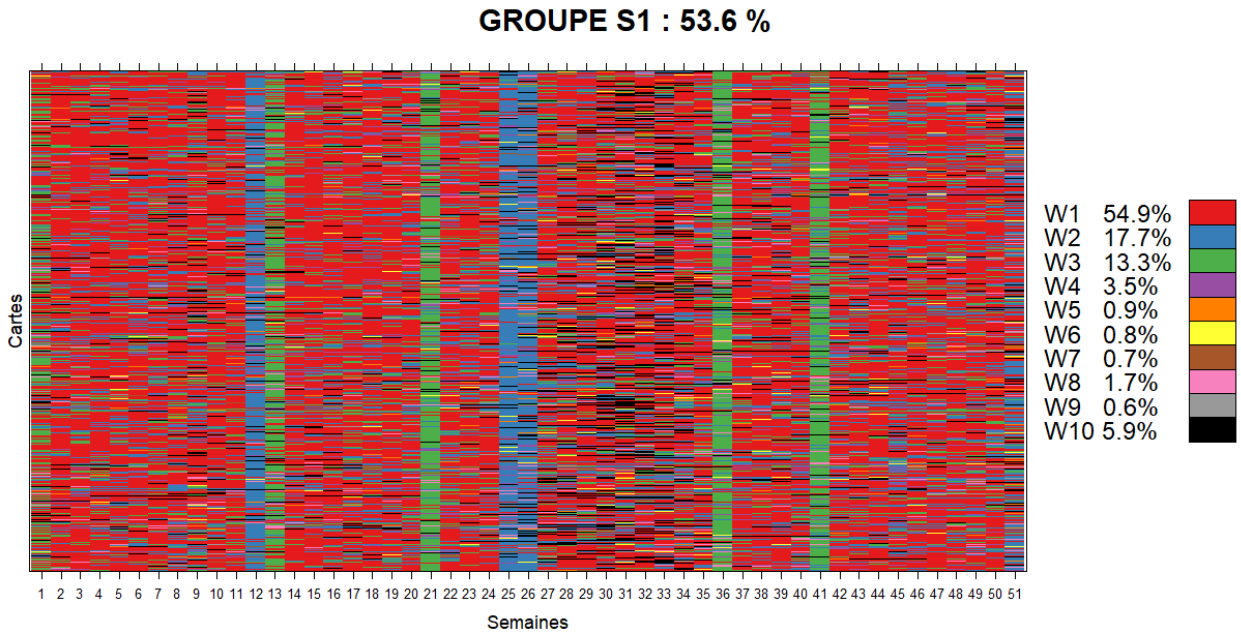


Figure 7.3 Séquences et distribution des types de semaines du groupe S1

Le groupe S2 est assez similaire à S1, mais avec une proportion plus élevée de semaines appartenant au type de semaine W4, ce qui implique que plus de déplacements sont faits en fin de semaine. Les usagers de ce groupe utilisent donc le transport en commun pour des motifs plus diversifiés. De plus, l'effet des jours fériés est moins net dans S2 que dans S1, indiquant que certains usagers continuent de se déplacer pendant les lundi et vendredi non ouvrables. Comparé au groupe S1, le groupe S2 est encore assez important (20% des séquences), mais tous les groupes suivants sont plus minoritaires (moins de 10% des séquences). Toutefois, ces groupes révèlent l'existence de comportements plus atypiques dans l'utilisation du transport en commun. Dans le groupe S3, plus de la moitié des semaines appartiennent au type de semaine W4 (en violet). Les déplacements des usagers de ce groupe sont donc principalement réalisés le samedi et le dimanche. Ce sont des utilisateurs assidus du transport en commun pour leurs activités de fin de semaine. Le groupe S4 rassemble quant à lui des utilisateurs qui se déplacent peu, du moins en transport en

commun, puisque le type de semaine W10 (semaine sans déplacements) l'emporte largement dans la distribution des semaines du groupe. Dans le groupe S5, le type de semaine W2 (en bleu) est celui qui ressort le plus. Les usagers de ce groupe ont donc tendance à ne pas se déplacer le vendredi, probablement parce que leur semaine de travail n'est composée que de quatre jours. On remarque également un plus grand nombre de semaines appartenant au type W10 pendant la période estivale : les comportements des usagers de ce groupe sont donc influencés par les vacances d'été. Le groupe S6 est composé d'utilisateurs ayant des comportements hebdomadaires assez hétérogènes et imprévisibles. La répartition des semaines de ce groupe parmi les 10 types de semaines est assez équilibrée, la plus grande proportion de semaines étant néanmoins rapportée pour le type W10 (semaine sans déplacements). Par ailleurs, c'est dans ce groupe qu'on a les proportions de semaines les plus importantes pour les types W5 à W9 (déplacements pendant un seul jour de la semaine), témoignant d'une utilisation plus parcellaire du transport en commun. Enfin, le groupe S7 contient des utilisateurs qui ressemblent à ceux du groupe S1 en début d'année, mais qui se déplacent moins de juin à septembre et de manière plus irrégulière dans le dernier trimestre de l'année.

Comme le premier groupe S1 renfermait plus de la moitié des utilisateurs étudiés, sa décomposition a été forcée en réappliquant l'algorithme hiérarchique agglomératif expliqué plus haut uniquement sur cet ensemble. Le dendrogramme résultant est donné par la Figure 7.4 ci-dessous. Le groupe S1 est ainsi divisé en 5 sous-groupes de tailles relativement homogènes, même si l'un d'entre eux (l'avant-dernier) se démarque encore par sa plus grande densité d'usagers très similaires.

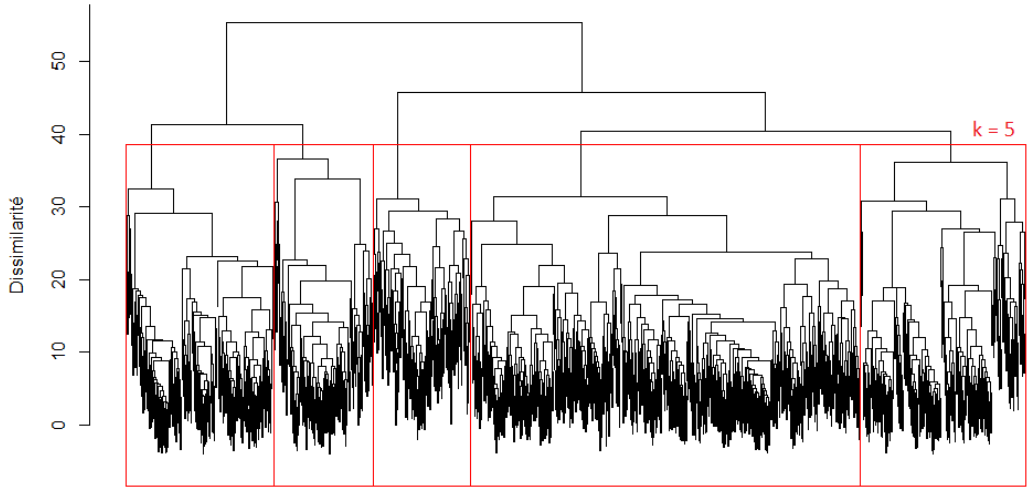


Figure 7.4 Dendrogramme - Segmentation des utilisateurs du groupe S1

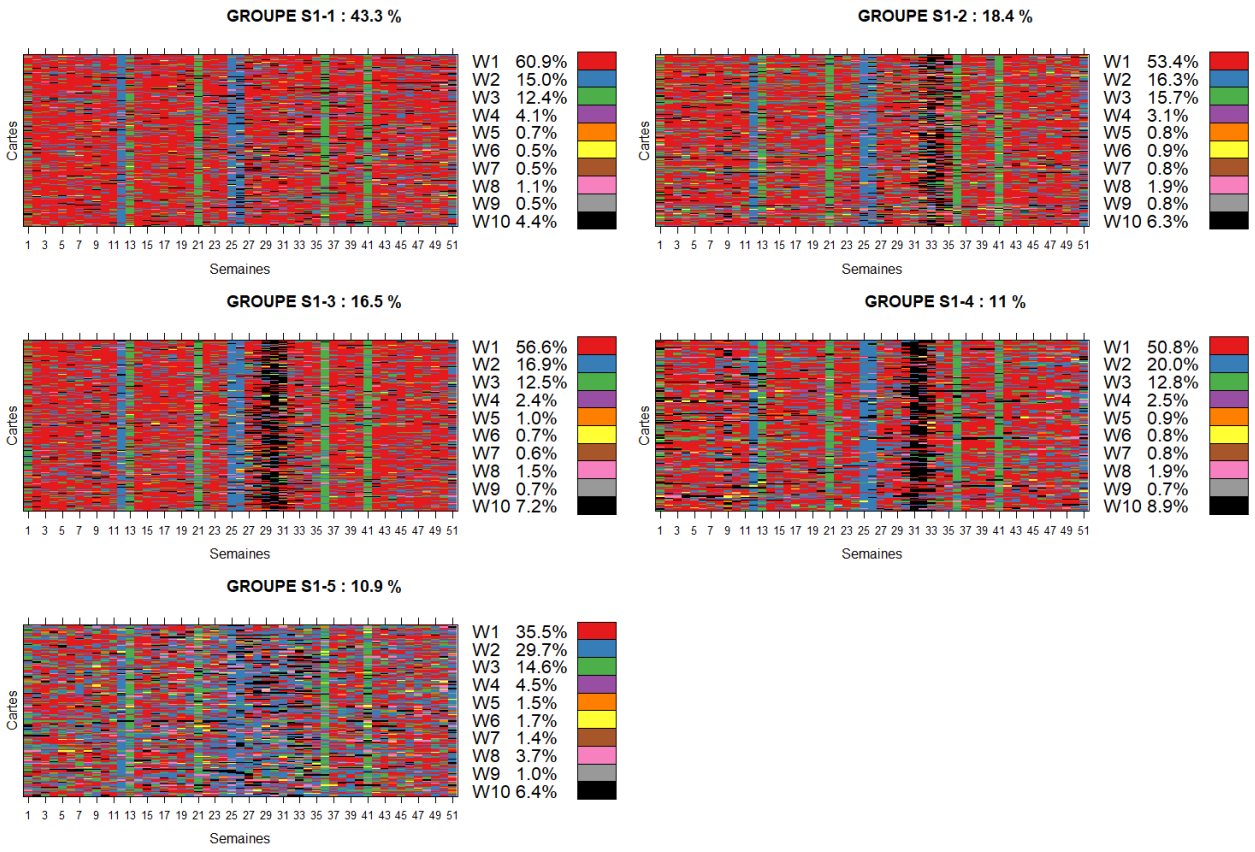


Figure 7.5 Séquences et distribution des types de semaines dans chacun des 5 sous-groupes de S1

De même que précédemment, les 5 sous-groupes formés sont présentés dans la Figure 7.5 dans l'ordre décroissant de leur taille. Ils sont nommés en utilisant le code S1-x où x est le $x^{\text{ème}}$ sous-groupe de S1. Cette nouvelle segmentation a séparé les utilisateurs du groupe S1 en fonction de leurs comportements pendant la saison estivale. En effet, les séquences observées dans les 5 sous-groupes se distinguent principalement par leur composition en été (rappel : d'après le calendrier de 2016 en ANNEXE D, les mois de juillet et août correspondent aux semaines 26 à 35). Différentes périodes de vacances, délimitées par les semaines appartenant au groupe W10 en noir, sont choisies par les usagers des groupes S1-2 à S1-4 : de fin juillet (groupe S1-3) à début août (groupe S1-4) voire fin août (groupe S1-2). Ces vacances sont généralement posées sur 2 semaines. Néanmoins, la proportion de semaines sans déplacements (type W10) est légèrement plus élevée dans le groupe S1-3. On observe effectivement quelques autres périodes d'arrêt dans les séquences de ce groupe. Les usagers du groupe S1-1 semblent quant à eux ne pas prendre de congés, à moins qu'ils ne continuent tout simplement de se déplacer en transport en commun pendant cette période pour d'autres motifs. Plus de 60% des semaines de ce groupe sont ainsi concentrées dans le type W1,

attestant la grande uniformité des comportements hebdomadaires des usagers de ce groupe. Le dernier sous-groupe S1-5 rassemble en revanche des utilisateurs qui ont des vacances plus éparpillées dans l'année. La part des semaines de type W2 (semaines sans déplacements le vendredi) est également plus élevée dans ce groupe et par conséquent, celle des semaines de type W1 est plus faible. Quelques-unes de ces semaines de 4 jours (en bleu) sont disséminées tout au long de l'année, mais la plupart sont accumulées pendant la période estivale : les usagers de ce groupe ont donc tendance à moins se déplacer le vendredi pendant l'été. Ce comportement peut être lié au fait que plusieurs sociétés montréalaises ferment le vendredi durant l'été.

Ainsi, les deux segmentations réalisées conduisent finalement à une typologie composée de 11 groupes. Ces 11 groupes montrent que des différences dans l'utilisation du transport en commun peuvent être décelées au niveau hebdomadaire parmi les utilisateurs d'abonnements annuels. En effet, les séquences diffèrent d'un groupe à l'autre par leur composition (en types de semaines), mais aussi par leur organisation. Des microphénomènes (jours fériés, périodes de vacances) sont notamment mis en évidence. La détection de ces événements est la preuve que la méthodologie proposée prend en compte la position des comportements hebdomadaires.

Par ailleurs, pour aider à l'interprétation de la typologie obtenue, la distribution tarifaire des déplacements de chaque groupe est présentée horizontalement et verticalement dans les Tableau 7.3 et Tableau 7.4. Dans chaque groupe, les abonnements annuels utilisés sont combinés avec un des trois types de tarifs suivants : ordinaire, réduit ou gratuit. Le tarif ordinaire est largement priorisé dans tous les sous-groupes de S1 (S1-1 à S1-5) et dans le groupe S2 ; en effet, plus de 90% des déplacements faits dans ces groupes sont réalisés avec un tarif ordinaire. Dans les groupes S3, S5 et S7, les déplacements effectués avec un tarif ordinaire sont également majoritaires. Néanmoins, on remarque un plus grand pourcentage d'abonnements gratuits dans S5 et la plus forte proportion de déplacements faits avec un tarif réduit est rapportée pour S7. Ce dernier groupe pourrait donc bien contenir quelques utilisateurs d'abonnements de 4 mois à tarif réduit 18-25 ans: cela justifierait le grand nombre de semaines sans déplacements observées de juin à septembre dans ce groupe. En effet, les abonnements de 4 mois ne sont pas nécessairement renouvelés pendant la session d'été et, même s'ils sont renouvelés, ils sont généralement moins utilisés que pendant les autres sessions (plus chargées pour les étudiants). Les usagers des groupes S4 et S6 ont réalisé quant à eux plus de 30% de leurs déplacements gratuitement. Une grande partie des abonnements annuels utilisés dans ces groupes pourraient donc être des abonnements offerts par les entreprises,

ce qui expliquerait la plus faible mobilité des usagers de ces groupes, en particulier pour S4, et la plus grande hétérogénéité des comportements rassemblés dans S7.

Tableau 7.3 Distribution des déplacements de chacun des 11 groupes par type de tarifs – 5% des utilisateurs d'abonnements annuels

	TYPES DE TARIFS					
Groupe	**Gratuité**	**Tarif étudiant**	**Tarif réduit**	**Tarif ordinaire**	**Tarif spécial**	**Total**
S1-1	3.3%	0.0%	1.0%	95.7%	0.0%	100.0%
S1-2	5.9%	0.0%	0.0%	94.1%	0.0%	100.0%
S1-3	5.7%	0.0%	0.5%	93.8%	0.0%	100.0%
S1-4	5.3%	0.0%	0.5%	94.1%	0.0%	100.0%
S1-5	6.2%	0.0%	1.3%	92.4%	0.0%	100.0%
S2	5.9%	0.0%	1.3%	92.9%	0.0%	100.0%
S3	8.4%	0.0%	4.6%	87.0%	0.0%	100.0%
S4	34.5%	0.0%	3.0%	62.6%	0.0%	100.0%
S5	15.9%	0.0%	1.2%	82.9%	0.0%	100.0%
S6	32.0%	0.0%	1.7%	66.3%	0.0%	100.0%
S7	12.1%	0.0%	5.7%	82.2%	0.0%	100.0%
Tous	6.5%	0.0%	1.4%	92.1%	0.0%	100.0%

Tableau 7.4 Distribution des déplacements par type de tarifs dans chacun des 11 groupes – 5% des utilisateurs d'abonnements annuels

	TYPES DE TARIFS					
Groupe	**Gratuité**	**Tarif étudiant**	**Tarif réduit**	**Tarif ordinaire**	**Tarif spécial**	**Tous**
S1-1	13.4%	0.0%	17.4%	27.2%	0.0%	26.2%
S1-2	9.1%	0.0%	0.0%	10.3%	0.0%	10.1%
S1-3	7.8%	0.0%	2.8%	9.0%	0.0%	8.9%
S1-4	4.6%	0.0%	2.1%	5.7%	0.0%	5.6%
S1-5	5.0%	0.0%	4.9%	5.3%	0.0%	5.3%
S2	22.7%	0.0%	22.1%	25.4%	0.0%	25.2%
S3	14.4%	0.0%	36.0%	10.6%	0.0%	11.2%
S4	3.5%	0.0%	1.4%	0.5%	0.0%	0.7%
S5	8.4%	0.0%	2.9%	3.1%	0.0%	3.4%
S6	6.9%	0.0%	1.7%	1.0%	0.0%	1.4%
S7	4.1%	0.0%	8.7%	2.0%	0.0%	2.2%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Dans la distribution inverse (Tableau 7.4), les déplacements sont concentrés dans les groupes les plus volumineux. En particulier, 93.5% des déplacements avec tarif ordinaire ont été effectués dans les trois premiers groupes S1, S2 et S3. À l'inverse, la part des déplacements réalisés par les usagers du groupe S4 est très faible quel que soit le tarif utilisé puisque ces usagers sont particulièrement peu actifs (la majorité de leurs semaines appartiennent au type W10). De même, les groupes S5, S6 et S7 cumulent peu de déplacements du fait de leur plus petite taille et de leur plus faible intensité d'utilisation du transport en commun.

7.2.2 Calcul d'indicateurs séquentiels

De plus, deux nouveaux indicateurs basés sur l'analyse des séquences construites précédemment sont proposés et évalués à l'intérieur de chacun des 11 groupes obtenus. Le premier indicateur mesure la variabilité des séquences au sein de chaque groupe en calculant une dissimilarité intragroupe moyenne. Saneinejad et Roorda (2009) ont utilisé le même type d'indice pour estimer au contraire une similarité intragroupe moyenne. Cet indicateur est calculé pour un groupe donné en additionnant toutes les distances (par paires) entre tous les usagers de ce groupe, puis en divisant par le nombre total de membres du groupe. Sa formulation mathématique est fournie par l'équation suivante.

$$V_k = \frac{1}{2n_k} \sum_{i,j \in S_k} d(i,j) \qquad \text{(Éq. 29)}$$

avec n_k la taille du groupe S_k et $d(i,j)$ la distance de Hamming pondérée entre deux cartes i et j donnée par l'équation 27 dans la section 7.1.2. Le facteur $\frac{1}{2}$ est ajouté pour ne pas répéter deux fois les mêmes distances (car, par symétrie, $d(i,j) = d(j,i)$ pour toutes cartes i et j).

Le deuxième indicateur appartient à la même famille que ceux qui ont été proposés précédemment dans la section 6.2.1 : il est ainsi appliqué pour quantifier la variabilité intrapersonnelle moyenne à l'intérieur de chaque groupe. Cet indicateur, défini à l'échelle individuelle par l'équation 30 ci-dessous, mesure la distance euclidienne moyenne entre deux semaines successives d'un même usager u. Il est similaire à la distance moyenne D^u_{moy} présentée précédemment, mais, au lieu de regarder toutes les semaines de l'usager, il compare seulement celles qui se suivent (la semaine i avec la semaine $i+1$). Ce nouvel indicateur permet donc d'examiner les semaines dans leur ordre d'apparition. Plus la valeur de cet indicateur est petite, plus la semaine qui suit a tendance à

ressembler à la semaine précédente (en termes d'appartenance aux types de semaines) et donc plus le comportement hebdomadaire de l'utilisateur est prévisible et stable dans le temps.

$$D^{u}_{succ,moy} = \frac{1}{N-1}\sum_{i=1}^{N-1} d_E(W_{(u,i)}, W_{(u,i+1)}) \qquad \text{(Éq. 30)}$$

avec N = 51 semaines, $d_E(.,.)$ la distance euclidienne et $W_{(u,i)}$ le groupe de semaines auquel appartient la $i^{\text{ème}}$ semaine de l'usager u. La somme est divisée par le nombre de comparaisons possibles entre deux semaines successives, soit $N-1$. Cet indicateur est calculé pour chaque utilisateur u puis une moyenne est évaluée à l'intérieur de chacun des 11 groupes. Un coefficient de variation de cet indicateur à l'intérieur de chaque groupe est également rapporté.

Les résultats de l'application de ces deux indicateurs ainsi que la proportion de cartes dans chaque groupe sont donnés dans le Tableau 7.5. Après la décomposition du groupe S1, on aboutit à 11 groupes de taille assez homogènes, même si les groupes S1-1 et S2 restent un peu plus importants. Comme on pouvait s'y attendre, les dissimilarités moyennes intragroupe obtenues pour les 11 groupes sont toutes largement inférieures à celle calculée avec le total des cartes: l'objectif de la segmentation était en effet de regrouper des séquences qui se ressemblent. Cependant, on observe une plus grande variabilité entre les séquences qui appartiennent aux deux plus gros groupes S1-1 et S2 (malgré le fait que l'indicateur ait été pondéré par la taille de chaque groupe). Ces deux groupes rassemblent donc un plus grand nombre de séquences, mais aussi des séquences plus diversifiées. Ils auraient pu être segmentés de nouveau pour faire ressortir différents types de séquences. En dépit de leur plus petite taille, les groupes S4 et S6 présentent également une dissimilarité moyenne intragroupe très élevée. Les séquences du groupe S4 sont essentiellement composées de semaines appartenant au type W10 (sans déplacements), mais quelques semaines appartenant aux autres types de comportements hebdomadaires sont réparties de manière assez disparate entre les séquences. De même, les séquences du groupe S6 observées sur la Figure 7.2 sont très hétéroclites et désordonnées : tous les types de semaines sont présents, mais leur organisation est confuse, aucune tendance temporelle dominante n'étant visible. Ces deux groupes réunissent donc une grande variété de séquences, d'où la valeur importante de l'indicateur proposé. Au contraire, les séquences les plus homogènes se trouvent dans les groupes S1-4 et S7.

La variabilité intrapersonnelle moyenne, mesurée par la distance euclidienne moyenne entre deux semaines successives, est de loin supérieure pour le groupe S6. Le coefficient de variations associé est également très faible. Les usagers de ce groupe sont effectivement peu prévisibles et il est très rare que deux semaines du même type se suivent dans leurs séquences. D'après l'indicateur proposé, le groupe S6 est donc le groupe le plus irrégulier au niveau intrapersonnel. Il est suivi par les groupes S4 et S7, qui rassemblent des utilisateurs avec des périodes d'inactivité plus longues. La valeur de l'indicateur évalué dans ces deux groupes est élevée, car les distances euclidiennes calculées dans le Tableau 6.3 entre le type W10 (semaine sans déplacements) et les autres types de semaines sont généralement grandes, surtout celles entre W10 et les types W5 à W9 (déplacements concentrés sur un seul jour), comportements atypiques dont la présence est non négligeable dans ces deux groupes. À l'inverse, la plus petite variabilité intrapersonnelle moyenne est rapportée pour le groupe S1-1. Le comportement hebdomadaire des usagers de ce groupe est similaire d'une semaine à l'autre puisque la totalité ou presque de leurs semaines appartiennent au même type W1. Seuls les jours fériés viennent interrompre cette continuité. La stabilité des semaines des groupes S1-2, S1-3 et S1-4 est également assez comparable (distance moyenne environ égale à 0.20), car seule la position de la période de vacances d'été change entre ces groupes. Toutefois, parmi les sous-groupes de S1, le cinquième S1-5 est le plus irrégulier, car les séquences de ce groupe sont composées majoritairement de deux types de semaines (W1 et W2) plutôt que d'un seul, conduisant à un plus grand nombre de distances euclidiennes non nulles, c'est-à-dire différentes de $d(W_1, W_1) = 0$. Le comportement des usagers de ce groupe est ainsi plus variable (en moyenne) que celui des groupes S2 et S3, mais un peu plus régulier que celui du groupe S5, ce dernier étant plus diversifié dans la composition de ses séquences.

Tableau 7.5 Résultats du calcul des deux indicateurs séquentiels dans chacun des 11 groupes

GROUPES	**S1-1**	**S1-2**	**S1-3**	**S1-4**	**S1-5**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**Total**
Taille (% cartes-année)	23.2%	9.9%	8.8%	5.9%	5.8%	19.7%	7.5%	6.1%	5.1%	4.6%	3.4%	100.0%
$\boldsymbol{V_k}$	2930	1490	1191	903	1233	4176	1370	1665	1428	2029	951	22943
$\boldsymbol{D^u_{succ,moy}}$ *Coeff de variation*	0.176 *(31.9%)*	0.205 *(45.8%)*	0.192 *(44.8%)*	0.205 *(44.0%)*	0.272 *(30.4%)*	0.254 *(48.3%)*	0.219 *(49.9%)*	0.334 *(42.7%)*	0.310 *(47.9%)*	0.550 *(19.6%)*	0.330 *(31.9%)*	0.245 *(53.5%)*

Contrairement au premier indicateur proposé, le deuxième indicateur est calculé comme une moyenne et peut donc être vérifié statistiquement. Les résultats de l'indice de taille d'effet mesuré (coefficient de corrélation r) sont résumés sous forme d'une matrice dans le Tableau 7.6. Cette

matrice est symétrique puisque le test appliqué est non directionnel (on teste si chaque groupe est différent de l'autre). Les couleurs attribuées correspondent encore une fois au critère de Cohen énoncé dans le Tableau 4.6. D'après les valeurs obtenues, le groupe S6 est très différent de tous les autres (tailles d'effet >> 0.5). La plus grande irrégularité des usagers de ce groupe est donc ici confirmée. Les groupes S4 et S7 sont fortement ou moyennement éloignés des autres, mais assez analogues entre eux (taille d'effet = 0.01). Le groupe S5 leur ressemble également. Les plus longues périodes d'inactivité des usagers de ces trois groupes ainsi que les plus grandes proportions de comportements hebdomadaires atypiques W5 à W9 les rendent comparables pour l'indicateur mesuré. De mêmes, les sous-groupes de S1, constitués d'utilisateurs presque « mono-type » W1, sont assez similaires entre eux, à l'exception du sous-groupe S1-5. La variabilité intrapersonnelle des séquences de ce dernier sous-groupe est néanmoins plus proche de celle du groupe S5. En effet, ces deux groupes se recoupent par leur forte proportion de semaines appartenant au type W2 (sans déplacements le vendredi). Par ailleurs, on remarque que la distance moyenne évaluée pour le groupe S3 est assez semblable à celle des quatre premiers sous-groupes de S1 et à S2. Malgré leurs différentes caractéristiques d'utilisation du transport en commun (jours ouvrables versus jours non ouvrables), ces groupes ont donc une régularité intrapersonnelle assez équivalente.

Tableau 7.6 Mesure de la taille d'effet pour l'indicateur mesurant la distance moyenne entre deux semaines successives

	S1-1	S1-2	S1-3	S1-4	S1-5	S2	S3	S4	S5	S6	S7
S1-1		0.08	0.02	0.10	0.46	0.37	0.15	0.47	0.37	0.64	0.47
S1-2	0.08		0.07	0.02	0.39	0.24	0.07	0.46	0.36	0.78	0.49
S1-3	0.02	0.07		0.09	0.46	0.29	0.13	0.51	0.41	0.80	0.55
S1-4	0.10	0.02	0.09		0.41	0.20	0.05	0.47	0.38	0.83	0.55
S1-5	0.46	0.39	0.46	0.41		0.16	0.32	0.22	0.10	0.82	0.26
S2	0.37	0.24	0.29	0.20	0.16		0.14	0.26	0.16	0.60	0.27
S3	0.15	0.07	0.13	0.05	0.32	0.14		0.41	0.31	0.79	0.45
S4	0.47	0.46	0.51	0.47	0.22	0.26	0.41		0.09	0.65	0.01
S5	0.37	0.36	0.41	0.38	0.10	0.16	0.31	0.09		0.68	0.10
S6	0.64	0.78	0.80	0.83	0.82	0.60	0.79	0.65	0.68		0.73
S7	0.47	0.49	0.55	0.55	0.26	0.27	0.45	0.01	0.10	0.73	

7.2.3 Comparaison avec la typologie d'usagers du chapitre 5

Dans cette dernière partie les deux typologies d'utilisateurs d'abonnements annuels créées précédemment dans ce mémoire sont comparées. Ces deux typologies sont le résultat de deux

processus de segmentation distincts appliqués sur des cartes-année décrites de deux manières différentes. La première typologie, développée dans le Chapitre 5 (section 5.3 Cas des utilisateurs d'abonnements annuels), est basée sur des indicateurs d'utilisation mensuelle du transport en commun. Les groupes de cette première typologie sont notés $C_i, i \in \{1, \dots, 6\}$ et sont rangés dans l'ordre décroissant de leur intensité d'utilisation. La deuxième typologie, basée sur des séquences de types de semaines, est celle qui vient d'être présentée dans ce chapitre (section 7.2.1). Les groupes de cette seconde typologie sont nommés $S_j, j \in \{1,2, \dots, 7\}$ et triés dans l'ordre décroissant de leur taille, le groupe S1 étant néanmoins divisé en 5 sous-groupes $S_{1-k}, k \in \{1,2, \dots, 5\}$ du fait de son plus gros volume.

Premièrement, deux matrices de confusion, l'une avec une distribution horizontale (Tableau 7.7) et l'autre avec une distribution verticale (Tableau 7.8), sont construites afin de croiser l'appartenance des 2850 cartes échantillonnées aux 6 et 11 groupes respectifs de chacune des deux typologies. Ainsi, une typologie est distribuée dans l'autre et vice versa (uniquement pour les 5% d'utilisateurs d'abonnements annuels étudiés dans ce chapitre). On s'aperçoit tout d'abord que les cartes appartenant au groupe S1 (S1-1 à S1-5) se retrouvent pour la plupart dans le groupe C4. Cette grande concentration s'explique probablement par le fait que les groupes C4 et S1 sont respectivement les deux plus gros groupes dans chaque typologie. Cependant, dans la distribution inverse, les membres du groupe C4, ainsi que les membres de C3, sont principalement représentés dans le sous-groupe S1-1. Les membres de ces deux groupes (C3 et C4) étant, d'après le Tableau 5.13, des utilisateurs fréquents du transport en commun qui effectuent la grande majorité de leurs déplacements pendant les jours ouvrables, leur comportement est cohérent avec celui de la semaine typique de travail caractéristique du groupe S1-1. Les travailleurs réunis dans ces deux groupes ont donc un comportement hebdomadaire très régulier dans le temps, uniquement perturbé par les jours fériés. Une bonne partie des utilisateurs contenus dans le groupe C4 (40.6% au total) se répartissent également dans les sous-groupes S1-2, S1-3 et S1-4 en fonction de la période choisie entre juillet et août pour prendre leurs vacances d'été. De plus, les cartes de C1 et C2 sont principalement classées dans les groupes S2 et S3, composés d'utilisateurs assez réguliers qui se déplacent aussi en fin de semaine. Cette correspondance fait écho aux proportions plus élevées de déplacements effectués en dehors des jours ouvrables relevées précédemment pour les groupes C1 et C2 (voir Tableau 5.13). Plus précisément, 56.1% des utilisateurs en C1 (utilisateurs les plus fréquents) sont dans S3 (variabilité intrapersonnelle moyenne: 0.22), alors que 51.9% des cartes en C2 (utilisateurs

les deuxièmes plus fréquents) se retrouvent en S2 (variabilité intrapersonnelle moyenne: 0.25 > 0.22). Les utilisateurs les plus fréquents en moyenne sont donc également les plus réguliers au niveau intrapersonnel puisque leur distance moyenne entre deux semaines successives est plus faible.

Tableau 7.7 Distribution des 11 groupes de la typologie basée sur des séquences dans les 6 groupes de la typologie basée sur des indicateurs d'utilisation

% de cartes		Typologie basée sur des indicateurs d'utilisation						
		C1	C2	C3	C4	C5	C6	Total
Typologie basée sur des séquences	S1-1	0.3%	4.5%	28.0%	57.9%	9.2%	0.0%	100.0%
	S1-2	0.4%	1.8%	12.5%	64.8%	18.9%	1.8%	100.0%
	S1-3	0.0%	0.4%	11.5%	67.5%	19.0%	1.6%	100.0%
	S1-4	0.0%	1.2%	8.3%	61.3%	26.2%	3.0%	100.0%
	S1-5	0.6%	0.6%	4.8%	52.4%	39.2%	2.4%	100.0%
	S2	2.3%	22.1%	34.2%	22.3%	14.1%	5.0%	100.0%
	S3	10.7%	35.0%	22.0%	13.6%	12.1%	6.5%	100.0%
	S4	0.0%	0.0%	0.0%	1.1%	4.0%	94.9%	100.0%
	S5	0.0%	0.7%	2.8%	17.4%	59.7%	19.4%	100.0%
	S6	0.0%	0.0%	0.8%	0.0%	15.3%	84.0%	100.0%
	S7	1.0%	0.0%	3.1%	15.5%	61.9%	18.6%	100.0%

Tableau 7.8 Distribution des 6 groupes de la typologie basée sur des indicateurs d'utilisation dans les 11 groupes de la typologie basée sur des séquences

% de cartes		Typologie basée sur des indicateurs d'utilisation					
		C1	C2	C3	C4	C5	C6
Typologie basée sur des séquences	S1-1	4.9%	12.6%	35.7%	34.2%	11.1%	0.0%
	S1-2	2.4%	2.1%	6.8%	16.2%	9.7%	1.3%
	S1-3	0.0%	0.4%	5.6%	15.2%	8.7%	1.0%
	S1-4	0.0%	0.8%	2.7%	9.2%	8.0%	1.3%
	S1-5	2.4%	0.4%	1.5%	7.8%	11.8%	1.0%
	S2	31.7%	51.9%	37.1%	11.2%	14.4%	7.3%
	S3	56.1%	31.4%	9.1%	2.6%	4.7%	3.7%
	S4	0.0%	0.0%	0.0%	0.2%	1.3%	43.5%
	S5	0.0%	0.4%	0.8%	2.2%	15.7%	7.3%
	S6	0.0%	0.0%	0.2%	0.0%	3.6%	28.8%
	S7	2.4%	0.0%	0.6%	1.3%	10.9%	4.7%
	Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Les cartes des groupes S4 et S6 (les deux groupes les moins stables au niveau intrapersonnel selon l'indicateur précédent) coïncident quant à elles principalement avec celles de C6 : les usagers appartenant à ces groupes sont donc à la fois peu fréquents et irréguliers. De même, les cartes de S5 et S7, caractérisées par une période estivale relativement longue et un assez haut niveau de variabilité intrapersonnelle (valeurs moyennes respectives de 0.31 et 0.33), sont classées majoritairement dans le groupe C5, donc au $5^{ème}$ niveau d'intensité d'utilisation. Dans la répartition inverse, le groupe C6 concorde également avec les groupes S4 et S6, et le groupe C5 se recoupe surtout avec le groupe S5. Ces distributions confirment que les utilisateurs les moins fréquents et les moins actifs au niveau mensuel tendent également à être moins réguliers au niveau hebdomadaire et vice versa. Cette conclusion révèle finalement les mêmes tendances que celles mises en évidence plus tôt avec les différents indicateurs de variabilité intrapersonnelle présentés à la section 6.2. Le tout valide ainsi les indicateurs et la typologie précédemment obtenus.

En outre, deux mesures internes de validation sont également calculées pour comparer la qualité des deux segmentations : l'indice de Dunn (D) et l'indice de silhouette (S). Ces deux indices ont été sélectionnés car ils sont basés sur les distances individuelles (mesurées entre paires d'observations) et non sur les centres des groupes formés. En effet, ces centres ne sont pas explicites dans le cas de la deuxième typologie réalisée à partir de séquences. Les deux indices choisis sont expliqués entre autres par Liu, Y. et al. (2010) ou encore Arbelaitz et al. (2013). D'une part, l'indice de Dunn est calculé comme le rapport entre la distance minimale entre deux cartes qui ne sont pas regroupées ensemble (mesure de séparation) et la distance maximale entre deux cartes d'un même groupe (mesure de compacité ou de cohésion). D'autre part, l'indice de silhouette (S) est une somme normalisée de toutes les différences entre les dissimilarités inter et intragroupes calculées pour chaque observation de chaque groupe. Les expressions mathématiques de ces indices sont données dans les références fournies. Le principe même de la segmentation étant de maximiser les distances entre les groupes (pour une meilleure séparation) et au contraire de minimiser les distances à l'intérieur de chaque groupe (pour une meilleure cohésion), la typologie optimale est celle pour laquelle ces deux indices sont les plus grands. Les résultats du calcul de ces deux indices sont rapportés dans le Tableau 7.9 ci-dessous. La distance euclidienne et la distance de Hamming pondérée (équation 27) ont été utilisées respectivement pour la première et la deuxième typologie.

Tableau 7.9 Calcul de deux indices de validation interne pour comparer la qualité des deux typologies

	INDICES DE VALIDATIONS	
TYPOLOGIE	D	S
Basée sur des indicateurs d'utilisation	0.013	0.450
Basée sur des séquences	0.033	0.110

L'indice de Dunn conduit à des résultats similaires pour les deux typologies, même si la valeur de l'indice est légèrement supérieure pour la seconde typologie, basée sur les séquences de types de semaines (0.033 > 0.013). En revanche, l'indice de Silhouette est manifestement plus élevé pour la première typologie, basée sur les indicateurs d'utilisation du transport en commun (0.450 > 0.110). De plus, selon la littérature, cet indicateur donne souvent de meilleurs résultats (Arbelaitz et al., 2013). La première typologie serait donc meilleure en termes de compacité et de séparation des groupes. Cependant, ce résultat est à nuancer et la deuxième typologie ne doit pas être discréditée pour autant. En effet, la première typologie a été réalisée sur un plus grand échantillon d'entraînement et pour une granularité moins fine (mensuelle plutôt qu'hebdomadaire) : il est donc normal qu'elle performe mieux. Par ailleurs, il n'y a pas de bonne ou de mauvaise segmentation. En tant que méthode non supervisée, la segmentation produit des résultats qui ne peuvent pas être vérifiés : on ne sait pas vraiment à quel groupe devrait appartenir chaque observation puisque ces groupes n'existent pas dans la réalité. L'objectif de la segmentation est néanmoins de permettre une meilleure interprétation des données. Or, sur ce point, la deuxième typologie obtenue est préférable puisqu'elle permet une analyse plus visuelle et plus intuitive des groupes d'utilisateurs par la représentation de leurs séquences de types de semaines. De plus, contrairement à la première, cette deuxième typologie prend en compte l'ordre des événements de mobilité et apporte ainsi des informations supplémentaires : elle renseigne notamment sur l'organisation des semaines de déplacements de chaque individu (et donc sur sa régularité intrapersonnelle au niveau hebdomadaire) et elle permet de confirmer l'impact de facteurs exogènes comme les périodes de vacances et les jours fériés. Par conséquent, la typologie optimale est celle qui répond le mieux à la finalité recherchée, selon ce que l'on souhaite montrer avec les données.

CHAPITRE 8 CONCLUSION ET RECOMMANDATIONS

Contrairement à ce qui est parfois supposé dans les modèles de prévision de la demande, l'achalandage du transport en commun est loin d'être constant dans le temps et dans l'espace. Cette variabilité est notamment due à des variations individuelles, observées entre les usagers (variations interpersonnelles), mais aussi dans le comportement longitudinal de chaque personne (variations intrapersonnelles). L'objectif de ce mémoire était de mieux comprendre et de mesurer cette variabilité individuelle de l'utilisation du transport en commun à l'aide de données de cartes à puce. Une synthèse de la méthodologie mise en œuvre et des résultats obtenus est ici présentée afin de souligner les contributions apportées par ce mémoire. Quelques limites sont également pointées, mais des perspectives d'améliorations et de futures explorations sont ensuite proposées.

8.1 Synthèse de la recherche

Une revue de littérature a d'abord permis d'introduire la technologie de la carte à puce et de relever l'exploitation des données transactionnelles produites dans de nombreux travaux scientifiques, faisant ainsi valoir leur richesse. Plus particulièrement, ces données ont souvent servi à décrire temporellement et spatialement l'utilisation du transport en commun; elles ont ainsi permis de mettre en évidence sa fluctuation. Différents moyens pour mesurer la variabilité des comportements de mobilité ont ensuite été recensés, ceux-ci prenant la forme d'indicateurs, de modèles, ou encore de méthodes pour analyser des séquences ou segmenter des usagers. Néanmoins, cette revue a permis de remarquer l'absence d'une définition claire de la variabilité, souvent confondue avec l'intensité et la périodicité. Une certaine carence a également été révélée dans les outils proposés, ces derniers ne permettant généralement pas de traiter les deux types de variabilité (inter et intrapersonnelle). Par ailleurs, le focus des auteurs sur la variabilité quotidienne a été expliqué par un manque de données longitudinales et individualisées.

Ce projet a au contraire bénéficié d'une base de données exhaustive de cartes à puce. Fournie par la STM, opérateur du réseau de transport en commun de Montréal, cette base a permis d'analyser les comportements individuels de mobilité sur une longue période d'étude d'un an. Ces comportements ont notamment été examinés aux niveaux mensuels et hebdomadaires. Basé sur l'exploitation des dites données et sur des définitions clarifiées dès le début de la recherche, ce mémoire a proposé une méthode globale pour évaluer la variabilité d'utilisation du transport en

commun à l'échelle individuelle en combinant plusieurs outils analytiques. Pour commencer, les validations des cartes ont été converties en déplacements, objets d'étude propres au transport, et une méthode de prétraitement des données a été développée afin de les rendre manipulables. Des vecteurs « cartes-année » et « cartes-semaine », composés d'indicateurs mensuels ou quotidiens de dispersion et d'intensité, ont ainsi été créés.

Dans un objectif de quantification de la variabilité observée, quatre indicateurs ont été construits pour mesurer différents types de variations dans l'utilisation du transport en commun. Ces indicateurs ont traduit la distribution des déplacements parmi les usagers, l'hétérogénéité des fréquences d'utilisation, ainsi que les variabilités temporelle et spatiale. Ils ont été mis en pratique pour comparer dix groupes de cartes définis en fonction de leur composition tarifaire. De plus, ils ont été vérifiés statistiquement à l'aide d'indices de taille d'effet, cette notion permettant de quantifier l'importance des différences observées indépendamment de la grande taille des échantillons qui fausse généralement les tests statistiques traditionnels. La principale conclusion de cette application est que les utilisateurs d'abonnements annuels et mensuels ont des comportements de mobilité très éloignés de ceux des utilisateurs de carnets de tickets, quel que soit le tarif utilisé (ordinaire ou réduit). Les indicateurs proposés ont en effet révélé que les utilisateurs d'abonnements annuels et mensuels étaient des clients plus fiables puisqu'ils sont à la fois fréquents et stables. Les utilisateurs d'abonnements annuels ont même tendance à être plus réguliers sur les plans temporel et spatial que les utilisateurs d'abonnements mensuels, mais l'analyse de la taille de l'effet a indiqué que les différences n'étaient pas si importantes. Au contraire, les utilisateurs de carnets sont des usagers occasionnels (faible fréquence d'utilisation) et une grande hétérogénéité a été soulignée entre ces utilisateurs (indices de Pareto faibles, coefficients de variation élevés).

Pour compléter et étoffer cette évaluation de l'utilisation du transport en commun, les variabilités inter et intrapersonnelles ont ensuite été analysées spécifiquement et séparément. Premièrement, toutes les cartes-années ont été segmentées afin de créer une typologie de l'ensemble des usagers de la STM. L'algorithme appliqué a fait ressortir différents profils d'utilisation, à savoir deux groupes d'usagers actifs toute l'année avec un niveau d'intensité différent, versus un groupe d'usagers plus occasionnels néanmoins présents toute l'année, un groupe d'étudiants renouvelant leur carte en septembre, ainsi que deux potentiels groupes de cyclistes utilisant le transport en commun pendant l'hiver. Cette interprétation a été appuyée par une analyse tarifaire de la distribution des déplacements et par le calcul de plusieurs indicateurs caractérisant l'utilisation et

la variabilité temporelle et spatiale (ou modale) de chaque groupe. Des préférences dans les types de jours choisis, les modes empruntés et les lieux d'embarquement ont ainsi été mises en évidences puis testées statistiquement. Les résultats ont révélé que les usagers les plus fréquents et les plus réguliers au niveau mensuel étaient principalement des utilisateurs d'abonnements annuels et mensuels qui se déplacent pendant les jours ouvrables, mais aussi pendant les jours non ouvrables. De plus, les déplacements de ces usagers sont assez bien répartis entre les modes métro et bus, ceci leur conférant une plus grande acquisition du réseau.

Le cas particulier des abonnements annuels avec une amplitude de 12 mois a également été examiné. Malgré leur grande régularité, différents niveaux d'intensité d'utilisation et d'activité ont été trouvés parmi ces usagers, leur fréquence diminuant pendant la période estivale à un degré plus ou moins élevé. De même, les indicateurs ont rapporté différentes caractéristiques d'utilisation dans chaque groupe. La variabilité intrapersonnelle des usagers de ces groupes a ensuite été étudiée au niveau hebdomadaire en créant une typologie de semaines puis en définissant des indicateurs qui mesuraient la répétition des mêmes types de semaines dans les comportements individuels. Cette même typologie de semaines a également servi à décrire chaque usager par une séquence de comportements hebdomadaires rangés dans l'ordre chronologique de leur apparition. Une typologie de séquences a ainsi été produite à partir d'une distance de Hamming pondérée, des indicateurs basés sur ces séquences ont été calculés, puis tous les résultats ont été croisés et validés. Ce long processus méthodologique a montré que les utilisateurs se déplaçant majoritairement pendant les jours ouvrables étaient les plus stables au niveau intrapersonnel, car leur comportement hebdomadaire est typiquement une semaine régulière de travail. En outre, les usagers les plus fréquents et actifs au cours de l'année, réguliers au niveau mensuel, ont également tendance à être plus réguliers au niveau hebdomadaire. Inversement, les utilisateurs occasionnels présentent un comportement plus variable et atypique. Néanmoins, ce dernier type d'utilisateurs représente une petite minorité des utilisateurs d'abonnements annuels, et une bonne partie d'entre eux sont probablement des employés qui ont reçu un abonnement gratuit avec leur entreprise.

Finalement, l'ensemble des résultats de ce mémoire confirme que la STM devrait poursuivre la fidélisation de ses clients avec un abonnement annuel. En effet, la majorité d'entre eux sont très fréquents et cette forte intensité d'utilisation du transport en commun est constante dans le temps. D'après l'analyse tarifaire faite au début de ce mémoire, l'abonnement annuel n'est que le troisième type de produit le plus utilisé à Montréal, après les carnets et les abonnements mensuels; sa

popularité peut donc encore être améliorée. Par ailleurs, quelques subtiles différences ont été détectées entre les utilisateurs d'abonnements annuels. Des usagers qui se déplacent davantage en fin de semaine ou qui ne se déplacent pas le vendredi (semaine de 4 jours) ont par exemple été identifiés. En dépit de leur comportement moins commun, ces usagers se sont aussi révélés très réguliers. Dans un contexte de tarification intégrée, ces résultats pourraient inciter la création de produits personnalisés (exemples : un titre « week-end » et un titre « 4 jours »), plus adaptés à ces types particuliers d'utilisateurs.

8.2 Contributions

Les contributions de ce mémoire sont multiples. D'une part, au niveau méthodologique, ce travail présente un exemple de traitement et de valorisation des données de cartes à puce pour l'étude des comportements de mobilité. Divers outils ont été mis à disposition afin de qualifier et de quantifier la variabilité individuelle d'utilisation du transport en commun. Un grand inventaire d'indicateurs simples et reproductibles a notamment été fourni : ceux-ci peuvent être exploités pour comparer différents groupes de cartes, mais aussi différentes années ou villes si les données correspondantes sont disponibles. De plus, des prototypes d'analyse des variabilités inter et intrapersonnelles ont été développés à partir d'algorithmes de segmentation et de la construction de séquences. Ces dernières ont en particulier permis la prise en compte de l'organisation temporelle des déplacements de chaque usager pendant l'année. Enfin, la notion de taille d'effet a été révélée dans le but de mieux interpréter les résultats des tests statistiques appliqués à de gros échantillons. Ce concept est particulièrement intéressant dans l'ère du Big Data.

D'autre part, en termes de résultats, plusieurs phénomènes ont été mis en évidence, confirmant ainsi l'utilité des données de cartes à puce pour analyser longitudinalement les comportements individuels de mobilité. L'existence de variations dans l'utilisation du transport en commun a largement été prouvée, que ce soit entre les usagers ou pour une même personne. Cette variabilité ne peut donc plus être niée et devrait être considérée par les différents acteurs du transport en commun. L'intérêt de prendre en compte l'ordre des événements de mobilité a également été confirmé puisque la deuxième typologie basée sur des séquences a apporté des informations supplémentaires par rapport à celle basée sur des vecteurs d'indicateurs agrégés avec la distance euclidienne. L'effet des congés (vacances, jours fériés) et du type de jour (semaine versus fin de semaine) a notamment pu être vérifié. Les services offerts par la STM sont déjà modifiés en

fonction de ces circonstances particulières mais pourraient sûrement être optimisés. De surcroît, une relation entre les titres de transport utilisés et les comportements observés a été montrée. Cette connaissance est notamment importante pour guider la politique de tarification des réseaux de transport, désormais assurée par l'ARTM dans le cas de Montréal. Cependant, la relation cause/conséquence induite par ce rapprochement n'est pas claire et la question suivante mériterait plus d'investigations: est-ce qu'un certain type de comportement individuel entraîne le choix d'un certain titre de transport ou bien est-ce l'achat de ce titre qui influence la mobilité de l'utilisateur ?

En conclusion, le développement des analyses amorcées dans ce projet pourrait bénéficier à différents protagonistes. Pour le planificateur, une meilleure compréhension des comportements de mobilité pourrait permettre une meilleure modélisation de la demande. La prise en compte de la variabilité individuelle aiderait notamment à obtenir des prévisions plus justes que celles basées sur un jour moyen de semaine. Pour l'opérateur, cette amélioration de la prédiction de la demande permettrait d'ajuster l'offre à une granularité plus fine, réduisant ainsi les coûts d'opération et permettant une meilleure allocation des véhicules sur le réseau. La réalisation de typologies comme celles présentées dans ce mémoire fournit également une meilleure connaissance de la clientèle desservie. Des services personnalisés et adaptés aux besoins de chaque type d'usagers pourraient ainsi être projetés. De plus, des études plus poussées seraient capables de déceler des opportunités de fidélisation et d'orienter la stratégie tarifaire. Des tarifs spéciaux et intégrés pourraient notamment être créés afin de retenir les usagers les plus réguliers. Enfin, du point de vue de l'utilisateur, la personnalisation des services augmenterait sa satisfaction.

8.3 Limites

Toutefois, les travaux réalisés dans le cadre de ce mémoire comportent quelques limitations. Celles-ci sont principalement liées aux données utilisées et à la méthodologie proposée.

8.3.1 Limites relatives aux données utilisées

La première limitation de cette catégorie est fondamentale et provient de la nature même des données : dans cette analyse prétendue de la variabilité « individuelle » de l'utilisation du transport en commun, des cartes et non des individus sont étudiées (même si les deux ont parfois été confondus dans ce mémoire par abus de langage). En effet, dans le système de la STM, une carte peut être utilisée par plusieurs usagers si elle est prêtée et un même usager peut également utiliser

plusieurs cartes s'il perd ou renouvelle sa carte durant l'année. Cette différence d'objet d'analyse a notamment généré quelques problèmes dans les segmentations réalisées précédemment, un même usager ayant possédé deux cartes pendant l'année pouvant se retrouver dans deux groupes distincts. Les étudiants ont par exemple été réunis dans le même groupe du fait de leur renouvellement obligatoire en septembre : cet événement ponctuel a participé à leur regroupement peut-être même plus que la similarité de leur comportement.

Par ailleurs, les données de cartes à puce exploitées dans ce projet ne sont pas sans défauts. Des anomalies liées à des dysfonctionnements du système ou à des erreurs de manipulation humaines peuvent s'être introduites dans les données. Quelques-unes des sources de problèmes possibles ont été répertoriées dans le premier chapitre de ce mémoire. De même, le projet pilote d'embarquement par toutes les portes initié par la STM en 2016 a entraîné une perte d'informations pour les validations de bus, impactant ainsi la répartition des déplacements de chaque usager par mode, et les arrêts d'embarquement du bus n'ont pas pu être géolocalisés ni utilisés dans les calculs d'entropie spatiale car l'information est manquante dans les données OPUS. En outre, comme aucun filtre n'a été appliqué en amont pour pouvoir analyser de manière exhaustive tous les types de comportements, de nombreuses valeurs aberrantes sont présentes dans les données. Ces usagers extrêmes ont notamment pu fausser les résultats des indicateurs appliqués, ceux-ci étant pour la plupart calculés comme une moyenne. Les grands coefficients de variation rapportés à l'intérieur de certains groupes dans les typologies précédemment produites ont parfois signalé ce problème. Les valeurs moyennes obtenues n'étaient donc pas toujours représentatives de l'ensemble des membres du groupe.

La méthode de prétraitement des données mise en œuvre au début du projet n'est pas non plus parfaite. Les validations ont notamment été transformées en déplacements à partir d'une logique tarifaire et non à partir d'une logique de mobilité : de « fausses correspondances » peuvent ainsi avoir été créées. En effet, les usagers de la STM sont capables de maîtriser et de contourner les règles d'affaires du système OPUS. À titre d'exemple, la même ligne de bus ne pouvant pas être empruntée deux fois dans le même déplacement, certains usagers (détenteurs de passages et non de titres à durée) utiliseront un autre trajet et donc une autre ligne de bus pour faire leur déplacement retour. Si elles sont réalisées dans un délai de 120 minutes, les deux validations faites à bord des deux lignes de bus seront considérées comme un même déplacement alors que deux déplacements ont en fait été effectués : un déplacement aller et un déplacement retour. En outre, la section 4.1.4

du mémoire de Giraud (2016) prouve que le calcul des déplacements est très sensible au seuil temporel tarifaire appliqué (ici, 120 minutes).

8.3.2 Limites méthodologiques

La méthodologie proposée comporte également quelques imperfections. Tout d'abord, les combinaisons de cartes définies dans la section 4.2 ne sont pas toutes très pertinentes. En effet, il était difficile de tirer des conclusions pour les combinaisons regroupant des utilisateurs ayant plus de diversité dans l'achat de leurs titres de transport (utilisation de plus d'un seul type de produit ou de tarif). Ces groupes étant composés d'usagers très diversifiés, allant de l'utilisateur de billet unitaire à l'utilisateur d'abonnement annuel, ils présentaient des similitudes avec de nombreuses autres combinaisons. Des analyses supplémentaires devraient donc être entreprises pour explorer ces groupes plus en détail. Cet approfondissement permettrait également de mieux comprendre les raisons qui poussent ces usagers à changer souvent de titre de transport.

De plus, les indicateurs développés se sont révélés très dépendants de la fréquence d'utilisation des usagers. Ainsi, une plus grande variabilité temporelle et spatiale a souvent été rapportée pour les groupes ayant une forte intensité d'utilisation du transport en commun. Cette tendance était notamment flagrante pour l'indicateur d'entropie : les usagers les plus occasionnels cumulant un plus grand nombre de probabilités nulles (probabilité de valider sa carte à une certaine station), leur entropie individuelle moyenne était nettement diminuée et leur faible intensité d'utilisation du transport en commun impactait ainsi l'évaluation de leur régularité. Cette dépendance a rendu difficile la comparaison des groupes d'usagers avec des niveaux d'utilisation différents, par exemple la comparaison entre les utilisateurs d'abonnements et les utilisateurs de carnets. Une normalisation des indicateurs serait donc nécessaire pour s'affranchir de cette fréquence. Par ailleurs, les indices de taille d'effet calculés pour vérifier statistiquement ces indicateurs ont peut-être mal été traduits, leur interprétation étant basée sur un critère arbitraire et non adapté au contexte du transport.

Enfin, la principale faiblesse des premières segmentations réalisées (première typologie d'usagers et typologie de semaines) est liée à l'application de la distance euclidienne pour mesurer les dissimilarités entre les usagers. Les inconvénients de cette distance peuvent notamment être mis en évidence en redécomposant les plus gros groupes de la typologie de tous les usagers de la STM (typologie de la section 5.2). Les résultats de ces nouvelles segmentations sont brièvement donnés

en ANNEXE H. Ils montrent que des profils mensuels très différents voire complètement opposés ont parfois été réunis dans un même groupe. C'est particulièrement le cas pour les groupes C1 et C2. De même, les usagers des sous-groupes de C6 sont actifs à différentes périodes de l'année et leur somme seulement est présente toute l'année sur le réseau. Cette faille est due au calcul de la distance euclidienne qui, en sommant les différences au carré de toutes les proportions de déplacements par mois, a masqué la position de chacun de ces mois dans le vecteur. La méthode des K-moyennes et en particulier l'application de la distance euclidienne sont donc peu adaptées à des séries chronologiques, ce qui confirme les critiques rapportées dans la revue de littérature. En outre, cette distance est également sensible aux unités utilisées et donc à la méthode de normalisation des vecteurs « cartes-année » et « cartes-semaine ».

De même, dans la typologie de séquences, la distance euclidienne a été utilisée pour évaluer les dissimilarités entre les types de semaines (Tableau 6.3) et l'ordre des jours de la semaine n'a donc pas été considéré. Des distances entre certains types de comportements hebdomadaires ont peut-être ainsi été sur ou sous-évaluées. Ces valeurs ont impacté le calcul de la distance de Hamming pondérée et donc ensuite la segmentation basée sur cette distance. Une deuxième limitation peut être relevée pour cette typologie de séquences : celle-ci concerne sa forte dépendance à la définition choisie pour la variabilité intrapersonnelle. En effet, cette variabilité a été examinée ici en fonction d'un cycle hebdomadaire, mais il a été rappelé par des auteurs cités dans la revue de littérature que la régularité des comportements ne suivait pas nécessairement un cycle préfini.

8.4 Perspectives

Pour répondre à ces limites, quelques pistes de solution peuvent être avancées. De plus, ce sujet de recherche étant très large, beaucoup d'autres pistes pourraient encore être explorées. Ce mémoire est une première contribution sur la variabilité individuelle d'utilisation du transport en commun analysée à partir des données de cartes à puce. Tout n'a donc pas pu être fait, laissant ainsi beaucoup de possibilités pour de futurs étudiants sur ce sujet.

8.4.1 Quelques pistes de solutions aux limites

Parmi les limites relatives aux données utilisées, la distinction entre carte et usager est problématique pour analyser la mobilité à l'échelle individuelle. Néanmoins, la régularité des comportements peut justement être exploitée pour développer des techniques de reconnaissance

d'usagers sur différentes périodes temporelles (Espinoza et al., 2017). Concernant la qualité des données, quelques logiques de validation à vérifier peuvent être trouvées dans le mémoire de maîtrise de Gagné (2006). De plus, la géolocalisation des arrêts d'embarquement du bus sera bientôt possible grâce aux données GPS du projet iBUS de la STM (Société de transport de Montréal, 2018c). Par ailleurs, les cartes aberrantes peuvent être étudiées séparément afin d'examiner leur influence (Cui, Z. et al., 2015). Les usagers très peu mobiles sont notamment assez difficiles à analyser avec les méthodes d'exploration de données traditionnelles puisqu'ils sont très peu observés sur les réseaux de transport. Leurs déplacements pourraient peut-être néanmoins être modélisés comme des événements rares tels que les accidents routiers. Enfin, en ce qui a trait au prétraitement des données, Exo est en train de développer un nouvel algorithme de calcul des déplacements à partir de toutes les validations OPUS et non seulement celles du réseau de la STM (Nazem & Fortin, 2018). Cet algorithme devrait être basé sur une logique de mobilité et non plus sur une logique tarifaire, les règles d'affaires étant différentes d'un réseau à l'autre. Des méthodes géométriques de calcul existent également dans la littérature (Munizaga & Palma, 2012).

Au niveau méthodologique, les indicateurs proposés dans ce mémoire peuvent être améliorés et normalisés avec la fréquence d'utilisation. Pour le calcul d'entropie, il faudrait probablement considérer uniquement les stations utilisées par chaque usager (c'est-à-dire normaliser chaque entropie par $\log(n_i)$ plutôt que $\log(n)$, où n_i est le nombre de stations utilisées par l'usager i). On pourrait également sélectionner les x stations les plus visitées de chaque usager et calculer l'entropie correspondante. Pour la variabilité temporelle, une variance individuelle pourrait être mesurée pour chaque usager pour les mois où il s'est déplacé. Une moyenne serait ensuite évaluée dans chaque groupe. Les autres indicateurs pourraient peut-être simplement être divisés par la fréquence d'utilisation. Ces propositions sont des idées qu'il faudrait vérifier et la sensibilité des indicateurs à ces différentes définitions devrait également être contrôlée. En outre, ces indicateurs et les indices de taille d'effet associés pourraient ensuite être estimés pour d'autres modes de transport. Cette plus large application permettrait de mieux définir les valeurs limites attribuées aux adjectifs « petit », « moyen » et « grand » du critère de Cohen dans un contexte de mobilité. L'incidence des effets mesurés sur le réseau opéré pourrait notamment être considérée dans la détermination de cette échelle. Enfin, d'autres métriques pourraient être testées dans les processus de segmentation d'usagers, de semaines et de séquences développés. Des distances adaptées à des séries temporelles comme la distance de déformation temporelle dynamique (He et al., 2018)

pourrait peut-être permettre de prendre en compte le caractère chronologique des variables de chaque vecteur.

8.4.2 Futures explorations

Par ailleurs, beaucoup d'autres idées restent à explorer ; ce sujet semble illimité. Tout d'abord, il serait intéressant d'effectuer le travail inverse de ce qui a été fait dans ce mémoire, c'est-à-dire d'utiliser les indicateurs proposés pour segmenter les usagers. Une typologie pourrait être réalisée avec les indicateurs temporels, une autre avec les indicateurs spatiaux, puis les résultats pourraient être croisés afin de déterminer les usagers réguliers sur deux niveaux (temporel et spatial), sur un seul niveau (temporel ou spatial) ou sur aucun niveau. De plus, de nombreuses autres segmentations d'usagers sont possibles d'après le schéma de la Figure 5.1. Seuls deux « chemins » ont été empruntés dans ce mémoire. La séparation des déplacements effectués en jours ouvrables et en jours non ouvrables pourrait notamment permettre d'affiner les résultats obtenus. Toutes ces typologies ne feraient qu'augmenter la connaissance de la clientèle de la STM et, plus généralement, la compréhension des comportements de mobilité.

Bien entendu, la méthodologie ici développée et validée dans le cas des abonnements annuels pourrait être appliquée à d'autres titres de transport. Les utilisateurs d'abonnements annuels étant en grande majorité des passagers fréquents et réguliers qui empruntent principalement le transport en commun pour se rendre au travail pendant les jours ouvrables, les résultats obtenus n'ont pas été extraordinaires et les groupes plus atypiques se sont révélés très minoritaires. Néanmoins, une plus grande variété de titres pourrait conduire à une plus grande diversité de comportements. Cela n'a pas été fait par manque de temps, mais la création d'une typologie de séquences par titre de transport utilisé auraient sûrement permis la mise en évidence de différentes organisations dans les comportements hebdomadaires des usagers. En particulier, il pourrait être lucratif de comparer la mobilité des utilisateurs d'abonnements annuels (amplitude de 12 mois) avec celle des utilisateurs d'abonnements mensuels (amplitude de 12 mois) : la possession d'un abonnement annuel (soit 12 mois d'abonnements mensuels) confère-t-elle des particularités que des usagers achetant chaque mois leur abonnement n'auraient pas ?

L'analyse des séquences produites peut également être poussée. La recherche de sous-séquences communes est une tâche qui pourrait notamment être explorée et automatisée. De plus, la répétition de ces sous-séquences ouvertes dans les comportements pourrait être quantifiée à l'aide d'une

entropie dite symbolique (Ruiz et al., 2012). Cette mesure permettrait d'évaluer la variabilité intrapersonnelle à partir de la structure organisationnelle des séquences en autorisant des décalages temporels. Par ailleurs, la méthodologie proposée dans ce mémoire a couvert les niveaux mensuels et hebdomadaires, mais elle pourrait être étendue à l'échelle quotidienne. Des séquences de types de jours pourraient ainsi être construites à partir des heures d'embarquement des usagers. Cet élargissement permettrait notamment de tester la sensibilité de la définition de la régularité intrapersonnelle. Plusieurs années pourraient également être étudiées pour analyser la stabilité des comportements et la fidélité des usagers sur une plus longue durée. L'idéal serait de générer une typologie exhaustive et durable dans le temps, de manière à pouvoir prédire l'appartenance à des groupes d'une année à l'autre sans avoir à recréer une nouvelle typologie. En outre, la composante spatiale pourrait être mieux intégrée et documentée à plusieurs échelles (arrêt, ligne, corridor). Les différentes analyses longitudinales produites pourraient également être croisées avec l'utilisation du sol grâce à la géolocalisation des principaux lieux d'embarquement des usagers.

Pour finir, un volet de modélisation manque à ce projet. Les indicateurs présentés dans ce mémoire pourraient être ainsi intégrés à un modèle en tant que variable dépendante ou indépendante. D'une part, la variabilité (ou régularité) individuelle des usagers du transport en commun pourrait peut-être être prédite à partir des titres de transport utilisés pendant l'année. En effet, cette information tarifaire est la seule vraie variable d'entrée dont on dispose, les autres données de cartes à puce faisant référence à des comportements d'utilisation observés (à moins que l'observation de ces comportements sur une courte période puisse permettre de prévoir leur variabilité sur une plus longue durée). D'autre part, un modèle de prévision de l'achalandage pourrait être développé afin de prendre en compte non seulement l'effet de la variabilité individuelle (intra et interpersonnelle), mais aussi l'impact de facteurs exogènes tels que la météo, les perturbations du service, ou encore l'interaction avec les autres modes de transport.

BIBLIOGRAPHIE

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

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

Agard, B., Partovi Nia, V., & Trépanier, M. (2013). *Assessing public transport travel behaviour from smart card data with advanced data mining techniques*. Communication présentée à 13th World Conference on Transport Research, Rio de Janeiro, Brazil.

Agence métropolitaine de transport. (2016). *Rapport annuel 2016*. Tiré de http://www.bv.transports.gouv.qc.ca/per/1133185/05_2016.pdf

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

Anand, S. S., & Büchner, A. G. (1998). *Decision support using data mining*: Financial Times Pitman Publishers.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243-256. doi:10.1016/j.patcog.2012.07.021

Attoh-Okine, N. O., & Shen, L. D. (1995). *Security issues of emerging smart cards fare collection application in mass transit* (p. 523-526). doi:10.1109/VNIS.1995.518887

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

Bagchi, M., & White, P. R. (2005). The potential of public transport smart card data. *Transport Policy, 12*(5), 464-474. doi:10.1016/j.tranpol.2005.06.008

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

Barry, J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board, 1817*, 183-187. doi:10.3141/1817-24

Batagelj, V. (1988). Generalized Ward and related clustering problems. *Classification and related methods of data analysis*, 67-74.

Bhat, C. R., Srinivasan, S., & Axhausen, K. W. (2005). An analysis of multiple interepisode durations using a unifying multivariate hazard model. *Transportation Research Part B, 39*(9), 797-823. doi:10.1016/j.trb.2004.11.002

Blythe, P. T. (2004). Improving public transport ticketing through smart cards. *Proceedings of the Institute of Civil Engineers - Municipal Engineer, 157*(1), 47-54. doi:10.1680/muen.2004.157.1.47

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

Chapleau, R., & Chu, K. K. A. (2007). *Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach.* Communication présentée à 11th World Conference on Transport Research.

Chikaraishi, M., Fujiwara, A., Zhang, J. Y., & Axhausen, K. W. (2009). Exploring Variation Properties of Departure Time Choice Behavior by Using Multilevel Analysis Approach. *Transportation Research Record: Journal of the Transportation Research Board, 2134*(2134), 10-20. doi:10.3141/2134-02

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

Chu, K., & Chapleau, R. (2008). Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board, 2063*(2063), 63-72. doi:10.3141/2063-08

Chu, K., & Chapleau, R. (2010). Augmenting Transit Trip Characterization and Travel Behavior Comprehension: Multiday Location-Stamped Smart Card Transactions. *Transportation Research Record: Journal of the Transportation Research Board*(2183), 29-40. doi:10.3141/2183-04

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

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), 997-1003.

Conover, W. J., Johnson, M. E., & 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), 351–361.

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The Handbook of Research Synthesis and Meta-Analysis*: Russell Sage Foundation.

Cui, C. L., Zhao, Y. L., & Duan, Z. Y. (2014). Research on the Stability of Public Transit Passenger Travel Behavior Based on Smart Card Data. Dans A. S. o. C. Engineers (édit.), *CICTP 2014: Safe, Smart, and Sustainable Multimodal Transportation Systems* (p. 1318-1326).

Cui, Z., Long, Y., Ke, R., & Wang, Y. (2015). *Characterizing evolution of extreme public transit behavior using smart card data.* Communication présentée à IEEE First International Smart Cities Conference (ISC2) (p. 1-6). doi:10.1109/ISC2.2015.7366217

De Oña, J., De Oña, R., Diez-Mesa, F., Eboli, L., & Mazzulla, G. (2016). A Composite Index for Evaluating Transit Service Quality across Different User Profiles. *Journal of Public Transportation, 19*(2), 128-153. doi:10.5038/2375-0901.19.2.8

Deakin, E., & Kim, S. (2001). Transportation Technologies: Implications for Planning. UC Berkeley: University of California Transportation Center.

Dempsey, P. S. (2008). Privacy issues with the use of smart cards. *Legal Research Digest*.

Descoimps, E. (2011). *Analyse des données issues d'un système de perception par carte à puce d'une société de transport en commun : Normalité des déplacements et influence des conditions météorologiques.* (École Polytechnique de Montréal).

Deza, M. M., & Deza, E. (2013). *Encyclopedia of Distances* (2e éd.): Springer.

Dharmowijoyo, D. B. E., Susilo, Y. O., & Karlström, A. (2016). Day-to-day variability in travellers' activity-travel patterns in the Jakarta metropolitan area. *Transportation, 43*(4), 601-621. doi:10.1007/s11116-015-9591-4

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

El Mahrsi, M. K., Côme, E., Baro, J., & Oukhellou, L. (2014). Understanding passenger patterns in public transit through smart card and socioeconomic data. *UrbComp'*.

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

Elango, V., Guensler, R., & Ogle, J. (2007). Day-to-Day Travel Variability in the Commute Atlanta, Georgia, Study. *Transportation Research Record: Journal of the Transportation Research Board, 2014*, 39-49. doi:10.3141/2014-06

Eliazar, I. (2016). Harnessing inequality. *Physics Reports - Review Section of Physics Letters, 649*, 1-29. doi:10.1016/j.physrep.2016.07.005

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR), 45*(1), 1-34. doi:10.1145/2379776.2379788

Espinoza, C., Munizaga, M., Bustos, B., & Trépanier, M. (2017). *Assessing the public transport travel behavior consistency from smart card data*. Communication présentée à ISCTSC 2017, 11th International Conference on Transport Survey Methods, Québec, Canada.

Eurosmart. (2018). Eurosmart's secure element shipment forecasts confirm a global market increase in all main market sectors.

Exo. (2018). Organismes de transport du Grand Montréal, qui fait quoi ? Tiré de https://rtm.quebec/fr/actualites/nouvelles-evenements/nouvelles/exo-organismes-transport-qui-fait-quoi2

Foell, S., Kortuem, G., Rawassizadeh, R., Phithakkitnukoon, S., Veloso, M., & Bento, C. (2013). *Mining temporal patterns of transport behaviour for predicting future transport usage.*

Communication présentée à Proceedings of the 2013 ACM conference on pervasive and ubiquitous computing adjunct publication (p. 1239-1248). doi:10.1145/2494091.2497354

Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M., & Bento, C. (2014). *Catch me if you can: Predicting mobility patterns of public transport users.* Communication présentée à 17th International IEEE Conference on Intelligent Transportation Systems (ITSC) (p. 1995-2002). doi:10.1109/ITSC.2014.6957997

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

Gagné, S. (2006). *Validation par logiques et traitement de données cartes à puce dans un contexte de transport en commun.* (École Polytechnique de Montréal).

Ghaemi, M. S., Agard, B., Trépanier, M., & Partovi Nia, V. (2017). A visual segmentation method for temporal smart card data. *Transportmetrica A: Transport Science, 13*(5), 381-404. doi:10.1080/23249935.2016.1273273

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

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

Goulet-Langlois, G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C, 64*, 1-16. doi:10.1016/j.trc.2015.12.012

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

Habib, K. M. N., & Hasnine, S. (2017). *Econometric Investigation of the Influence of Transit Passes on Transit Users' Behavior in Toronto, Canada*. Communication présentée à 96th Annual Meeting of the Transportation Research Board Washington, D.C.

Habib, K. M. N., & Miller, E. J. (2008). Modelling daily activity program generation considering within-day and day-to-day dynamics in activity-travel behaviour. *Transportation, 35*(4), 467-484. doi:10.1007/s11116-008-9166-8

Hägerstrand, T. (1970). *What about people in regional science?* Communication présentée à Papers of the Regional Science Association (vol. 24, p. 6-21).

Hay, B., Wets, G., & Vanhoof, K. (2004). Mining Navigation Patterns Using a Sequence Alignment Method. *Knowledge and Information Systems, 6*(2), 150-163. doi:10.1007/s10115-003-0109-6

He, L., Agard, B., & Trépanier, M. (2018). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. *Transportmetrica A: Transport Science*, 1-20. doi:10.1080/23249935.2018.1479722

Hofmann, M., & O'Mahony, M. (2005). *Transfer journey identification and analyses from electronic fare collection data.* Communication présentée à Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE (p. 34-39).

Huang, J., Xu, L., & Ye, P. (2015). Exploring Transit Use Regularity Using Smart Card Data of Students. Dans Q. PengK. C. P. WangX. Liu & B. Chen (édit.), *ICTE 2015* (p. 617 - 625). Dalian, China.

Ibrahim, M. F. (2003). Car ownership and attitudes towards transport modes for shopping purposes in Singapore. *Transportation, 30*(4), 435-457. doi:10.1023/A:1024701011162

Inoue, J.-i., Ghosh, A., Chatterjee, A., & 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), 184-204.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR), 31*(3), 264-323. doi:10.1145/331499.331504

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Unsupervised Learning. Dans *An introduction to statistical learning: with applications in R* (vol. 103). New York, NY: Springer.

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

Joh, C.-H., & 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*(2231), 10-17. doi:10.3141/2231-02

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

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

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

Kitamura, R., Yamamoto, T., Susilo, Y. O., & 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), 259-279. doi:10.1016/j.tra.2005.07.002

Kruskal, J., & Liberman, M. (1983). The symmetric time-warping problem: from continuous to discrete. Dans *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (p. 125 -161): D Sankoff, J Kruskal (Addison-Wesley, Reading, MA).

Lathia, N., Froehlich, J., & Capra, L. (2010). *Mining Public Transport Usage for Personalised Intelligent Transport Systems* (p. 887-892). doi:10.1109/ICDM.2010.46

Li, Z., Wang, J., & Han, J. (2015). ePeriodicity: Mining Event Periodicity from Incomplete Observations. *IEEE Transactions on Knowledge and Data Engineering, 27*(5), 1219-1232. doi:10.1109/TKDE.2014.2365801

Liu, L., Hou, A., Biderman, A., Ratti, C., & Chen, J. (2009). *Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen*.

Communication présentée à 2009 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, Missouri.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). *Understanding of internal clustering validation measures.* Communication présentée à Data Mining (ICDM), 2010 IEEE 10th International Conference on (p. 911-916).

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems, 53*, 19-35. doi:10.1016/j.compenvurbsys.2015.02.005

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

Machado, J. L., de Oña, R., Diez-Mesa, F., & de Oña, J. (2018). Finding service quality improvement opportunities across different typologies of public transit customers. *Transportmetrica A: Transport Science*, 1-23. doi:10.1080/23249935.2018.1434257

Madhulatha, T. S. (2012). An overview on clustering methods. *IOSR Journal of Engineering, 2*(4), 719-725. doi:arXiv:1205.1117v1

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

McDonald, N. (2000). Multipurpose smart cards in transportation: benefits and barriers to use. *University of California Transportation Center*.

Meloche, J.-P., Perreault, S., & Martel-Castonguay, J.-F. (2012). *Le financement du transport en commun dans la région métropolitaine de Montréal - Pour un meilleur équilibre entre la ville et ses banlieues*. Tiré de http://www.obsmobilitedurable.umontreal.ca/recherche/pdf/Note01-2012_JPMeloche.pdf

Moiseeva, A., Tinnmermans, H., Choi, J., & 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*(2412), 49-56. doi:10.3141/2412-06

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

Morency, C., Trepanier, M., Frappier, A., & Bourdeau, J. S. (2017). *Longitudinal Analysis of Bikesharing Usage in Montreal, Canada*. Communication présentée à 96th Annual Meeting of the Transportation Research Board, Washington, D.C.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies, 24*, 9-18. doi:10.1016/j.trc.2012.01.007

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification, 31*(3), 274-295. doi:10.1007/s00357-014-9161-z

Nassir, N., Hickman, M., & Ma, Z.-L. (2015). Activity detection and transfer identification for public transit fare card data. *Transportation, 42*(4), 683-705.

Nazem, M., & Fortin, P. (2018). *Variabilité des typologies de déplacement en transport collectif dérivées de données passives.* Communication présentée à 53e Congrès de l'AQTr : l'innovation, ça nous transporte!, Québec.

Nishiuchi, H., King, J., & Todoroki, T. (2013). Spatial-Temporal Daily Frequent Trip Pattern of Public Transport Passengers Using Smart Card Data. *International Journal of Intelligent Transportation Systems Research, 11*(1), 1 - 10. doi:10.1007/s13177-012-0051-7

Ortega-Tong, M. A. (2013). *Classification of London's public transport users using smart card data.* (Massachusetts Institute of Technology).

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

Pareto, V. (1896-97). *Cours d'économie politique professé à l'Université de Lausanne*: Lausanne, F. Rouge.

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

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

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). *Activity-aware map: Identifying human daily activity pattern using mobile phone data.* Communication présentée à International Workshop on Human Behavior Understanding (p. 14-25).

Loi modifiant principalement l'organisation et la gouvernance du transport collectif dans la région métropolitaine de Montréal : RLRQ, chapitre O-7.3 (2016), à jour au 1er juin 2018.

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

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

Reddy, A., Lu, A., Kumar, S., Bashmakov, V., & Rudenko, S. (2009). *Application of entry-only automated fare collection (AFC) system data to infer ridership, rider destinations, unlinked trips, and passenger miles.* Communication présentée à 88th Annual Meeting of the Transportation Research Board, Washington.

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

Ruiz, M., López, F., & Páez, A. (2012). Comparison of thematic maps using symbolic entropy. *International Journal of Geographical Information Science, 26*(3), 413-439. doi:10.1080/13658816.2011.586327

Saneinejad, S., & Roorda, M. J. (2009). Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. *Transportation Letters-The International Journal Of Transportation Research, 1*(3), 197-211. doi:10.3328/TL.2009.01.03.197-211

Schlich, R. (2003). *Homogenous groups of travellers*. Communication présentée à 10th International Conference on Travel Behaviour Research (IATBR 2003), Lucerne, Switzerland.

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

Seaborn, C., Attanucci, J., & Wilson, N. (2009). Using Smart Card Fare Payment Data To Analyze Multi-Modal Public Transport Journeys in London. *Transportation Research Record: Journal of the Transportation Research Board, 2121*, 55-62. doi:10.3141/2121-06

Secure Technology Alliance. (2018a). About Smart Cards : Introduction : Primer. Tiré de http://www.smartcardalliance.org/smart-cards-intro-primer/

Secure Technology Alliance. (2018b). Smart Cards Applications. Tiré de http://www.smartcardalliance.org/smart-cards-applications/

Shelfer, K., & Procaccino, J. (2002). Smart card evolution (vol. 45, p. 83-88). New York: ACM.

Smart Card Basics. (2018a). Smart Card Overview. Tiré de http://www.smartcardbasics.com/smart-card-overview.html

Smart Card Basics. (2018b). Types of Smart Card. Tiré de http://www.smartcardbasics.com/smart-card-types.html

Société de transport de Montréal. (2008). Lancement de la carte OPUS dans les transports collectifs au Québec.

Société de transport de Montréal. (2016a). Embarquement par toutes les portes sur la ligne de bus 121.

Société de transport de Montréal. (2016b). Embarquement par toutes les portes sur les lignes 139 et 439.

Société de transport de Montréal. (2017a). *Budget annuel 2018*. Tiré de http://stm.info/sites/default/files/pdf/fr/budget2018.pdf

Société de transport de Montréal. (2017b). *Plan stratégique organisationnel 2025*. Tiré de http://stm.info/sites/default/files/pso-2025.pdf

Société de transport de Montréal. (2018a). Découvrez la STM et son histoire > Histoire. Tiré de http://stm.info/fr/a-propos/decouvrez-la-STM-et-son-histoire/histoire

Société de transport de Montréal. (2018b). En savoir plus sur votre carte OPUS. Tiré de http://www.stm.info/fr/infos/titres-et-tarifs/carte-opus-et-autres-supports/carte-opus

Société de transport de Montréal. (2018c). iBUS : En route vers le temps réel. Tiré de http://www.stm.info/fr/a-propos/grands-projets/ibus

Société de transport de Montréal. (2018d). Pour une bonne utilisation de votre carte OPUS. Tiré de http://www.stm.info/sites/default/files/pdf/fr/stminfo/180103-stminfo.pdf

Société de transport de Montréal. (2018e). *Rapport annuel 2017*. Tiré de http://stm.info/sites/default/files/pdf/fr/ra2017.pdf

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

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*(1), 1 - 34. doi:10.1348/000711005X48266

Strauss, T., & von Maltitz, M. J. (2017). Generalising Ward's Method for Use with Manhattan Distances. *PLoS ONE, 12*(1). doi:10.1371/journal.pone.0168288

Tao, S., Rohde, D., & Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *Journal of Transport Geography, 41*, 21-36. doi:10.1016/j.jtrangeo.2014.08.006

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), 139–152. doi:10.1007/s13547-011-0019-z

Trépanier, M., Barj, S., Dufour, C., & Poilpré, R. (2004). *Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain.* Communication présentée à Congrès de l'Association des transports du Canada.

Trépanier, M., & Morency, C. (2010). *Assessing transit loyalty with smart card data.* Communication présentée à 12th World Conference on Transport Research, Lisbon, Portugal.

Trépanier, M., Morency, C., & Agard, B. (2009). Calculation of Transit Performance Measures Using Smartcard Data. *Journal of Public Transportation, 12*(1), 79-96. doi:10.5038/2375-0901.12.1.5

Trépanier, M., Morency, C., & Blanchette, C. (2009). *Enhancing household travel surveys using smart card data.* Communication présentée à 88th Annual Meeting of the Transportation Research Board

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems, 11*(1), 1-14. doi:10.1080/15472450601122256

Trépanier, M., & Vassiviere, F. (2008). *Democratized Smartcard Data for Transit Operator.* Communication présentée à 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting.

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning. *Transportation Research Record: Journal of the Transportation Research Board, 1971*, 119-126. doi:10.3141/1971-16

Van der Laan, M., Hsu, J.-P., Peace, K. E., & 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), 38-39.

Van Oort, N., Brands, T., & de Romph, E. (2015). Short term ridership prediction in public transport by processing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*(2015).

Vogel, M., Hamon, R., Lozenguez, G., Merchez, L., Abry, P., Barnier, J., . . . Robardet, C. (2014). From bicycle sharing system movements to users: a typology of Vélo'v cyclists in Lyon based on large-scale behavioural dataset. *Journal of Transport Geography, 41*, 280-291. doi:10.1016/j.jtrangeo.2014.07.005

Vuchic, V. R. (2007). History and role of public transportation in urban development. Dans *Urban Transit: Systems and Technology*. Hoboken, N.J: John Wiley & Sons.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236-244. doi:10.1080/01621459.1963.10500845

Westphal, C., & Blaxton, T. (1998). *Data Mining Solutions*: John Wiley & Sons, Inc., New York.

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

Williams, M. J., Whitaker, R. M., & 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), 1017-1038. doi:10.1068/a301017

Xianyu, J., Rasouli, S., & 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), 533-553. doi:10.1007/s11116-015-9666-2

Zhang, F., Yuan, N. J., Wang, Y., & Xie, X. (2015). Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach. *Knowledge and Information Systems, 44*(2), 299-323. doi:10.1007/s10115-014-0763-x

Zhao, J., Tian, C., Zhang, F., Xu, C., & Feng, S. (2014). *Understanding temporal and spatial travel patterns of individual passengers by mining smart card data*. Communication présentée à 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China.

Zhong, C., Huang, X., Müller Arisona, S., Schmitt, G., & Batty, M. (2014). Inferring building functions from a probabilistic model using public transportation data. *Computers, Environment and Urban Systems, 48*, 124-137. doi:10.1016/j.compenvurbsys.2014.07.004

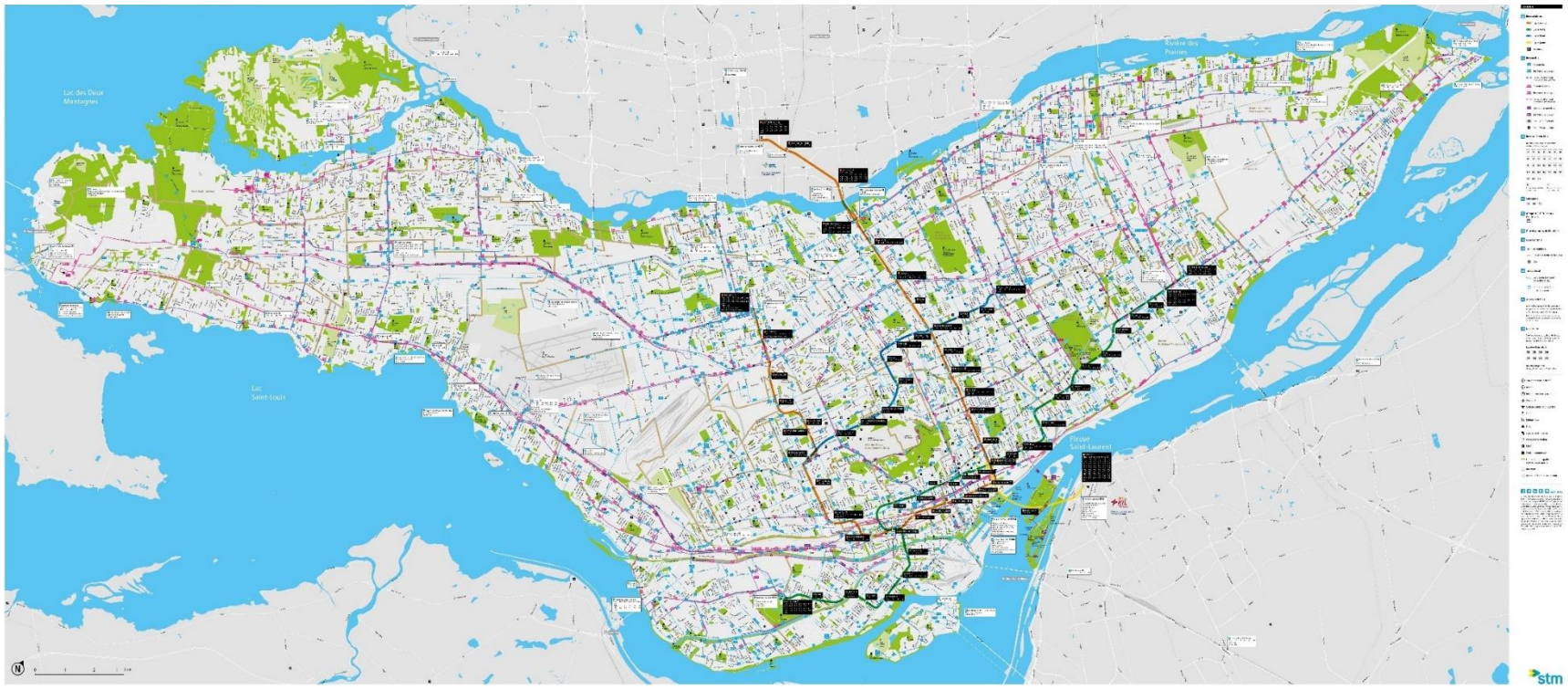
Zhong, C., Manley, E., Arisona, S. M., Batty, M., & Schmitt, G. (2015). Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science, 9*, 125 - 130. doi:10.1016/j.jocs.2015.04.021

Zhu, L., Gonder, J., & Lin, L. (2017). Prediction of Individual Social-Demographic Role Based on Travel Behavior Variability Using Long-Term GPS Data. *Journal of Advanced Transportation, 2017*, 1-13. doi:10.1155/2017/7290248

ANNEXES

ANNEXE A - PLAN DES RÉSEAUX DE MÉTRO ET DE BUS DE LA STM

(Plan tiré de http://www.stm.info/sites/default/files/pdf/fr/plan_reseau.pdf)



ANNEXE B - GRILLE TARIFAIRE EN VIGUEUR AU 1er JUILLET 2016

TARIFS EN VIGUEUR 1er juillet 2016

BUS ET MÉTRO

TITRES	TARIF ORDINAIRE	TARIF RÉDUIT (1)		VALIDE				VALIDE		EN VENTE
		6-17 ANS 65 ANS ET +	ÉTUDIANTS 18-25 ANS	À BORD DU BUS	OPUS (6)			747	VIA LAVAL • LONGUEUIL	DÉTAILLANTS AUTORISÉS
1 passage	3,25 $	2,25 $ (3)	------	● (3)	●	● (5)	● (5)		●	●
2 passages • *Un seul détenteur*	6,00 $	4,00 $ (3)	------		●	● (8)	● (8)		●	●
10 passages	27,00 $	16,50 $	------		●					●
Soirée illimitée • *18 h à 5 h* *Première utilisation avant minuit*	5,00 $	------	------		●	●			●	●
1 jour • *24 h*	10,00 $	------	------		●	●		●	●	●
Week-end illimité *Du vendredi 16 h au lundi 5 h*	13,75 $	------	------		●	●		●	●	●
3 jours • *Consécutifs* *Jusqu'à 23 h 59 le 3e jour*	18,00 $	------	------		●	●		●	●	●
Hebdo *Du lundi au dimanche 23 h 59*	25,75 $	15,75 $	------		●			●		●
Mensuel *Du 1er à la fin du mois*	83,00 $	49,75 $	49,75 $		●			●		●
4 mois • *Consécutifs*	------	197,00 $	197,00 $		●			●		●
Groupe (4) *1 adulte + 10 enfants de 6-13 ans*	------	16,50 $	------				●		●	
Aéroport Mtl-Trudeau (747) (10)	10,00 $	------	------	●				●	●	
Passage 747 ouest (9)	10,00 $	------	------	●				●		

Gratuit pour les enfants de 5 ans et moins en tout temps et pour les enfants de 6 à 11 ans à certaines périodes dans le cadre du programme Sorties en famille. (2)

TRANSPORT ADAPTÉ ET SERVICES PAR TAXI

TITRES	TARIF ORDINAIRE	TARIF RÉDUIT (1)		À BORD	OPUS		EN VENTE
		6-17 ANS 65 ANS ET +	ÉTUDIANTS 18-25 ANS				DÉTAILLANTS AUTORISÉS
Montant exact dans le véhicule (11)	2,70 $	1,65 $	------	●			
1 passage	3,25 $	2,25 $	------			●	
Hebdo *Du lundi au dimanche 23 h 59*	25,75 $	15,75 $	------		●		●
Mensuel *Du 1er à la fin du mois*	83,00 $	49,75 $	49,75 $		●		●
4 mois • *Consécutifs*	------	197,00 $	197,00 $		●		●

TRANSPORT ADAPTÉ — TARIFS DES DÉPLACEMENTS MÉTROPOLITAINS	
Vers les territoires du RTL et de la STL	Vers les autres destinations à l'intérieur du territoire délimité par l'AMT
Double tarification STM	Triple tarification STM

16020_01_Grille2016 - Web Fr

IMPORTANT	**Les titres de transport ne sont pas vendus dans les véhicules. Monnaie exacte requise pour les paiements au comptant à bord des bus. Visitez le centre de service à la clientèle ou un point de service pour remplacer une carte défectueuse ou rectifier une erreur d'achat (7).**

1- Le client sans carte OPUS avec photo, lorsque celle-ci est exigée, s'expose à une amende plus les frais administratifs.

2- Les enfants doivent être accompagnés en tout temps pour bénéficier d'une gratuité de transport. Le programme Sorties en famille permet à un maximum de 5 enfants de 6 à 11 ans de voyager gratuitement lorsqu'accompagnés d'un adulte détenant un titre de transport valide. Programme en vigueur la fin de semaine (du vendredi 16 h au dimanche à la fin du service), les jours fériés ainsi que du 21 juin au 28 août 2016, du 23 décembre 2016 au 8 janvier 2017, du 3 au 12 mars 2017, ainsi que du 22 juin au 27 août 2017. Programme non valide à bord des bus de la ligne 747 Aéroport Montréal-Trudeau.

3 - Titres 1 passage et 2 passages à tarif réduit :
- 6-11 ans : possibilité de payer au comptant dans le bus ou le métro sans carte OPUS avec photo. Carte photo requise pour l'achat chez un détaillant autorisé.
- 12-17 ans : tarif réduit sur carte OPUS avec photo seulement.
- 65 ans et plus : possibilité de payer au comptant dans le bus ou le métro sur présentation de la carte OPUS avec photo.

4- Ce titre donne droit à un passage simultané pour le groupe.

5- Le titre 1 passage au tarif ordinaire sur carte *L'occasionnelle* est disponible seulement chez les détaillants autorisés. Auprès du changeur et à la distributrice automatique du métro, ce titre est disponible sur carte magnétique.

6- Des frais d'émission de 6 $ sont exigés pour une carte OPUS avec photo pour les 6-11 ans et 15 $ pour les 12-25 ans et les 65 ans et +. La carte OPUS sans photo coûte 6 $ et un titre doit obligatoirement y être chargé. Détails et conditions à stm.info.

7- Vous devez acheter votre titre pour valider votre passage afin de vous rendre dans un point de service.

8- Le titre 2 passages au tarif ordinaire est disponible sur *L'occasionnelle* et au tarif réduit sur carte magnétique.

9- Paiement aux bornes Stationnement de Montréal désignées. Valide en direction de l'aéroport Montréal-Trudeau seulement.

10- En payant en argent comptant à bord de la 747, vous recevez un titre valide pour 24 heures.

11- Le comptant à bord est accepté pour les déplacements au transport adapté et les dessertes de navettes Or par taxi. Il est toutefois impossible de payer comptant à bord des dessertes par taxi collectif.

De plus, l'application des tarifs et l'utilisation des titres de transport y correspondant doivent s'effectuer conformément aux dispositions du Règlement R-105 et ses amendements concernant les conditions au regard de la possession et de l'utilisation de tout titre de transport émis par la STM et ses modifications.

stm

ANNEXE C - AVANTAGES ET INCONVÉNIENTS DE LA CARTE À PUCE SELON TROIS POINTS DE VUE

Cette annexe fournit un tableau récapitulatif qui résume la section 2.1.3 de la revue de littérature (Chapitre 2).

VISION	Avantages	N° Réf	Inconvénients	N° Réf
Usagers	- Commodité et facilité d'utilisation	3, 15	- L'usager doit prendre soin de sa carte et faire attention à la puce	-
	- Rapidité d'utilisation (réduction du temps d'embarquement)	3, 5, 15	- L'usager doit penser à valider son titre même s'il a un accès illimité, sous peine d'une contravention	2
	- Satisfaction et perception améliorées grâce à un meilleur service et à une image plus moderne du TC	8, 10	- Problèmes liés à la vie privée (confidentialité)	2, 11
	- Mise en place possible de services personnalisés et de tarifs réduits	9		
	- Aucun effort additionnel autre que la validation de la carte pour rapporter ses déplacements (moins pénible que de répondre à une enquête)	12		
Opérateurs	- Validation plus fiable : moins de fraude, meilleur contrôle	15	- Investissement important	6, 11, 13, 15
	- Politique tarifaire flexible et variée	1, 6, 10, 11, 15	- Complexité technique, notamment lors l'installation du système : personnel qualifié et nouveaux équipements requis	11, 15
	- Intégration tarifaire entre plusieurs réseaux facilitée ➔ meilleures gestion et allocation des revenus, partage d'informations	1, 4, 6, 13, 15	- Système figé : l'introduction de nouveaux composants ou processus dans le système est difficile	6
	- Réduction des coûts d'opération à long terme (réduction du personnel dans les stations de métro & réduction du nombre de véhicules/chauffeurs de bus)	10, 11, 15	- Acceptation sociale lente	6, 11
	- Augmentation de l'achalandage d'après le principe d'élasticité de la demande ➔ augmentation des revenus	10, 15	- Aucune garantie de profitabilité	6, 10

Planificateurs et chercheurs	- Grandes quantités de données longitudinales, disponibles sur de longues périodes	2, 7, 12, 13, 14	- Données partielles (informations non collectées : motif de déplacement, perception et satisfaction des usagers, origine et destination finales, temps d'accès, débarquement, données socio-démographiques)	2, 3, 7, 12, 13, 15
	- Données de grande précision dans le temps et dans l'espace	7, 13, 15	- Données non universelles : l'analyse ne tient pas compte des non titulaires de cartes et des autres modes de transport	13
	- Données individualisées, qui permettent études au niveau désagrégé	2, 7, 13	- Données pas toujours formatées pour la planification, nécessitant des traitements particuliers et des modèles plus puissants	7, 13
	- Données passives : réduction du rôle de l'usager dans le processus de collecte ➔ moins de biais introduits	2, 13	- Qualité des données pas toujours parfaite (défaillance des équipements, erreurs humaines ou non validation des usagers)	13, 15
	- Données peu coûteuses à collecter une fois que les systèmes de paiement sont installés	12, 13	- Une carte ne correspond pas toujours à un seul usager	13

N° Réf	Référence	N° Réf	Référence
1	(Attoh-Okine & Shen, 1995)	9	(Kieu et al., 2014)
2	(Bagchi & White, 2004)	10	(McDonald, 2000)
3	(Bagchi & White, 2005)	11	(Pelletier et al., 2011)
4	(Blythe, 2004)	12	(Spurr et al., 2015)
5	(Chira-Chavala & Coifman, 1996)	13	(Trépanier, 2012)
6	(Deakin & Kim, 2001)	14	(Utsunomiya et al., 2006)
7	(El Mahrsi et al., 2017)	15	(White et al., 2010)
8	(Ibrahim, 2003)		

ANNEXE D - CALENDRIER 2016 ET JOURS FÉRIÉS

[ ] Semaines considérées

Janvier 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
53					1	2	3
1	4	5	6	7	8	9	10
2	11	12	13	14	15	16	17
3	18	19	20	21	22	23	24
4	25	26	27	28	29	30	31

Février 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
5	1	2	3	4	5	6	7
6	8	9	10	11	12	13	14
7	15	16	17	18	19	20	21
8	22	23	24	25	26	27	28
9	29						

Mars 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
9		1	2	3	4	5	6
10	7	8	9	10	11	12	13
11	14	15	16	17	18	19	20
12	21	22	23	24	25	26	27
13	28	29	30	31			

Avril 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
13					1	2	3
14	4	5	6	7	8	9	10
15	11	12	13	14	15	16	17
16	18	19	20	21	22	23	24
17	25	26	27	28	29	30	

Mai 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
17							1
18	2	3	4	5	6	7	8
19	9	10	11	12	13	14	15
20	16	17	18	19	20	21	22
21	23	24	25	26	27	28	29
22	30	31					

Juin 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
22			1	2	3	4	5
23	6	7	8	9	10	11	12
24	13	14	15	16	17	18	19
25	20	21	22	23	24	25	26
26	27	28	29	30			

Juillet 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
26					1	2	3
27	4	5	6	7	8	9	10
28	11	12	13	14	15	16	17
29	18	19	20	21	22	23	24
30	25	26	27	28	29	30	31

Août 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
31	1	2	3	4	5	6	7
32	8	9	10	11	12	13	14
33	15	16	17	18	19	20	21
34	22	23	24	25	26	27	28
35	29	30	31				

Septembre 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
35				1	2	3	4
36	5	6	7	8	9	10	11
37	12	13	14	15	16	17	18
38	19	20	21	22	23	24	25
39	26	27	28	29	30		

Octobre 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
39						1	2
40	3	4	5	6	7	8	9
41	10	11	12	13	14	15	16
42	17	18	19	20	21	22	23
43	24	25	26	27	28	29	30
44	31						

Novembre 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
44		1	2	3	4	5	6
45	7	8	9	10	11	12	13
46	14	15	16	17	18	19	20
47	21	22	23	24	25	26	27
48	28	29	30				

Décembre 2016

	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Dim.
48				1	2	3	4
49	5	6	7	8	9	10	11
50	12	13	14	15	16	17	18
51	19	20	21	22	23	24	25
52	26	27	28	29	30	31	

Jours fériés le vendredi Jours fériés le lundi

2016

Calendrier-365.fr

Jours fériés du Québec

1er janvier : *Jour de l'An*	23 mai : *Journée nationale des patriotes*	5 septembre : *Fête du travail*
25 mars : *Vendredi Saint*	24 juin : *Fête nationale du Québec*	10 octobre : *Action de Grâce*
28 mars : *Lundi de Pâques*	1er juillet : *Fête du Canada*	25 décembre : *Noël*

ANNEXE E - DISTRIBUTIONS DES CARTES POUR TOUTES LES COMBINAISONS DE NOMBRE ET DE TYPE DE PRODUITS ET DE TARIFS UTILISÉS EN 2016

% cartes		TARIFS																				
	Nombre	1					2									3					4	
PRODUITS	Nombre	Type	A	B	C	D	F	A-B	A-C	A-D	B-C	B-D	B-F	C-D	C-F	D-F	A-C-D	B-C-D	B-C-F	B-D-F	C-D-F	B-C-D-F
	1	a	0,003%	0,000%	0,347%	0,628%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,003%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b	1,889%	0,000%	9,673%	17,591%	0,001%	0,000%	0,000%	0,242%	0,000%	0,000%	0,000%	0,076%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		c	0,000%	0,000%	0,094%	4,246%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,009%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		d	0,872%	0,131%	0,973%	5,446%	0,004%	0,000%	0,000%	0,000%	0,006%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		e	0,009%	1,668%	9,097%	10,509%	0,000%	0,000%	0,000%	0,090%	0,242%	0,008%	0,000%	0,015%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		f	0,007%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
	2	a-b	0,000%	0,000%	1,490%	2,187%	0,000%	0,000%	0,000%	0,040%	0,000%	0,000%	0,000%	0,047%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-c	0,000%	0,000%	0,015%	0,439%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,017%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-d	0,000%	0,000%	0,029%	0,186%	0,000%	0,000%	0,000%	0,001%	0,000%	0,003%	0,000%	0,073%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-e	0,000%	0,000%	0,486%	0,969%	0,000%	0,000%	0,000%	0,006%	0,004%	0,113%	0,000%	0,579%	0,000%	0,000%	0,000%	0,055%	0,000%	0,000%	0,000%	0,000%
		b-c	0,000%	0,000%	0,135%	2,992%	0,001%	0,000%	0,000%	0,069%	0,000%	0,000%	0,000%	0,122%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b-d	0,000%	0,000%	0,086%	0,201%	0,000%	0,000%	0,000%	0,002%	0,000%	0,006%	0,000%	0,135%	0,000%	0,002%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b-e	0,000%	0,000%	2,227%	5,251%	0,000%	0,000%	0,000%	0,066%	0,021%	0,586%	0,000%	1,397%	0,000%	0,000%	0,000%	0,117%	0,000%	0,000%	0,000%	0,000%
		b-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		c-d	0,000%	0,000%	0,007%	0,024%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,055%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		c-e	0,000%	0,000%	0,251%	2,364%	0,000%	0,000%	0,000%	0,009%	0,000%	0,056%	0,000%	0,377%	0,000%	0,000%	0,000%	0,011%	0,000%	0,000%	0,000%	0,000%
		c-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		d-e	0,000%	0,016%	0,447%	0,095%	0,000%	0,000%	0,000%	0,001%	0,011%	0,000%	0,000%	0,001%	0,000%	0,006%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		d-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
	3	a-b-c	0,000%	0,000%	0,068%	1,413%	0,000%	0,000%	0,000%	0,016%	0,000%	0,000%	0,000%	0,122%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-b-d	0,000%	0,000%	0,021%	0,041%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,038%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-b-e	0,000%	0,000%	0,759%	1,274%	0,000%	0,000%	0,000%	0,009%	0,003%	0,229%	0,000%	0,476%	0,000%	0,000%	0,000%	0,087%	0,000%	0,000%	0,000%	0,000%
		a-c-d	0,000%	0,000%	0,001%	0,003%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,014%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-c-e	0,000%	0,000%	0,050%	0,457%	0,000%	0,000%	0,000%	0,002%	0,000%	0,024%	0,000%	0,112%	0,000%	0,000%	0,000%	0,010%	0,000%	0,000%	0,000%	0,000%
		a-d-e	0,000%	0,000%	0,020%	0,007%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,050%	0,000%	0,000%	0,000%	0,003%	0,000%	0,000%	0,000%	0,000%
		a-d-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b-c-d	0,000%	0,000%	0,007%	0,029%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,035%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b-c-e	0,000%	0,000%	0,347%	2,840%	0,000%	0,000%	0,000%	0,016%	0,000%	0,133%	0,000%	0,474%	0,000%	0,000%	0,000%	0,029%	0,000%	0,000%	0,000%	0,000%
		b-c-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b-d-e	0,000%	0,000%	0,138%	0,047%	0,000%	0,000%	0,000%	0,000%	0,000%	0,003%	0,000%	0,123%	0,000%	0,002%	0,000%	0,007%	0,000%	0,000%	0,000%	0,000%
		b-d-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		b-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		c-d-e	0,000%	0,000%	0,006%	0,013%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,040%	0,000%	0,001%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%
		c-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		d-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
	4	a-b-c-d	0,000%	0,000%	0,003%	0,010%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,024%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-b-c-e	0,000%	0,000%	0,199%	1,522%	0,000%	0,000%	0,000%	0,007%	0,000%	0,135%	0,000%	0,415%	0,000%	0,000%	0,000%	0,045%	0,000%	0,000%	0,000%	0,000%
		a-b-c-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-b-d-e	0,000%	0,000%	0,042%	0,010%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,044%	0,000%	0,000%	0,000%	0,005%	0,000%	0,000%	0,000%	0,000%
		a-b-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-c-d-e	0,000%	0,000%	0,001%	0,002%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,012%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%
		b-c-d-e	0,000%	0,000%	0,020%	0,017%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,046%	0,000%	0,001%	0,000%	0,002%	0,000%	0,000%	0,000%	0,000%
		b-c-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,001%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
	5	a-b-c-d-e	0,000%	0,000%	0,009%	0,009%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,035%	0,000%	0,000%	0,000%	0,002%	0,000%	0,000%	0,000%	0,000%
		a-b-c-d-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%
		a-b-c-e-f	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%	0,000%

Dictionnaire des produits

a	Billets unitaires
b	Carnet
c	Abon. jours/hebdo
d	Abon. longs
e	Abon. mensuels
f	Titres spéciaux

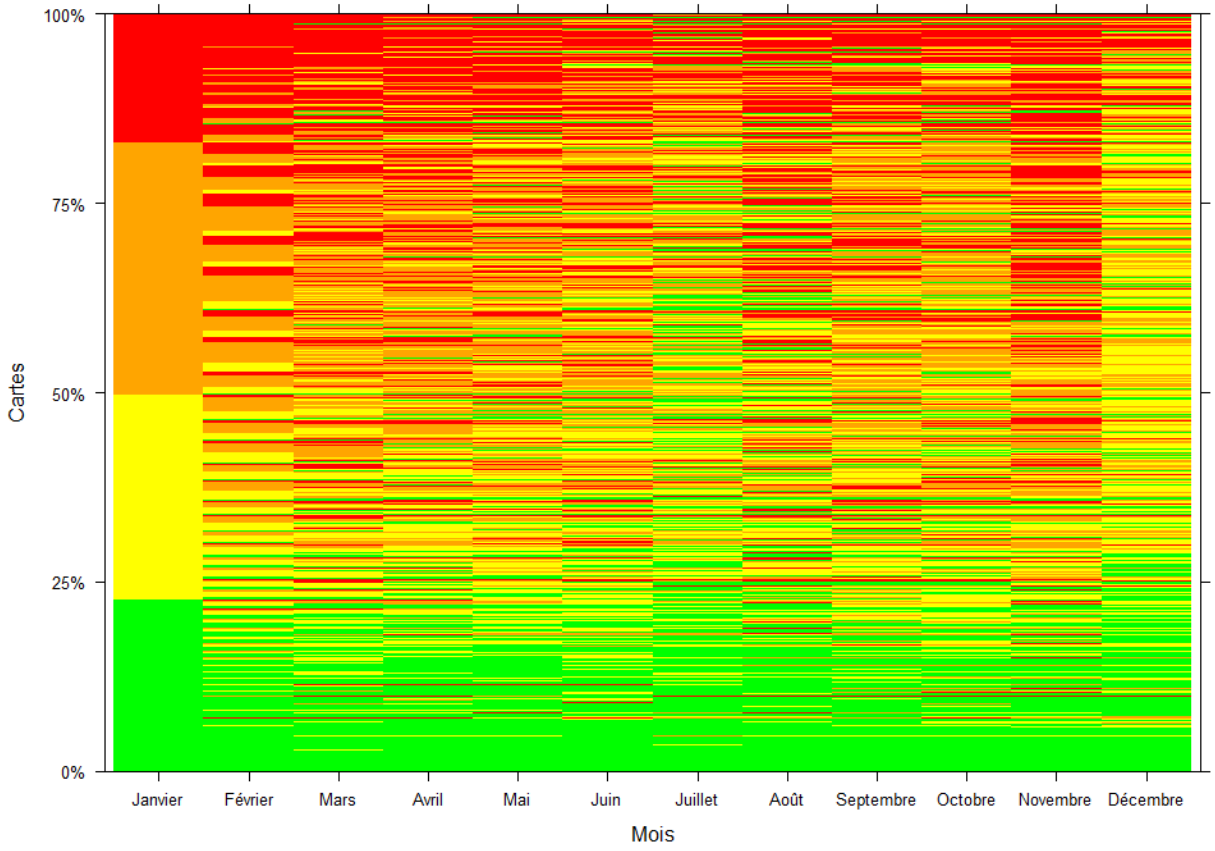
Dictionnaire des tarifs

A	Gratuité
B	Tarif étudiant
C	Tarif réduit
D	Tarif ordinaire
E	Tarif spécial

ANNEXE F - RÉSULTATS DES TESTS STATISTIQUES APPLIQUÉS À CHAQUE INDICATEUR DE VARIABILITÉ POUR LES 10 COMBINAISONS DE CARTES

Test	Dispersion des déplacements	Fréquence d'utilisation	Variabilité temporelle	Variabilité spatiale – métro	Variabilité spatiale – bus
CO1 v CO2	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO1 v CO3	1.44E-06	0.00E+00	1.57E-02	0.00E+00	1.51E-204
CO1 v CO4	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO1 v CO5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO1 v CO6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.39E-265
CO1 v CO7	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO1 v CO8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO1 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO1 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO2 v CO3	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO2 v CO4	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.64E-46
CO2 v CO5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO2 v CO6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO2 v CO7	0.00E+00	0.00E+00	0.00E+00	3.91E-138	0.00E+00
CO2 v CO8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO2 v CO9	0.00E+00	0.00E+00	8.64E-43	0.00E+00	0.00E+00
CO2 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	5.87E-01
CO3 v CO4	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO3 v CO5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO3 v CO6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.46E-07
CO3 v CO7	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO3 v CO8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO3 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO3 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO4 v CO5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO4 v CO6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO4 v CO7	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO4 v CO8	0.00E+00	0.00E+00	0.00E+00	5.90E-11	0.00E+00
CO4 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO4 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO5 v CO6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO5 v CO7	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.11E-298
CO5 v CO8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO5 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO5 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO6 v CO7	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO6 v CO8	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO6 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO6 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO7 v CO8	0.00E+00	0.00E+00	0.00E+00	1.14E-45	9.28E-01
CO7 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO7 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	1.30E-27
CO8 v CO9	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
CO8 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	3.98E-41
CO9 v CO10	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00

ANNEXE G - VARIABILITÉ DU NOMBRE DE DÉPLACEMENTS PAR MOIS PAR CARTE EN JOURS OUVRABLES (CAS DES UTILISATEURS D'ABONNEMENTS ANNUELS AVEC UNE AMPLITUDE DE 12 MOIS)



Code couleur, avec N le nombre de déplacements par mois

Rouge : $3^{ème}$ quartile $\leq N \leq$ max

Orange : $2^{ème}$ quartile (médiane) $\leq N \leq 3^{ème}$ quartile

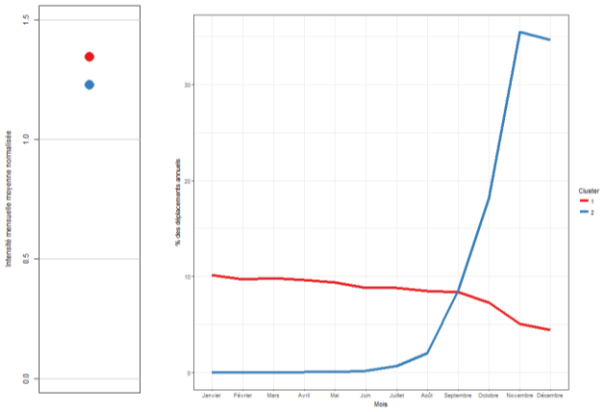
Jaune : 1^{er} quartile $\leq N \leq 2^{ème}$ quartile (médiane)

Vert : min $\leq N \leq 1^{er}$ quartile

Les cartes, représentées horizontalement sur l'axe des y, sont triées en ordre décroissant selon le nombre de déplacements réalisés en janvier, puis février, puis mars, etc. jusqu'à décembre. Une couleur est attribuée à chaque nombre de déplacements par mois par carte en fonction de sa position par rapport aux trois quartiles (quartiles de la distribution de tous les nombres de déplacements par mois, tous mois confondus). La valeur de ces quartiles n'est pas précisée ici par souci de confidentialité.

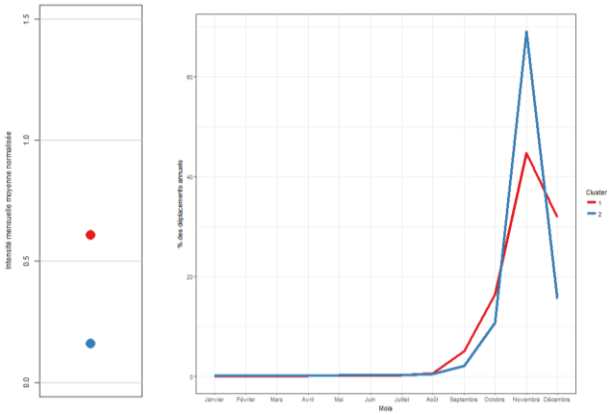
ANNEXE H - REDÉCOMPOSITION DES GROS GROUPES DE LA TYPOLOGIE OBTENUE (CAS DE TOUS LES USAGERS)

GROUPE C1



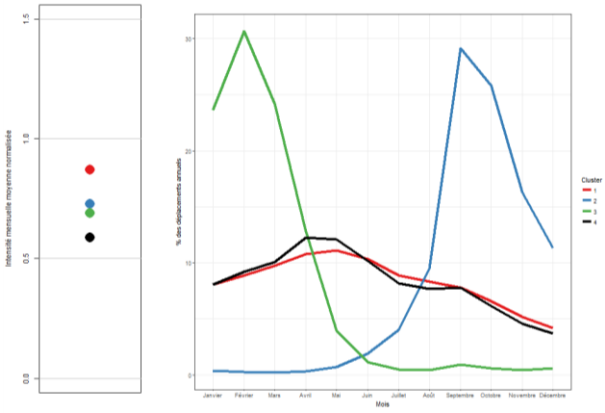
Groupe	Proportion C1	Proportion TOT
C1-1	68.2%	8.2%
C1-2	31.8%	3.8%
TOTAL	100.0%	12.0%

GROUPE C3



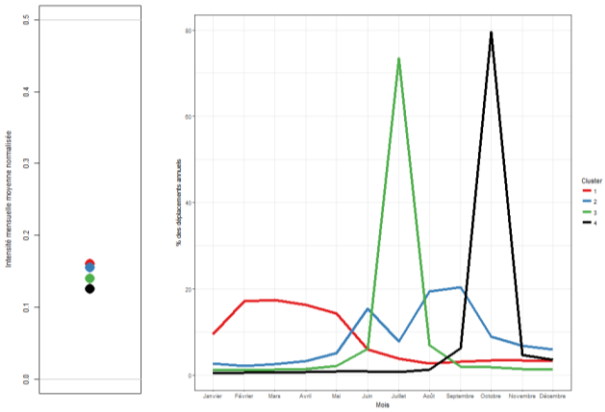
Groupe	Proportion C3	Proportion TOT
C3-1	57.9%	7.2%
C3-2	42.1%	5.2%
TOTAL	100.0%	12.4%

GROUPE C2



Groupe	Proportion C2	Proportion TOT
C2-1	36.1%	9.6%
C2-2	14.1%	3.7%
C2-3	16.6%	4.4%
C2-4	33.3%	8.8%
TOTAL	100.0%	26.6%

GROUPE C6



Groupe	Proportion C6	Proportion TOT
C6-1	47.9%	18.5%
C6-2	39.3%	15.2%
C6-3	6.4%	2.5%
C6-4	6.4%	2.5%
TOTAL	100.0%	38.6%