Titre: Title:	Méthodologie pour l'étude de l'évolution des comportements des voyageurs de transport collectif urbain
Auteur: Author:	Alexis Viallard
Date:	2018
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Viallard, A. (2018). Méthodologie pour l'étude de l'évolution des comportements des voyageurs de transport collectif urbain [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/3721/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/3721/
Directeurs de recherche: Advisors:	Martin Trépanier, & Catherine Morency
Programme: Program:	Maîtrise recherche en génie industriel

UNIVERSITÉ DE MONTRÉAL

MÉTHODOLOGIE POUR L'ÉTUDE DE L'ÉVOLUTION DES COMPORTEMENTS DES VOYAGEURS DE TRANSPORT COLLECTIF URBAIN

ALEXIS VIALLARD

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION

DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(GÉNIE INDUSTRIEL)

DÉCEMBRE 2018

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

MÉTHODOLOGIE POUR L'ÉTUDE DE L'ÉVOLUTION DES COMPORTEMENTS DES VOYAGEURS DE TRANSPORT COLLECTIF URBAIN

présenté par : VIALLARD Alexis

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, Doctorat, président

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

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

Mme CHANDESRIS Maguelonne, Doctorat, membre

DÉDICACE

À ma famille,

À mes amis,

À ma muse…

REMERCIEMENTS

J’aimerais, dans un premier temps, remercier Martin Trépanier, directeur de recherche appartenant au département de mathématiques et de génie industriel, ainsi que Catherine Morency, codirectrice de recherche du département des génies civil, géologique et des mines, de m’avoir assisté, conseillé et subventionné le long de cette maîtrise recherche à l’École Polytechnique de Montréal. Leur maîtrise du sujet m’a notamment permis de découvrir et apprécier le monde du transport tout en approfondissant mes connaissances en analyse de données.

Je tiens également à remercier la Société de Transport de l’Outaouais (STO) pour l’accès à leurs données, sans qui ce projet de recherche n’aurait pas vu le jour. Je tiens particulièrement à ce que ces travaux puissent servir aux futures études dans leur amélioration visant une mise en place en milieu industriel.

Mes remerciements vont aussi à la Société Nationale des Chemins de fers Français (SNCF) pour l’accès à leurs données, permettant ainsi de tester ce projet à une échelle plus importante, mais également pour le financement réalisé par le biais de l’organisme MITACS.

Je suis également reconnaissant envers l’École Nationale Supérieure d’Arts et Métiers pour m’avoir donné l’opportunité enrichir mon expérience universitaire et humaine à travers ce double diplôme à l’École Polytechnique de Montréal.

Enfin merci à mes parents, ma famille et mes amis pour leur support continu au fil des années.

RÉSUMÉ

Démocratisé depuis déjà plusieurs années les systèmes tarifaires automatisés, relatifs à l'accès aux transports en commun, génèrent des masses de données encore trop peu exploitées. Ces données issues de cartes à puce sont devenues si volumineuses que leur analyse représente un véritable défi pour l'homme, mais également un immense potentiel pour la planification en transport en commun.

Ce mémoire s'inscrit dans le cadre de la valorisation de volumétries importantes de données quotidiennes. Ouvrant un projet commun avec des exploitants de transports, il s'agit de s'intéresser à l'analyse de la demande. L'ensemble des méthodes seront développées à partir de trois ans de données de transaction issues de l'utilisation du transport par bus à Gatineau.

L'objectif principal de la recherche est de présenter une méthodologie simple et complète, relative à l'étude longitudinale des comportements d'usage des cartes à puces sur long terme en utilisant différentes techniques d'exploration de données. À terme, cette méthode d'analyse fournit des résultats aidant le travail d'un planificateur de réseau.

Les sous-objectifs de l'étude sont les suivants :

- Développer un algorithme permettant une analyse comportementale des usagers.
- Développer un algorithme expérimental améliorant la méthode précédente, afin que l'analyste puisse suivre l'évolution des comportements des usagers à travers le temps.
- Proposer une méthode de prévision des évolutions, enrichissant ainsi les connaissances apportées à la planification.

Ce mémoire débute par une revue de littérature présentant l'intérêt de l'utilisation des cartes à puces en analyse. Il s'agit de s'intéresser aux diverses études réalisées, notamment dans le cadre d'analyses comportementales. Une partie de la littérature s'intéresse aux techniques d'exploration de données, particulièrement dans le cas de segmentations et de prévisions. La section méthodologie présente les raisonnements répondant aux trois sous-objectifs, et la dernière partie les résultats des diverses expérimentations effectuées sur les données fournies par la STO.

Les contributions apportées par ce mémoire sont :

- La présentation d'une méthode classique d'analyse comportementale des cartes à puce à partir de leurs utilisations. Un travail de segmentation est effectué sur l'ensemble des déplacements hebdomadaires en transports en commun afin de repérer les similarités entre comportements.
- La conception et la critique d'une méthode expérimentale basée sur une segmentation hebdomadaire visant à montrer l'évolution des comportements des cartes à travers le temps.
- Différents indicateurs de qualité et de stabilité de segmentation sont proposés afin de comparer les diverses méthodes engagées, et de caractériser la population de cartes étudiée.
- Jouant sur une possible évolution comportementale des cartes, une critique sur la fiabilité de l'utilisation de méthodes de prévision est réalisée. Les prévisions sont appliquées sur l'évolution comportementale des groupes ainsi que l'évolution de la taille de leur population.

En conclusion, ce projet présente une méthode classique, fonctionnelle et applicable en industrie permettant l'analyse comportementale des usagers. Prenant comme entrée un jeu de données de cartes à puce, la méthode exporte les résultats de segmentation liés à l'utilisation des transports en commun. Par cela, elle définit 6 groupes d'identifiants aux comportements similaires dont les caractéristiques propres permettent l'aide à la décision en planification des transports. Il s'agit de trois groupes dont les déplacements récurrents en semaine ressemblent à ceux de travailleurs à temps plein et à mi-temps. Deux groupes représentent les comportements de déplacements occasionnels et le dernier contient l'ensemble des cartes qui produisent le plus de déplacement. Le tout est réalisé en un temps relativement long : 11 minutes pour la segmentation de 10 millions de déplacements. La méthode expérimentale, quant à elle, se concentre sur l'évolution comportementale possible des cartes. Sans pour autant être parfaite, elle admet un potentiel énorme. En effet, elle permet une analyse des comportements des 6 groupes de cartes en un temps de calcul très court (38 secondes pour une qualité similaire). Il s'agit de groupes dont les caractéristiques sont très proches de ceux issus de la méthode classique, mais le principe incrémental de la segmentation rend possible l'étude de l'évolution comportementale, jugée fixe dans la méthode classique.

ABSTRACT

For several years automated fare systems related to public transport access are generating an, not enough, exploited massive volume of data. These smart card data became so voluminous they represent a challenge for humans and a huge potential to public transport planning too.

This work aims to value massive volumes of daily data. Opening a common project with transit services, the analysis is based on studying demand. All methods were developed thanks to the three years of transactions from the usage of Gatineau's bus network.

Presenting a simple and a complete methodology to apply a longitudinal analysis on smart card usage behavior using different data mining techniques, represent the main purpose of the research. At the end, the analysis methodology gives the results helping to do the transit planners job.

The sub-objectives are the following ones:

- Develop an algorithm allowing users' behavior analysis
- Develop an improved algorithm (experimental), allowing to follow the users' behavior evolution through time.
- Propose an evolution prevision methodology, enhancing transit planning knowledge

This works starts with a literature review presenting smart card data usage in analysis, particularly through different studies done in behavior analysis. The second part of the literature review is about data mining techniques like clustering and forecasting. The methodology section describes the three sub-objectives, and the final section presents the different applications on STO's data.

The main achievements of this project are:

- The presentation of a classical methodology allowing to analyze smart cards' behavior through their activities. A clustering technique is applied on all weekly usage of public transit to find similarity between behaviors.
- The conception and critic of an experimental method based on week-to-week clustering aiming to show the users' behavior evolution through time.
- Different quality and stability indicators are proposed to compare the methods applied and to characterize the population.

- Knowing that users' behavior can evolve, a critic on prevision technique viability is applied. Forecasts methods are used on clusters' behavior evolution and their population size evolution.

Finally, this project presents an industrially applicable methodology on transit users' behavior. Taking smart card data as input, the algorithm exports the transit usage clustering results. This way it defines 6 groups of IDs with similar behavior which the proper characteristics help the transit planner to take decisions. There are three groups which the trips patterns look like full time and part-time workers trip patterns. Two of the groups represent occasional trip behavior and the last one holds the cards with the most trips. The computation time is relatively high: 11 minutes for the clustering of 10 million transactions. The experimental method focuses on the possible smart cards' behavior evolution. Without being perfect, it shows a huge potential. Indeed, the method allows a behavior analysis of 6 groups with a shorter computation time (only 38 seconds for a similar quality). These groups present the same characteristics as those from the traditional method, but the way the algorithm works makes the behavior evolution analysis possible in this case.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

Cette liste présente dans l'ordre alphabétique, les sigles et abréviations utilisés dans le mémoire ainsi que leur signification.

ARMA	Autoregressive-Moving-Average Model – Modèle autorégressif et moyenne mobile
CSV	Comma-Separated Values
DBSCAN	Density-Based Spatial Clustering of Application with Noise – Classification spatiale basée sur la densité, avec bruit
EUC	Indicateur de qualité intragroupe euclidien
GTFS	General Transit Feed Specification
HAC	Hierarchical Agglomerative Clustering – Classification ascendante hiérarchique
K	Nombre de groupes
MAPE	Mean Absolute Percentage Error – Pourcentage moyen d'erreur absolue
OD	Origine-Destination
PCA	Principal Component Analysis – Analyse en composantes principales
REL	Indicateur de qualité intragroupe relatif
RFID	Radio-Frequency Identification – Identification par radio-fréquences
RMSE	Root-Mean-Square Error – Erreur moyenne quadratique
SNCF	Société Nationale des Chemins de fers Français
STM	Société de transport de Montréal
STO	Société de transport de l'Outaouais
TAZ	Traffic Analysis Zone – Zone d'analyse de trafic
WSI	Weighted Sequential Instability - Instabilité séquentielle pondérée

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

1.1 Mise en contexte

Aujourd'hui, plus que jamais, l'ère est aux données massives (*Big Data*). Au cours des deux dernières années, le volume de données produites a dépassé celui de toute l'histoire de l'humanité. Dans notre monde toujours plus numérique, l'humain est devenu un fournisseur ambulant de données. Tout est enregistré et conservé afin d'être étudié. On cherche ainsi à tout connaître, tout analyser, tout optimiser dans le but de réduire les coûts ou de produire le meilleur service possible. Les experts et grandes entreprises du monde entier s'accordent pour dire que les données sont en train de transformer l'industrie. Véritable mine d'or pour l'économie digitale, l'analyse de données massives représente l'une des grandes problématiques informatiques de la décennie. S'intéressant sur ce point, de nombreuses technologies ont émergé permettant l'accumulation de toujours plus de données et apportant chacune leur lot d'opportunités.

Dans le domaine du transport en commun, ces technologies sont déjà bien implantées dans les réseaux. En effet, les systèmes de cartes à puce, permettant l'accès aux véhicules, enregistrent depuis les années 1990 des données relatives à l'utilisation du réseau. Visant, dans un premier temps, à remplacer le système financier de l'époque (jetons ou ticket), cette technologie permet aujourd'hui l'accumulation continue de nombreuses informations spatio-temporelles sur l'ensemble des déplacements effectués par les usagers. En effet, les sociétés exploitantes ont rapidement pris conscience de l'énorme potentiel lié à l'analyse des transactions pour la prise de décisions, montrant ainsi un intérêt certain pour l'enrichissement des données récoltées.

Depuis plusieurs décennies, de nombreux auteurs se sont intéressés à l'analyse de l'utilisation du réseau. S'intéressant peu à peu aux données récoltées par cartes à puce, ils se retrouvent vite face à une impasse technologique liée à une problématique de données volumineuses. Mais possédant aujourd'hui des puissances de calcul informatiques bien supérieures, le potentiel d'exploitation de ces données est enfin révélé, permettant notamment l'application d'études à l'échelle individuelle. Plus récemment, les travaux s'intéressent principalement à l'enrichissement de ces données et à l'étude comportementale de l'usage des cartes.

Ce mémoire s'inscrit dans une problématique de valorisation de volumétries importantes de données quotidiennes, s'intéressant particulièrement à extraire les habitudes comportementales des usagers. L'étude ouvre un projet commun avec la Société Nationale des Chemins de fers Français (SNCF) et la société Keolis Canada, relatif à l'étude de l'offre et de la demande. Le présent mémoire est basé sur des données mises à disposition par la Société de transport de l'Outaouais (STO) et vise le développement d'algorithmes d'analyse de l'utilisation du réseau de Gatineau.

Enfin, les données de cartes à puce de la STO ont l'avantage de ne s'intéresser qu'à des déplacements de bus. Par cela, les transactions sont obligatoirement effectuées lors de correspondances, permettant ainsi de transformer les informations de transactions en des déplacements d'usagers. Les documents mis en place par la STO représentent l'ensemble des transactions effectuées entre janvier 2012 et décembre 2014, comprenant un total de plus de 35,4 millions de transactions.

1.2 Objectifs

L'objectif principal est de proposer une méthode universelle d'étude comportementale des usagers du transport en commun à partir de l'exploitation de données issues de cartes à puce. Il s'agit ainsi de programmer des algorithmes permettant l'analyse des données de transactions (« données brutes ») récoltées quotidiennement par la STO.

Pour y parvenir, il s'agit de s'intéresser aux sous-objectifs suivants :

- Mettre en place d'un algorithme fonctionnel, issu de la littérature, permettant l'analyse comportementale des usagers.
- Développer un algorithme expérimental fonctionnel améliorant la méthode précédente, afin de pouvoir suivre l'évolution des comportements des usagers à travers le temps.
- Proposer une méthode de prévision des évolutions, enrichissant ainsi les connaissances apportées à la planification.

1.3 Structure du mémoire

Le mémoire commence dans un premier temps par une revue de littérature présentant divers travaux effectués dans le domaine de l'exploitation des données de cartes à puce en transport en commun. En effet, il est proposé de montrer l'intérêt de l'analyse de ces données, notamment dans le domaine de l'étude comportementale des usagers à long terme. De plus, s'intégrant dans une optique de compréhension complète de la problématique, une seconde partie de la revue est consacrée à la présentation d'outils de fouille de données massives sur les thèmes de segmentation ainsi que de prévision.

La méthodologie présente ensuite les raisonnements employés dans le but de traiter les différentes problématiques de la maîtrise. Elle s'organise suivant les phases chronologiques nécessaires à l'analyse des comportements des usagers, présentant donc successivement les différentes étapes d'importation, de manipulation et de segmentation des données, mais également d'analyse et de prévision des résultats. Ce mémoire propose d'ailleurs deux processus de segmentation, un premier, plus traditionnel, dont l'intérêt a été maintes fois prouvé dans la littérature, et un second, expérimental, dont le potentiel reste à établir. Cette méthode expérimentale a été développée dans le cadre de ce projet et vise à améliorer la qualité et le temps de calcul de la segmentation. S'inscrivant dans une volonté d'analyser des données produites en continu, le processus se veut incrémental, permettant l'évaluation de comportements temporellement évolutifs. De plus, deux algorithmes de prévision sont appliqués et comparés sur les résultats de segmentation afin de comprendre le choix du processus le mieux adapté.

Finalement, la dernière partie présente les expérimentations et résultats d'application de la méthodologie sur les données issues de l'exploitation des cartes à puce de Gatineau. Une analyse traditionnelle complète des comportements des usagers adultes réguliers est produite dans un premier temps. Il y est montré une division de la population en des groupes distincts d'activités différentes. L'analyse expérimentale, quant à elle, s'intéresse à l'évolution des comportements de ces groupes à travers le temps. Produisant un résultat rapide et précis, il y est présenté le potentiel que représente ce nouveau processus.

La conclusion répond aux diverses problématiques et objectifs énoncés, montre les limites des méthodes et propose des perspectives d'application et d'amélioration en lien avec le présent sujet.

CHAPITRE 2 REVUE DE LITTÉRATURE

La revue de littérature s'attarde sur les différents travaux effectués en lien avec le sujet de recherche. En effet, elle commence par une description générale des cartes à puce et de leur utilité dans l'étude des comportements des usagers. Dans un second temps, il s'agit de s'intéresser aux démarches d'analyses usuelles appliquées en planification, puis aux différentes méthodes de segmentation couramment utilisées dans le but d'étudier les réseaux de transport et les comportements des individus. Finalement, ce chapitre se conclut par une partie réservée aux méthodes de prévisions usuelles en analyse dans le domaine des transports.

2.1 Cartes à puce

2.1.1 Définition

Connaissant une utilisation de plus en plus importante depuis les années 1990, les cartes à puce désignent des systèmes de transport intelligent dont le fonctionnement diffère très peu selon la ville qui les met en place. Chaque usager possède une carte à identifiant unique, et lorsqu'il veut se déplacer par le biais du transport en commun, il ne lui suffit que de valider son titre sur la borne automatique qui accepte ou non l'accès au véhicule. Ces lecteurs de cartes se situent généralement aux entrées de stations de métro et à la montée de chacun des bus. On peut parfois les retrouver à la sortie des moyens de transport, lorsque, par exemple, la ville établit une politique tarifaire liée à la distance effectuée. Visant à remplacer le système de tickets jugé obsolète, les cartes à puce se présentent sous la forme de cartes de crédit, proposant les avantages d'être portables et durables (Lu, 2007). D'un point de vue plus technique, une puce RFID est installée dans chacune des cartes permettant ainsi de stocker et transmettre des informations, par exemple l'identifiant du voyageur. Le lecteur envoie généralement une onde électromagnétique, ce qui active le marqueur de la carte à proximité qui lui renvoie de l'information, assurant ainsi l'identification et l'autorisation de transport. Pour des raisons de sécurité et de respect de la vie privée les cartes à puces sont toutes anonymisées. À chaque autorisation, une transaction numérique est enregistrée et diverses informations sont relevées puis stockées afin de percevoir l'usage du réseau.. Chaque carte est rechargeable suivant des conditions tarifaires dépendant des caractéristiques des usagers (étudiant, adulte, personne âgée...), mais également de leurs besoins (1 trajet, abonnement mensuel…). À Gatineau, l'usager peut effectuer le rechargement de sa carte en ligne, sur le site Web de la STO.

2.1.2 Intérêt international

Enregistrant en continu les transactions effectuées par les usagers, la technologie des cartes à puce présente ainsi de nouvelles opportunités d'analyses avec son lot de contraintes. Le risque et défi principal réside dans le fait de parvenir à extraire l'information utile de la masse de données sans la noyer. Pour cela, depuis le début du 21e siècle, de nombreux auteurs se sont penchés sur le problème, visant ainsi à analyser divers réseaux de transport en commun urbain à travers le monde. La littérature retiendra Barry, Newhouser, Rahbee, et Sayeda (2002), véritables pionniers dans l'analyse de données issues de cartes à puce. Ils proposent une méthode d'exploitation des données ayant pour but d'estimer les destinations des voyageurs du métro de New York, ville dont la technologie ne produit pas d'elle-même ce type d'information. Connaissant depuis une expansion internationale, on retrouve l'utilisation des cartes à puce et l'analyse de leurs données dans les plus grandes villes mondiales : Shanghai (Sun, Shi, et Schonfeld, 2016), Londres (Ceapa, Smith, et Capra, 2012), Santiago au Chili (Munizaga, Palma, et Mora, 2010), mais également dans des villes de plus petite taille comme Gatineau au Québec (Briand, Côme, Trépanier, et Oukhellou, 2017).

2.1.3 Limites de la technologie

Traitant du potentiel d'exploitation des cartes à puces dans les réseaux de bus, Bagchi & White (2004) montrent que cette technologie fournit des échantillons bien plus importants que les sources existantes. Par cela, on a la possibilité d'étudier les comportements des voyageurs sur de longues périodes. Des limites matérielles apparaissent: en effet, il est pensable que le réseau ne fournisse pas de données relatives à l'arrêt de destination, diminuant ainsi considérablement le champ d'action d'analyse des cartes. De plus, en considérant que chaque transaction est continuellement enregistrée par les automates et que la taille du réseau est généralement proportionnelle aux dimensions de la ville, l'étude de zones urbaines importantes peut présenter des complications matérielles. Toute analyse relève alors d'un problème de mégadonnées et ainsi, les volumétriques deviennent si importantes que les outils informatiques actuels peuvent être dépassés. À titre d'exemple, en Île-de-France, c'est une dizaine de millions de transactions qui sont effectuées chaque jour. Finalement, la fraude, qui est estimée à plus de 10% à Paris, ne peut qu'être supposée homogène sur tout le réseau. Elle biaise donc les statistiques d'analyse de données de cartes à puce et représente une limite à leur exploitation.

2.2 Analyse de la demande

Il s'agit, dans cette partie, de montrer l'intérêt de l'utilisation des données de cartes à puce dans l'analyse de la demande. Seront présentées ici deux approches d'étude, l'approche par modélisation de la demande et les approches longitudinales.

2.2.1 Modélisation de la demande

2.2.1.1 La Procédure séquentielle classique

La Procédure séquentielle classique (PSC) représente l'approche traditionnelle de la modélisation en transport. S'appuyant principalement sur les données de recensement, elle a été développée comme outil de base pour la prévision et l'analyse des performances d'un réseau. Le modèle trouve son application dans l'étude de tous les modes de transport, notamment le transport en commun. Cette méthode permet d'estimer la demande avec une bonne précision, cependant on peut lui reprocher une non-prise en compte de contraintes extérieures complexes influant également sur la demande. Citons par cela le management des infrastructures existantes ou même les phénomènes météorologiques influant sur le comportement des usagers.

La Figure 2-1 décrit le fonctionnement de la méthode : T représente le réseau, A l'ensemble des données sociodémographiques et N le flux de voyageurs, résultat de la PSC. La méthode passe par quatre étapes successives (Bonnel, 2002) :

- La génération de la demande consiste à déterminer le nombre de trajets effectués par zone, différenciant les déplacements produits par les zones et ceux attirés par les zones.
- La distribution des déplacements correspond au choix de destination de l'individu, estimant ainsi les déplacements intrazonaux et interzonaux. Cette étape génère la matrice OD.
- La répartition correspond au choix du mode de déplacement de l'individu. Elle vient définir le moyen de transport associé à chaque voyage, séparant ainsi la matrice OD générale en des matrices OD par mode.
- Finalement l'étape d'affectation assigne les itinéraires choisis par les individus minimisant un indicateur de désutilité.

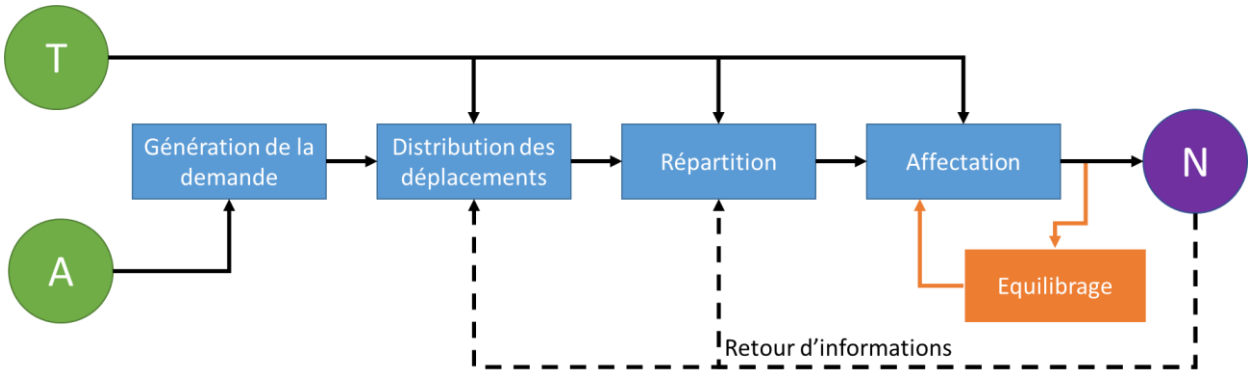


Figure 2-1: Schéma de compréhension méthode PSC – tiré de Mcnally (2000)

2.2.1.2 Le calibrage des modèles de distribution

Dans l'étape de distribution de la PSC, la méthode nécessite un calibrage à partir de matrices OD de références, représentatives de l'utilisation du réseau. Il s'agit d'un concept aussi bien utilisé dans le domaine du transport de personnes que de marchandises, servant notamment à identifier géographiquement les besoins en mobilité.

Il s'agit ici de proposer au modèle une matrice OD suffisamment précise pour des fins de modélisation de bonne qualité. Parmi les sources usuelles utilisées par les planificateurs de réseaux, nous nous intéressons aux données d'enquêtes ménages déplacements. Généralement sous la forme d'enquêtes sur une importante portion de la population, elles permettent d'obtenir des informations détaillées sur les habitants et d'estimer la demande. Elles ne sont, cependant, que rarement mises en place à la vue de leur coût d'application très important. Prenons comme exemple l'enquête OD conduite à Montréal. Il s'agit d'une source d'information générée tous les cinq ans, sondant 5% de la population dans le but de dresser un portrait de la mobilité des personnes. Elle permet notamment de reconnaître des comportements similaires entre les usagers. Cette étude se présente sous la forme d'un sondage complet fournissant des informations détaillées sur les ménages et leurs déplacements. La taille du ménage, l'âge, le genre, le revenu des habitants et leur niveau de motorisation sont des questions types associées à cette étude permettant d'estimer les habitudes et déplacements des habitants montréalais. L'étude, se concentrant principalement sur une analyse quinquennale de la demande, ne peut révéler les variations de comportements des usagers entre les applications d'enquêtes.

C’est pourquoi les données extraites des cartes à puce présenteraient une solution alternative à ce type de données. Outre le fait que cette technologie propose un coût d’application bien moins onéreux, elle permet un suivi détaillé spatio-temporel de l’utilisation des transports en commun. Les matrices OD présentent les flux de déplacements en fonction des origines et des destinations. Puisque les cartes à puce fournissent des données complètes et en temps réel, il est donc possible de générer des diagrammes de flux permettant de comprendre la demande des usagers. Ali, Kim, et Lee (2016), par exemple, construisent des matrices OD à partir des données de cartes à puce de Séoul en Corée. Ce réseau de transport est composé de 400 lignes de bus et 19 lignes de trains urbains. Puisqu’il s’agit d’un réseau de grande dimension, les auteurs définissent des TAZ (Traffic Analysis Zone) qui vont regrouper les stations proches dans le but de réduire considérablement les dimensions de la matrice et ainsi simplifier les calculs. Finalement leurs résultats les amènent à définir des zones d’activités dont la demande évolue au cours du temps.

2.2.1.3 Limites à l’utilisation des cartes à puces

D’un autre côté, Bagchi et White (2005) présentent une critique à l’utilisation des cartes à puce, visant à exposer leurs limites. En effet, l’absence de certaines informations relative à la ville étudiée prévient d’une analyse efficace du réseau. Pour la STO, par exemple, le fait qu’aucune information sur la destination des déplacements des usagers n’est récupérée détériore l’analyse des habitudes de trajets des usagers. Pour pallier ce problème, la littérature propose d’estimer les destinations en fonctions de certains paramètres. Dans l’une des premières études des données de cartes à puce, Barry et al. (2002) décrivent une méthode d’estimation des arrêts de destinations des usagers du métro de New York. Ils supposent que l’arrêt de départ d’un trajet a une très forte probabilité d’être le même que l’arrêt de destination du trajet précédent. Plus tard, Trépanier, Tranchant et Chapleau (2007) proposent une démarche pour déterminer les destinations des usagers à Gatineau. Fonctionnant de manière assez similaire, elle consiste à supposer que l’usager est susceptible de choisir comme départ l’arrêt géographiquement le plus proche de sa destination précédente. Leur algorithme va donc estimer la destination d’un trajet comme étant l’arrêt dont la distance (en mètres) est la plus courte par rapport au trajet suivant. S’appuyant sur ces travaux, Munizaga et Palma (2012) définissent une distance temporelle et sélectionnent donc l’arrêt dont le trajet prend le moins de temps. Ils montrent notamment un taux d’estimations correctes de 80% pour un réseau de transport en commun multimodal, basé à Santiago au Chili.

2.2.2 Analyse longitudinale

Une analyse longitudinale résulte d'une étude de suivi d'une population à travers le temps. Les données de cartes à puce, se présentant sous la forme de transactions temporelles régies par les mêmes attributs, coïncident parfaitement avec ce type d'étude. Largement documenté par Pelletier, Trépanier et Morency (2011), on remarque un engouement certain pour la compréhension des comportements des voyageurs. En effet, les auteurs ne cessent de promouvoir de nouveaux outils de fouilles de données dans le but d'enrichir l'analyse traditionnelle des comportements. Il est à noter que l'on parle communément de comportements de voyageurs alors que les études se basent sur l'usage de cartes à puces. En réalité, l'anonymisation fait qu'on ne puisse pas savoir qui voyage avec la carte étudiée. Pour la suite du mémoire, on adoptera la terminologie de comportements de voyageurs.

2.2.2.1 Les modèles d'activité

La plupart du temps, l'étude comportementale cherche à définir des activités pour les déplacements afin de comprendre les motivations de l'individu. Ainsi Devillaine, Munizaga et Trépanier (2012) ont travaillé sur un modèle d'activité à partir des données issues des cartes à puce de Santiago (Chili) et Gatineau (Canada). Ils basent l'assignation de l'information « activité » sur une série de règles prenant en compte les trajets sélectionnés et les temps entre les transactions des cartes. Puis ils segmentent l'ensemble des informations afin de repérer des patrons similaires entre les usagers et concluent en comparant les comportements des usagers des deux villes en question. Plus récemment Goulet-Langlois, Koutsopoulos, et Zhao (2016) ont proposé une méthode d'analyse à partir de données longitudinales enregistrées sur quatre semaines continues. Par cela, ils créent des séquences d'activités hebdomadaires propres à chacun des voyageurs du réseau londonien. Chaque activité est représentée par le déplacement d'une zone à l'autre et chaque zone est définie préalablement de manière à maximiser le nombre de déplacements entre les zones. À partir de ce point, les auteurs partitionnent leur échantillon en onze groupes aux patrons similaires. Reliant les informations spatiales et temporelles, ils étudient ainsi l'hétérogénéité des séquences parmi les voyageurs, et expérimentent une analyse de stabilité des groupes. Pour conclure, ils effectuent une analyse croisée avec des données d'enquêtes afin de retrouver les connexions significatives entre les attributs sociodémographiques et les patrons de déplacements.

2.2.2.2 Les trajets individuels

Les études doivent aussi se porter sur le suivi des individus. En effet, les lecteurs de cartes collectent les données d'identification de cartes à puce, ce qui permet de suivre le comportement d'un usager par le biais de création de chaînes de déplacements. En se concentrant sur l'analyse de l'individu, Ma, Wu, Wang, Chen, et Liu (2013) ont développé une approche de fouille de données au niveau microscopique, croisant données spatiales et temporelles. Cette démarche offre une réelle alternative pour mesurer la similarité et variabilité des comportements des voyageurs. Dans leur étude, ils reconstruisent les chaînes de déplacements quotidiennes des usagers, et utilisent notamment des algorithmes de segmentation à réduction de bruit (DBSCAN) pour identifier les comportements réguliers. Ces informations sont très importantes dans la compréhension des patrons journaliers des habitants de Beijing en Chine, trouvant notamment une application dans la segmentation de marché.

Cependant, il est possible que le comportement quotidien soit influencé par les autres jours de la semaine en question. C'est pourquoi d'autres auteurs comme Agard, Morency, et Trépanier (2006) ou plus récemment, Mahrsi, Côme, Oukhellou, et Verleysen (2017) s'intéressent à la création de comportements de semaines. Ce dernier papier compare une méthode plus classique d'analyse de patrons d'activités avec une nouvelle méthode basée sur le regroupement d'individus dont les trajets sont temporellement similaires. Ils reconstruisent donc, ici aussi, des chaînes de déplacements individuelles relatives à l'utilisation hebdomadaire des transports en commun à Rennes en France.

De leur côté, Lathia, Smith, Froehlich, et Capra (2013) cherchent à se dédouaner des problèmes d'analyse que peuvent représenter les jours de fin de semaine. Ils séparent ainsi les profils individuels de déplacements en semaine, des profils individuels de déplacements en fin de semaine. Leur étude propose notamment l'utilisation d'algorithmes de segmentation par agglomération hiérarchique (HAC) dans le but de révéler les différents comportements. Finalement ces résultats aspirent à montrer l'importance de l'utilisation des données de cartes à puce, particulièrement dans les analyses à l'échelle individuelle, dans le but de créer des systèmes d'information ciblant l'usager.

2.2.2.3 Le long terme

Rappelons que la technologie des cartes puce enregistre toutes les transactions effectuées sur un réseau. Fournissant ainsi une gigantesque masse de données continuellement grandissante, il est donc possible d'étudier les motifs d'activités et leurs variabilités sur le long terme (Agard et al., 2006). Morency, Trepanier, et Agard (2006) ont notamment travaillé sur une représentation journalière du comportement de l'individu et son analyse sur le long terme comparant diverses méthodes de fouilles de données. L'étude est appliquée sur la base de données issue des transactions de cartes à puce de Gatineau, couvrant une période de dix mois. Leurs expériences montrent finalement que les comportements des usagers évoluent en fonction du temps, aussi bien en fréquence qu'en heures de départ. Il est très intéressant de constater ces évolutions saisonnières et chercher à comprendre ces comportements en fonction des types tarifaires. Ces analyses permettent de montrer, par exemple, que le comportement d'un travailleur à temps plein ne va pas évoluer de la même manière qu'un retraité, aidant ainsi le planificateur à mieux comprendre l'évolution des comportements des voyageurs.

D'autres travaux cherchent à pousser l'analyse en mettant ces évolutions comportementales en lien avec des événements météorologiques. Dans son mémoire, Descoimps (2011) cherche à montrer l'influence météorologique sur l'utilisation des transports en commun, en analysant les données de cartes à puce de Gatineau sur du long terme. Dans la continuité de l'étude précédente, l'article met en évidence les différences comportementales des types de titres, cherchant notamment à montrer les diminutions de déplacements provoquées par des conditions météo extrêmes. Par exemple, un aîné sera bien plus influencé par une forte pluviosité ou des chutes de neige que des travailleurs réguliers se déplaçant obligatoirement en semaine. L'auteur cherche, ici aussi, à montrer le potentiel énorme que permet la technologie des cartes à puce dans l'analyse de la variabilité des comportements.

Il est à noter que les périodes d'étude peuvent être raccourcies par l'organisme gérant l'anonymisation des cartes. En effet, sur le réseau Transilien de Paris en France, les identifiants des usagers SNCF sont anonymisés tous les trois mois, réduisant donc la période de suivi des individus. Il s'agit d'une limite extérieure aux analyses de données de cartes à puce. Mais les récents travaux cherchent souvent à montrer l'intérêt de l'utilisation des cartes à puce, notamment dans le but de faire tomber ces limitations.

2.3 Méthodes de segmentation

Il s'agit de présenter dans cette partie diverses méthodes de fouilles de données applicables dans l'extraction de patrons de déplacements similaires chez les usagers. On tient, tout d'abord, à montrer une distinction entre classification et segmentation, deux outils de fouilles de données souvent confondus. La classification sert à assigner des attributs aux données en fonction des arrangements récurrents des autres attributs. Elle commence par reconnaître les patrons dans une base d'apprentissage, pour enfin les deviner dans une base de test. La segmentation, quant à elle, va simplement subdiviser une population dans le but de maximiser l'homogénéité des individus dans les groupes et maximiser l'hétérogénéité entre les groupes. On dit donc que la segmentation est une méthode de classification à apprentissage non supervisée, puisqu'elle apprend par observation.

2.3.1 Caractéristiques

Dans leur livre, Han, Pei, et Kamber (2011) définissent différentes caractéristiques à la segmentation et à ses résultats. On définit ici une liste d'exigences nécessaires au choix d'un algorithme de segmentation applicable au domaine des transports en commun :

- La scalabilité : un bon algorithme de segmentation doit être capable de s'adapter à un changement d'ordre de grandeur du volume de données. Beaucoup d'algorithmes fournissent d'excellents résultats en traitant de petits volumes de données, mais proposent des résultats biaisés lorsqu'il s'agit de traiter des millions d'informations. Un bon outil de segmentation doit donc savoir maintenir ses fonctionnalités et performances avec les grosses volumétries, qui sont notamment très courantes dans le domaine des transports en commun.

- La flexibilité est la caractéristique de s'adapter à différents types d'attributs. On peut effectivement retrouver aussi bien des attributs discrets, continus ou encore asymétriques dans les données de cartes à puce.

- La forme des groupes : la plupart des algorithmes se basent sur des calculs de distances à partir de mesures euclidiennes ou de Manhattan, fournissant généralement des résultats sphériques de densités similaires. La limite réside dans le fait que les groupes devraient

pouvoir avoir des formes et des densités différentes, dans le but de se rapprocher au mieux du réel.

- La connaissance du domaine pour l'entrée des paramètres de l'algorithme : les algorithmes nécessitent souvent des paramètres d'entrée plus ou moins difficiles à estimer. Le nombre de groupes par exemple est une donnée d'entrée nécessitant parfois un jugement humain dans le but d'optimiser la compréhension et la viabilité des résultats.
- Le bruit fait partie intégrante des données issues de cartes à puce. Il peut notamment apparaître sous la forme d'erreurs matérielles ou encore d'erreurs humaines. Par cela plusieurs déplacements peuvent ne pas avoir de sens réel. Même si on effectue généralement un nettoyage des données au préalable, l'algorithme doit ne pas être trop sensible au bruit restant.
- Segmentation incrémentale : dans le domaine des transports en commun, les données arrivent continuellement. Certains algorithmes nécessitent de simuler sur l'ensemble des données à chaque ajout de nouvelles données. Un bon algorithme se doit de pouvoir résoudre ce problème.
- Grandes dimensions : dans le but de fournir un maximum de détails à l'analyse, il est possible de travailler sur des données comprenant un grand nombre d'attributs. Il s'agit d'un véritable enjeu puisque l'analyse en grandes dimensions est très facilement biaisée.
- Contraintes extérieures : dans l'analyse de la demande, lorsqu'on cherche à améliorer notre étude en la complétant d'informations spatiales, beaucoup de contraintes extérieures apparaissent. En effet, les configurations du réseau et de la ville jouent un rôle important dans le choix d'une station par un utilisateur. Très peu de méthodes actuelles sont capables d'implanter ce type de contraintes, proposant un axe de réflexion d'amélioration des méthodes de segmentation.
- L'interprétabilité désigne la facilité d'utilisation des résultats de segmentation. Souvent dépendant de la forme des vecteurs d'entrée, le résultat doit être clair et simple à analyser.

2.3.2 Algorithmes de segmentation courants

On présente ici une liste des algorithmes de segmentation les plus courants dans la littérature notamment dans l'analyse comportementale des individus en transports en commun.

2.3.2.1 Méthode agglomérative hiérarchique (HAC)

La segmentation agglomérative (ou classification ascendante hiérarchique) fait partie de la catégorie des méthodes hiérarchiques algorithmiques, en opposition aux méthodes hiérarchiques probabilistes et bayésiennes qui sont bien moins utilisées en transports. Lorsque l'on cherche à organiser une population en un nombre fixe de groupes exclusifs, on cherche à partitionner cette population à différents niveaux, et c'est sur ce principe que fonctionne l'algorithme.

La HAC est donc une méthode de décomposition hiérarchique de la population. Pour fonctionner, elle commence par définir que chaque objet constitue un groupe, puis elle vient regrouper itérativement ces groupes en conservant l'ordre de fusion. Les fusions ne concernent que les deux groupes les plus proches lors de l'itération. L'algorithme s'arrête lorsqu'il ne reste plus qu'un seul groupe. Ce groupe devient donc la racine de l'arbre de regroupement formé par la méthode, souvent représenté par un dendrogramme. La littérature définit également différentes métriques afin de s'adapter au mieux aux problèmes à analyser.

Fonctionnant suivant un principe très simple, sa complexité est pourtant mauvaise. Puisqu'une seule fusion n'est possible par itération, la méthode nécessite donc n-1 itérations. Ayant une mauvaise capacité à s'adapter à un changement d'ordre de grandeur, une application sur de grosses volumétries est donc impossible. De plus, fonctionnant suivant un principe itératif fixe, la méthode ne peut revenir en arrière pour corriger ses erreurs de fusion. Le bruit biaise donc la méthode. Elle trouve néanmoins son utilité dans les domaines de généralisation de bases de données ou dans le cadre de visualisations, à la vue de sa très bonne interprétabilité.

Dans leurs récents travaux, Ghaemi, Agard, Trépanier, & Nia (2017) proposent une méthode de visualisation de résultats de la HAC. Par cela ils cherchent à regrouper des individus dont le comportement temporel est similaire, puis à visualiser sur un graphique en trois dimensions la proximité comportementale entre les usagers. La méthode du HAC, en plus de proposer une division sur 18 groupes, fournit les fusions futures de ces groupes, permettant donc de visualiser les proximités entre les groupes. L'étude permet ainsi une meilleure compréhension des comportements des groupes et des similarités entre eux.

2.3.2.2 Algorithme des K-Moyennes

Lorsque l'on vient à travailler avec des données massives, l'algorithme de segmentation par excellence est la méthode des K-Moyennes. Elle fait partie de la catégorie des méthodes de partition de données. Soit une population de n objets, l'algorithme vient séparer cet ensemble en k groupes non vides. Les allocations des individus dans les groupes se réalisent par la méthode des plus courtes distances. À chaque itération, l'algorithme vient recalculer la position des centres des groupes et réajuster l'allocation des individus, augmentant par cela la qualité des partitions. L'algorithme s'arrête généralement lorsqu'il trouve une convergence, c'est-à-dire que les itérations n'apportent plus de modification aux allocations. Il s'agit donc d'une méthode heuristique, qui s'approche rapidement d'un optimum local.

Il s'agit de l'algorithme le plus populaire et le plus simple à implémenter. Sa faible complexité lui permet de traiter rapidement des bases de données de très grandes tailles, on lui accorde notamment une excellente efficacité pour les petites et moyennes volumétries. Cet outil permet de trouver des groupes strictement différents en taille et densité, mais qui sont de forme sphérique puisque la méthode de calcul des distances se base sur la mesure euclidienne. La méthode fournit des centres de groupes et les populations respectives appartiennent à l'hypersphère formée par l'habitant le plus éloigné. Étant un algorithme très simple, il présente néanmoins beaucoup d'inconvénients. Plusieurs paramètres, dont le nombre de groupes, sont à fixer par l'utilisateur. Son principe heuristique se base sur le fait de simuler plusieurs fois l'algorithme à partir de noyaux aléatoires, trouvant à chaque fois un minimum local, et concluant par celui qui présente la meilleure qualité. Cette méthode basée sur l'étude de centroïdes ne cherche pas à identifier l'optimum général; son utilisation ne devrait pas servir dans ce but.

Même dans le domaine des transports en commun, la littérature montre une volonté d'améliorer ce type d'algorithme. On peut notamment citer la méthode des K-médoïdes qui se concentre sur l'étude des médoïdes des groupes (Trasarti, Pinelli, Nanni, & Giannotti, 2011), ou encore une amélioration plus connue, la méthode des K-moyenne ++, qui cherche à optimiser le choix des noyaux de départs des itérations (Dzikrullah, Setiawan, & Sulistyo, 2016).

2.3.2.3 Algorithme de densité (DBSCAN)

Les méthodes précédentes se concentrent sur un partitionnement basé sur une métrique générant ainsi des groupes dits « sphériques ». Dans certaines analyses, spatiales notamment, il est impossible de concevoir que les résultats soient strictement sphériques. Pour résoudre ce problème, les méthodes de densités sont mises en place.

Comme son nom l'indique, il s'agit de méthodes se concentrant sur la densité d'individus afin d'effectuer son partitionnement. Le DBSCAN va scanner un à un les individus de la population en passant de voisin en voisin. La densité d'un individu est mesurée par le nombre d'individus proches de ce dernier. On définit cette notion de proximité par un paramètre $\epsilon > 0$, choisi par l'utilisateur, qui représente le rayon de densité. Ainsi, tout objet se trouvant dans l'hypersphère formée par le centre d'un individu et ce rayon se fait appeler voisin de cet individu. Plus la valeur de ce paramètre est faible, mieux le résultat de l'algorithme sera détaillé, enlevant un maximum de bruit, mais risque de fournir trop de groupes. L'algorithme cherche dans un premier temps des noyaux : il s'agit d'individus dont le voisinage est très dense. On définit un individu au voisinage dense s'il possède plus de voisins que le paramètre $MinPts$, lui aussi choisi au préalable par l'utilisateur. Plus la valeur de ce paramètre est élevée, plus les groupes seront denses, mais il est plus probable que certains « bons » individus seront considérés comme du bruit. Lorsque l'algorithme ne trouve plus de voisin non visité, il a deux choix : soit le groupe de voisins visités comporte au moins un noyau et le groupe est donc considéré comme un groupe. Soit le groupe de voisins visités ne comporte pas de noyau et sera donc considéré comme du bruit. La méthode reprend en choisissant un noyau aléatoire non visité et ne s'arrête que lorsque tous les individus ont été visités.

Il s'agit d'un algorithme assez simple ne nécessitant pas d'information sur le nombre de groupes à trouver. Il traite également facilement le bruit en l'enlevant du processus de partitionnement. Pourtant sa forte complexité ($o(n\log(n))$), rend la méthode non scalable et limite une utilisation sur de grosses volumétries. De plus, le choix des paramètres est indispensable à son bon fonctionnement. Pour trouver une combinaison efficace de ces valeurs, il est nécessaire de simuler l'algorithme avec diverses combinaisons de paramètres et de comparer leurs résultats. Finalement, la présence même de ces paramètres empêche l'algorithme de traiter des groupes de densités différentes. On tient notamment à citer les travaux de Kieu, Bhaskar, & Chung (2014), dont l'analyse comportementale des individus s'exécute à partir de cet algorithme de segmentation.

2.3.3 Algorithmes de segmentation dynamique

Motivés par une volonté de produire des segmentations incrémentales au sens de Han et al. (2011), divers auteurs se sont concentrés sur la réception fragmentée et continue des données, et ont introduit une possibilité d'évolution temporelle des groupes. Il s'agit donc de méthodes capables de traiter les données de façon itérative à chaque disponibilité de nouvelles données. Sur chacune des itérations, les algorithmes génèrent des paramètres qui seront utilisés dans le traitement des données suivantes. Ces méthodes permettent notamment de traiter des volumes importants en des temps de traitements raisonnables puisqu'elles ne s'intéressent qu'à des échantillons (Assaad, 2014). Telle que définie par Assaad, la classification dynamique peut se présenter sous trois modèles : les modèles de mélanges avec régressions polynomiales, les modèles gaussiens avec a priori sur la dynamique d'évolution des centres et les modèles dynamiques à espace d'état.

2.3.3.1 Les modèles de mélanges avec régressions polynomiales

Il s'agit d'un modèle de mélange gaussien dédié au partitionnement de données temporelles, dans lequel l'évolution temporelle de chaque classe est modélisée par une fonction polynôme. On suppose donc que la distribution des individus suit une loi polynomiale, en opposition au modèle de mélange gaussien traditionnel qui ne possède généralement aucune information sur cette distribution. Pour résumer, ce modèle permet une modélisation de l'évolution non linéaire des classes.

Plus d'informations sont disponibles dans les travaux de DeSarbo et Cron (1988) qui sont les premiers à définir ce type de modélisation, et plus récemment chez Antoniadis, Bigot, et von Sachs (2009).

2.3.3.2 Les modèles Gaussien avec a priori sur la dynamique d'évolution des centres

Il s'agit d'un modèle de mélange, analogue au précédent. Seule la modélisation de l'évolution des groupes est sensiblement améliorée. Dans cette méthode, on émet l'hypothèse que les centres de groupes sont dépendant du temps.

Plus d'informations sont disponibles dans les travaux de Calabrese et Paninski (2011) qui ont proposé cette méthode.

2.3.3.3 Modèle dynamique à espace d'état

Souvent utilisés dans le traitement de séries temporelles, ils ont été développés dans le but de suivre les trajectoires des satellites à partir des lois physiques. La méthode sert notamment à prédire les prochains mouvements du satellite, prenant compte des données arrivant en continu.

Plus d'informations sont disponibles dans les travaux de Swerling (1958) et, plus récemment, dans le traitement de séries temporelles (Bentoglio, Fayolle, & Lemoine, 2001; Harvey, 1990)

2.3.3.4 Conclusion sur les modèles

Ces trois méthodes se basent sur une modélisation de l'évolution des centres des groupes. Certains, comme le modèle à dynamique d'état, permettent notamment de respecter des contraintes extérieures comme les lois physiques. Cette liste sert essentiellement à montrer que des applications existent et les travaux cités permettent un approfondissement sur le fonctionnement propre des algorithmes, qui sont inutiles dans notre étude.

Il est à noter que le domaine du transport en commun ne semble pas encore s'être penché sur l'application de ces modélisations.

2.3.4 Choix du nombre de groupes, méthodes usuelles

D'habitude complètement arbitraire, le choix du nombre de groupes pour effectuer les segmentations K-Moyennes porte souvent à discussion. Classiquement, plusieurs outils sont mis en place pour « optimiser » combien de groupes choisir. D'un côté, un nombre trop élevé génère beaucoup trop de détails dans l'analyse et l'on perd le but même de la segmentation qui consiste à regrouper les profils similaires. De l'autre côté, un nombre trop petit ne fournit pas assez de détails pour caractériser la population. Afin de conserver une segmentation de qualité, on chercher souvent à maximiser l'hétérogénéité entre les groupes et à maximiser l'homogénéité à l'intérieur de ces groupes.

On présente plusieurs méthodes d'aide à la décision pouvant être utilisées en fonction de diverses volumétries.

2.3.4.1 Méthode du coude

Il s'agit d'une des méthodes les plus anciennes, probablement développée par Thorndike (1953), une méthode visuelle nécessitant de simuler la segmentation par le biais de la méthode des K- Moyennes, pour des K successifs allant généralement de 1 à 10. Après ces 10 simulations, on compare la distorsion de chacune des simulations. La distorsion, aussi appelée inertie intragroupe, est un indicateur de qualité de segmentation défini par la somme des distances au carré entre les vecteurs et leur centre associé. En traçant la courbe de la distorsion en fonction de K, on cherche ensuite à déterminer une rupture dans cette courbe, synonyme d'une forte hétérogénéité dans les groupes considérés. Il est classique de choisir K supérieur au point de rupture, on prend généralement le point le plus bas de la chute.

On dit que la courbe forme un « coude », selon la Figure 2-2 on choisirait $K = 3$.

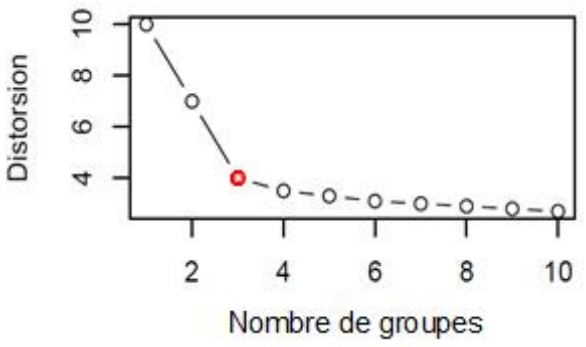


Figure 2-2: Indentification du « coude »

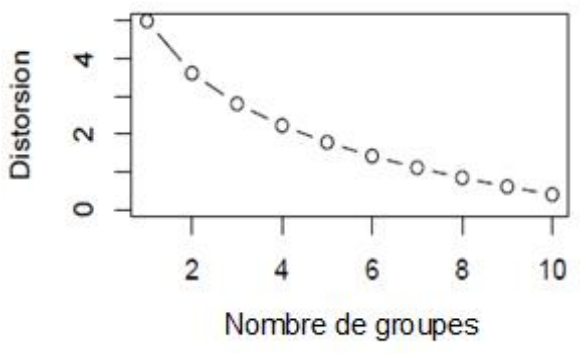


Figure 2-3 : Ambiguïté d'identification par la méthode du « coude »

Cette méthode est malheureusement souvent sujette à des ambiguïtés. En effet, le « coude », principal indicateur de la méthode, peut, dans certains cas, ne pas être distinctement visible, comme pour la Figure 2-3, ou même apparaître plusieurs fois (Ketchen et Shook, 1996). Ces ambiguïtés ne nous permettent pas de conclure sur la valeur de K et nous devons utiliser une autre méthode de détermination du nombre de groupes optimal.

2.3.4.2 Méthode des silhouettes

Il s'agit d'une méthode également visuelle post-segmentation introduite par Rousseeuw (1987), prenant $S(i)$ de l'Équation 2.1 comme indicateur principal. Dans cette expression, $a(i)$ correspond à la dissimilitude moyenne intragroupe, c'est-à-dire la moyenne des distances entre un identifiant lambda i et chacun des autres éléments du même groupe, tandis que $b(i)$ correspond au minimum des moyennes des distances avec les autres groupes. C'est-à-dire que l'on vient calculer une moyenne des distances entre i et tous les éléments pour chacun des autres groupes, et que l'on choisit uniquement la moyenne la plus basse. On dit généralement qu'il s'agit du groupe voisin le plus proche de i. On vient ainsi calculer l'indicateur $S(i)$ pour chacun des éléments de la base de données. Chaque $S(i)$ a une valeur comprise entre -1 et 1. Par cela, plus l'indicateur $S(i)$ a une valeur faible moins l'individu i est sensé se trouver dans son groupe et inversement.

Équation 2.1 : Indicateur Silhouette

$$S(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

Ici aussi, la méthode nécessite de simuler la segmentation pour plusieurs valeurs de K et ensuite de comparer la moyenne des $S(i)$ de chaque simulation. On conserve celle dont l'indicateur $S(i)$ est le plus grand et donc la valeur de K associée.

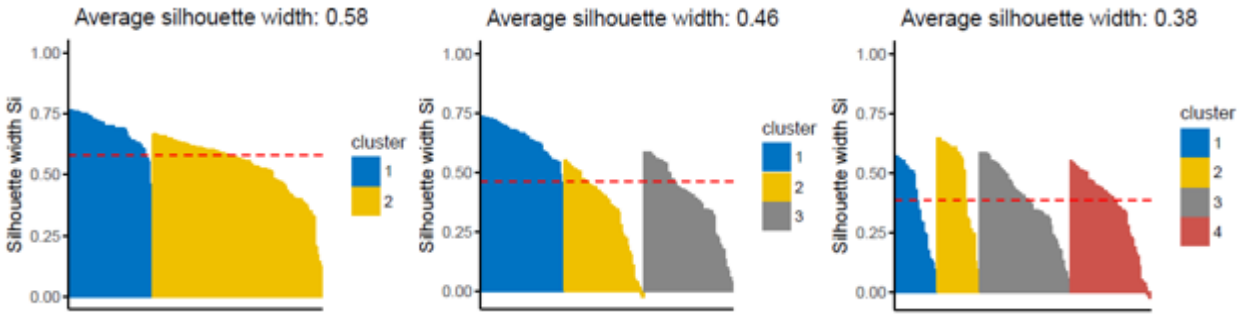


Figure 2-4 : Comparaison des silhouettes pour $K = 2, 3, 4$, pour les données « iris »

Afin de mieux comprendre la méthode visuelle, on traite l'exemple de l'utilisation de celle-ci sur une base de données fournie par R : « iris ». En simulant successivement la segmentation pour différentes valeurs de K, on est capable de représenter l'ensemble des éléments par des barres de couleur sur la Figure 2-4. Chaque élément i est donc de taille proportionnelle à la valeur de $S(i)$ et de même couleur que les autres éléments du groupe. Triées par ordre décroissant suivant l'indicateur $S(i)$, ces figures nous donnent un bon aperçu de la qualité de segmentation pour chacun des groupes. Puisque les valeurs de $S(i)$ sont comprises entre -1 et 1, avec 1 désignant une excellente qualité de segmentation. On peut dire que pour $K = 2$, le groupe 1 contient des éléments qui ont été très bien placés, tandis que pour le groupe 2 plusieurs points perdent en qualité. On pourrait se demander ainsi si ce groupe peut se séparer en 2. En ce qui concerne le choix de K, on se réfère à la Figure 2-5, et l'on choisit celui dont le $S(i)$ moyen est le plus haut, ici $K = 2$.

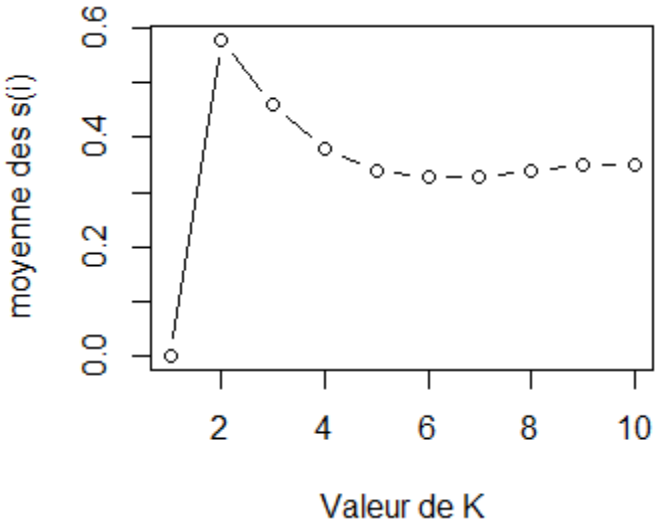


Figure 2-5 : Comparaison des moyennes de $S(i)$ pour Kallant de 1 à 10, données « iris »

Cette méthode fournit donc des informations détaillées quant à la qualité de segmentation relative à chacun des individus de la base de données. Cependant, lors de traitement de bases de données de très grandes tailles, le temps de calcul devient extrêmement long par rapport à la méthode du « coude », et nécessite une mémoire bien plus importante (création d'un vecteur par individu, à chaque K simulé). Il peut donc arriver que les ressources de l'ordinateur ne suffisent pas à réaliser le calcul.

2.3.4.3 Méthode de la Statistique *Gap*

Il s'agit d'une méthode essentiellement calculatoire post-segmentation introduite par Tibshirani, Walther, et Hastie (2001), visant à simuler l'algorithme des K-Moyennes sur une plage de données de K, généralement de 1 à 10, et à comparer un indicateur défini.

Tout d'abord, les auteurs définissent une variable W_K représentant les distances intragroupes combinées. Dans l'Équation 2.2, r est le groupe sélectionné, n le nombre d'éléments et d signifiant la distance euclidienne entre les éléments choisis.

Équation 2.2 : Variable W_k

$$W_K = \sum_{r=1}^{K} (\frac{1}{2n_r} \sum_{i,i' \in C_r} d_{i,i'})$$

L'idée principale de la méthode est de standardiser la comparaison entre $\log(W_K)$ et une distribution des données dite de référence. Cette distribution de référence est générée de manière à ce qu'il n'y ait pas de segmentation évidente de ses valeurs. Pour ce faire, il suffit de prendre le rectangle des valeurs limites des données réelles et de générer aléatoirement B échantillons de mêmes valeurs limites comportant autant de données que la base observée. Pour standardiser cette comparaison, on cherche à maximiser l'indicateur $Gap(K)$ défini par l'Équation 2.3, où E^* correspond à cette distribution de référence.

Équation 2.3 : Indicateur Gap

$$Gap_n(K) = E_n^*\{\log(W_K)\} - \log(W_K)$$

Afin de comprendre visuellement cette méthode, on la simule sur la base de données fournie par R : « USArrests ». Sur la Figure 2-6, on a tracé les valeurs du $\log(W_K)$ observé ainsi que celle du $\log(W_K)$ de référence. On remarque que les valeurs d'observation sont inférieures à celle de références, ce qui est généralement vrai puisque la distribution de référence est censée donner une très mauvaise qualité de segmentation. On cherche à maximiser l'écart entre les deux courbes et donc, selon la Figure 2-7, on choisit $K = 4$ puisque l'écart est maximal en ce point.

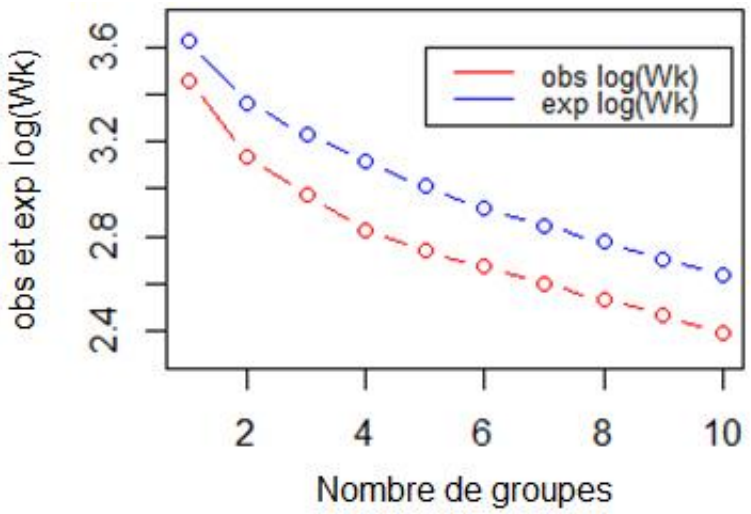


Figure 2-6 : Évolution du $\log(W_K)$ pour la distribution observée et de référence « UsArrests »

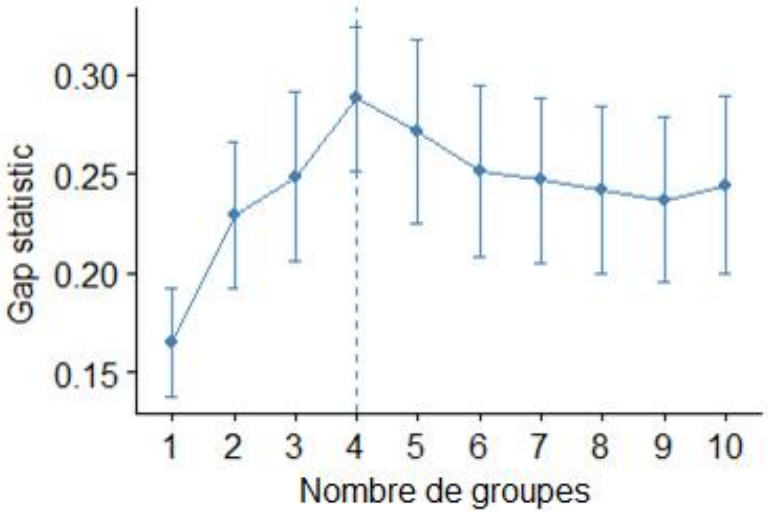


Figure 2-7 : Évolution de l'indicateur Gap pour les données « USArrests »

Le principal inconvénient de la méthode est que le calcul de W_K revient à générer des matrices de distances intragroupes pour chacun des groupes, et ce, pour chacune des valeurs de K simulées. Lorsque l'on travaille avec des données de masse, il peut arriver que les ressources de l'ordinateur ne suffisent pas à réaliser le calcul.

2.4 Méthodes de prévision

Diverses méthodes de prévision ont été développées lorsque les données socioéconomiques représentaient les principales sources. Les méthodes, comme la Procédure séquentielle classique, permettaient notamment de prévoir la demande de transport à différents horizons (Bonnel, 2002). Fournissant des informations plus complètes et continues, les cartes à puce constituent des séries temporelles, pouvant donc être analysées de manière analogue aux traitements de signaux. On présentera ici deux méthodes courantes et simples, développées pour le traitement de signal et trouvant leurs applications dans l'étude des données de transports.

2.4.1 Les modèles ARMA

Les modèles autorégressifs et à moyenne mobile (ARMA) sont les modèles les plus courants utilisés en traitement de séries temporelles. Il s'agit de processus stationnaires fournissant un très bon outil de prévision ainsi qu'une méthode automatique d'estimation des paramètres optimaux. Souvent noté sous la forme ARMA(p,q), ils sont composés d'un modèle autorégressif AR d'ordre p et d'un modèle moyenne mobile MA d'ordre q. Le processus fonctionne selon le principe que toute valeur de la série temporelle à l'instant t doit pouvoir s'écrire sous la forme d'une addition entre une fonction du passé de la série (AR(p)) et une fonction des perturbations de la série (MA(q)) (Bougas, 2013). L'estimation des paramètres p et q s'applique, soit manuellement à partir d'une analyse de corrélation, soit automatiquement en minimisant les critères AIC d'Akaike (1974) et BIC de Schwarz (1978).

Des améliorations sont apportées au modèle de base dans le but de traiter des données non stationnaires, comme le sont les données de carte à puce. On les appelle les modèles autorégressifs intégrés et à moyenne mobile (ARIMA). La méthode définit généralement un opérateur de différentiation Δ qui, après répétition, va supprimer les tendances des séries. La Figure 2-8 présente le fonctionnement général de la méthode.

Le test de stationnarité peut s'appliquer à partir d'un test de Dickey-Fulley (DF), de Phillips-Perron (PP) ou de Kwiatkowski-Phillips-Schmidt-Shin (KPSS). Si le résultat donne une non-stationnarité des données, il est nécessaire d'appliquer la fonction de différenciation. S'il est concluant on exécute le processus ARMA, puis on effectue des tests d'adéquation et de normalité afin de définir si la modélisation est acceptable ou non. Pour les tests d'adéquation des résidus, les tests de Ljung-

Box et de Box-Pierce sont les plus courants et pour le test de normalité, on pourra utiliser le test de Kolmogorov-Smirvnov.

Une dernière amélioration de la méthode permet de s'occuper des données non stationnaires avec présences de saisonnalité. Il s'agit de la méthode SARIMA (Seasonal ARIMA), sa seule différence réside dans l'étape transformant la série en une série stationnaire, en ajoutant un opérateur éliminant la périodicité.

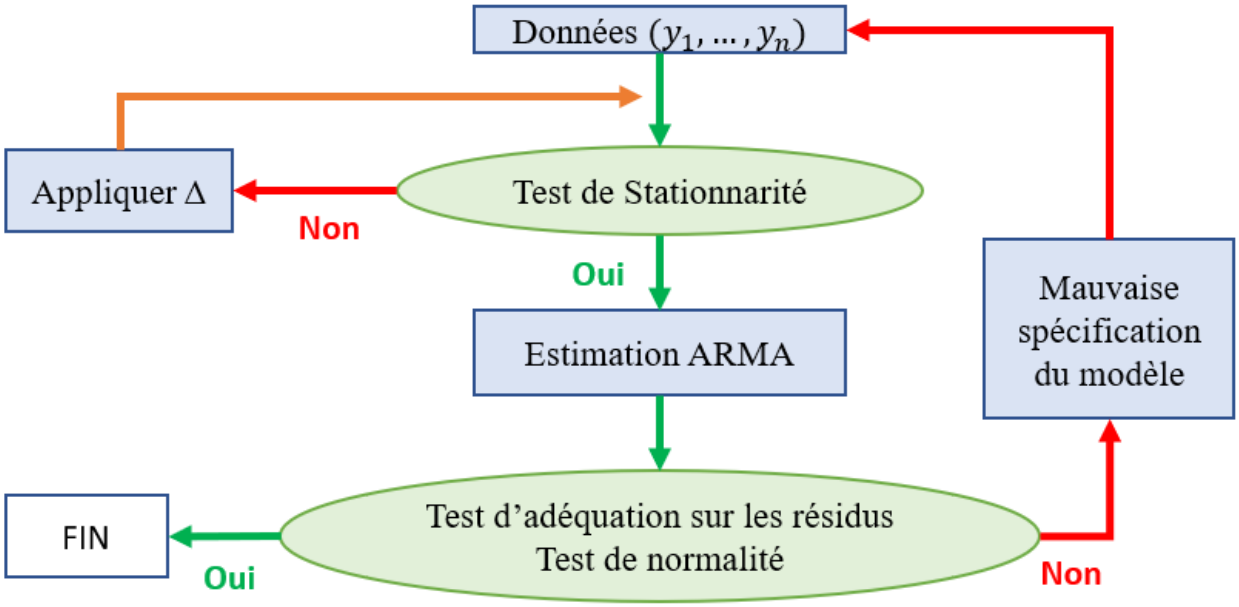


Figure 2-8: Schéma modélisation d'une série temporelle par ARIMA – tiré de Boutahar (2007)

Ni, He, & Gao (2017) proposent un algorithme de détection d'événements afin de repérer et prévoir les flux de voyageurs de métro à partir de données issues de médias sociaux. Ils appliquent notamment le processus SARIMA afin d'établir un modèle d'utilisation du réseau à partir des occurrences de mots clés. Mais ils montrent rapidement l'incapacité de la méthode à fournir des résultats concluants sur le long terme.

Très développés dans le domaine du trafic automobile et aérien, les modèles ARMA ne fournissent qu'une très faible littérature d'application dans le transport en commun. Les données de cartes à puce s'apparentant à des séries temporelles non stationnaires avec saisonnalité génèrent pourtant un énorme potentiel d'application du processus SARIMA.

2.4.2 Holt-Winters

Les procédures récursives de Holt-Winters ont principalement été développées dans le but de traiter des séries temporelles présentant une tendance et une saisonnalité. Comme pour une décomposition de courbe, il existe deux méthodes associées à cette procédure :

- Le Holt-Winters additif définissant l'estimation de la valeur de la série à l'instant t, comme étant la somme des termes de saisonnalité, de bruit, du niveau et de la tendance. Cette méthode est généralement préférée lorsque l'amplitude de la saisonnalité reste constante.
- Le Holt-Winters multiplicatif définissant l'estimation comme étant un produit des termes de la prévision additive. Cette méthode est préférée lorsque le patron de saisonnalité est caractérisé par une amplitude variable en fonction du niveau de la série.

Les termes de tendance, de niveau et de saisonnalité sont trois fonctions du temps qui seront, dans un premier temps, estimées pour l'ensemble des valeurs réelles de la série temporelle. On appelle cette étape le lissage exponentiel. La prévision à un horizon h, passe donc par la prise en compte de tous les termes précédents.

Les méthodes de prévision par lissage exponentiel fournissent des prévisions très peu coûteuses et souvent de bonne qualité. Les erreurs de prévisions sont souvent faibles que ce soit à court, moyen et long terme. On remarque cependant une limite majeure à l'utilisation de la méthode, ne s'intéressant qu'à l'historique de la série temporelle, elles ne prennent pas en compte les contraintes extérieures comme les possibles modifications dans la configuration du réseau. Une simulation sur plusieurs années affine fortement le modèle, une fois réalisé, l'analyse des erreurs de prévision est un bon moyen de reconnaître les changements comportementaux des usagers.

Dans la littérature, Bougas (2013) établit une comparaison de ces processus de prévisions en analysant les flux de voyageurs aériens. Par cela, il conseille l'utilisation des modèles ARIMA et SARIMA à la vue des très faibles hypothèses en jeu. Il s'agit ici de deux algorithmes très souvent utilisés en prévision de la demande de transport aérien. Malgré des erreurs de prévisions généralement plus faibles pour les modélisations, les méthodes de lissage exponentiel de courbe permettent une meilleure prévision à long terme. Le calcul des erreurs de prévision se réalise principalement à partir des indicateurs MAPE et RMSE.

2.5 Conclusion sur la revue de littérature

Cette étude cherchait tout d'abord à montrer l'énorme potentiel des cartes à puce en les comparant à des sources plus traditionnelles. Puisque l'ensemble des transactions des voyageurs sont enregistrées grâce à cette technologie, les données de cartes sont exhaustives, fournissant également des informations spatio-temporelles. Cela rend possible des analyses longitudinales et montre ainsi l'intérêt de telles données dans la compréhension du comportement des usagers. Pouvant aussi être considérées comme une limite, d'immenses volumétries incomplètes (manque d'information sur les destinations, activités) sont générées, nécessitant des capacités de calculs toujours plus conséquentes.

Dans un second temps, cette revue de littérature visait à dresser un état des lieux sur l'avancement des recherches dans les différents domaines d'analyse de données de cartes à puce. Il apparaît clairement que les études longitudinales des individus sur le long terme représentent, encore aujourd'hui, un domaine à approfondir. Dans le but de se concentrer sur ce type d'analyse, on a défini un ensemble de critères importants pour le choix de méthodes de segmentation à appliquer. Après avoir comparé les algorithmes les plus courants, il apparaît qu'une utilisation de la méthode des k-moyennes convient le mieux à notre étude traditionnelle. En effet, il s'agit d'un processus alliant simplicité, scalabilité, flexibilité et interprétabilité, mais qui nécessite l'introduction du nombre de groupe par l'utilisateur de la méthode. De plus, de nombreux défauts liés au bruit et à l'impossibilité de traitement incrémental nous poussent à chercher à améliorer la méthode. L'introduction de l'idée d'un comportement évolutif des usagers s'allie parfaitement à l'apparition continuelle de nouvelles données. Ainsi on cherchera, dans une méthode expérimentale, à compenser les limites traditionnelles dans le but d'effectuer une analyse de l'évolution des comportements. Prévoir la demande constitue finalement une dernière composante indispensable dans le travail du planificateur. Des prévisions simples et de bonnes qualités notamment à long terme sont à privilégier, réduisant donc notre choix aux processus de lissages exponentiels.

CHAPITRE 3 MÉTHODOLOGIE

On vient présenter ici une méthodologie complète sur l'étude comportementale des usagers des transports en commun. Cette méthode d'étude est applicable à tout système de transport et sur n'importe ville tant que les données nécessaires au bon fonctionnement sont fournies. Ces données en question doivent provenir de cartes à puce pour l'utilisation des transports en commun. Cette méthode est appliquée dans le chapitre suivant sur un jeu de données provenant de la STO.

3.1 Structure de l'étude

3.1.1 Méthode classique

Suivant la méthode du diagramme SADT, introduite par (Marca & McGowan, 1987), on explique le fonctionnement global de la méthode classique et de ses différentes étapes majeures.

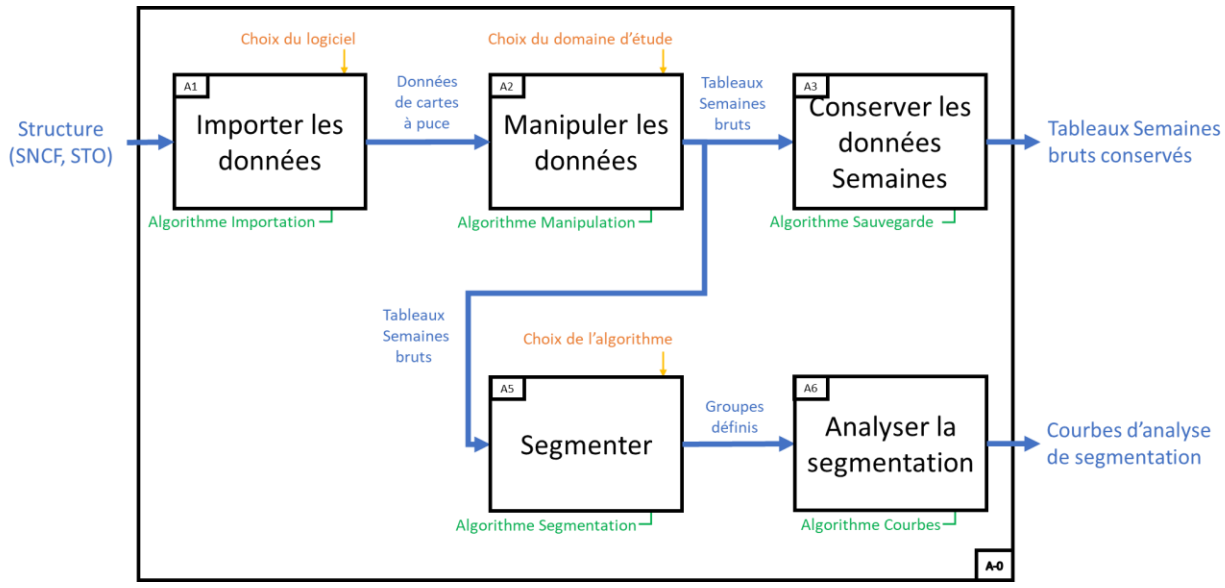


Figure 3-1 : Diagramme SADT fonctionnement méthode classique

Selon la Figure 3-1, l'étude par la méthode classique de segmentation prend en entrée les données provenant des différents demandeurs de l'étude que ce soit la STO ou encore la SNCF. Les deux premières étapes, l'Importation et la manipulation des données, mettent en forme les données d'entrée afin qu'elles deviennent exploitables par le module de Segmentation. On choisit également de conserver les données de sortie de l'étape de manipulation des données afin de ne pas avoir à

réitérer les deux premières étapes, jugées comme lentes au vu de la masse de données. Après avoir défini les différents groupes, l'étude finit par une étape d'analyse dans laquelle on veillera à interpréter les résultats et définir des indicateurs de qualité de la segmentation ainsi que des indicateurs de stabilité du comportement des cartes.

3.1.2 Méthode expérimentale

Toujours suivant la méthode SADT le fonctionnement global de la méthode expérimentale et de ses différentes étapes majeures est décrit.

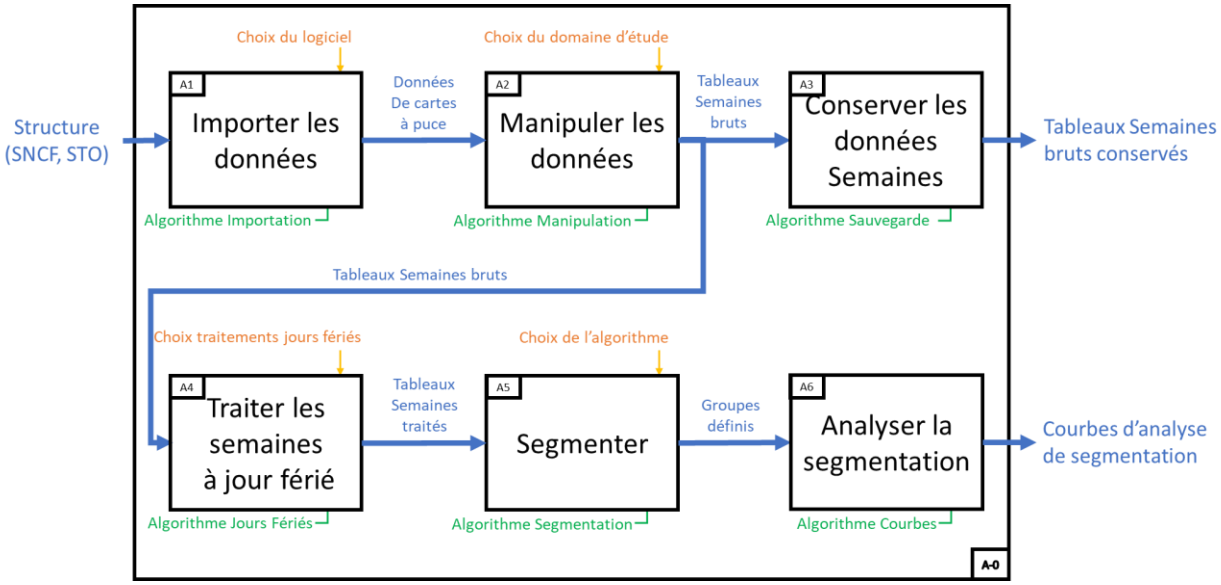


Figure 3-2 : Diagramme SADT fonctionnement méthode expérimentale

Selon la Figure 3-2, le schéma global ressemble fortement à celui de la méthode classique. On fait remarquer qu'un organe supplémentaire apparaît puisqu'il est indispensable de traiter les semaines à jour férié pour le bon fonctionnement de la méthode. On peut ajouter que l'étape d'analyse de la segmentation comporte un élément supplémentaire, la prévision, argument important de la méthode expérimentale.

3.2 Importation et manipulation des données

Les données d'entrée doivent être sous un format CSV, puisqu'il s'agit d'un format de données économe en mémoire. Le logiciel utilisé tout au long de l'étude est R, un logiciel gratuit et un langage de programmation destiné aux statistiques et à la science des données. Grâce à son impressionnante communauté de contributeurs, le logiciel profite d'un très grand nombre de paquetages permettant l'introduction de méthodes aussi bien statistique que de visualisation.

Il est nécessaire dans un premier temps d'importer les données brutes puis de les transformer en tableaux utilisables par l'algorithme de segmentation. Ici, le choix se porte sur la création de tableaux retraçant le nombre de déplacements par jour des identifiants-semaine.

Les données brutes proviennent généralement des transactions des usagers. En effet, à chaque passage de cartes à puce aux bornes, une ligne est créée dans la base de données. Cette ligne comporte un grand nombre d'informations comme : le numéro d'identifiant, le type de carte, le jour, l'heure de montée, le numéro de service et de ligne, la direction… La fraude sera supposée nulle et donc, aucune modification ne sera apportée à la base de données brute.

Pour le bon déroulement de la méthode, on décide de ne conserver que le numéro d'identifiant, le type de carte et le jour de passage de la carte. Les autres informations pourraient permettre de pousser l'étude descriptive ou même de partir sur d'autres axes de recherche sur le comportement des usagers. Par exemple, on pourrait s'interroger sur la fidélité de l'usager ou sur la régularité temporelle de ses déplacements (Trépanier et Morency, 2010).

3.2.1 Choix du traitement des données : le déplacement

Les données brutes fournies par l'entreprise sont sous la forme de transactions, or un individu peut effectuer plusieurs transactions sur le même déplacement, on dit qu'il effectue des correspondances. En ne payant qu'une seule fois, l'usager pourrait donc utiliser plusieurs bus ou métros sur un même trajet. De manière générale, en mobilité urbaine, on dit qu'un déplacement est l'action pour se rendre d'un point de départ à un point d'arrivée, par l'utilisation d'un ou plusieurs modes de transport (Orfeuil, 2001).

Pour comprendre le comportement de l'usager, il ne faudrait donc s'intéresser qu'à ses déplacements et non à ses transactions. Pour cela, il existe certaines règles, variables selon la ville

en question, qui définissent un déplacement. Par exemple, pour Montréal comme le définit le site de la STM, les différentes règles sont : 120 minutes pour compléter le trajet, un aller-retour compte comme deux déplacements, prendre deux fois la même ligne de bus compte comme deux déplacements et finalement une réutilisation du métro après sortie de station compte comme un second déplacement.

D'un point de vue méthodologique, pour transformer l'ensemble des transactions en déplacements, la méthode classique proposée consiste à développer un algorithme qui filtre les données selon les différentes règles de la ville en question afin de distinguer les déplacements des transactions. Cette méthode, plus fiable est cependant plus contraignante à mettre en place.

Certains auteurs ont traité le problème différemment. Au lieu de s'intéresser à la logique tarifaire mise en place par la ville, ils se sont concentrés sur une logique de mobilité, d'activités. En effet, Ma et al. (2013) estiment que 94% des voyageurs à Beijing réalisent leur déplacement en moins de 60 min et donc considèrent que deux validations séparées de moins de 60 min ne forment qu'un seul et même trajet. Tandis que Kieu et al. (2014) définissent un logigramme de génération de déplacements éliminant le bruit, récupérant les informations spatiales et séparant les déplacements d'une logique de 60 min également. Quant à eux, Seaborn, Attanucci, et Wilson (2009) définissent un temps de transfert maximal différent s'il s'agit d'une correspondance bus - bus, métro - bus ou métro - métro. Même si l'approche est différente, l'indicateur principal de la génération de déplacement, reste le temps.

3.2.2 Choix de la population étudiée : adultes réguliers

Dans l'optique de mieux faire ressortir les résultats de la segmentation, on décide de ne sélectionner que les usagers ayant pour type de carte : adultes réguliers. Ce choix fait office d'une première segmentation arbitraire de la population. Il est justifié par le fait que ce type de carte représente une très grande partie de la population et qui est, au niveau du comportement, beaucoup plus stable qu'une carte de type 10 trajets. De plus, ce choix permet d'éliminer de la population un grand nombre de déplacements qui pourraient être considérés comme du bruit.

3.2.3 Création des identifiants-semaines

Le but est de transformer cette masse de données en des tableaux ciblés sur la carte afin de mieux comprendre son comportement. On vient pour cela compter le nombre de déplacements effectués par jour puis trier par semaine, et ce, pour chaque identifiant à l'aide d'un tableau de contingence. Chaque tableau est associé à un identifiant, Figure 3-3. Le nombre de lignes est égal au nombre de semaines étudiées et chaque ligne correspond à un vecteur de 7 composantes témoignant du nombre de déplacements effectués chaque jour. On l'appelle par la suite le vecteur comportement, puisqu'il témoigne du comportement d'une carte sur une semaine.

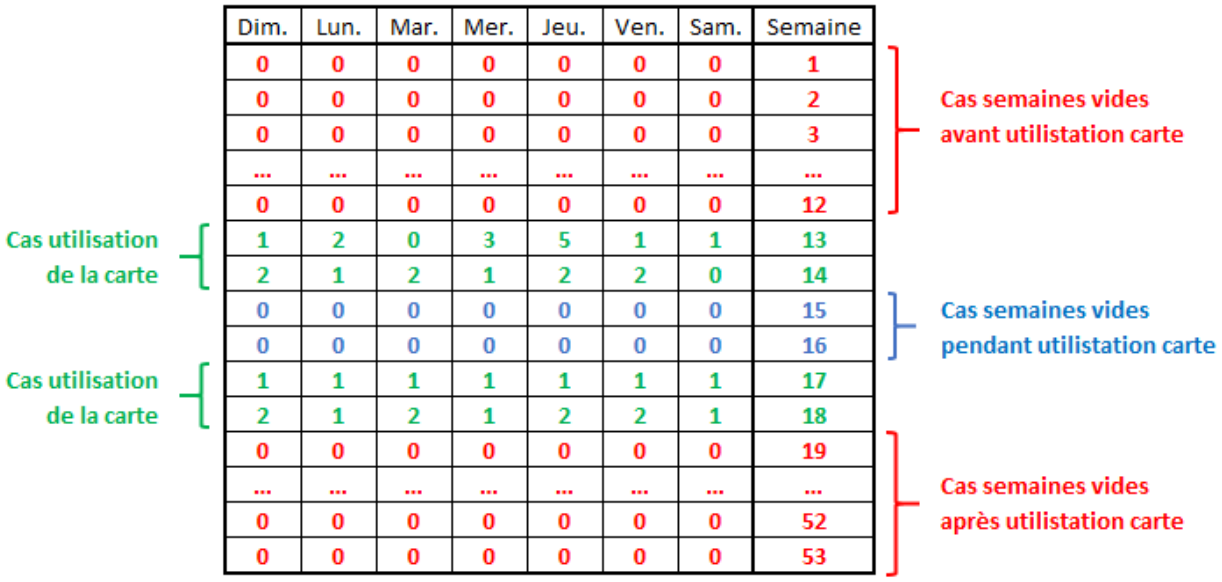


Figure 3-3 : Exemple de tableau identifiant-semaine lié à l'usage d'une carte à puce sur un an

À partir de ce moment, un filtre est appliqué pour supprimer les comportements aberrants. En effet, si sur un jour, un utilisateur a effectué plus de 10 déplacements, le vecteur comportement associé est supprimé et apparaît comme une semaine vide.

Il s'agit ici de générer volontairement des vecteurs comportements simples, afin de comprendre et montrer le potentiel des deux mécanismes d'études comportementale en jeu. La génération de ces vecteurs peut être sujette à divers changements suivant le niveau de détail attendu en analyse de résultats.

3.2.4 Création des tableaux semaines

Le but est de créer les listes de vecteurs utilisables par l'algorithme de segmentation. Pour cela, il suffit de regrouper l'ensemble des vecteurs comportement et de les trier par numéro de semaine.

Ainsi sur la Figure 3-3, il s'agit de récupérer uniquement les vecteurs en vert et de les regrouper en tableaux par semaine comme sur la Figure 3-4. Il existe un tableau de cette sorte pour chacune des semaines de l'étude, comportant les vecteurs comportements associés. Ainsi, chaque identifiant ayant effectué au moins un déplacement dans une semaine s, voit apparaître son vecteur comportement dans le tableau de la semaine s. On l'appelle dans l'algorithme le tableau « semaine ».

	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	id	week	total_semaine
41	0	2	0	2	2	1	0	4206204915	4	7
42	0	0	2	3	1	2	0	4206251506	4	8
43	0	2	0	3	3	4	0	4206251762	4	12
44	0	2	2	2	2	2	0	4206271475	4	10
45	2	1	1	2	2	1	0	4206272737	4	9
46	0	3	3	2	4	3	0	4206317810	4	15
47	0	2	2	2	2	2	0	4206337761	4	10
48	0	2	2	2	2	2	0	4206342375	4	10
49	0	2	2	2	2	4	0	4206373221	4	12
410	0	2	2	2	2	2	0	4206382834	4	10
411	0	0	2	2	2	0	0	4206383602	4	6

Figure 3-4 : Exemple d'un tableau semaine pour la semaine 4

La colonne « total_semaine » ci-dessus est un outil de statistique descriptive très pratique permettant, d'une part, de témoigner du nombre de déplacements par semaine les diverses cartes et d'autre part, de conserver uniquement les semaines non vides, indispensables au reste de l'étude.

3.3 Segmentation, méthode classique

La segmentation est l'étape principale de l'étude. Elle permet de regrouper les comportements similaires des usagers. Pour cela, on définit une méthode dite classique, qui reprend les algorithmes les plus courants pour effectuer la segmentation.

L'algorithme de segmentation le plus courant dans la littérature est, sans aucun doute, l'algorithme des K-Moyennes qui a la particularité d'être d'une très faible complexité et peut donc être utilisé pour les données de masses. Algorithme de segmentation par méthode de distance, la méthode des K-Moyennes jouit pourtant de plusieurs défauts vus précédemment dans la revue de littérature. Il est extrêmement dépendant de son initialisation et nécessite de lui fournir un nombre de groupes fixes pour fonctionner.

La méthode classique se décompose donc en deux parties, le choix du nombre de groupes à obtenir à la fin du processus et le fonctionnement de l'algorithme des K-Moyennes.

Pour la suite de l'étude, le nombre choisi de groupes est nommé K.

3.3.1 Choix du nombre de groupes, la méthode du dendrogramme

On a vu, dans la littérature, que la méthode des silhouettes et la méthode du Gap raisonnent à partir de matrices de distances entre de nombreux points et donc, ne sont pas réalisables lorsqu'on utilise des bases de données conséquentes. C'est pourquoi il existe une autre méthode notamment utilisée par Morency, Trepanier, Frappier, et Bourdeau, (2017) qui permet de travailler sur les données massives.

Il s'agit d'une méthode entièrement visuelle et arbitraire visant à effectuer une segmentation hiérarchique sur un échantillon représentatif des données. On vient tout d'abord réaliser une segmentation des données par la méthode des K-Moyennes, pour un K grand (généralement $K > 30$) : les centres des groupes obtenus font office d'échantillons représentatifs des données. Il s'agit finalement d'effectuer une HAC (Hierarchical Agglomerative Clustering) sur ces centres et de choisir K en fonction de la granularité désirée. La méthode fournit généralement un résultat satisfaisant et donne un aperçu global des données.

Ward Jr. (1963) suggère une méthode générale pour l'utilisation de la HAC, de complexité $o(n^3)$, de mémoire nécessaire $o(n^2)$. Cette méthode fournit une décomposition hiérarchique groupale de l'ensemble des données.

La HAC, autrement appelée classification hiérarchique ascendante en français, suppose comme initialisation que chaque individu de la base donnée forme son propre groupe. Il vient ensuite un calcul de la matrice de proximité entre chaque paire de points suivant une métrique et un critère de dissimilitude préalablement définis. On propose généralement la distance de Manhattan ou la distance euclidienne comme métrique pour le calcul de la matrice de proximité. Il existe de nombreux critères de dissimilitude, on ne s'intéresse qu'à quatre d'entre eux : « Average » qui définit la proximité comme la distance entre les centres de deux groupes, « Single » comme la plus courte distance qui existe entre les points de deux groupes, « Complete » comme la plus longue, et finalement « Ward » comme égale à la variation d'inertie intragroupe due à un regroupement.

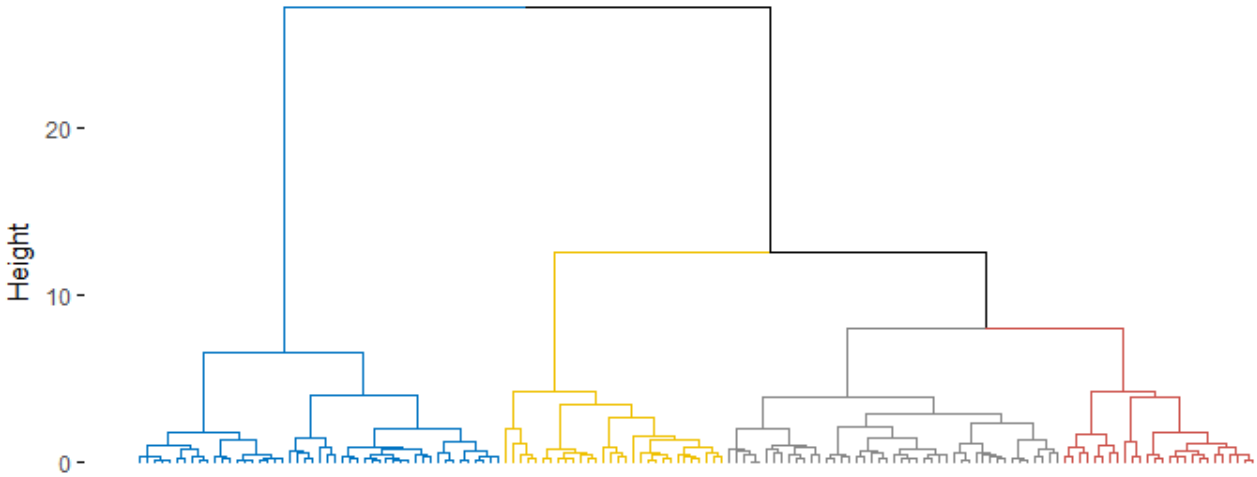


Figure 3-5 : Choix du nombre de groupes à partir d'un Dendrogramme post HAC

L'étape de fusion est une boucle consistant à regrouper les deux groupes les plus proches afin d'obtenir un nouveau groupe, puis de mettre à jour la matrice de proximité. Cette boucle ne s'arrête que lorsqu'il ne reste qu'un seul groupe. Petit à petit, la méthode construit un dendrogramme, ayant pour ordonnée la dissimilitude (distance) entre les groupes à chaque regroupement.

Le principal inconvénient de la HAC réside dans le fait qu'une fois qu'une fusion a été effectuée, on ne puisse pas revenir en arrière pour corriger les erreurs de groupement. En contrepartie, il n'est pas nécessaire d'imputer un nombre de groupes fixe, car le dendrogramme nous permet de décider d'un K optimal.

Par souci de clarté visuelle, on a tracé en Figure 3-5 le dendrogramme correspondant à la segmentation hiérarchique issue des données « iris » fournies par R. On peut deviner que choisir $K = 4$ fournit une granularité acceptable. On retrouve ainsi des groupes de tailles similaires et un niveau de détail suffisant. On peut également juger les fusions sur la hauteur qui symbolise la similarité entre les groupes à fusionner. Quatre groupes sont donc mis en avant par couleur sur le dendrogramme : chaque regroupement final s'applique à niveau dissimilitude important, démontrant une hétérogénéité conséquente entre les groupes, tandis que les sous-regroupements, se font à des dissimilitudes bien plus faibles, synonyme d'une forte homogénéité intragroupe. Dans un souci de choisir un K optimal, on décide de simuler la HAC avec les 4 critères de dissimilitude différents et de ne conserver que celui dont la qualité de segmentation est la meilleure.

3.3.2 Fonctionnement de la méthode des K-Moyennes

La méthode des K-Moyennes est un algorithme de segmentation dont la forme la plus courante a été suggérée par Lloyd (1982). D'une complexité linéaire lorsque le nombre de groupes et la dimension sont fixés, il convient ainsi parfaitement à l'exploitation de données de masses.

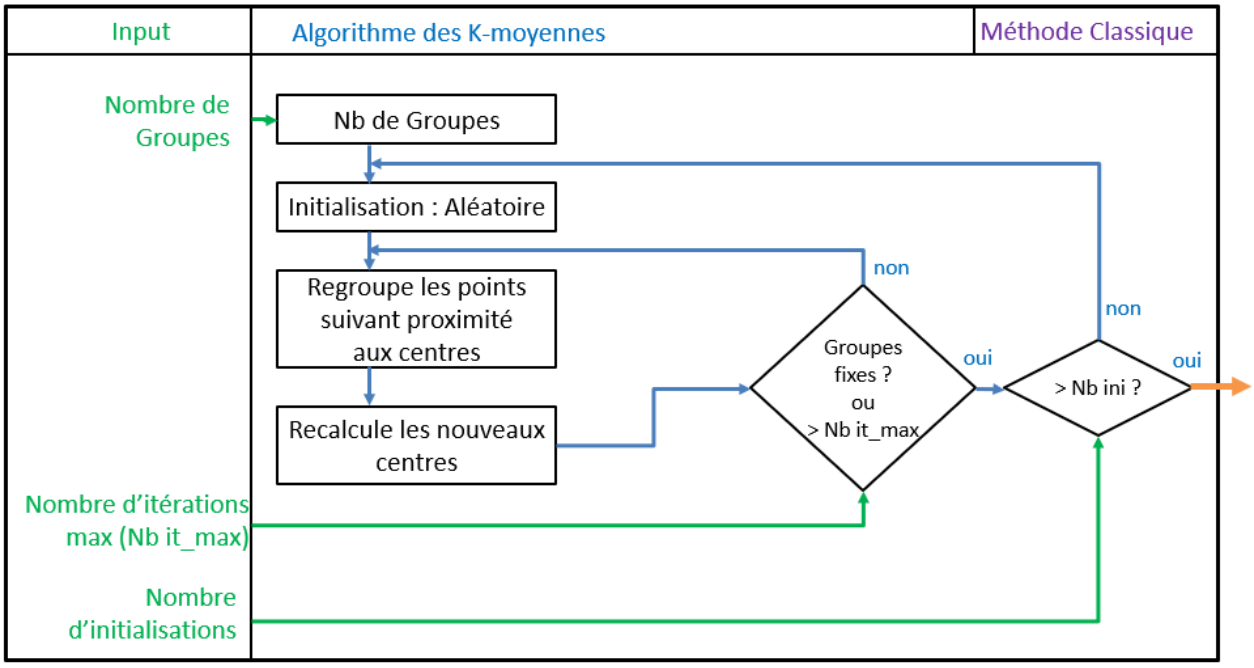


Figure 3-6 : Schéma de fonctionnement de l'algorithme des K-Moyennes

Pour comprendre l'algorithme, on définit le centre d'un groupe comme le barycentre des coordonnées des éléments de ce même groupe. Par cela, la méthode des K-Moyennes est une technique de segmentation qui utilise les centres pour représenter ces mêmes groupes. On cherche à maximiser la distance intergroupe et à minimiser la distance intragroupe pour chacun des groupes. La distance intra- groupe correspond à la somme des distances euclidiennes entre chaque point et le centre de ce même groupe tandis que la distance intergroupe correspond à la distance euclidienne entre les centres des groupes.

Selon le schéma de fonctionnement de l'algorithme des K-Moyennes en Figure 3-6, l'algorithme ne peut débuter sans l'information sur le nombre de groupes à simuler. L'initialisation est appliquée de manière aléatoire, c'est-à-dire que l'algorithme choisit K points dans la base de données et leur donne le rôle de noyau initial pour chaque groupe. S'en suit l'étape de fusion : l'algorithme vient regrouper chaque point avec le noyau le plus proche. On obtient donc K groupes de points, dont les barycentres deviennent les nouveaux noyaux. Puis l'on répète l'étape de fusion jusqu'à convergence, c'est-à-dire jusqu'à ce qu'il n'y ait plus de modifications dans les coordonnées des noyaux après fusion. Les noyaux finaux deviennent donc les centres des groupes en question et chaque point est associé à un groupe. Montrant ici un inconvénient de la méthode : tout bruit est associé à un groupe et donc modifie les coordonnées du centre associé.

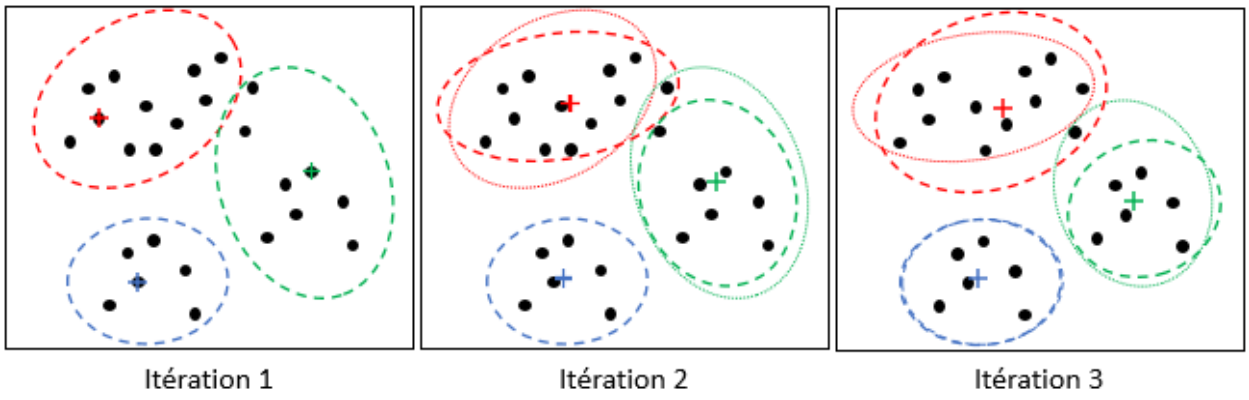


Figure 3-7 : Premières itérations de segmentation K-Moyennes sur un exemple de données

Par souci de clarté, on a représenté, en Figure 3-7, les premières itérations d'une segmentation par la méthode des K-Moyennes d'un exemple de données, pour $K = 3$. Lors de l'itération 1, on peut observer l'initialisation sur l'ensemble des données (croix de couleurs), ainsi que l'étape de fusion regroupant les points les plus proches aux noyaux aléatoirement générés. Sur l'itération 2, il y a

déplacement des noyaux puisque l'on recalcule les nouvelles coordonnées des centres de groupe. On peut observer qu'il y a transfert d'un point du groupe vert jusqu'au groupe rouge. Sur l'itération 3, on recalcule une nouvelle fois les nouvelles coordonnées des centres et l'on remarque à nouveau un transfert de point entre groupe vert et rouge. Finalement, on arrête la boucle lorsqu'il n'y a plus de variation entre les centres de l'itération i et $i + 1$.

Pour notre étude, nous utilisons les tableaux Semaines, présentés en 3.2.4. Il s'agit de la réunion de l'ensemble des vecteurs comportements triés par semaine. En rappelant qu'un vecteur comportement possède 7 composantes, nous travaillons dans un espace vectoriel euclidien de dimension 7, où chaque identifiant-semaine et centre de groupe est localisé par ses coordonnées correspondant à son utilisation TC suivant chacun des 7 sept jours de la semaine. Ces tableaux Semaines sont considérés comme des listes de vecteurs comportements et constituent les données d'entrée pour l'algorithme. En sortie, l'algorithme fournit ses résultats de segmentation, que sont : les coordonnées des centres des groupes, ainsi que l'information du groupe d'appartenance pour chaque vecteur comportement. À partir de ces informations, nous sommes capables d'établir des indicateurs quant à la qualité de la segmentation, ainsi qu'à l'analyse des résultats.

Plusieurs paramètres de réglage de l'algorithme doivent cependant être fixés, voir Figure 3-6. Le nombre de groupes, dont la détermination est définie dans la sous-partie précédente, est un paramètre indispensable pour le bon fonctionnement de l'algorithme; il fait fortement diminuer la qualité de la segmentation s'il est mal choisi. Lorsque la boucle ne converge pas rapidement, il est également nécessaire de fixer un nombre d'itérations maximum afin de limiter le temps de calcul de l'algorithme. Plus sa valeur est importante, meilleure est la qualité de segmentation, mais en contrepartie augmenter le nombre d'itérations signifie rallonger le temps de calcul. Finalement, le nombre d'initialisations définit le nombre de boucles à initialiser. Ici aussi, plus le nombre de boucles est grand, meilleure est la qualité de segmentation en dépit du temps de calcul.

L'algorithme des K-Moyennes converge la plupart du temps vers un optimum local, c'est pourquoi on dit souvent que son principal point faible est sa forte dépendance à son initialisation. En effet, si l'on simule deux initialisations aléatoires complètement différentes, nous ne sommes pas certains de trouver le même résultat final. Voilà pourquoi on décide dans cette démarche d'initialiser plusieurs fois et de ne conserver que la segmentation de meilleure qualité.

3.4 Segmentation, méthode expérimentale

À la lumière des différents défauts de la méthode classique, on propose ici une méthode expérimentale visant à être plus rapide et à produire une segmentation de meilleure qualité. Dans l'optique d'accélérer le processus sans perte de qualité, la plupart des travaux, voir 2.3, se concentrent sur l'optimisation de l'initialisation à chaque itération de l'algorithme des K-Moyennes. Il est vrai que diminuer le nombre d'itérations accélère fortement l'algorithme, mais ce n'est pas le seul facteur pouvant influer sur la complexité. La méthode suivante se compose de segmentations successives, de type K-Moyennes, sur des échantillons réduits et temporellement choisis. Réduire la taille des données d'entrées entraîne une diminution du temps de calcul de l'algorithme. Cette méthode s'adresse principalement à des données qui s'inscrivent dans le temps, plusieurs années de données sont donc recommandées.

3.4.1 Hypothèse de la méthode

Il est vrai que dans la méthode classique, la segmentation s'applique sur l'ensemble des données, et donc que l'on obtient des centres fixes sur toute la longueur temporelle de l'étude. Malheureusement, cela revient à comparer des comportements de semaines très différentes. En effet, d'un côté, les déplacements de la dernière semaine de novembre sont principalement liés à des activités de type travail ou école, qui correspondent surtout à des déplacements en semaine. La météo n'étant pas favorable à cette période de l'année, moins de déplacements sont à remarquer en fin de semaine (Zhou et al., 2017). D'un autre côté, les déplacements de la dernière semaine de juillet sont dépourvus d'activités de type école, limités en type travail, et la météo favorise une augmentation des déplacements en fin de semaine. On ne pourrait pas vraiment comparer ces deux périodes de l'année puisqu'elles sont fortement différentes. Et pourtant c'est sur quoi la méthode classique se concentre. Excellente sur courte période cette méthode est également très pratique pour comparer la même semaine sur plusieurs années en montrant les variations de la population d'un même groupe.

On a vu précédemment que l'algorithme des K-Moyennes faisait en sorte d'utiliser les centres pour représenter les groupes associés. Ainsi, la méthode classique définit un groupe par le comportement moyen des identifiants qui le constitue. Cependant, on a vu ci-dessus que le comportement était variable dans le temps, suivant divers facteurs qu'ils soient météorologiques, événementiels, ou

même temporels. Prenons donc comme hypothèse que les centres des groupes puissent se déplacer au cours du temps. Remarquer une variation saisonnière ou non dans le comportement d'un même groupe serait fortement intéressant pour les entreprises de planification de transports. En plus d'obtenir une information sur l'évolution de la population du groupe, l'entreprise pourrait en apprendre plus sur les variations comportementales de ces groupes.

3.4.2 Fonctionnement de la méthode des K-Moyennes séquentielles

Suivant le schéma à la Figure 3-8, la méthode raisonne sur le fait de simuler des segmentations K-Moyennes en forçant l'initialisation sur les semaines successives du domaine d'étude. Ainsi, au lieu de générer plusieurs fois des noyaux aléatoires pour l'ensemble des semaines, on décide de ne générer qu'une seule fois des noyaux préalablement choisis, à chaque traitement d'une nouvelle semaine. Le choix des noyaux se base sur le fait que, malgré l'hypothèse des centres de groupes variables, la variation des coordonnées des centres sur deux semaines successives est faible, et que, la semaine *s* de l'année *a* ressemblé à la semaine *s* de l'année *a-1*. Par cela, pour traiter la semaine s, on vient chercher l'optimum local autour de la position des noyaux générés à partir des résultats de la semaine *s-1* et de celles des années précédentes.

Pour fonctionner, l'algorithme passe par une étape d'apprentissage, une initialisation consistant à simuler un minimum d'un an de données suivant la méthode classique et à imputer les résultats (centres) en tant que noyau de la première semaine d'étude. Cette initialisation a pour avantage d'être plus rapide qu'une simulation sur l'ensemble du jeu de données et permet d'obtenir un suivi des groupes comparable entre les deux méthodes (classique et expérimentale).

Une limite à cette méthode se trouve dans l'échantillonnage. En effet le choix a été effectué de filtrer la base de données en fonction du numéro de semaine en cours. Les jours fériés et périodes de vacances forment des semaines où l'achalandage diminue fortement et où les comportements des usagers diffèrent significativement. C'est pourquoi une fonction d'apprentissage, définie en 3.4.3, s'occupe de générer les noyaux, tout en prenant en compte la présence ou non de congés.

En sortie, l'algorithme fournit ses résultats de segmentations pour chaque semaine de la plage d'étude, que sont les coordonnées des centres des groupes, ainsi que l'information du groupe d'appartenance pour chaque vecteur comportement.

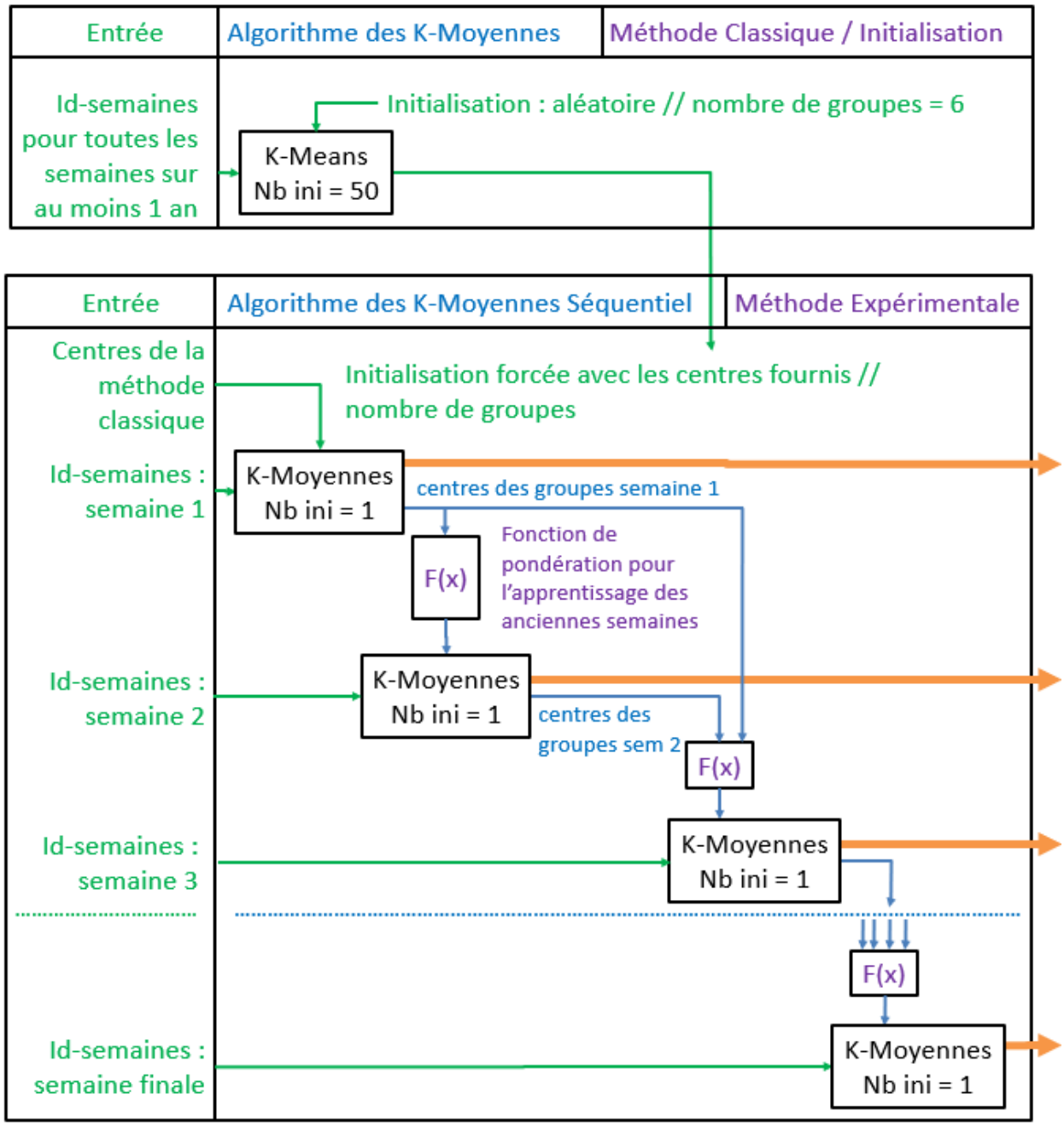


Figure 3-8 : Schéma de fonctionnement de la méthode expérimentale

3.4.3 Fonctions d'apprentissage

On pourrait définir une semaine à congé, comme une semaine qui comporte au moins un jour férié ou au moins un jour de vacances scolaires. Pendant ces semaines à congé, les utilisateurs ont un comportement complètement différent de leur comportement habituel. Ainsi, on peut parfois remarquer une baisse d'activité de plus de 80% sur le réseau. Il est à noter que pour simplifier l'étude, la dernière semaine de l'année *a-1* et la première semaine de l'année *a* contiennent généralement moins de sept jours et sont donc considérées comme des semaines à congés.

L'étude par segmentation séquentielle est suffisamment dépendante des semaines précédentes pour être biaisé par la proximité de semaine à congé. Il est donc d'importance capitale de traiter ces semaines à congé. En effet lors d'une segmentation séquentielle par la méthode des K-Moyennes, il s'agit normalement d'imputer les centres des groupes de la semaine *s-1*, comme noyaux pour la semaine *s*. Ainsi, si la semaine *s-1* fournie un résultat non cohérent avec les données de la semaine *s*, la segmentation de la semaine *s* est biaisée et fausse les segmentations des semaines à venir.

La méthode de traitement se résume en deux parties, dans un premier temps il faut déterminer les semaines à congé et dans un second temps générer des noyaux cohérents.

Pour déterminer les semaines à congé, une analyse des 53 semaines annuelles du calendrier est requise, puisqu'il est différent selon le pays dans lequel s'applique l'étude.

La génération des noyaux s'exécute selon les équations suivantes :

Équation 3.1 - Noyau en cas normal

$$Noyau_s = \frac{centre_{s-1} + \sum_{a=1}^{A} centre_{s-(a\times 53)}}{A+1} \text{ Où } A = \lfloor s/53 \rfloor$$

Équation 3.2 - Noyau si la semaine *s-i* est à congé où $i \in [1 ; 2]$

$$Noyau_s = \frac{centre_{s-i-1} + \sum_{a=1}^{A} centre_{s-(a\times 53)}}{A+1} \text{ Où } A = \lfloor s/53 \rfloor$$

Équation 3.3 - Noyau si la semaine *s* est la 1ère ou 2e semaine de l'année

$$Noyau_s = \frac{centre_{A\times 53-2} + \sum_{a=1}^{A} centre_{s-(a\times 53)}}{A+1} \text{ Où } A = \lfloor s/53 \rfloor$$

3.5 Analyse

Que la segmentation soit effectuée à partir de la méthode classique ou de la méthode expérimentale, les résultats sont de mêmes types et donc peuvent être analysés de manière similaire. On passe par trois étapes : une analyse des résultats pour comprendre les fondements généraux de la population étudiée, une analyse des indicateurs de qualité de segmentation afin de prouver la véracité des résultats et enfin une analyse des indicateurs de stabilité pour caractériser le comportement des usagers.

3.5.1 Analyse des résultats

3.5.1.1 Centres

Les centres des groupes déterminés correspondent aux déplacements moyens des individus du groupe sur chacun des jours de la semaine en question. Il s'agit d'effectuer une analyse descriptive afin de déterminer les différents paramètres caractérisant ces groupes pour se faire une idée de leur comportement.

Si l'on a traité la segmentation avec la méthode expérimentale, ces centres ont des coordonnées variables au cours du temps, il est également intéressant d'étudier ces variations pour mieux comprendre le comportement des usagers.

3.5.1.2 Population

La population des groupes est le second résultat que fournit directement la segmentation. Étudier son évolution au cours du temps nous en apprend plus sur les différents flux de changements de groupe et donc de changement de comportement. Cette analyse temporelle nous permettrait ainsi de relier une date à un changement de comportement majeur dans la population.

De plus, en mobilité, on dénote que l'utilisation des transports en commun est très souvent saisonnière, et l'étude de la décomposition de la courbe de population fait ressortir la tendance générale ainsi que la saisonnalité associée. Ces informations donnent un aperçu général du comportement de la population et de son évolution au cours du temps.

3.5.2 Indicateurs de qualité

On définit ici divers indicateurs permettant de juger la qualité de la segmentation effectuée. De manière générale, les indicateurs relatifs à ce sujet se concentrent sur la similarité à l'intérieur des groupes et la dissimilitude entre les groupes. Ces indicateurs sont calculés pour chacune des semaines du domaine et leur variation au cours du temps est analysée.

3.5.2.1 Indicateur Intergroupes

L'indicateur Inter propose de juger la dissimilitude entre les groupes. Par cela, on vient calculer la matrice des distances euclidiennes entre chacun des centres de groupes, selon l'Équation 3.4. Soit p et q deux centres de groupe, la distance euclidienne entre les deux correspond à la racine de la somme des carrés des différences sur chacune des composantes des centres. Dans notre cas, $n = 7$ puisqu'on a considéré des vecteurs comportement à 7 variables.

Équation 3.4 : Équation de la distance euclidienne entre p et q

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Tableau 3-1 : Exemple de matrice Inter pour une segmentation à 6 groupes

	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5	Groupe 6
Groupe 1	X	1.5	2	3	1,8	1,1
Groupe 2		X	4	2,5	1,4	2
Groupe 3			X	1.1	3	1,6
Groupe 4				X	1.2	1,8
Groupe 5					X	3
Groupe 6						X

Cette matrice nous donne une information importante quant à la proximité ou non des groupes entre eux. Une faible valeur de l'indicateur signifie une proximité significative, c'est pourquoi on les représente en rouge dans le Tableau 3-1. Une segmentation de qualité est une méthode qui permet d'obtenir une forte hétérogénéité entre les groupes (valeur d'indicateur haute); maximiser les

valeurs dans la matrice serait donc appréciable. Grâce à la méthode expérimentale, en calculant la matrice sur chacune des semaines de l'étude, on serait à même de repérer le rapprochement où l'éloignement de groupes.

3.5.2.2 Indicateurs intragroupes

Deux indicateurs visant à caractériser la qualité de segmentation à l'intérieur des groupes sont mis en lumière. Ces indicateurs sont calculés pour chacun des groupes, sur chacune des semaines à l'étude.

Dans un premier temps, l'indicateur Euc, défini par l'Équation 3.5, correspond à la moyenne des distances euclidiennes entre chaque point d'un groupe et son centre. Cet indicateur, strictement positif, proche de l'inertie intra- groupe (Lebart et al, 1982), propose une logique telle que plus la valeur est faible, plus le groupe est homogène et meilleure est la segmentation. On peut ainsi traiter de l'évolution de la qualité de chacun des groupes au cours du temps.

Équation 3.5 : Indicateur euclidien de qualité de la segmentation

$$Euc_{c,s} = \frac{1}{Nb(Id_s)} \sum d(Id_{i,s}, c)$$

Dans un second temps, l'indicateur Rel, défini par l'groupe. On est donc capable de caractériser cette erreur synonyme de mauvaise qualité de segmentation.

Équation 3.6, correspond à la moyenne des distances relatives entre chaque point d'un groupe et son centre. Cet indicateur propose une logique telle que plus la valeur est proche de 0, plus le centre est représentatif de la population en jeu. En effet, puisque la population à l'intérieur d'un même groupe n'est pas la même d'une semaine à l'autre, le centre, défini pour l'ensemble des identifiants semaine dans la méthode classique de segmentation, ne forme pas forcement le barycentre des points d'un même groupe. On est donc capable de caractériser cette erreur synonyme de mauvaise qualité de segmentation.

Équation 3.6 : Indicateur relatif de qualité de la segmentation

$$Rel_{c,s} = \frac{1}{Nb(Id_{i,s})} \sum (Id_{i,s} - c)$$

Afin de porter un regard critique sur la segmentation en général, on peut effectuer une moyenne pondérée par la population des groupes pour ces deux indicateurs.

3.5.2.3 Indicateur mixte

Au vu de la littérature, il existe plusieurs indicateurs permettant de rejoindre ces indices de distances intragroupe et intergroupe. Par exemple, le critère de Dunn, défini par l'Équation 3.7, (Dunn, 1974), caractérise le ratio des pires valeurs d'indices. En effet, on divise la plus petite distance intergroupe par la plus grande distance intragroupe. Ce critère permet ainsi de joindre les deux prérequis d'une bonne segmentation du point de vue de l'analyse interne. Les distances peuvent être calculées suivant différentes métriques, mais on choisit de ne s'intéresser qu'aux distances euclidiennes. Une valeur haute du critère est donc appréciable.

Équation 3.7 : Critère de Dunn

$$D = \frac{min_{1 \leq i \leq j \leq n}(d(i,j))}{max_{1 \leq k \leq n}(d'(k))}$$

On pourrait retrouver l'indicateur Silhouette défini en 2.3.4.2, puisqu'il traite également de la qualité de la segmentation, mais d'une manière légèrement différente. Fournissant plus de détails, il risque, ici aussi, de nécessiter trop de ressources de l'ordinateur pour pouvoir être estimé.

3.5.3 Indicateurs de stabilité comportementale

Il est intéressant de chercher une certaine stabilité dans le comportement des usagers afin de mieux le comprendre. Nous basons cette étude de stabilité comportementale sur la méthode de stabilité décrite par Leskovec, Rajaraman, et Ullman (2014).

Pour mettre en avant la stabilité dans le comportement d'un utilisateur, on s'intéresse à la distance que sa carte parcourt entre les groupes auxquels il appartient au cours du temps, comme le décrit l'équation ci-contre. Cette équation est valable pour chaque semaine s, pendant une durée N_i propre à chaque utilisateur i. Le $cluster$ ($ID_{i,s}$) représente donc le groupe dans lequel se situe l'identifiant de l'utilisateur i pendant la semaine s.

Équation 3.8 : Instabilité séquentielle pondérée

$$WSI_i = \left(\sum_{s=1}^{N_i} distance(cluster(ID_{i,s}), cluster(ID_{i,s+1}))\right)/N_i$$

Comme décrit précédemment, une segmentation est effectuée sur chacune des semaines de l'année. Ainsi chaque utilisateur est associé à un groupe, et ce, pour chacune des semaines de la période étudiée. Il s'agit de cibler les utilisateurs un à un pour comprendre leurs comportements. Un utilisateur avec un comportement stable est donc facile à prédire.

Pour cela, on vient tout d'abord récupérer la suite d'appartenance aux groupes au fil des semaines d'un même utilisateur. Puis, en l'associant à une matrice de distances intergroupes d'une semaine à l'autre, on est à même de calculer la distance totale entre les groupes que visite l'usager. Classiquement, on parle de distance intergroupe lorsqu'on qu'on calcule la distance entre les centres de deux groupes. Il est à noter qu'au vu de l'hypothèse de mouvement des groupes au cours du temps, cette matrice de distance intergroupe évolue également au cours du temps. Finalement, on vient diviser cette somme par le nombre de semaines où l'utilisateur s'est déplacé, permettant par cela d'obtenir un indice qui est comparable peu importe le nombre de semaines étudiées.

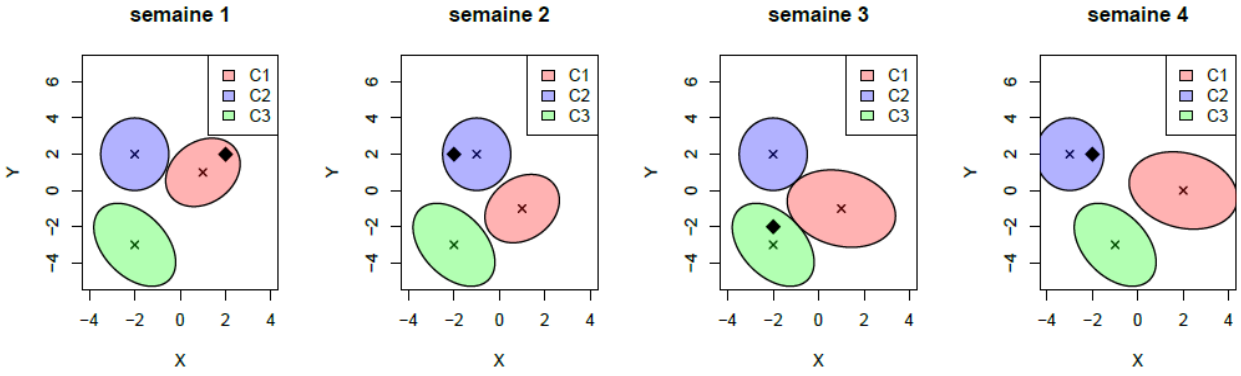


Figure 3-9 : Exemple de résultat de segmentation sur 4 semaines

Par exemple, imaginons dans un premier temps qu'un utilisateur (repéré par un symbole losange sur la Figure 3-9), se situe dans des groupes différents au fil des semaines. On y remarque que l'utilisateur appartient au groupe 1, la première semaine, puis le groupe 2 la 2^{e} et le groupe 3 la 3^{e} semaine. La 4^{e} semaine, il appartient de nouveau au groupe 2. L'indice d'instabilité associé à son comportement est donc calculé à partir de l'Équation 3.9. Dans ce cas, l'individu a changé de

nombreuses fois de groupes au fil des semaines : son indice d'instabilité est élevé et l'on peut conclure sur le fait que l'individu a un comportement instable.

Équation 3.9 : Instabilité séquentielle pondérée appliquée à l'exemple

$$WSI_1 = (dist(C1_{sem1}, C2_{sem2}) + dist(C2_{sem2}, C3_{sem3}) + dist(C3_{sem3}, C2_{sem4}))/3$$

Dans un souci de clarté explicative, l'exemple de la Figure 3-9 a été simplifié à un problème à deux dimensions. L'étude entière de segmentation se réalise, quant à elle, à partir de 7 variables.

Par cette somme, on comprend que si un utilisateur a un comportement stable, c'est-à-dire qu'il reste principalement dans le même groupe pendant la durée étudiée, alors l'indice d'instabilité est très faible. Par opposition, un utilisateur qui change beaucoup de groupes pendant la durée étudiée, et donc qui a un comportement fortement instable a une grande chance d'obtenir un indice de stabilité beaucoup plus conséquent.

3.6 Prévisions

Il s'agit de s'intéresser ici aux méthodes de prévision par lissage exponentiel mises en place dans un premier temps par Holt (1957, réimprimé 2004) puis améliorées par Winters (1960). Utilisées sur des séries chronologiques saisonnières, elles permettent généralement d'établir des prévisions mensuelles ou trimestrielles sur ces séries. Un des avantages de la méthode expérimentale était de pouvoir suivre l'évolution des centres des groupes à travers le temps. L'utilisation de ces méthodes pourrait donc permettre d'établir des prévisions sur cette évolution pour les différentes semaines à venir. Ces méthodes de prévision sont également appliquées à l'évolution de la population des groupes, et ce, tant pour la méthode classique que la méthode expérimentale.

L'utilisation des transports en commun suit généralement une saisonnalité propre à la ville étudiée. En effet, lorsque l'on étudie les comportements des usagers sur plusieurs années, on remarque des similitudes de déplacements d'année en d'année sur les mêmes périodes. Une décomposition des courbes de l'étude permet généralement d'identifier la saisonnalité de la tendance de la courbe. Une décomposition des courbes de population peut donc être calculée afin d'obtenir, dans un but strictement informatif, la tendance générale d'évolution de l'utilisation des transports en commun.

3.6.1 Méthode Holt-Winters saisonnière additive

Il s'agit d'une méthode que l'on privilégie en lissage exponentiel lorsque l'on travaille avec des séries temporelles, à décomposition additive, comportant une tendance et une saisonnalité. Soit une série d'observations $x_1, x_2, \dots, x_N$, on cherche à prédire les valeurs $x_N, x_{N+1}, \dots, x_{N+h}$ où h correspond à l'horizon de la prévision. Pour cela, la méthode fonctionne selon le principe suivant : on vient effectuer un lissage simultané de l'estimation de trois termes que sont : le niveau *L* de la série désaisonnalisée, la pente *b* de la tendance et *S* la saisonnalité.

Équation 3.10 : Équations d'estimation du niveau, de la pente et de la saisonnalité (additive)

$$L_t = \alpha(x_t - S_{t-s}) + (1-\alpha)(L_{t-1} + b_{t-1})$$
$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$
$$S_t = \gamma(x_t - L_t) + (1-\gamma)S_{t-s}$$

Équation 3.11 : Formule de prévision Holt-Winters saisonnière additive

$$\hat{x}(t,h) = L_t + hb_t + S_{t-s+h}$$

Équation 3.12 : Équation des moindres carrés de l'erreur

$$f(\alpha,\beta,\gamma) = \sum_{t=2}^{N} (\hat{x}(t,1) - x_{t+1})^2$$

Équation 3.13 : Formules d'initialisation pour le niveau, la pente et la saisonnalité

$$L_t = \frac{x_1 + \cdots + x_s}{s}$$
$$b_t = \frac{1}{s}\left(\frac{x_{1+s} - x_1}{s} + \cdots + \frac{x_{2s} - x_s}{s}\right)$$
$$S_t = x_t / L_t$$

Ces trois termes sont régis par des équations propres à la méthode (
Équation 3.10), et se recoupent suivant une formule de prévision à l'horizon *h*, selon l'Équation 3.11. Les paramètres α, β et γ sont compris entre 0 et 1 et sont choisis de sorte que *f(α, β, γ)* soit minimal (Équation 3.12), tandis que l'initialisation du niveau, de la pente et de la saisonnalité est appliquée sur 3 périodes *s* et se calcule selon Équation 3.13.

3.6.2 Méthode Holt-Winters saisonnière multiplicative

Il s'agit d'une méthode que l'on privilégie en lissage exponentiel lorsque l'on travaille avec des séries temporelles, à décomposition multiplicative, comportant une tendance et une saisonnalité. Elle fonctionne de manière analogue à la méthode additive. En effet, le niveau, la pente et la saisonnalité sont régis par des équations propres à la méthode (Équation 3.14), et se recoupent suivant la formule de prévision à l'horizon *h* (Équation 3.15). Les paramètres α, β et γ sont également compris entre 0 et 1 et sont choisis de sorte que *f(α, β, γ)* soit minimal (Équation 3.12), tandis que l'initialisation du niveau, de la pente et de la saisonnalité est appliquée sur 3 périodes *s* et se calcule selon Équation 3.13.

Équation 3.14 : Équations d'estimation du niveau, de la pente et de la saisonnalité (multiplicative)

$$\begin{cases} L_t = \alpha \dfrac{x_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1}) \\ b_t = \beta(L_t - L_{t-1}) + (1-\beta) b_{t-1} \\ S_t = \gamma \dfrac{x_t}{L_t} + (1-\gamma) S_{t-s} \end{cases}$$

Équation 3.15 : Formule de prévision Holt-Winters saisonnière multiplicative

$$\hat{x}(t,h) = (L_t + h b_t) S_{t-s+h}$$

L'avantage des méthodes de prévision par lissage exponentiel réside dans le fait qu'elles fournissent une prévision peu coûteuse et souvent de bonne qualité. Cependant, elles ne travaillent que sur l'historique de la série temporelle et ainsi, ne prennent pas en compte les diverses informations en parallèle. Par cela, les changements majeurs dans la configuration du réseau ne sont pas pris en compte et ainsi ne peuvent être prédits.

3.6.3 Calcul de l'erreur de prévision

Il existe dans la littérature de très nombreux indicateurs permettant de relever de la qualité de la prévision. La plupart d'entre eux se concentrent sur les erreurs, leurs moyennes et médianes. La littérature indique que l'indicateur le plus utilisé reste le MAPE (Pourcentage d'erreur absolue moyen) (O'Connell et Koehler, 2005), il définit le pourcentage de la moyenne des erreurs absolues, selon l'Équation 3.16.

Équation 3.16 : Indicateur MAPE

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{x_i}\right| \times 100$$

Cependant, Makridakis, Wheelwright, et Hyndman (1998) soulèvent plusieurs contraintes à cet indicateur, notamment lorsque les valeurs réelles sont nulles, le MAPE devient infini. Plusieurs déclinaisons à cet indice existent, traitant de la médiane ou même de l'écart type. Pour notre étude, on utilise l'indicateur MAPE pour l'évaluation des prévisions de population puisque les valeurs sont largement supérieures à 0.

Équation 3.17 : Indicateur RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$$

D'une autre manière, pour pouvoir évaluer les prévisions relatives aux nombres de déplacements journaliers des différents groupes, on s'attarde sur l'indicateur RMSE (Erreur moyenne quadratique), comprendre racine de la moyenne des erreurs au carré selon l'Équation 3.17. En effet, vu que le nombre de déplacements varie entre 0 et 10, on ne peut utiliser l'indice MAPE jugé beaucoup trop contraignant. L'indice RMSE trouve sa popularité dans son intérêt en modélisation statistique et sa sensibilité plus importante que le MAPE (Hyndman et Koehler, 2006).

CHAPITRE 4 EXPÉRIMENTATIONS ET RÉSULTATS (STO)

Ce chapitre expose les expérimentations et les résultats de l'étude des données provenant de la Société de Transport de l'Outaouais. Ce chapitre commence par une analyse descriptive du réseau et de la base de données étudiée. Puis, tout en suivant la méthodologie expliquée précédemment, il présente l'application de la méthode classique, de la méthode expérimentale et enfin une comparaison des deux méthodes et de leurs résultats.

4.1 Analyse descriptive

D'après les informations fournies dans le plan stratégique 2017-2026 publié en 2017, la Société de Transport de l'Outaouais est un service de transport en commun affilié à la ville de Gatineau. Il comprend les secteurs de Buckingham, Masson-Angers, Hull et Aylmer, représentant une superficie de 342 km² pour 278 589 résidents (Figure 4-1). Le réseau permet également la liaison avec le centre-ville d'Ottawa, la capitale du Canada. En 2017, la STO a mis en circulation 310 autobus sur 67 lignes assurant le transport de 70% d'adultes, de 27% d'étudiants et de 3% autres (enfants ou sénior).

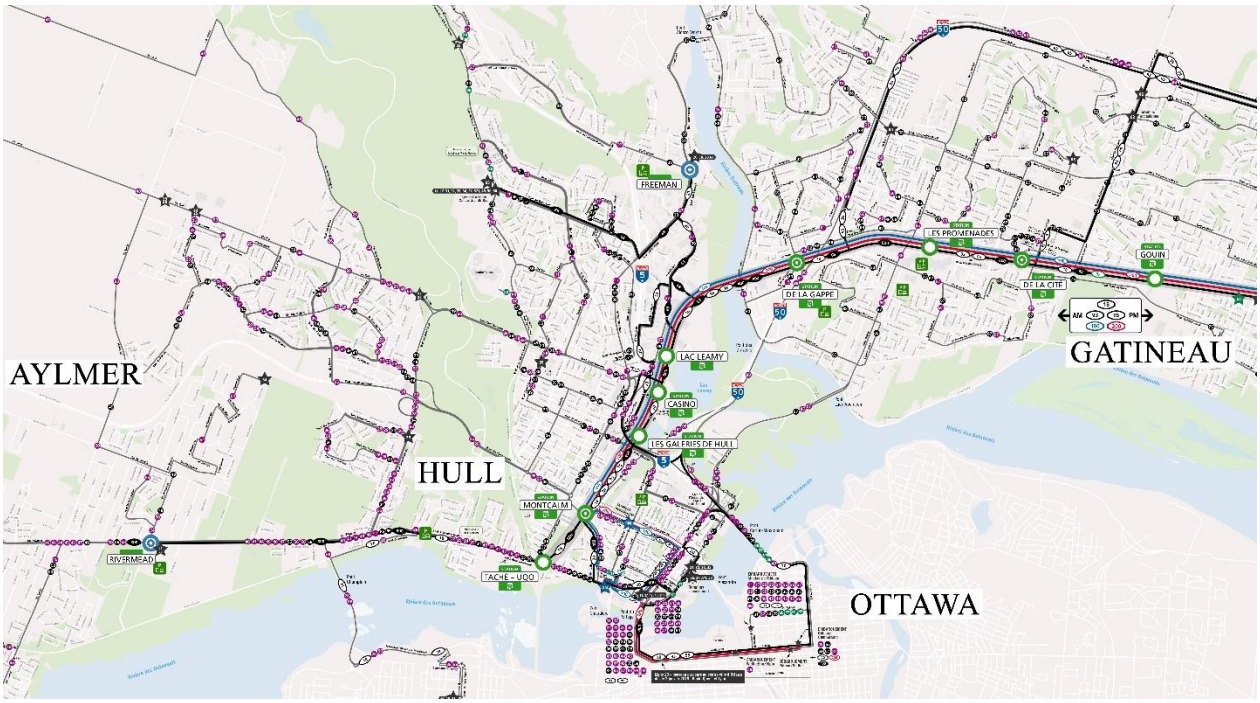


Figure 4-1 : Plan du réseau - tiré de STO

L'analyse descriptive des données s'applique en traçant l'achalandage des usagers en fonction du jour de la semaine. La Figure 4-2 représente donc l'évolution de l'utilisation des transports en commun de la STO, en transactions et en déplacements, tout au long de l'année 2012. Dans un premier temps, on remarque que, quelle que soit la saison étudiée, l'achalandage en fin de semaine suit une constante. En effet, les activités de fin de semaine constituées principalement d'activités de type « magasinage » ne sont pas régulières et n'attirent pas autant d'usagers que les activités de type « travail ». En opposition, les jours de semaine présentent, entre eux, des patrons similaires ainsi qu'une saisonnalité. On dénote un fort achalandage entre janvier (semaine 1) et mai (semaine 18), puis une baisse progressive d'activités jusqu'à fin août (semaine 35), suivi d'une nouvelle hausse entre septembre et fin décembre (semaine 53). Cette évolution suit parfaitement le calendrier de travail de la population active et étudiante. De plus, les creux d'utilisation en semaine correspondent à des jours fériés, se trouvant principalement les lundis et vendredis au Canada; l'utilisation du transport en commun ressemble donc fortement à des jours de fin de semaine.

Finalement, on vient récupérer les jours non travaillés et les périodes de relâche, afin de générer les semaines à congé décrites en 3.4.3 et nécessaires au fonctionnement de la méthode expérimentale.

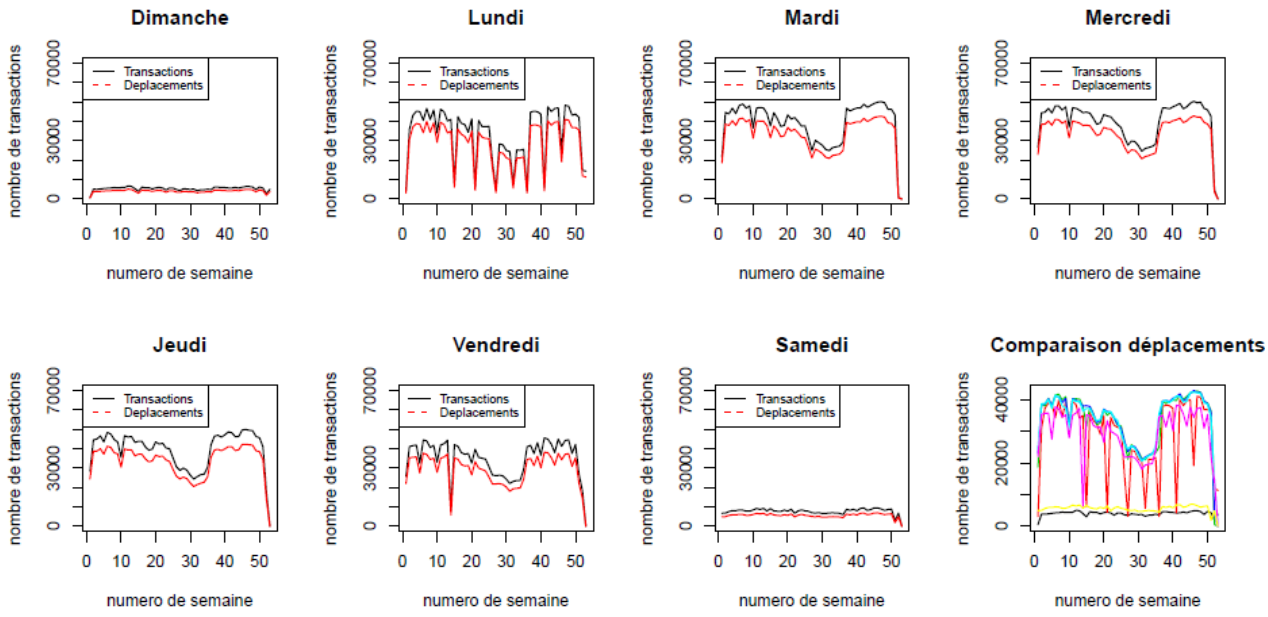


Figure 4-2 : Évolution du nombre de transactions et déplacements en fonction du jour – 2012

4.2 Application de la méthode classique

4.2.1 Importation et manipulation de données

La base de données étudiée constitue l'ensemble des données de transactions de bus effectuées par les adultes réguliers et enregistrées par la STO entre janvier 2012 et décembre 2014. Chaque transaction est caractérisée par 12 informations présentant entre autres : l'identifiant de la carte (anonymisé), le type de carte, le jour, l'heure de montée, le numéro de service ou de ligne, la direction, l'heure de départ de la ligne, comme présenté sur la Figure 4-3.

	id	type	jour	x	Htransaction	Nservice	Nligne	Direction	Hdepartligne	Njour	Narret	Nbus
1	10038611	3	2012-05-29 ...	1	60551	110	298	Sud	54154	2	4643	9704
2	10038611	3	2012-05-29 ...	1	153322	490	298	Nord	151913	2	5042	12
3	10038611	3	2012-05-30 ...	1	60550	110	298	Sud	54235	3	4643	9704
4	10038611	3	2012-05-30 ...	1	153010	490	298	Nord	152052	3	5042	10
5	10038611	3	2012-06-04 ...	1	55650	106	298	Sud	52741	1	4643	7
6	10038611	3	2012-06-04 ...	1	152733	490	298	Nord	150650	1	5069	10
7	10038611	3	2012-06-05 ...	1	60518	110	298	Sud	54140	2	4643	9702
8	10038611	3	2012-06-05 ...	1	153433	490	298	Nord	151802	2	5042	9704
9	10038611	3	2012-06-06 ...	1	152858	490	298	Nord	152133	3	5014	10
10	10038611	3	2012-06-07 ...	1	55656	106	298	Sud	53057	4	4569	8

Figure 4-3 : Échantillon de la base de données de la STO

Seules les quatre premières colonnes sont utiles au bon fonctionnement de l'algorithme : l'identifiant anonymisé de la carte, son type tarifaire, le jour de la transaction et enfin, s'il s'agit ou non d'une correspondance.

Dans le cas de la STO, l'automate génère automatiquement une information relative au fait qu'il s'agisse d'une première validation de carte ou d'une correspondance. En effet, la colonne x présente le résultat 1 s'il s'agit d'une première transaction et 2 s'il s'agit d'une correspondance. L'utilisation de cette colonne permet de respecter le principe tarifaire de la ville sans utiliser un algorithme de filtrage supplémentaire. Cependant, on fait entièrement confiance à la mesure de l'automate, qui peut être sujette à quelques erreurs.

Afin d'effectuer l'analyse comportementale, il s'agit de transformer la base de données brute en une base de données lisible par l'algorithme de segmentation. L'algorithme de manipulation de

données vient dans un premier temps générer des tableaux identifiant-semaines propres à chaque identifiant, comme présenté dans la méthodologie en 3.2.3. Il est question de récupérer uniquement les déplacements effectués par les types de cartes « adulte régulier » et « adulte régulier desfire ». En effet, à partir du 1er juillet 2013, la dénomination des cartes a changé dans la base de données, transformant ainsi les « adultes réguliers » en « adultes réguliers desfire ». Regrouper ces deux types tarifaires prend donc tout son sens. L'algorithme s'occupe finalement de calculer le nombre de déplacements chaque jour de la semaine pour chaque individu créant ainsi un tableau identifiant-semaines. Dans ce dernier, un filtre relatif au nombre de déplacements maximal par jour (fixé à 10) est appliqué permettant d'ignorer les comportements improbables.

Dans un second temps, l'algorithme de manipulation de données vient générer les tableaux semaines, véritables données d'entrée de l'algorithme de segmentation. Chaque ligne des tableaux identifiant-semaines comportant au minimum un déplacement, constitue un vecteur comportement propre au numéro de semaine et à l'identifiant en question. Les 159 tableaux semaines sont ainsi créés en regroupant les vecteurs comportements en fonction de leur numéro de semaine, méthodologie décrite en 3.2.4.

Tableau 4-1 : Résumé des tailles de bases de données

Année	2012	2013	2014
Nombre de Transactions	10 619 120	11 477 175	13 311 375
Nombre de Déplacements	8 906 390	9 276 198	9 988 357
Nombre de Déplacements Adulte Régulier + Desfire	3 039 420	2 973 399	3 586 654
Nombre de Vecteurs Comportements	390 777	394 330	472 332

Depuis la mise en place de Rapibus en octobre 2013, on remarque une fidélisation plus importante chez les usagers adultes réguliers, passant ainsi de 32% d'adultes réguliers en 2013 à 36% en 2014. L'algorithme de segmentation traite les vecteurs comportements qui comportent l'ensemble des informations nécessaires dans la base de données des déplacements. Il est ainsi plus rapide et moins onéreux en mémoire de travailler sur 1,26 million de tableaux comportements qu'avec 9,60 millions de tableaux déplacements.

4.2.2 Détermination du nombre de groupes optimal

La base de données comporte un très grand nombre de lignes. Il nous est impossible d'opérer avec les méthodes classiques pour connaître le nombre de groupes optimal à choisir pour nos segmentations K-Moyennes. On se tourne alors vers la méthode du dendrogramme.

La méthode consiste à simuler une segmentation K-Moyennes avec un grand nombre de groupes et d'ensuite appliquer une segmentation hiérarchique agglomérative (HAC) sur l'ensemble des centres de ces groupes. Chaque centre représentant une bonne approximation de la moyenne des occupants, la méthode permet par cela d'estimer une valeur pour K cohérente avec la base de données. Nous avons donc simulé des HAC sur un ensemble de 30 centres de groupes, suivant 4 méthodes HAC. Chacune de ces méthodes est paramétrée selon la méthode de calcul des distances entre les individus. Elles sont ici comparées suivant un critère de qualité. Sur le Tableau 4-2, il apparaît clairement que la méthode dite de « Ward » est de meilleure qualité que les autres sur la base de l'indicateur de qualité (l'indicateur de qualité fournit un résultat compris entre 0 et 1, où 1 est la meilleure qualité).

Tableau 4-2 : Comparaison de l'indicateur de qualité des différentes méthodes HAC

Average	Single	Complete	Ward
0.5892922	0.3098481	0.7213563	0.7957428

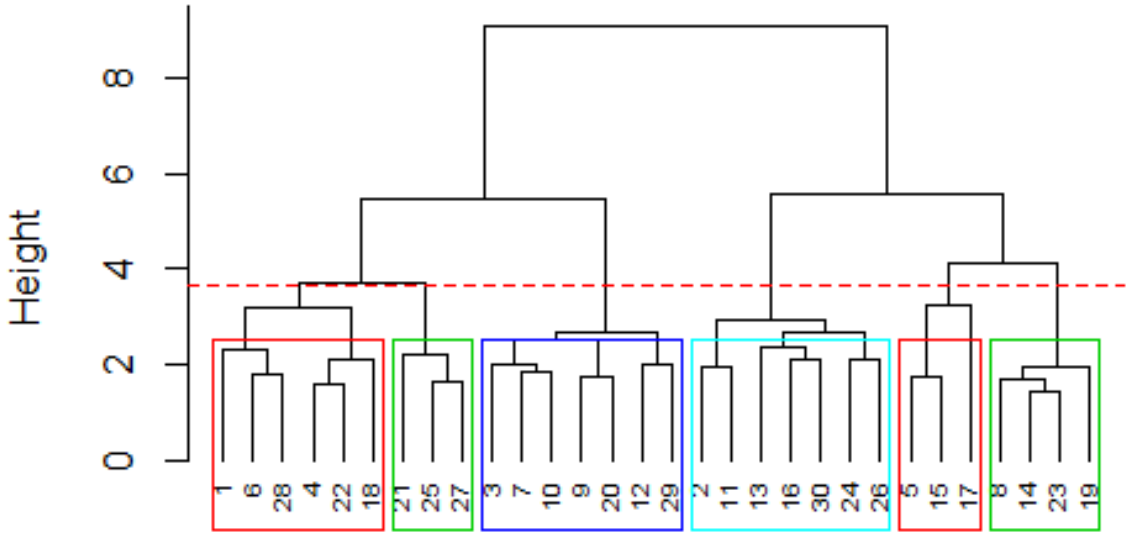


Figure 4-4 - Dendrogramme de l'HAC méthode Ward

Il s'agit finalement de tracer le dendrogramme associé à cette HAC par méthode de Ward (voir Figure 4-4). Dans ce cas, le choix du nombre de groupes est arbitraire et se base sur des critères de répartition de données et de dissimilitude entre les groupes. Les regroupements pour $K = 4$ et $K = 6$ présentent tous deux des avantages relatifs à ces derniers critères. Le choix final se porte sur 6 groupes pour garantir un résultat plus détaillé qu'avec 4 groupes.

4.2.3 Segmentation

Aux vues des sous-parties précédentes, divers paramètres ont été fixés permettant le fonctionnement de l'algorithme de segmentation. Elle s'applique donc sur l'ensemble des vecteurs comportements issus des déplacements effectués par les utilisateurs au long des 3 ans de données, avec $K = 6$.

Tableau 4-3 : Paramètres algorithme segmentation classique

Dénomination	Valeur	Paramètre dans R
Taille base de données	3 ans = 159 semaines	s = 159
Type d'information	Déplacement	x = 1
Nombre de groupes	6	K = 6
Nombre d'itération max	100	it_max = 100
Nombre d'initialisation	50	nb_ini = 50

D'autres paramètres, intrinsèques à l'algorithme de segmentation (Tableau 4-3), permettent de jouer sur la qualité et la rapidité d'exécution de la méthode. D'après la méthodologie en 363.3.2, il s'agit de définir une valeur pour le nombre d'itérations maximal et le nombre d'initialisations à effectuer. En effet, il est nécessaire de fixer un nombre d'itérations maximal pour le cas où la convergence vers des groupes fixes après itération serait trop lente à arriver. Le nombre d'initialisations quant à lui, sert à simuler la segmentation à partir de noyaux aléatoires différents. Théoriquement, plus les valeurs de ces paramètres sont importantes, meilleure est la qualité de segmentation en dépit d'un temps de calcul plus long. On privilégie plutôt la qualité que le temps de calcul, c'est pourquoi le nombre d'itérations maximal est fixé à 100 et le nombre d'initialisations à 50.

4.2.4 Analyse des résultats

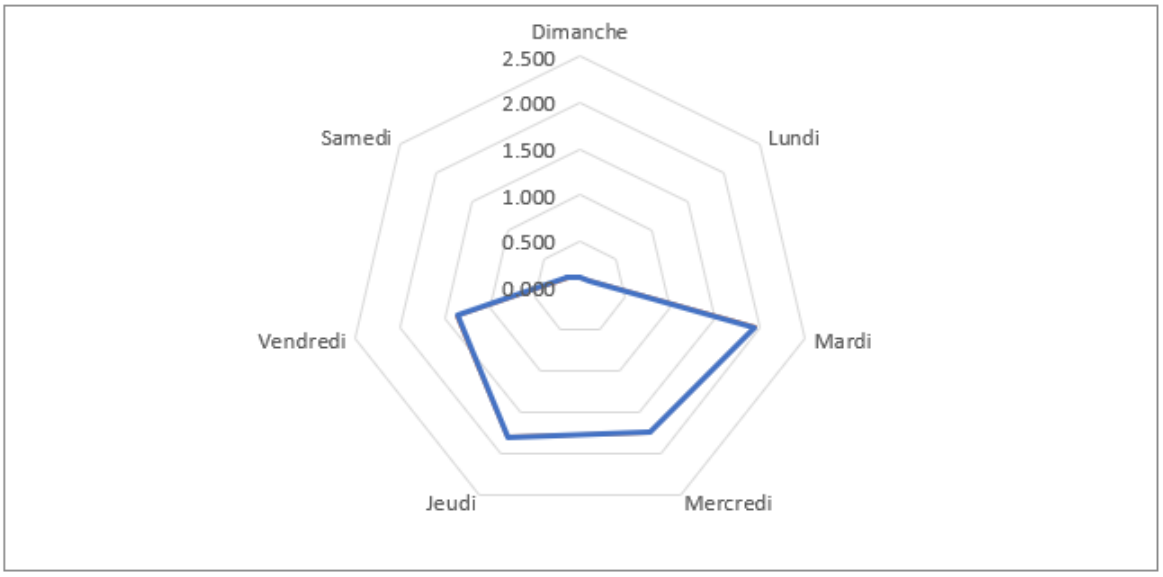
Cette partie propose une analyse des résultats de segmentations issus de la méthode classique, présentant les différents groupes ainsi que leurs caractéristiques intrinsèques.

Sur la Figure 4-5, le groupe 1 est essentiellement constitué d'usagers se déplaçant en moyenne deux fois les mardis, mercredis et jeudis, et d'une à deux fois le vendredi, mais très peu de déplacements sont à constater en fin de semaine. Concernant la population du groupe, le groupe conserve une taille constante autour de 1000 usagers (en moyenne : 15,0% de la population hebdomadaire étudiée) tout au long de la période d'étude malgré une saisonnalité identifiable par une légère baisse du nombre d'usagers pendant les périodes d'été. Ce groupe s'apparente à un groupe d'usagers dont le patron de déplacement ressemble à celui d'un travailleur à mi-temps où le lundi n'est pas travaillé. En effet, deux déplacements par jour désignent souvent un aller le matin et un retour le soir. Combiné à une population essentiellement constituée d'adultes, on peut deviner qu'il s'agit d'une activité de type travail. L'évolution de la population du groupe rejoint cette théorie, puisqu'en semaine à congé on remarque une forte hausse de la population du groupe. Ce groupe récupère certainement, la population de travailleurs à temps plein (groupe 2) lors des semaines où le lundi est férié.

Le groupe 2 (Figure 4-6) est principalement constitué d'usagers se déplaçant deux fois par jour en semaine et ayant très peu de mobilité en transport en commun en fin de semaine. Il s'agit du groupe le plus fourni puisqu'il varie entre 2000 et 4000 individus (en moyenne : 28,4% de la population hebdomadaire étudiée) en fonction de la période de l'année. L'évolution de la population suit effectivement une saisonnalité facilement identifiable, puisqu'on dénote une perte de plus de 50% des individus entre la deuxième semaine de janvier 2014 et la première semaine d'août 2014. Ces baissent d'activité apparaissent année après année aux périodes d'été. Ce groupe rassemble très certainement les travailleurs à temps plein aux vues du nombre d'usagers, de son évolution périodique et des caractéristiques du groupe. En effet, les travailleurs à temps plein effectuent généralement un aller-retour sur chacun des jours de la semaine, mais très peu de déplacement en fin de semaine. De plus, on remarque une quasi-disparition du groupe lors des semaines à congé; ces individus sont certainement récupérés par le groupe 1 vu précédemment, puisque la plupart des jours fériés sont des lundis au Canada.

a) Comportement moyen du groupe 1 (position du barycentre)

groupe 1	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	0.106	0.111	1.944	1.742	1.803	1.368	0.167



b) Evolution de la population dans le groupe 1

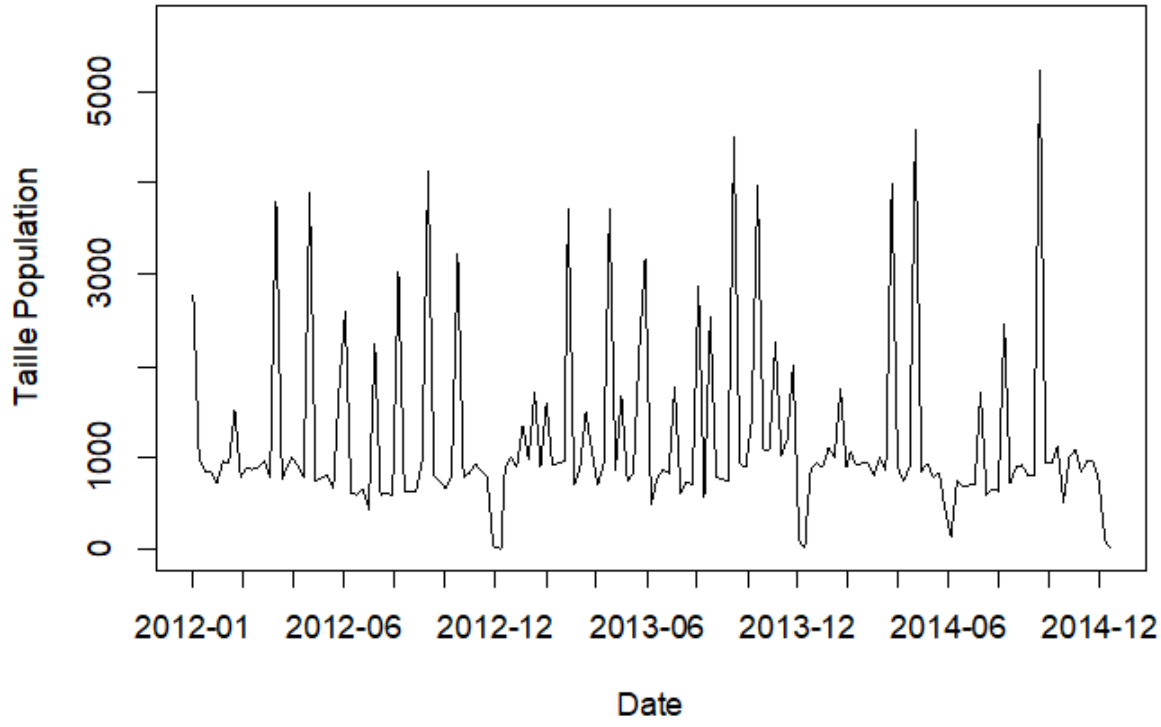
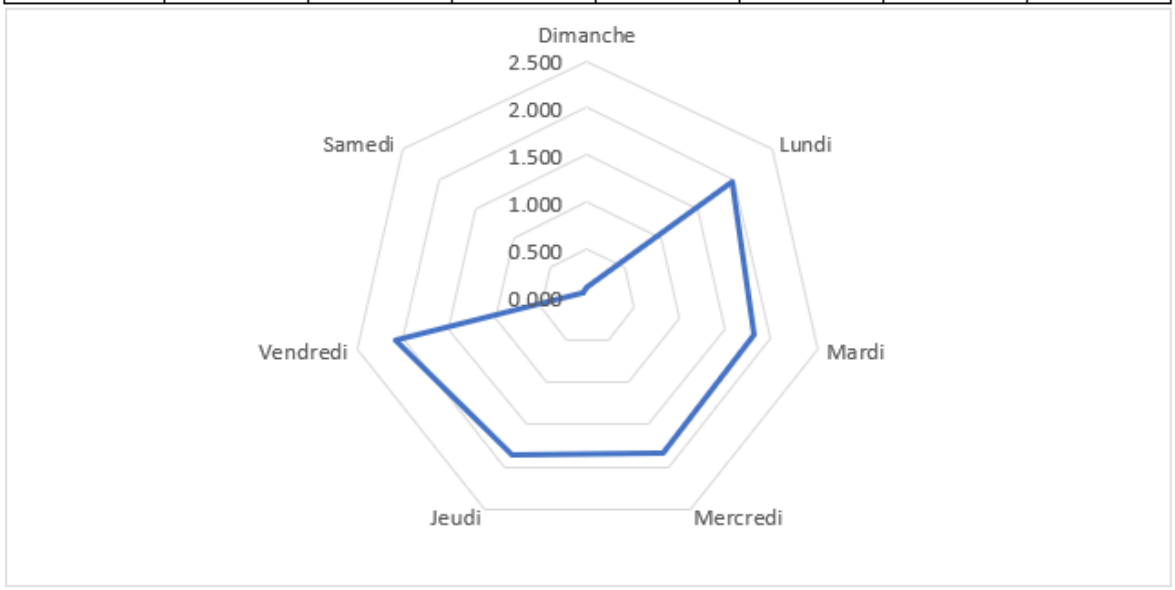


Figure 4-5: Résultats groupe 1 - méthode classique

a) Comportement moyen du groupe 2 (position du barycentre)

groupe 2	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	0.106	1.955	1.812	1.831	1.851	2.084	0.063



b) Evolution de la population dans le groupe 2

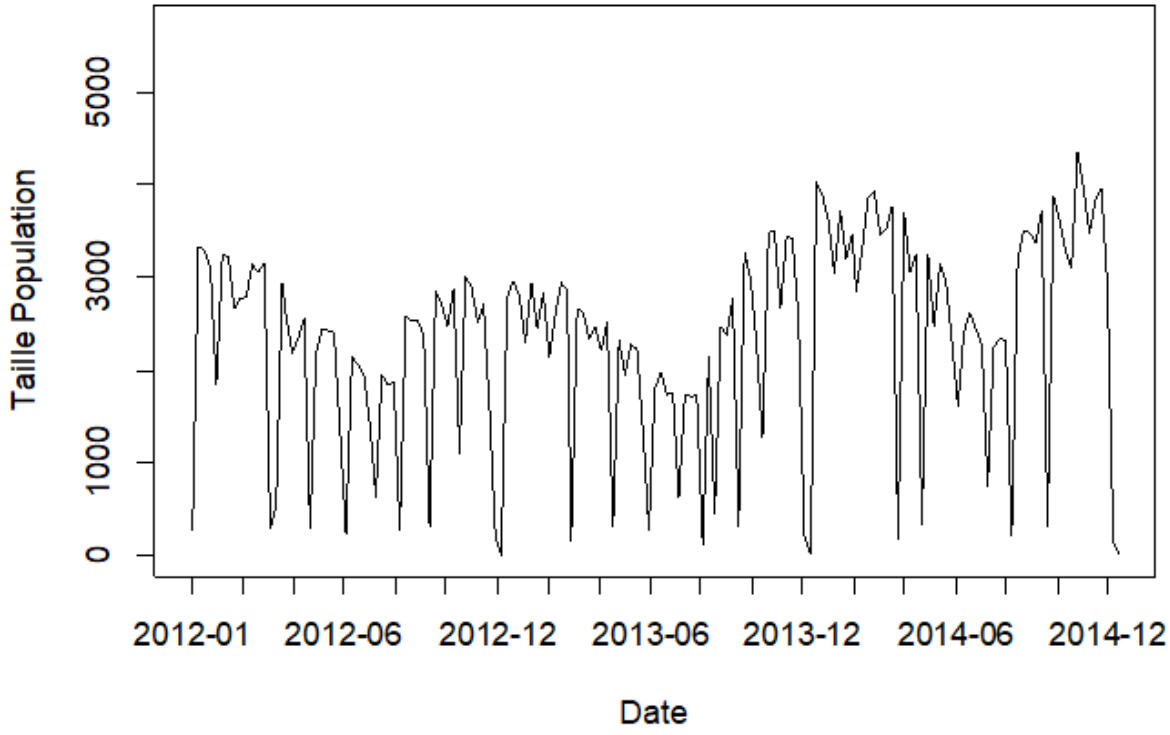
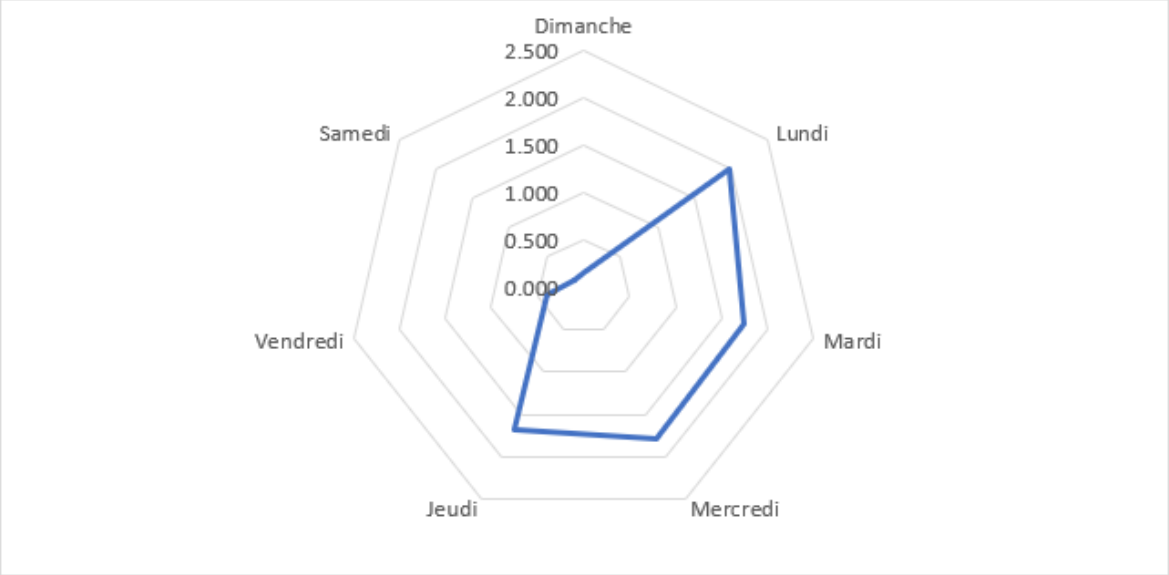


Figure 4-6: Résultat groupe 2 - méthode classique

a) Comportement moyen du groupe 3 (position du barycentre)

groupe 3	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	0.156	1.983	1.735	1.781	1.670	0.394	0.124



b) Evolution de la population dans le groupe 3

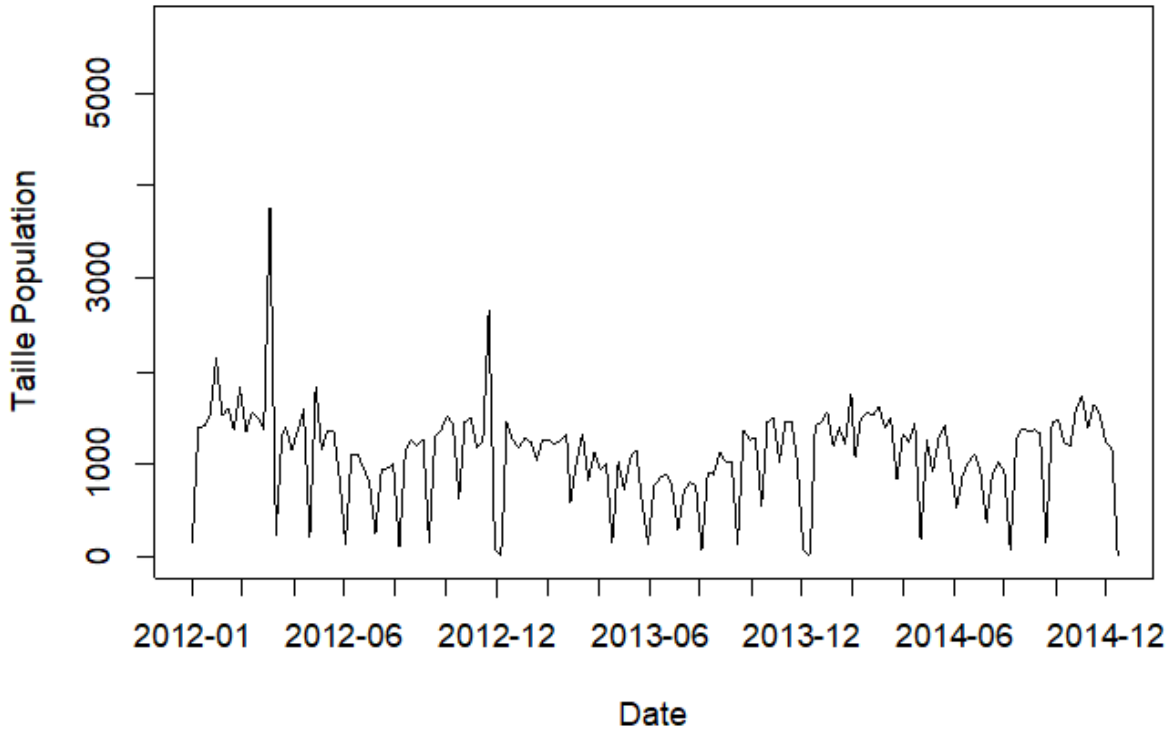


Figure 4-7: Résultat groupe 3 - méthode classique

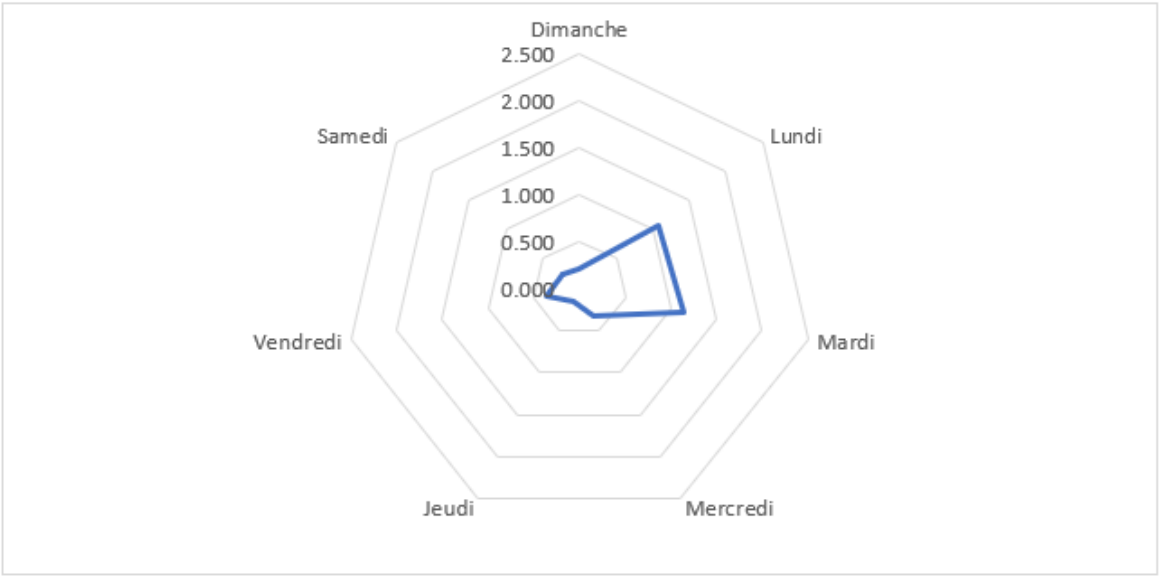
Caractéristiquement proche du groupe 2 (Figure 4-6), le groupe 3 (Figure 4-7) est constitué d'usagers qui effectuent deux déplacements sur chacun des jours de la semaine sauf le vendredi. Oscillant également autour de 1000 usagers (en moyenne : 13,5% de la population hebdomadaire étudiée), la saisonnalité est facilement identifiable puisqu'on remarque un écart type plus important. Montrant des périodes de variation similaires, le groupe 3 présente lui aussi une baisse périodique de sa population lors des périodes d'été. Aux vues des fluctuations relatives aux semaines à jours de congé, on peut supposer que le groupe 3 vient récupérer les usagers du groupe 2 dans les semaines où le vendredi est férié, mais vient perdre des usagers au profit du groupe 1, lors des semaines où le lundi est férié. Il y a, ici aussi, quasi disparition du groupe 3 pendant les semaines à lundi férié. À première vue, ce groupe représente les travailleurs à mi-temps effectuant des déplacements vers leurs lieux d'activité du lundi au jeudi compris.

Sur la Figure 4-8, le groupe 4 présente des caractéristiques sensiblement différentes. Ce groupe est constitué d'usagers réalisant peu de déplacements, et ce, uniquement le lundi et le mardi. Sa population reste constante autour de 1000 usagers (en moyenne : 15,0% de la population hebdomadaire étudiée), avec une très faible saisonnalité. La variance est, par contre, très importante. Semblable à un bruit très prononcé, la population peut augmenter ou diminuer de 500 usagers d'une semaine à l'autre. Ce groupe regroupe des usagers occasionnels qui ne nécessitent pas forcément d'utiliser les transports en commun pour aller travailler ou ne travaillant pas. Il est également à prendre en compte qu'un déplacement, suivant les conditions tarifaires d'une ville, peut signifier un aller-retour sur une période réduite, comme un aller-retour pour magasiner, accompagner ses enfants à l'école… On notera une forte hausse de la population pour les semaines de Noël et jours de l'an de chaque année, étant des semaines particulières et très peu achalandées il est logique que le groupe 4 récupère les usagers des autres groupes sur ces périodes.

D'apparence très proche du groupe 2 (Figure 4-6), les usagers du groupe 5 (Figure 4-9) se déplacent deux fois par jour sauf le dimanche où au minimum un déplacement est effectué. Cependant, l'étude de l'évolution de la population dans ce groupe présente une signification bien différente. En effet, on ne retrouve a priori aucune influence des semaines à congé ni de saisonnalité sur l'évolution. Cette population reste constante autour de 900 usagers (en moyenne : 11,1% de la population hebdomadaire étudiée) et disparaît presque lors des semaines de fin d'année. La non-influence des semaines à congé et la faible taille de cette population amènent à la réflexion que les usagers de ce groupe ne travaillent pas, mais effectuent tout de même des activités journalières.

a) Comportement moyen du groupe 4 (position du barycentre)

groupe 4	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	0.210	1.074	1.152	0.337	0.149	0.360	0.228



b) Evolution de la population dans le groupe 4

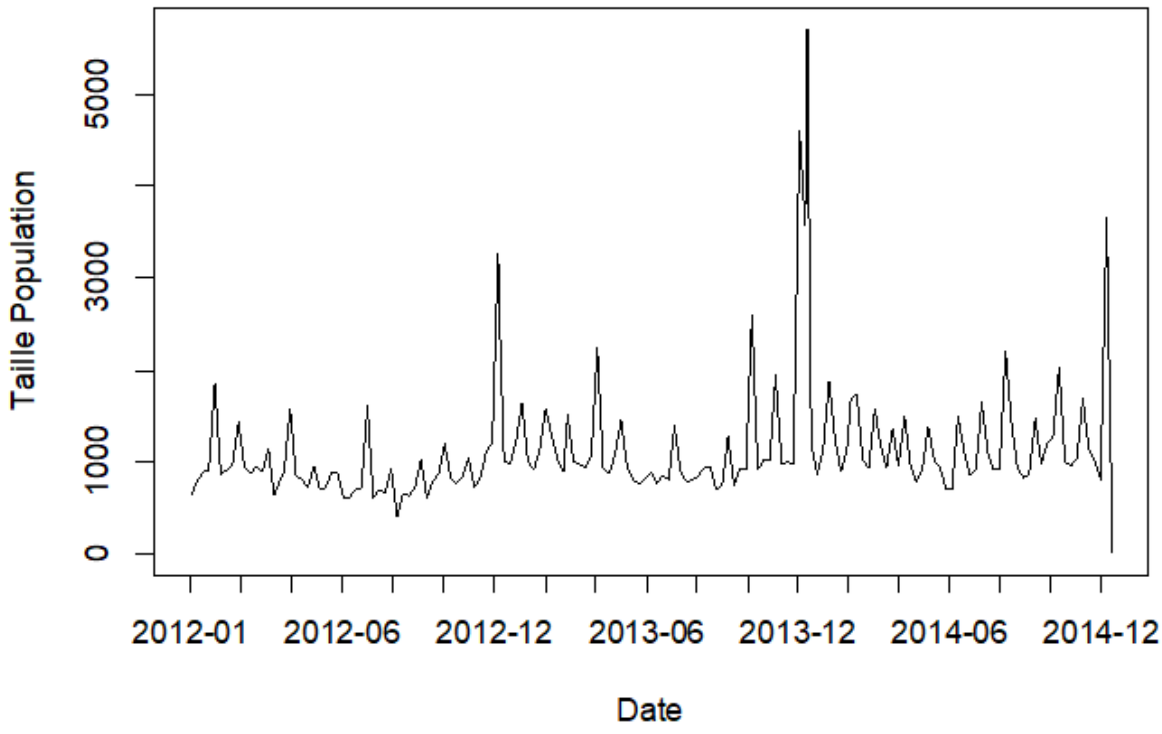
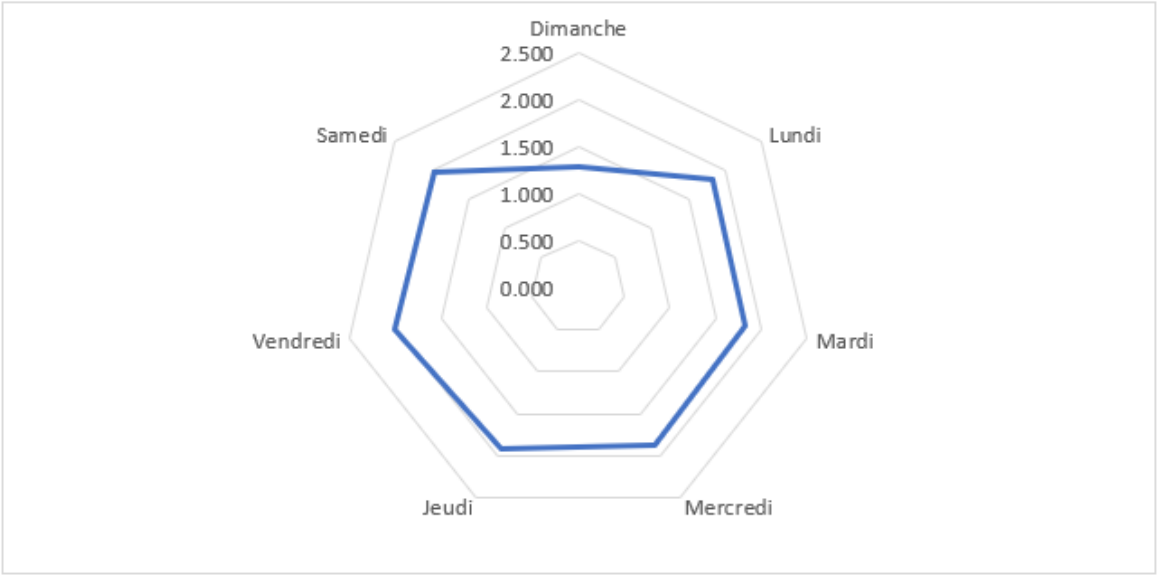


Figure 4-8: Résultat groupe 4 - méthode classique

a) Comportement moyen du groupe 5 (position du barycentre)

groupe 5	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	1.280	1.841	1.829	1.872	1.921	2.018	1.963



b) Evolution de la population dans le groupe 5

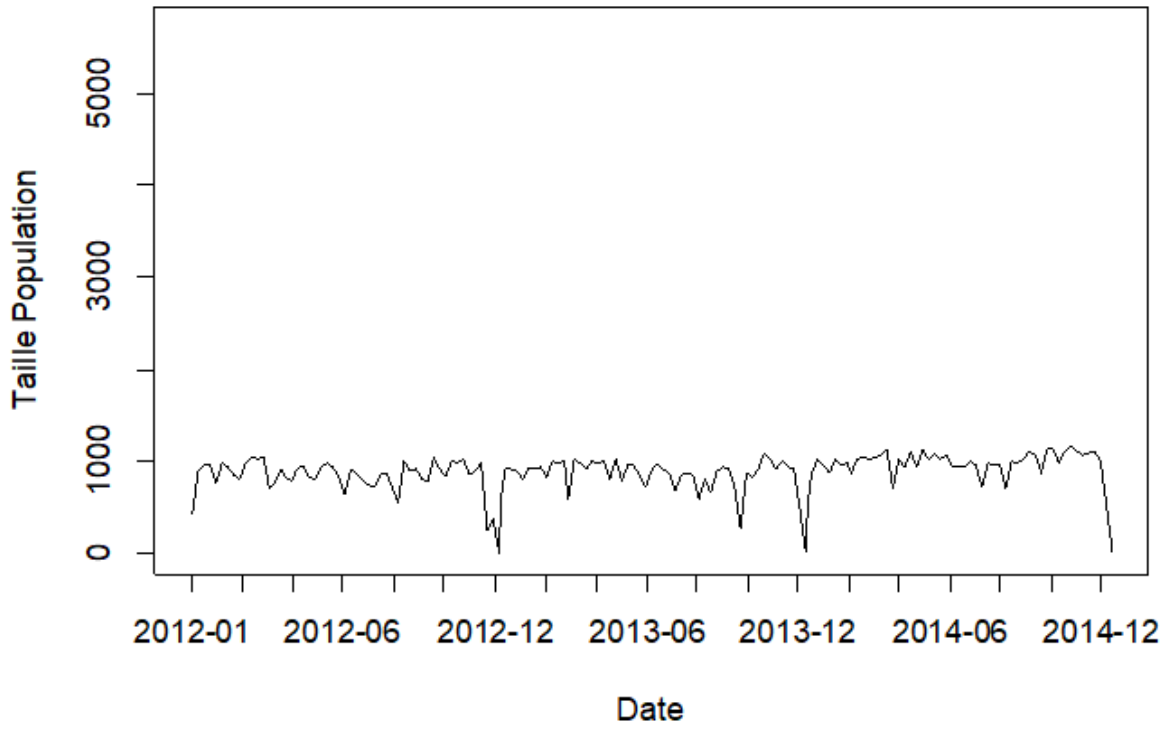
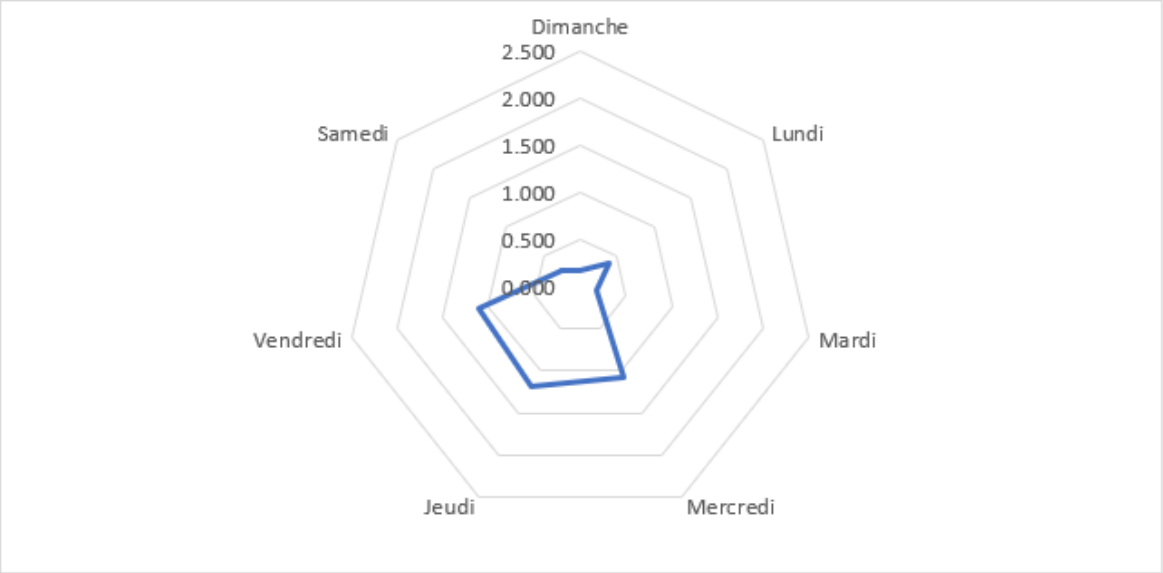


Figure 4-9: Résultat groupe 5 - méthode classique

a) Comportement moyen du groupe 6 (position du barycentre)

groupe 6	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	0.163	0.384	0.181	1.076	1.192	1.104	0.257



b) Evolution de la population dans le groupe 6

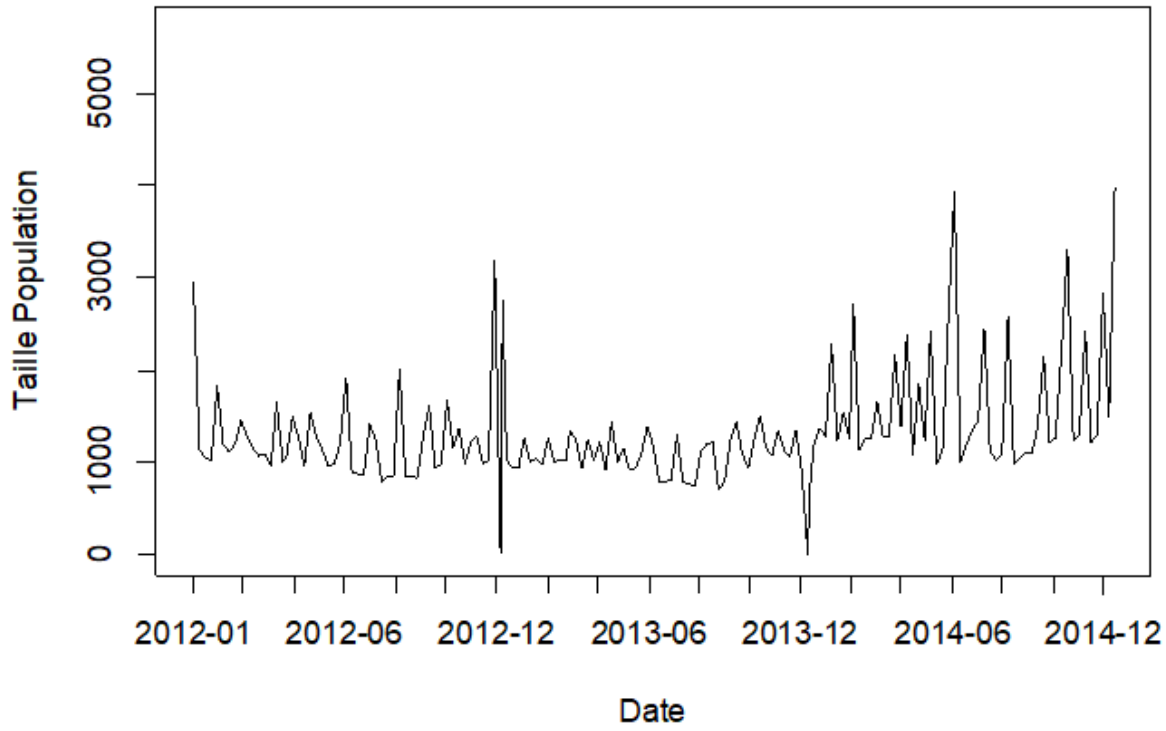


Figure 4-10: Résultat groupe 6 - méthode classique

Semblable au groupe 4 (Figure 4-8), le groupe 6 (Figure 4-10) présente les caractéristiques d'un groupe dont les usagers ne se déplacent que faiblement et seulement le mercredi, jeudi et vendredi. Sa population suit une constante autour de 1100 usagers (en moyenne : 17,0% de la population), avec une faible, mais visible saisonnalité. La variance est ici aussi très importante : la population peut augmenter ou diminuer de plus de 1000 usagers d'une semaine à l'autre. Ce groupe regroupe des usagers occasionnels qui ne nécessitent pas de se déplacer en transport en commun pour aller travailler ou ne travaillant pas.

La Figure 4-11 présente un condensé des résultats énoncés précédemment en combinant l'évolution des populations des différents groupes (en pourcentage) et la population totale (courbe noire). On peut voir les transferts de population de groupe à groupe. Par exemple à la semaine 15 (lundi de Pâques), à la semaine 21 (lundi de Pentecôte) ou à la semaine 27 (lundi 1er juillet : fête du Canada), le groupe 1 récupère les usagers du groupe 2 comme l'explique l'analyse précédente. Très pratique pour avoir une idée générale des problèmes engagés dans le réseau, ce graphique présente néanmoins l'inconvénient d'être surchargé et ainsi la moindre information se perd facilement dans la masse de données.

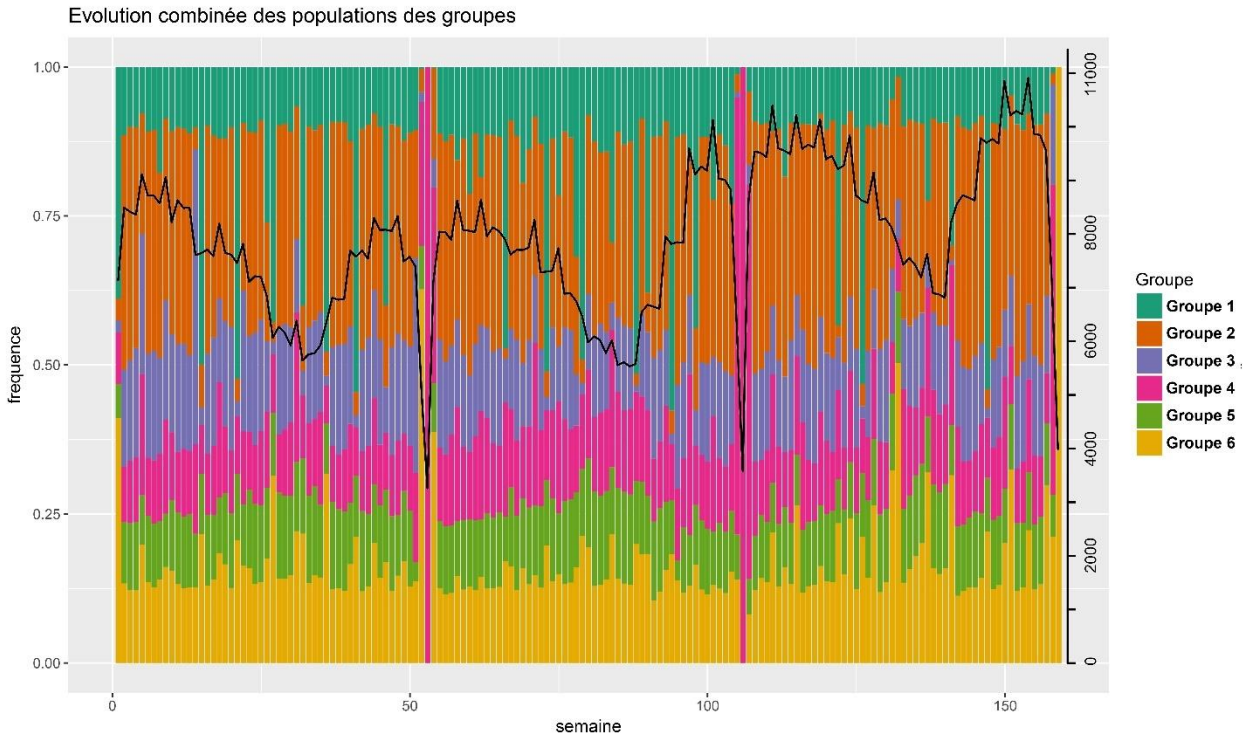


Figure 4-11 : Évolution combinée des populations des groupes

Les définitions de ces groupes ne sont qu'hypothèses. Il se peut que certains comportements, plus rares, se glissent dans les mauvais groupes et biaisent les résultats. On pense notamment aux cartes dont le comportement se situe en bordure de groupe. Généralement, les méthodes classiques prévoient de croiser ces informations à des données sociodémographiques propres à la ville afin d'enrichir la caractérisation des groupes.

4.2.5 Analyse de la qualité de segmentation

4.2.5.1 Indicateur de qualité intergroupe

La matrice Inter montre les distances entre les centres des groupes : plus les centres sont éloignés, plus la valeur de la distance est importante et donc plus les groupes présentent des caractéristiques différentes. Les centres des groupes étant fixes tout au long de la période d'étude, la matrice Inter fournit un résultat valable sur l'ensemble des 159 semaines. Un rapide coup d'œil à ce tableau témoigne de la proximité et de l'éloignement de certains groupes.

Tableau 4-4 : Matrice inter, méthode classique

	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5	Groupe 6
Groupe 1	X	1.988	2 126	2 701	2 839	2 020
Groupe 2		X	1.704	3 058	2 239	2 671
Groupe 3			X	2 362	2 717	2 493
Groupe 4				X	3 666	1.900
Groupe 5					X	3 316
Groupe 6						X

Le groupe 4 et le groupe 5 sont sensiblement différents. En effet, lorsque l'on revient à l'analyse des résultats dans la sous-partie précédente, on remarque que le groupe 4 représente, a priori, une partie de la population qui ne se déplace pas pour aller au travail ou qui ne travaille pas et donc dont le volume de déplacement est très faible. En revanche, le groupe 5, qui est constitué d'une population effectuant deux déplacements par jour sans être temporellement influencée, a un volume de déplacement très fort. Il est donc logique que l'éloignement de ces deux groupes soit maximisé pour une segmentation de bonne qualité.

À l'inverse, les groupes 2 et 3, sont les deux groupes les plus rapprochés d'après le Tableau 4-4. En effet à l'étude des résultats, on a montré que le groupe 2 est constitué de travailleurs réguliers tandis que le groupe 3 d'une population active que ne réalise que très peu de déplacement le vendredi. Ces deux groupes, ayant des caractéristiques très proches, se retrouvent avec un indice Inter très faible, et donc il est plus simple pour un usager de passer de l'un à l'autre au fil des semaines.

Un problème est mis en lumière : si l'on cherche à mieux détailler notre analyse, on a souvent tendance à augmenter le nombre de groupes. Pourtant on voit que plus K est grand plus la proximité entre groupes est faible et donc les individus peuvent plus facilement changer de groupes au fil des semaines Ainsi la population des groupes est sujette à de plus grosses variations tendant même à la disparition du groupe en fonction de la période étudiée. Il est donc d'ordre capital de choisir une bonne valeur pour K afin que l'analyse qui en découle soit cohérente.

4.2.5.2 Indicateur de qualité euclidien intragroupe

L'indicateur euclidien montre la moyenne euclidienne des distances entre les individus d'un groupe et son centre. Dans un espace euclidien de dimension 7, la valeur EUC définit le rayon de l'hypersphère moyenne autour de laquelle se trouve la population du groupe. Plus ce rayon est important, plus le groupe englobe des individus éloignés et moins la qualité de la segmentation est bonne.

Pour comprendre de manière moins théorique cette définition, on a représenté les différents usagers appartenant au groupe 2 pendant la semaine du 12 au 18 février 2012 (Figure 4-12). Ne pouvant pas représenter l'hypersphère facilement en dimension 7, il s'agit d'une projection sur deux dimensions : mardi et mercredi. Chaque carte est représentée par un point : par souci de superposition on a placé à droite des points le nombre d'individus de même comportement. Le centre du groupe 2 est défini par la croix rouge et est logiquement proche du point (2,2) à la vue des populations des différentes combinaisons. La valeur de l'indice EUC définit ainsi le rayon du cercle rouge, symbole de la distance moyenne euclidienne des points au centre. Ce cercle n'est donc que la projection de l'hypersphère sur les dimensions mardi et mercredi, dont on cherche à minimiser le rayon.

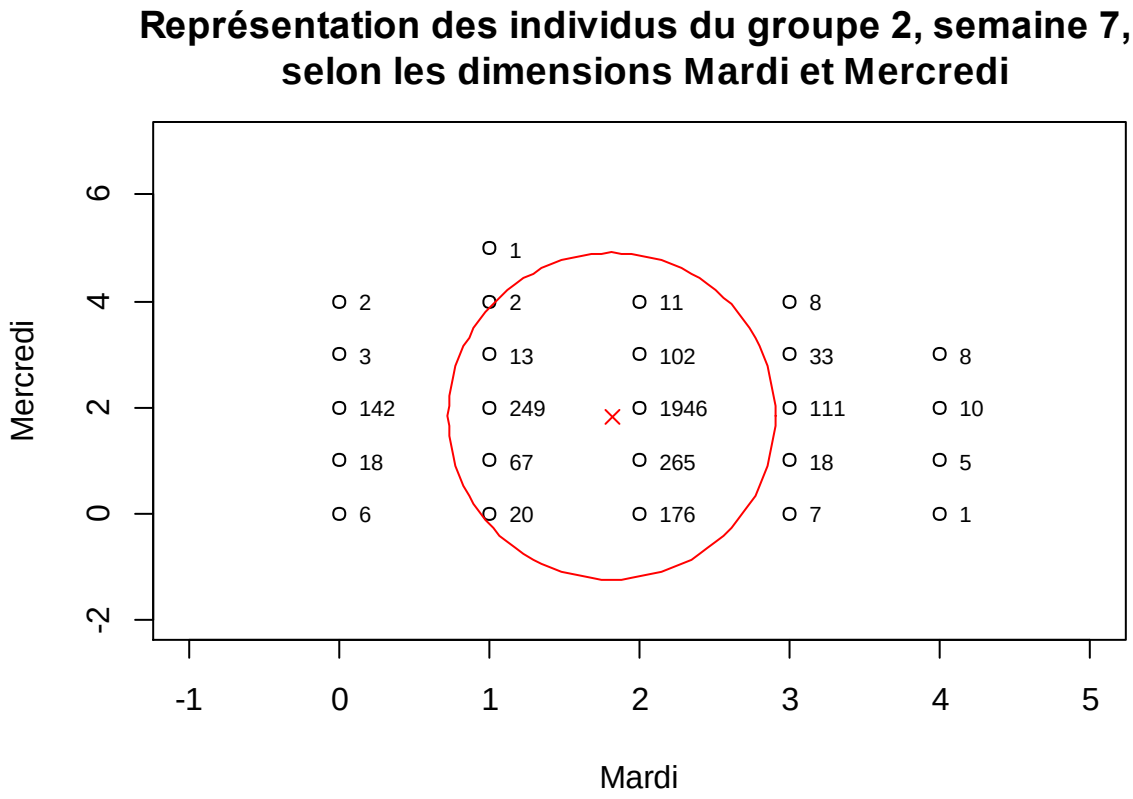


Figure 4-12 : Représentation des individus du groupe 2, semaine 7, sur mardi et mercredi

Sachant cela, il est donc plus intéressant d'obtenir des groupes compacts puisque le centre du groupe définit mieux les individus. La population variant le long de la période d'étude, l'indicateur EUC fournit donc des résultats différents sur chacune des semaines. La Figure 4-13 présente donc l'évolution comparée des indicateurs EUC de chacun des groupes. Toutes les évolutions ci-contre sont plus ou moins influencées par une saisonnalité annuelle ainsi que par les semaines à jour férié. On repère ici quatre types de courbes :

Le premier, représenté par le groupe 5, fournit l'indicateur le plus grand. Puisque ce groupe est défini comme le plus éloigné des autres, il vient logiquement chercher des individus beaucoup plus loin de son centre. On peut supposer qu'il englobe également les usagers dont le volume de déplacements est le plus grand. L'évolution de l'indicateur montre que ce groupe a tendance à être bien moins compact lors des périodes d'été et lors des semaines à congé. Pourtant, ces paramètres n'avaient aucune influence sur la taille de la population. Ceci signifie que les usagers s'éloignent du centre du groupe pendant ces périodes, et donc, que définir ce groupe par son centre perd en intérêt.

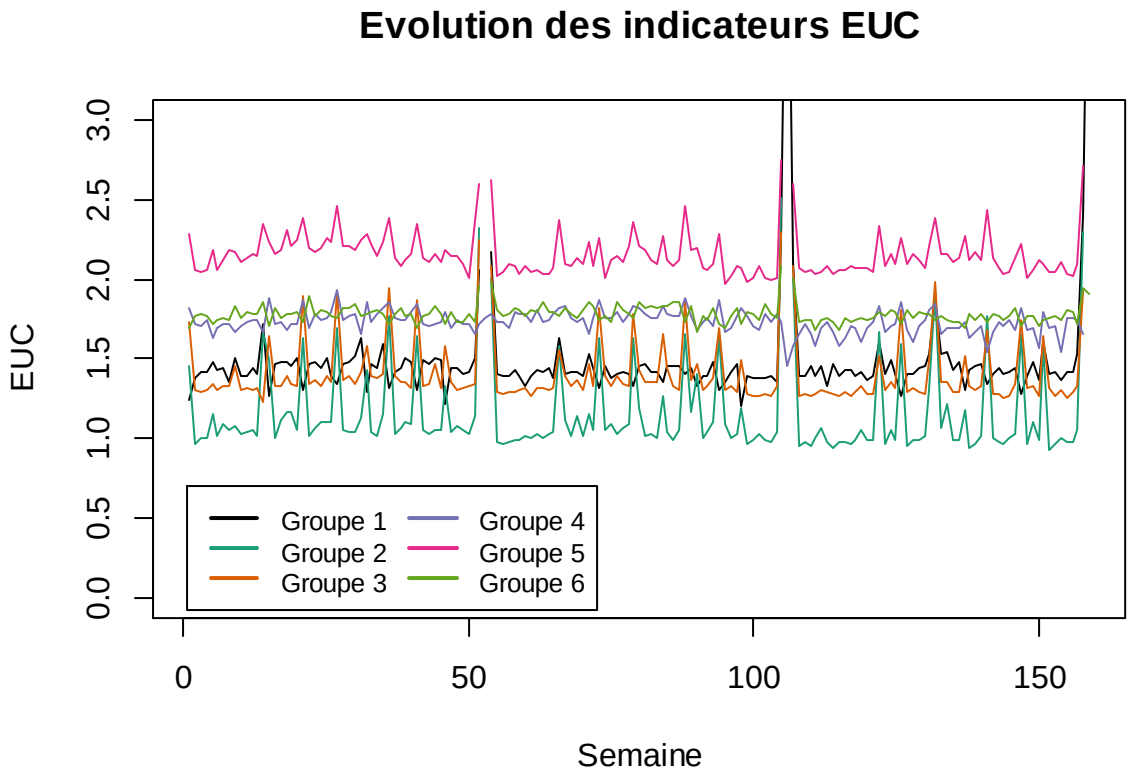


Figure 4-13 : Comparaison de l'évolution de l'indicateur EUC appliqué à chaque groupe

Le second type de courbe correspond aux groupes 4 et 6 qui se ressemblent fortement. En effet, ces deux groupes très proches (indicateur « inter » faible) présentent des similarités dans leurs variations de tailles de population, mais également dans l'évolution de leur indicateur EUC. Dans cette étude, la saisonnalité et les jours fériés ont une très faible incidence sur l'évolution de l'indicateur, qui oscille légèrement autour de 1,7. À titre de comparaison, ces deux groupes fournissent un résultat globalement inférieur à celui du groupe 5 : ils sont ainsi plus compacts et leur population ressemble davantage à leurs centres respectifs.

Le troisième type est représenté par le groupe 2, qui propose l'indice le plus faible en dépit d'une population la plus conséquente. Peu importe la saison, l'indicateur fournit globalement un résultat de qualité similaire (légèrement supérieur à 1). Cependant, il est sensiblement affecté par les semaines à jour férié. On retrouve ici les caractéristiques du groupe 2 puisque, représentant les semaines de travail à temps, il n'a presque plus lieu d'exister en semaine à congé et donc perd conséquemment de sa qualité.

Le dernier correspond aux courbes d'évolution des groupes 1 et 3. Existant tous deux comme groupes complémentaires au groupe 2, le groupe 1 est influencé par les semaines dont le vendredi est férié et le groupe 3 par celles dont le lundi est férié. Leurs valeurs d'indice relativement constantes et faibles leur permettent de conserver une qualité satisfaisante; les centres de ces groupes définissent donc bien mieux leur population que pour le 5.

4.2.5.3 Indicateur de qualité relatif intragroupe

L'indicateur relatif montre la moyenne des distances relatives entre les individus d'un groupe et son centre. Il s'agit d'un calcul d'erreur moyen pour chaque usager quant à l'approximation de cet usager par le centre du groupe auquel il appartient. La courbe présentée en Figure 4-14 représente l'évolution de la moyenne de l'indicateur REL des différents groupes, pondérée par le pourcentage de population que représente chacun des groupes.

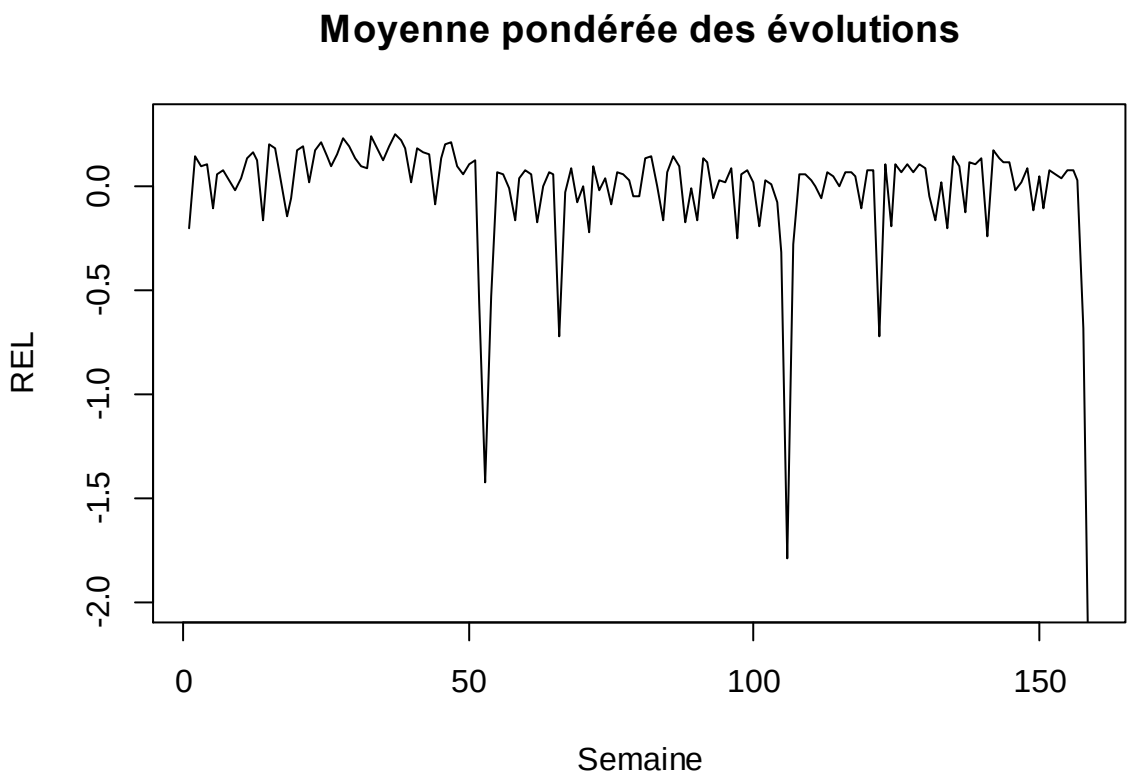


Figure 4-14 : Moyenne des évolutions de l'indicateur REL, pondérée par la population

Comme attendu, cette évolution s'approche de 0, puisque le centre d'un groupe est le barycentre des points de ce groupe. Cependant, comme la segmentation s'applique sur l'ensemble des semaines simultanément, elle est sujette à certaines erreurs lorsqu'on s'intéresse aux semaines une à une. En effet, puisque la population est strictement différente d'une semaine à l'autre, l'approximation globale ne peut parfaitement décrire chacune des semaines. C'est pourquoi on peut remarquer des erreurs plus ou moins conséquentes en fonction de la semaine étudiée. Par exemple, lorsque l'on s'intéresse aux semaines de Noël, l'utilisation des transports en commun est fortement dégradée par la période. L'utilisation globale des transports en commun sur trois ans est logiquement différente de celle des fêtes de fin d'année. On peut ainsi remarquer une baisse d'activité moyenne par usager de plus de 1,5 déplacement sur cette semaine chaque année par rapport à l'approximation globale effectuée. Deux autres baisses conséquentes d'activité apparaissent aux semaines 67 et 123, respectivement les lundis de Pâques des années 2013 et 2014. Une chute d'activité de plus de 0,5 déplacement par usager est à témoigner par rapport à l'approximation globale. Étonnement, cette chute n'est pas aussi significative pour l'année 2012. De manière générale, outre ces cas particuliers, l'approximation a une erreur entre -0,2 et 0,2 déplacement par usager, véritable limite de la méthode qu'il faut améliorer.

4.2.5.4 Critère de Dunn

Le critère de Dunn est égal au ratio des pires valeurs d'indices, c'est-à-dire qu'il divise la plus petite distance intergroupe par la plus grande distance intragroupe. Par cela, il définit un résumé des indicateurs quant à la qualité de la segmentation. Ayant des valeurs variables en distances intragroupes, on suit donc l'évolution du critère de Dunn sur les 159 semaines de la période d'étude (Figure 4-15). Rappelons que plus la valeur du critère est haute, plus le résultat de la segmentation est fidèle à la réalité (bonne qualité de segmentation).

Fortement influencée par la période de l'année, la valeur de l'indicateur suit une saisonnalité où les maximums se trouvent aux périodes d'été et les minimums en hiver. La variation globale du nombre de déplacements produit le résultat inverse. On peut conclure que plus il y a d'usagers, plus il y a de patrons de déplacements sensiblement différents de notre approximation par segmentation. Et donc moins bonne est la qualité de cette méthode.

Les semaines à congé ont également une forte influence sur la valeur de l'indicateur. En effet, les baisses significatives apparaissent sur les semaines où au moins un jour est férié. Ce résultat était attendu puisque l'évolution de l'indicateur EUC montrait déjà cette information.

Les trois années présentent certes un patron similaire, mais la moyenne du critère est sensiblement différente pour chacune. En effet, les résultats de segmentation tendent à s'améliorer entre 2012 et 2013, mais à diminuer plus fortement en 2014. On pourrait donc se demander si la mise en place de la ligne Rapibus, en fin 2013, n'a pas modifié les comportements des usagers par rapport aux deux autres années. Une autre hypothèse pourrait faire intervenir le nombre de déplacements sur les différentes années. En effet, on a vu précédemment que la population a une influence sur ce critère. Le nombre de déplacements adultes + desfire (Tableau 4-1 : Résumé des tailles de bases de données (Tableau 4-1) diminue légèrement entre 2012 et 2013 montrant une augmentation de qualité, puis augmente plus sensiblement de 2013 à 2014 résultant en une diminution de la qualité associée.

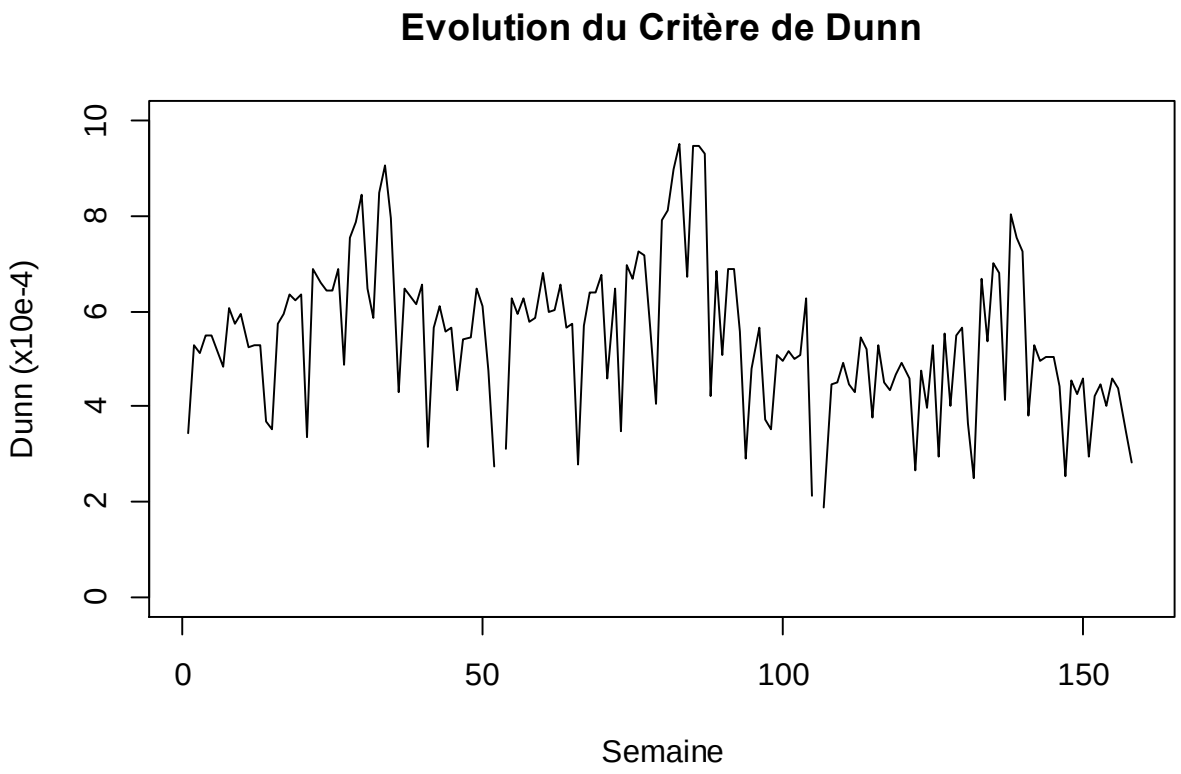


Figure 4-15 : Évolution du critère de Dunn

4.2.6 Analyse de la stabilité de la population

L'indicateur défini selon l'Équation 3.8 est appliqué sur chacune des cartes à puce. Il s'agit d'un critère d'instabilité où plus la valeur est faible plus l'individu en question a un comportement stable, c'est-à-dire un patron de déplacement qui change peu. Cette modification de comportement se traduit par un changement de groupe associé. Un individu qui change souvent de groupe obtient donc un WSI plus important.

La Figure 4-16 sert essentiellement de support visuel à la compréhension de l'analyse de la Figure 4-17. Il y est représenté l'ensemble des individus ayant effectué des déplacements sur au moins quatre semaines, triés par la valeur de leur WSI. Plus l'indice est faible, et donc plus l'usager a un comportement stable, plus la couleur se rapproche du vert, et inversement du rouge.

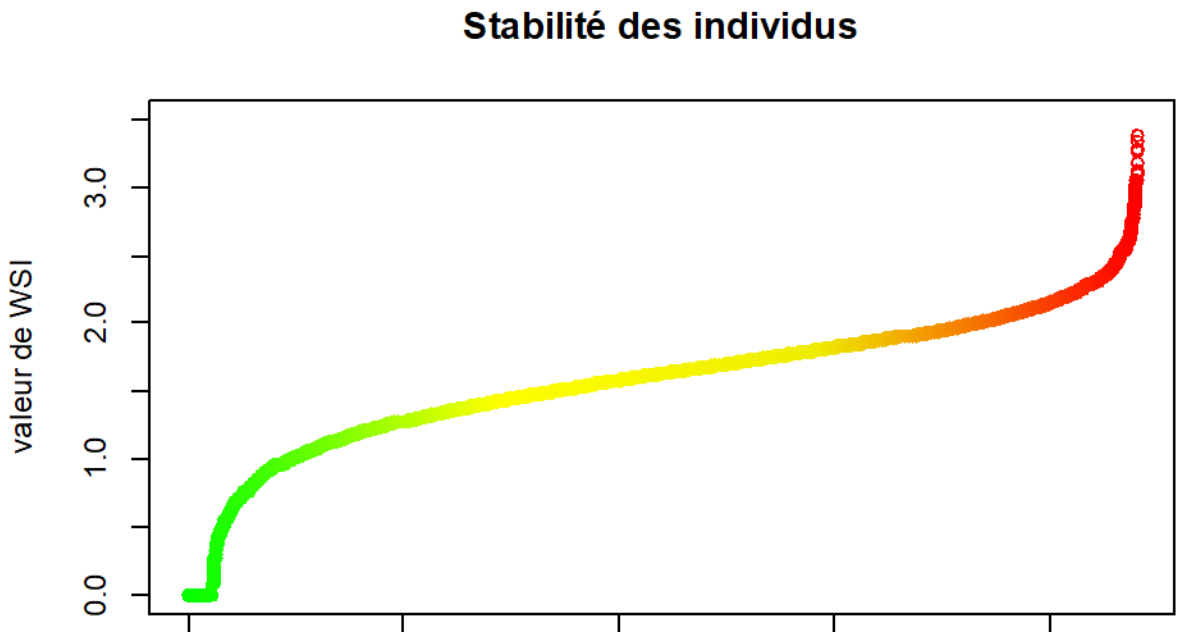


Figure 4-16 : Stabilité des individus par la méthode classique

La Figure 4-17 représente l'ensemble des individus ayant réalisé des déplacements sur au moins quatre semaines différentes. Un test de Kolmogorov Smirnov est appliqué pour comparer l'ensemble des WSI_i à une loi normale. Même si la répartition des valeurs semble visuellement suivre la loi normale sur l'histogramme (en rouge), l'hypothèse est rejetée par le test (p-val=0). En effet, d'apparence symétrique, la distribution des stabilités des usagers est en fait légèrement positive à gauche. Pour comprendre ce résultat, il s'agit d'effectuer l'analyse de la courbe quantile-quantile (Figure 4-17).

Chaque carte à puce ayant réalisé des déplacements sur plus de quatre semaines est représentée par un point. L'ordonnée présente les valeurs réelles des indicateurs WSI des individus. L'abscisse, du nom de quantile théorique, représente les différents quantiles d'une loi normale sur une échelle. On cherche ici à comparer la distribution réelle à celle d'une loi normale, pour cela il suffit de comparer l'ensemble des points à la droite.

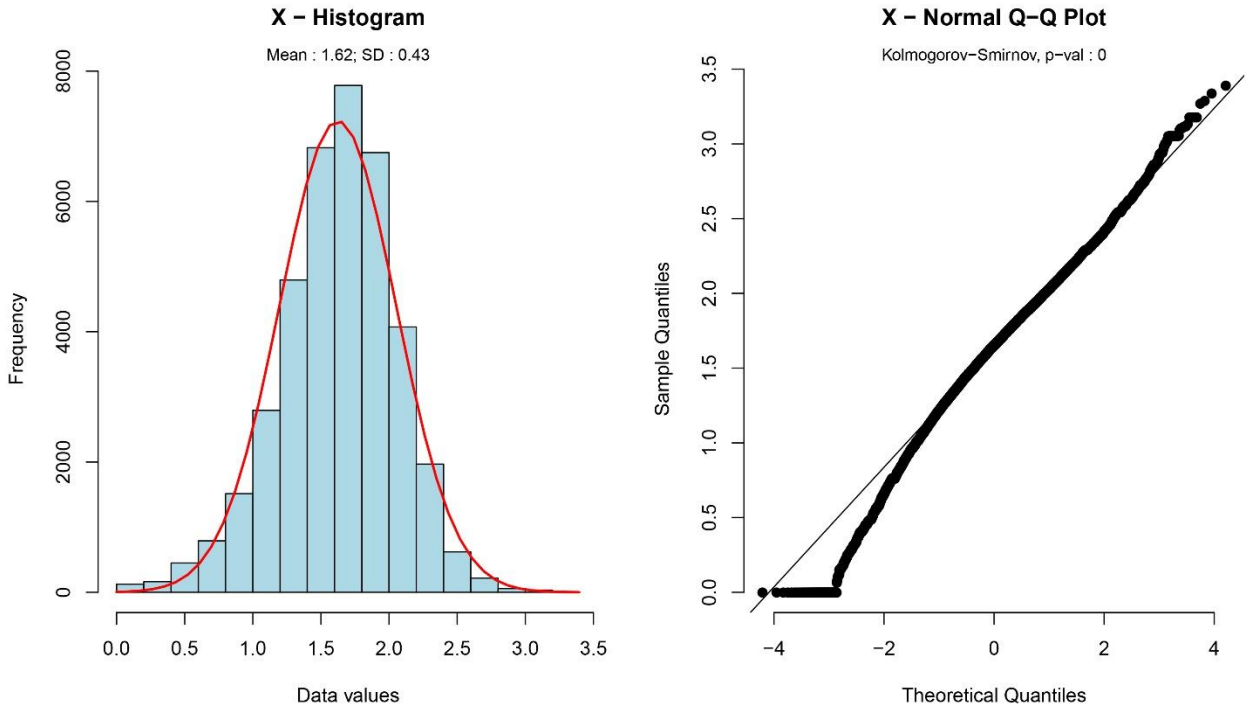


Figure 4-17 : Test Kolmogorov-Smirnov sur la stabilité des individus, méthode classique

Les individus dont le WSI est le plus faible (quantiles théoriques entre -4 et -1,5 donc des WSI entre 0 et 1) se situent en dessous de la droite normale. Ceci signifie que, sur cette plage, il y a beaucoup plus de faibles valeurs d'indicateur par rapport à une distribution normale. De plus, on remarque que les individus entre les quantiles théoriques -4 et -3 conservent un quantile réel de 0. Ceci signifie que cette partie de la population, dont la loi normale prévoyait un résultat entre 0 et 0,5, obtient un WSI réel de 0. On dit que la distribution forme un pic de valeurs identiques, empêchant la distribution de suivre une loi normale. Le reste des valeurs se situent relativement proches de la droite normale et donc n'ont pas de particularités différentes à une distribution normale. Seuls les individus de plus hauts WSI sont sous-estimés par la loi normale. Grâce à ces indices, tout porte à croire que la distribution des WSI suit une forme positive à gauche.

4.3 Application de la méthode expérimentale

4.3.1 Importation, manipulation de données et apprentissage

Ne nécessitant aucune modification par rapport à la méthode classique, l'importation et la manipulation des données sont appliquées de la même manière qu'en 4.2.1. On rappelle que la manipulation de données permet notamment la création de tableaux Semaine. Chacun de ses tableaux n'est utilisé qu'une seule fois en tant que données d'entrées à l'algorithme de segmentation.

Afin de permettre son démarrage, l'algorithme de segmentation nécessite un ensemble de noyaux, véritables centres de départ indispensable à l'application de l'algorithme sur la semaine 1. Il est donc nécessaire de simuler la méthode classique sur au moins un an de données afin de fournir cet ensemble de noyaux. On appelle cette étape, l'apprentissage. Dans le but de maximiser la qualité de la segmentation et donc produire des résultats au plus proches du réel, il est décidé d'effectuer l'étape d'apprentissage sur l'ensemble des données, soit 3 ans. Les centres issus de la méthode classique sont donc utilisés en tant que noyaux pour la segmentation de la semaine 1 (Tableau 4-5), impliquant ainsi que l'étape de détermination du nombre de groupes ait été réalisée au préalable dans la simulation de la méthode classique, et donc que $K = 6$.

Tableau 4-5 : Noyaux à imputer pour la segmentation semaine 1

	Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
Groupe 1	0,106	0,111	1.944	1.742	1.803	1.368	0,167
Groupe 2	0,106	1.955	1.812	1.831	1.851	2 084	0,063
Groupe 3	0,156	1.983	1.735	1.781	1.670	0,394	0,124
Groupe 4	0,210	1.074	1.152	0,337	0,149	0,360	0,228
Groupe 5	1.280	1.841	1.829	1.872	1.921	2 018	1.963
Groupe 6	0,163	0,384	0,181	1.076	1.192	1.104	0,257

Finalement, la liste des semaines à congé, déterminée en 4.1, est récupérée et est utilisée dans le calcul des noyaux de segmentation à partir de la fonction d'apprentissage.

4.3.2 Segmentation

Divers paramètres ont été fixés permettant le fonctionnement de l'algorithme de segmentation. Le regroupement s'applique donc sur les vecteurs comportements séparés en fonction de la semaine associée, pour un $K = 6$.

Comme pour la méthode classique, d'autres paramètres intrinsèques à l'algorithme sont nécessaires à son bon fonctionnement. Afin d'optimiser la qualité de la segmentation, le choix se porte sur le fait de maximiser les paramètres en dépit du temps de calcul. Les différentes valeurs sont ainsi reportées dans le Tableau 4-6.

Tableau 4-6 : Paramètres algorithme, segmentation expérimentale

Dénomination	Valeur	Paramètre dans R
Taille base de données	3 ans = 159 semaines	s = 159
Type d'information	Déplacement	x = 1
Nombre de groupes	6	K = 6
Nombre d'itérations max	100	it_max = 100
Nombre d'initialisations	50	nb_ini = 50
Nombre d'années d'apprentissage	3 ans = 159 semaines	S = 159
Liste des semaines à congé	/	P

Fonctionnant suivant le principe selon lequel les semaines sont segmentées chronologiquement une à une en prenant en compte les résultats des semaines précédentes, il nous faut déterminer des noyaux pour chacune de ces segmentations. La semaine 1 prend comme noyaux les centres issus de la méthode classique, tandis que les semaines 2 à 159 utilisent chacune à leur tour l'une des fonctions d'apprentissage décrites en 3.4.3.

4.3.3 Analyse des résultats

Cette partie propose une analyse des résultats de segmentations issus de la méthode expérimentale, présentant les différents groupes ainsi que leurs caractéristiques intrinsèques.

Sur la Figure 4-18, le groupe 1 est essentiellement constitué d'usagers se déplaçant en moyenne une fois et demie les mardis, mercredis, jeudis, et vendredis, mais très peu de déplacements sont à constater en fin de semaine et le lundi. Lors de la décomposition de la courbe du mardi, on se rend compte que la saisonnalité n'a qu'une faible influence sur l'évolution (légère baisse au printemps). Cependant, la tendance générale du nombre de déplacements par usager diminue nettement en 2014. Elle reste toutefois constante sur les autres dimensions ; seules des chutes pour les périodes de Noël sont à constater les mardis et mercredis, et pour la fin de semaine de Pâques le vendredi. Concernant la population du groupe qui représente en moyenne 18,3% de la population hebdomadaire étudiée, le groupe a tendance à gagner une centaine d'usagers par année avec une baisse de l'ordre de 400 usagers pendant l'été. Ce résultat peut s'expliquer par le fait que les déplacements de courtes distances sont plus souvent réalisés à pied en été qu'en hiver. Comme dans la méthode classique, ce groupe récupère les usagers du groupe 2 lors des semaines à lundi férié. Le groupe s'apparente donc, ici aussi, à un groupe d'usagers dont le patron de déplacement ressemble à celui d'un travailleur à mi-temps où le lundi n'est pas travaillé.

Le groupe 2 (Figure 4-19) est principalement constitué d'usagers se déplaçant deux fois par jour en semaine et ayant très peu de mobilité en transport en commun en fin de semaine. De manière générale, les volumes individuels de déplacements sur toutes les dimensions restent approximativement constants. Des augmentations apparaissent chaque année aux mêmes périodes sur certaines fins de semaine (Pâques, Pentecôte, fête du Canada, fête du Travail, Action de grâce). La population, qui représente en moyenne 29,1% de la population hebdomadaire étudiée, suit une saisonnalité facilement identifiable puisqu'on dénote une perte de plus de 50% des individus pendant les périodes d'été. La tendance confirme une augmentation de la population du groupe de 800 usagers entre 2013 et 2014 alors qu'il n'y en a pas eu depuis 2012. Comme pour la méthode classique, le groupe perd ses usagers au profit du groupe 1 lors des semaines où le lundi est férié et au profit du groupe 3lors des semaines où le vendredi est férié. Le groupe s'apparente donc, ici aussi, au groupe des travailleurs à temps plein.

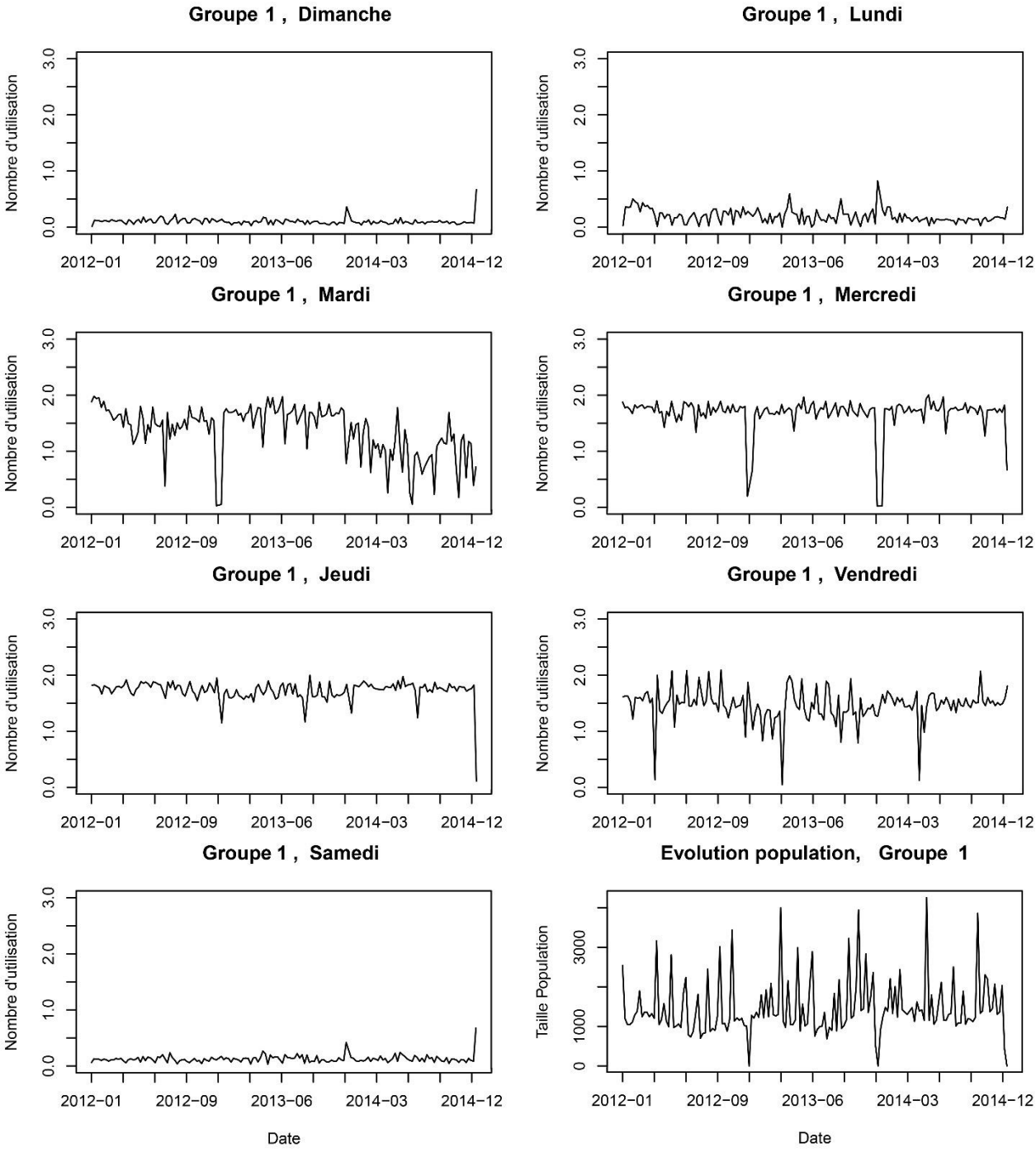


Figure 4-18: Résultats groupe 1 - méthode expérimentale

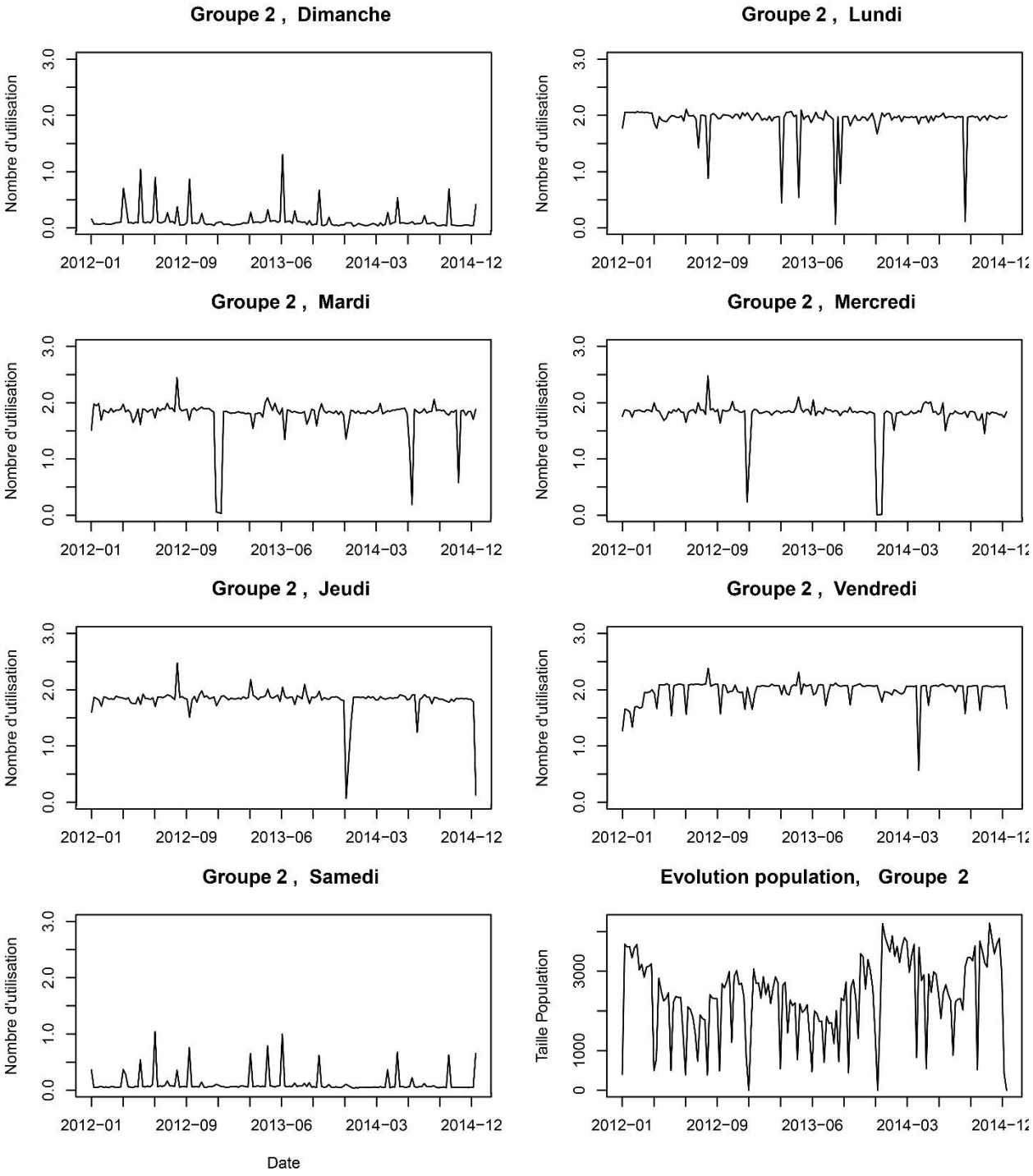


Figure 4-19: Résultats groupe 2 - méthode expérimentale

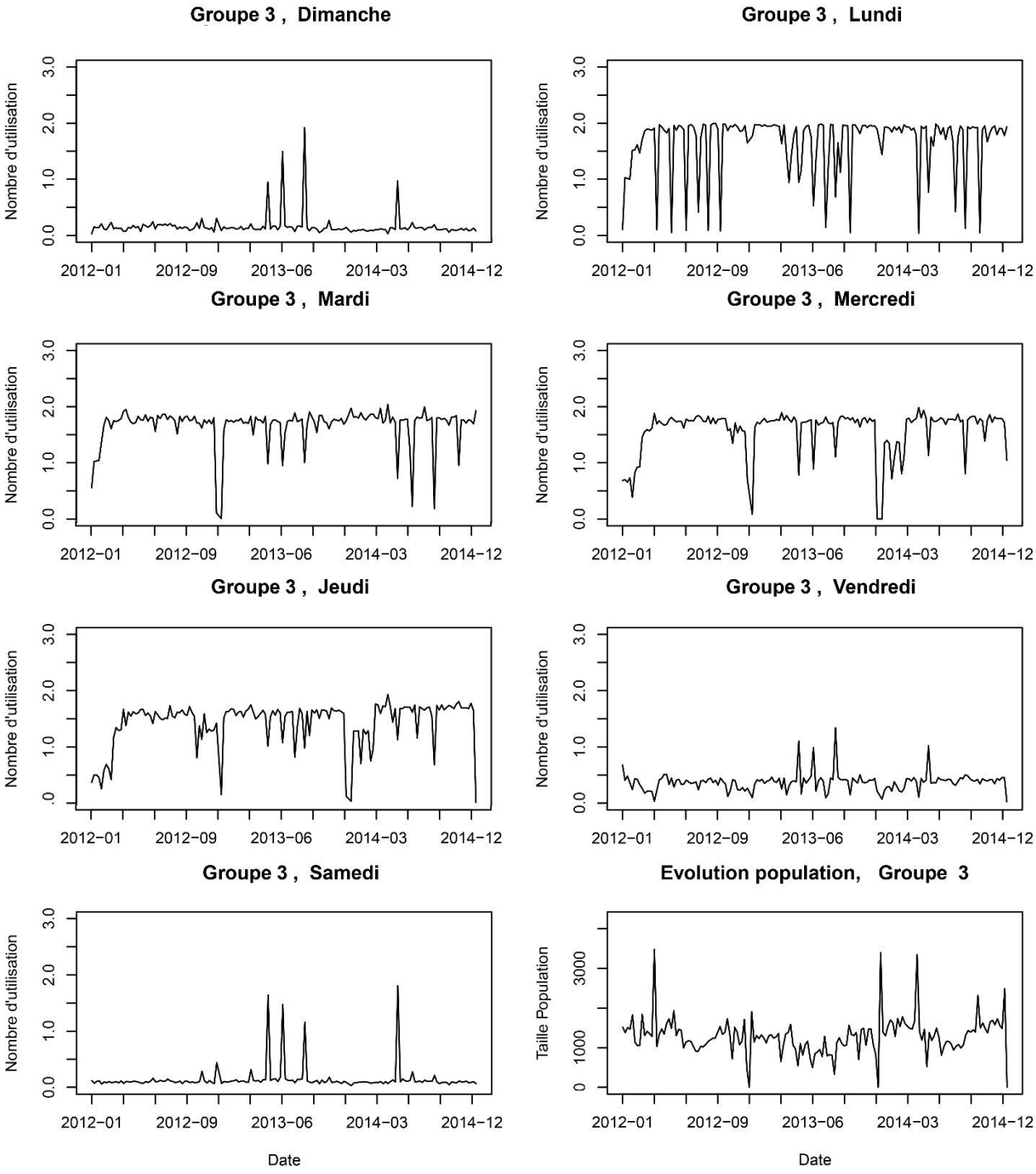


Figure 4-20: Résultats groupe 3- méthode expérimentale

Caractéristiquement proche du groupe 2 (Figure 4-19), le groupe 3(Figure 4-20) est constitué d'usagers qui effectuent deux déplacements sur chacun des jours de la semaine sauf le vendredi. Quelques soucis directement liés à l'algorithme surviennent dans les 4 premières semaines, mais il parvient à se stabiliser de lui-même par la suite. Très peu de variations sont à constater au niveau des déplacements par usager. On remarque un comportement étrange sur 3 semaines en 2013. En effet, lors de la semaine de l'Ascension, de la fête du Canada et de la fête du Travail, le groupe semble définir une autre partie de la population, comme s'il avait englobé le groupe 4 puisque toutes les composantes atteignent 1 déplacement par usager. Ces erreurs sont certainement dues à une imperfection de la méthode, notamment dans la fonction d'apprentissage. Concernant l'évolution de la population (en moyenne : 16,2% de la population hebdomadaire étudiée), on remarque une faible amplitude de la saisonnalité avec décroissance aux périodes d'été et sur l'année 2013. On peut supposer que le groupe 3vient récupérer les usagers du groupe 2 dans les semaines où le vendredi est férié. À première vue, ce groupe représente les travailleurs à mi-temps effectuant des déplacements vers leurs lieux d'activité du lundi au jeudi compris. Si la méthode classique montrait une disparition du groupe pendant les semaines à lundi férié, il n'en est rien ici. L'algorithme permet de suivre ces utilisateurs même s'ils arrêtent de se déplacer les lundis fériés.

Sur la (Figure 4-21), le groupe 4 présente des caractéristiques sensiblement différentes que dans la méthode classique. Dans ce cas, il s'agit d'un groupe aux comportements très variables à la population constante très faible (en moyenne : 7,7% de la population hebdomadaire étudiée). En effet, les usagers du groupe semblent effectuer un déplacement sur chacun des jours sauf le samedi (2 déplacements). Ses variations atteignent généralement une amplitude très grande (environ 1 déplacement). On peut donc supposer que ce groupe sert à récupérer les usagers à comportement instable. Habituellement, les groupes d'utilisateurs dans la tranche adulte régulier qui proposent des comportements assez aléatoires sont souvent constitués d'une population inactive (qui ne se déplace pas pour travailler, ou ne travaille pas).

D'apparence très proches du groupe 2 (Figure 4-19), les usagers du groupe 5 (Figure 4-22) se déplacent deux fois par jour sauf le dimanche où au minimum un déplacement est effectué. Cependant, l'étude précédente a montré qu'il s'agit en fait d'un groupe très différent qui suit une saisonnalité. En effet, sur chaque année, on remarque des comportements similaires (sur la plupart des composantes du vecteur) directement liés à la période de l'année et non aux semaines à congé. On peut d'ailleurs noter que les usagers du groupe ont tendance à diminuer leur nombre de

déplacements en semaine et à l'augmenter en fin de semaine. De plus, la population de ce groupe, qui représente en moyenne 8,6% de la population hebdomadaire étudiée, montre une croissance constante, mais très faible (gain d'environ 100 usagers en 3 ans) et une faible saisonnalité. On peut supposer qu'il s'agisse d'un groupe relativement proche du groupe 5 décrit en 584.2.4.

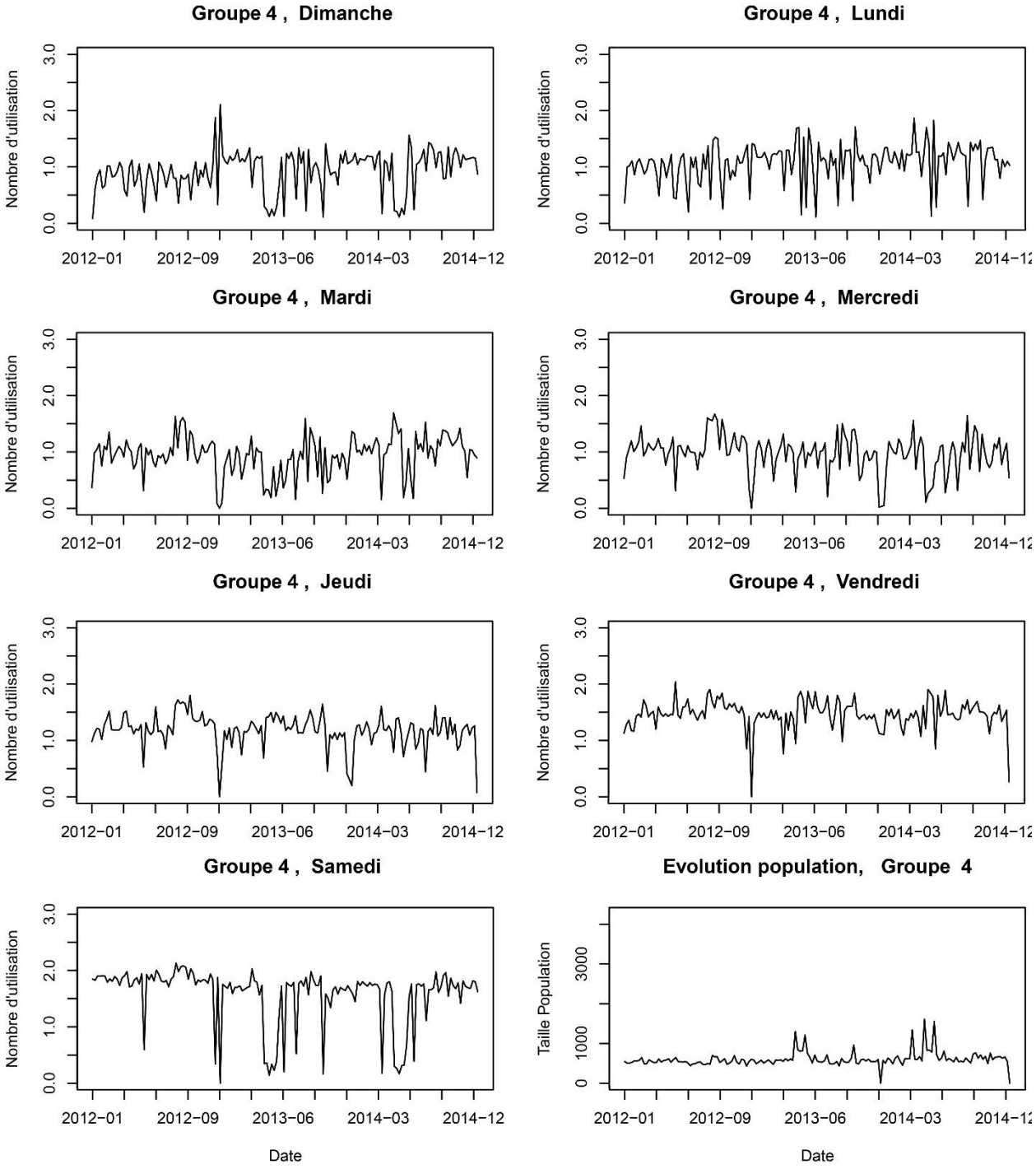


Figure 4-21: Résultats groupe 4 - méthode expérimentale

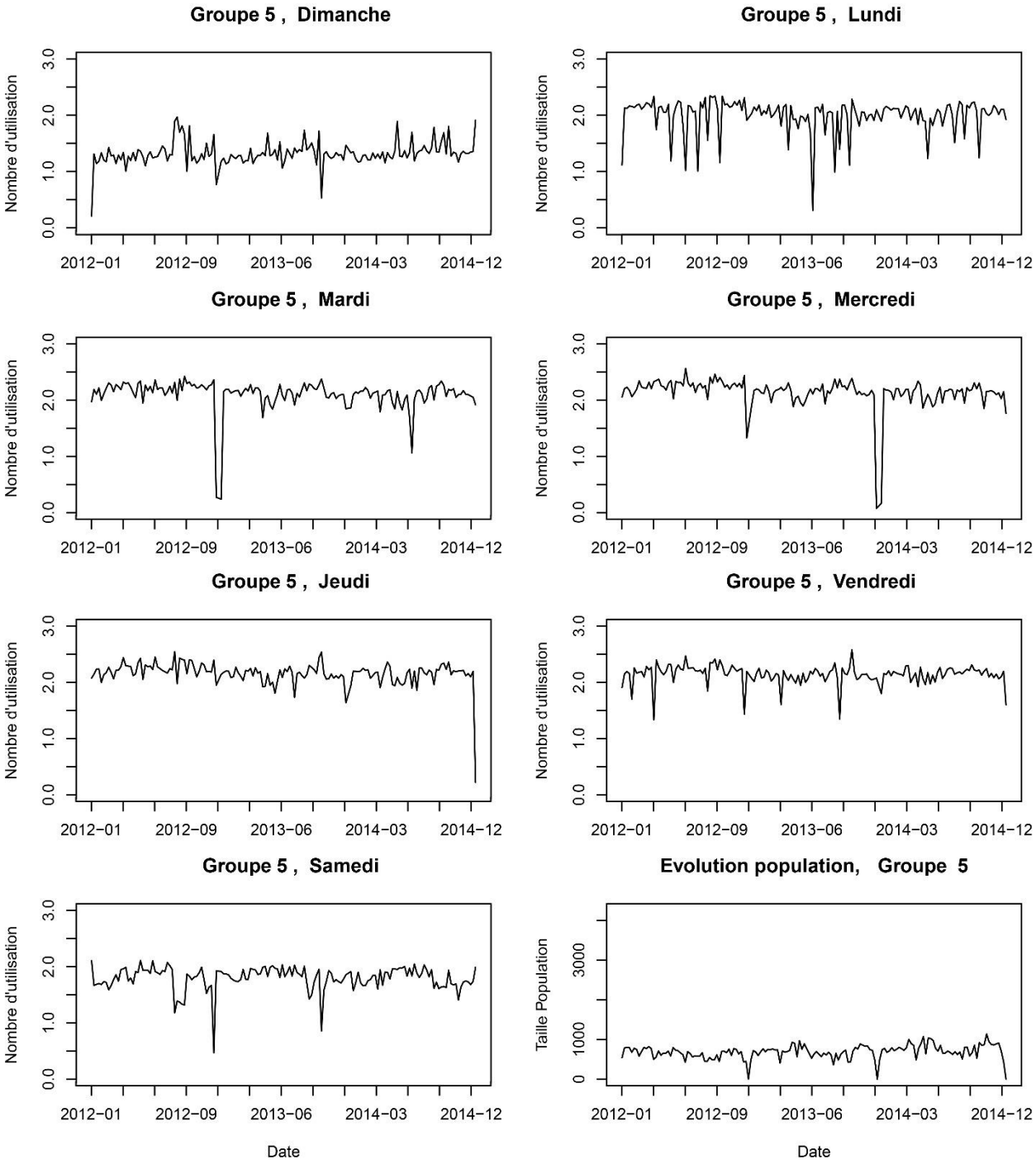


Figure 4-22: Résultats groupe 5 - méthode expérimentale

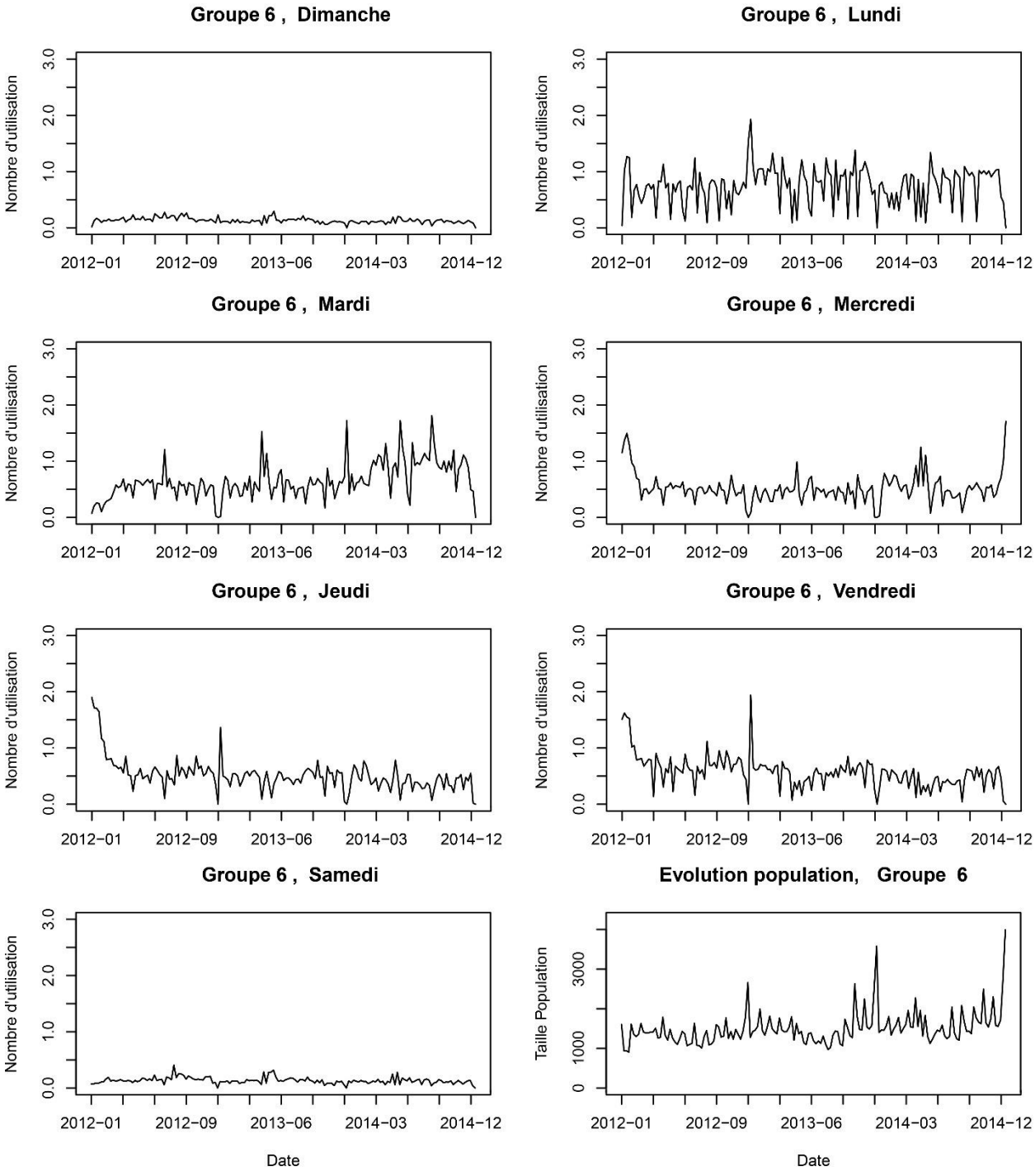


Figure 4-23: Résultats groupe 6 - méthode expérimentale

Les usagers du groupe 6 (Figure 4-23) se caractérisent par leurs faibles volumes de déplacements en semaine et inexistants en fin de semaine. Quelques soucis directement liés à l'algorithme surviennent dans les 4 premières semaines, mais il parvient à se stabiliser de lui-même par la suite comme pour le groupe3. Il semble que ces deux groupes se soient intervertis pendant ces 4 premières semaines. Une amélioration de la fonction d'apprentissage est donc à approfondir. Les caractéristiques font penser à une fusion des groupes 4 et 6 proposés par la méthode classique. En effet, on a vu que ces deux groupes présentaient des évolutions complémentaires en déplacements, on avait d'un côté, un groupe présentant des déplacements uniquement le lundi et le mardi, et de l'autre des déplacements le mercredi, jeudi et vendredi. L'étude de la population indique une forte croissance constante atteignant une augmentation de 300 usagers en trois ans, alors qu'elle représente 20% de la population hebdomadaire étudiée. Une saisonnalité relative à la période d'été est ici aussi présente, coïncidant aux périodes de baisse d'achalandage générale à Gatineau.

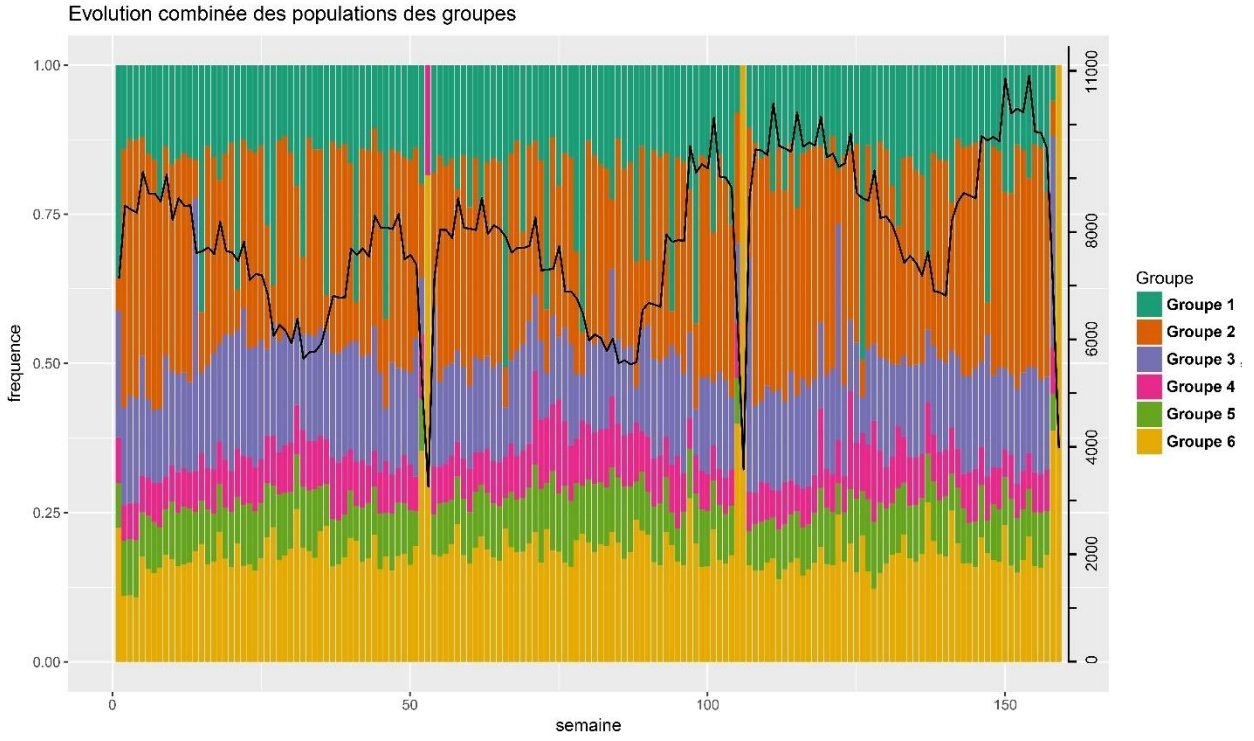


Figure 4-24 : Évolution combinée des populations des groupes

Comme pour la méthode classique, la Figure 4-24Figure 4-11 présente un condensé des résultats énoncés précédemment en combinant l'évolution des populations des différents groupes (en pourcentage) et la population totale (courbe noire). Vu que la méthode expérimentale fournit des résultats avec moins de variations, la figure est bien plus lisible et l'on distingue très clairement l'évolution comparée des populations des différents groupes.

4.3.4 Analyse de la qualité de segmentation

4.3.4.1 Indicateur de qualité « inter »

Les matrices « inter » montrent les distances entre les centres des groupes : plus les centres sont éloignés, plus la valeur de la distance est importante et donc plus les groupes présentent des caractéristiques différentes. Les centres des groupes étant variables sur le long de la période d'étude, les 159 matrices sont donc représentées sous forme de courbes montrant l'évolution des distances entre les centres. La Figure 4-25 révèle ainsi le rapprochement et l'éloignement des centres en fonction de la semaine en question. Un rapide coup d'œil permet d'établir que la tendance des évolutions reste constante au fil du temps, gage de stabilité de l'algorithme.

L'analyse précédente montre que ce groupe englobe une population au comportement très variable et dont le volume est d'environ un déplacement par jour. Il est donc attendu que la distance avec tous les autres centres soit haute. Cependant, le groupe 4 semble se trouver à équidistance de tous les autres centres. On peut donc supposer qu'il vient récupérer des individus situés en bordure des autres groupes, notamment les usagers dont les déplacements en fin de semaine sont sensiblement différents de ceux proposés par les autres groupes. Il formerait donc un groupe tampon principalement composé des usagers de bordure des groupes de la méthode classique, permettant une densification de ces groupes et donc une amélioration de la qualité d'approximation de la population.

Comme pour le résultat de la segmentation par méthode classique, le groupe 5 est le plus éloigné des autres centres puisqu'il est défini par un fort volume de déplacements. On confirme ici aussi que ce groupe contient des individus dont le comportement est très différent de celui des autres groupes. A priori, il joue aussi le rôle de tampon récupérant les individus qui se déplacent beaucoup plus que la moyenne, améliorant ainsi l'approximation faite sur les travailleurs temps plein et mi-temps.

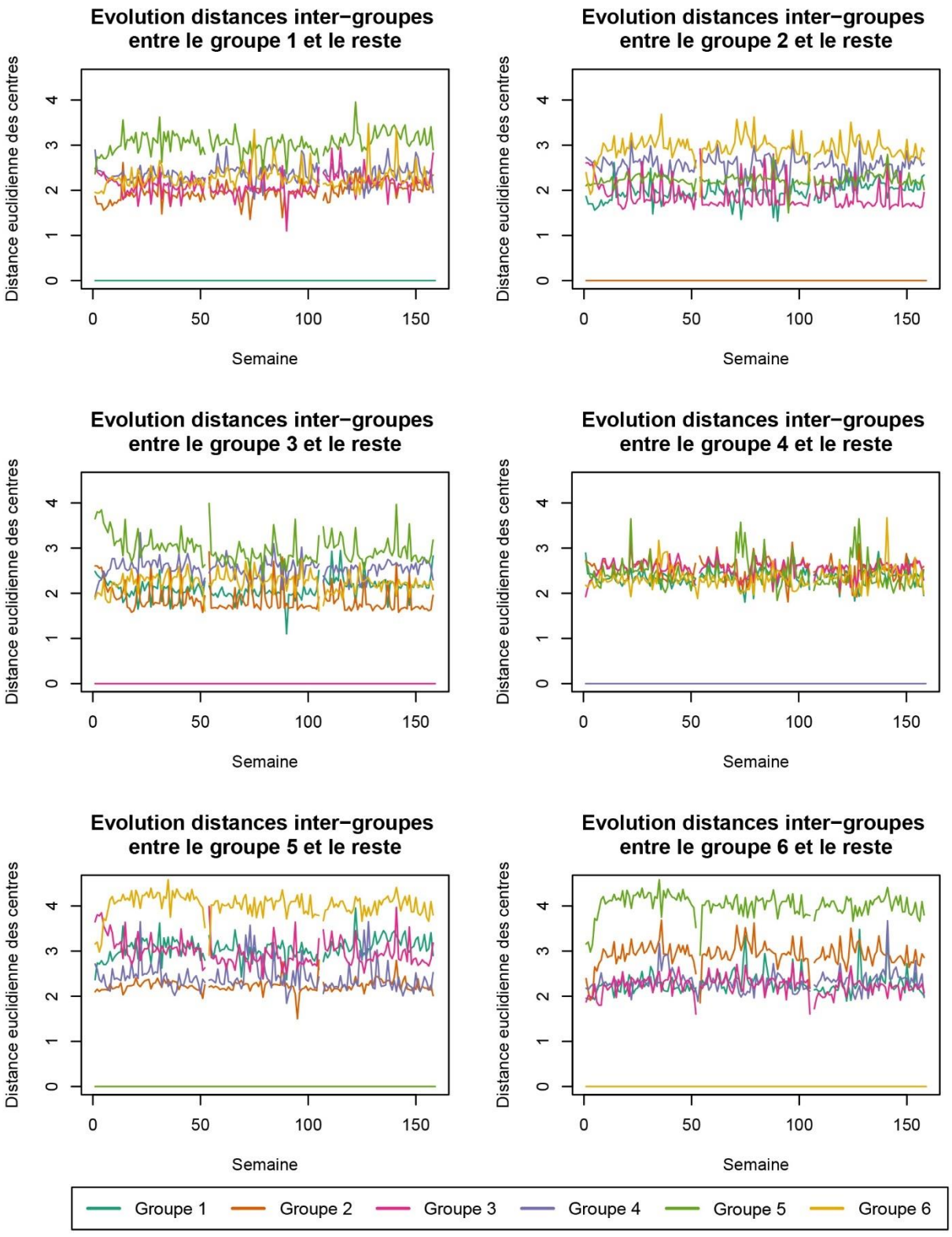


Figure 4-25 : Évolution des distances intergroupes

Les groupes 1, 2 et 3, semblent voisins comme le montrait déjà l'analyse de la méthode classique. Cette proximité s'explique par le fait que les comportements des trois groupes se ressemblent au point de s'échanger facilement des usagers en semaines à congé. Comme le volume de déplacements est plus faible dans les groupes 1 et 3, ils sont très éloignés du groupe 5 dont le volume est le plus haut. Le groupe 2 s'éloigne quant à lui fortement du groupe 6 dont le volume est le plus faible. En moyenne, ces distances euclidiennes maximales se rapprochent de 3, valeur bien plus importante que lors de l'étude de la méthode classique. Cela signifie donc une plus forte dissimilitude entre les groupes. Rappelons que maximiser les différences intergroupes augmente la qualité de la segmentation. On a supposé dans la partie précédente que le groupe 6 représentait la fusion des groupes 4 et 6 issus de la segmentation par méthode classique, générant ainsi un groupe dont les déplacements sont nuls en fin de semaine et faibles en semaine. Fortement différent du groupe 5, il en résulte que l'indicateur oscille autour de la valeur 4.

4.3.4.2 Indicateur de qualité intra euclidien

L'indicateur euclidien montre la moyenne euclidienne des distances entre les individus d'un groupe et son centre. Dans un espace euclidien de dimension 7, la valeur EUC définit le rayon de l'hypersphère moyenne, autour duquel se trouve la population du groupe. Plus ce rayon est important, plus le groupe englobe des individus éloignés et moins la qualité de la segmentation est bonne.

Afin d'effectuer une analyse analogue à la méthode classique, on a ici aussi représenté les différents usagers appartenant au groupe 2 pendant la semaine du 12 au 18 février 2012 (Figure 4-26). Ne pouvant pas représenter facilement l'hypersphère en dimension 7, il s'agit d'une projection sur deux dimensions : mardi et mercredi. Chaque carte est représentée par un point. Par souci de superposition, on a placé à droite des points le nombre d'individus de même comportement. Le centre du groupe 2 est défini par la croix rouge. On a vu que les individus aux faibles valeurs sur mardi et mercredi appartiennent désormais au groupe 4 : c'est pourquoi le centre se rapproche mieux du point (2,2) que dans la méthode classique. La valeur de l'indice EUC définit le rayon du cercle rouge, symbole de la distance moyenne euclidienne des points au centre. Ce cercle n'est donc que la projection de l'hypersphère sur les dimensions mardi et mercredi, dont on cherche à minimiser le rayon.

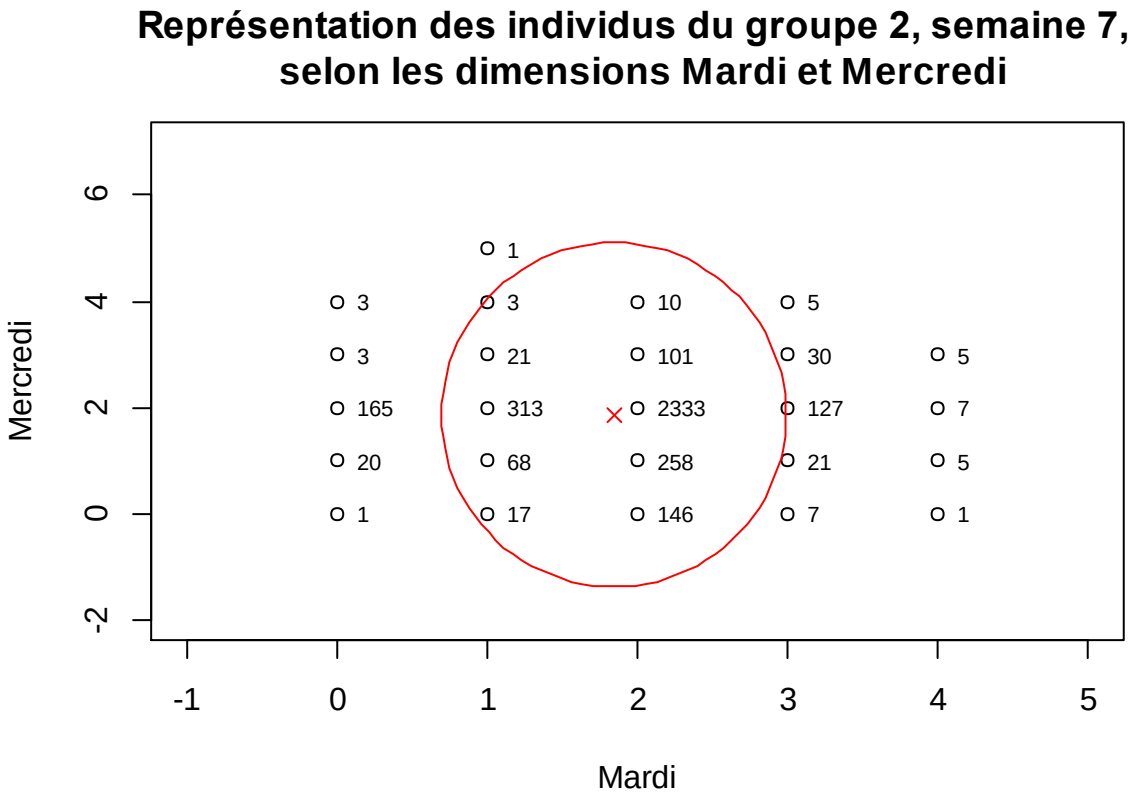


Figure 4-26 : Représentation des individus du groupe 2, semaine 7, sur mardi et mercredi

Sachant cela, il est donc plus intéressant d'obtenir des groupes compacts puisque le centre du groupe définit mieux les individus. La population variant le long de la période d'étude, l'indicateur EUC fournit des résultats différents sur chacune des semaines et la Figure 4-27 présente donc son évolution comparée sur chacun des groupes. Toutes les évolutions ci-contre sont plus ou moins influencées par une saisonnalité annuelle ainsi que par les semaines à jour férié. On repère trois types de courbes :

Le premier, représenté par les groupes 4 et 5, forment les courbes dont la valeur de l'indicateur est le plus haut. Ce résultat est attendu puisqu'on supposait que ces groupes récupéraient les individus très éloignés du comportement moyen. Le groupe 4 semble constant selon la saison, mais suggère des pics de meilleurs résultats sur les semaines à congé, tandis que le groupe 5 semble variable selon la saison. En effet, une valeur plus haute de l'indicateur dans les périodes d'été suggère une diminution de la qualité de ce groupe. Ce résultat s'explique par le fait que moins de déplacements sont à remarquer en période estivale et donc que les disparités à l'intérieur du groupe prennent plus d'importance.

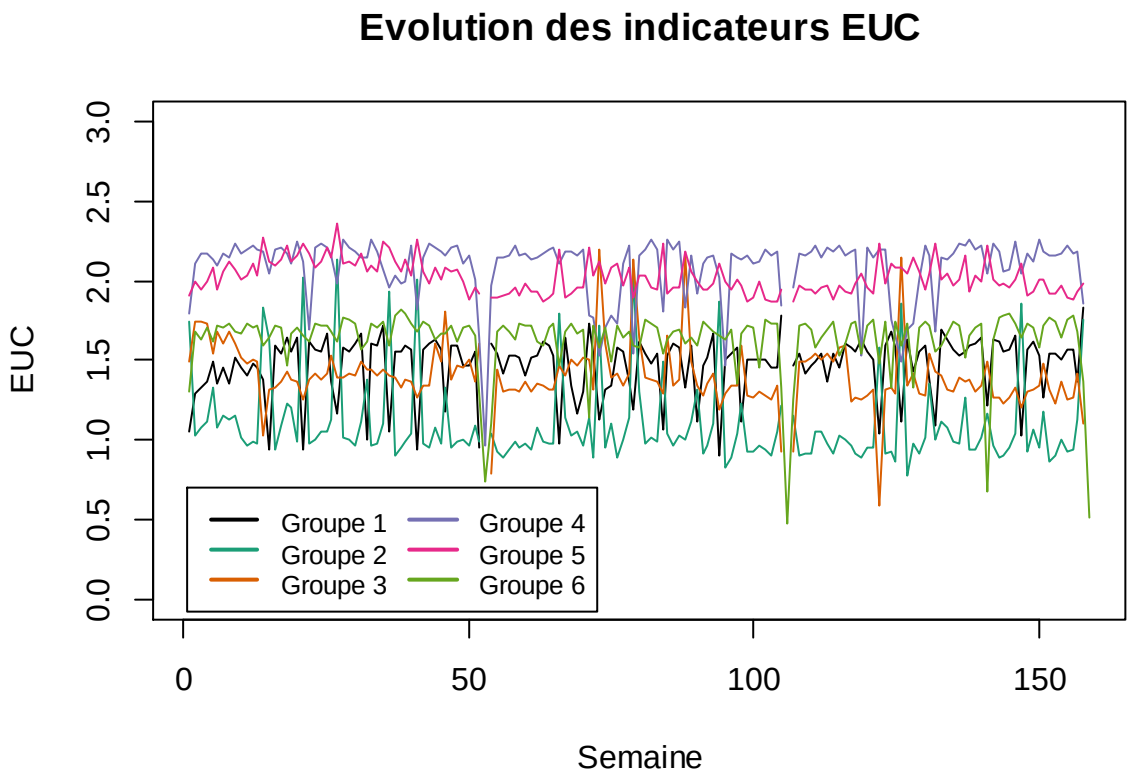


Figure 4-27 : Comparaison de l'évolution de l'indicateur EUC appliqué à chaque groupe

Le second type de courbe est composé des groupes 1, 3 et 6, groupes donc l'évolution de l'indicateur EUC dévoile une distance intra faible et donc une bonne qualité de segmentation. D'un côté, le groupe 1 suit une très faible saisonnalité où une amélioration de EUC survient lors des lundis fériés. En effet, considérablement augmenter la population de valeurs acceptables pour l'approximation permet ainsi de diminuer l'importance des disparités et donc d'améliorer la valeur de l'indicateur. On peut remarquer le même comportement pour le groupe 3 lors des vendredis fériés, tandis que le groupe 6 présente une qualité intragroupe quasi constante.

Le groupe 2 forme le type de courbe dont la valeur EUC est la plus faible. Inversement aux groupes 1 et 3 elle perd énormément en qualité puisqu'elle leur cède une grande partie de sa population.

4.3.4.3 Indicateur de qualité intra relatif

L'indicateur relatif montre la moyenne des distances relatives entre les individus d'un groupe et son centre. Il s'agit d'un calcul d'erreur moyen pour chaque usager quant à l'approximation de cet usager par le centre du groupe auquel il appartient. La courbe présentée en Figure 4-28Figure 4-14 représente l'évolution de la moyenne de l'indicateur REL des différents groupes, pondérée par le pourcentage de population que représente chacun des groupes.

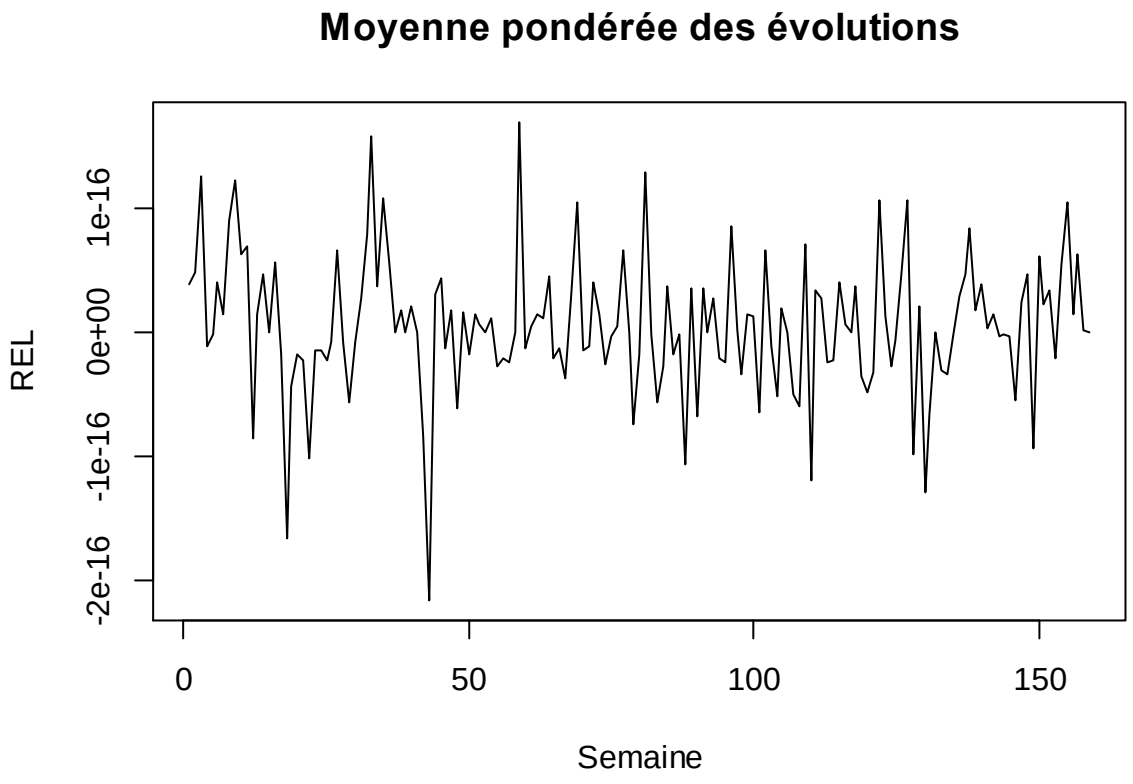


Figure 4-28 : Moyenne des évolutions de l'indicateur REL, pondérée par la population

Puisqu'une segmentation K-Moyenne est effectuée chaque semaine, de nouveaux centres sont définis chaque fois. Comme un centre d'un groupe est le barycentre des individus de ce groupe, on devine que les erreurs d'approximation sont extrêmement faibles. L'indicateur montre donc une valeur de REL constante à 0, définissant une qualité intra relative parfaite.

4.3.4.4 Critère de Dunn

Le critère de Dunn est égal au ratio des pires valeurs d'indices, c'est-à-dire qu'il divise la plus petite distance intergroupe par la plus grande distance intragroupe. Par cela, il définit un résumé des indicateurs quant à la qualité de la segmentation. Ayant des valeurs variables en distances intragroupes et intergroupes, on suit donc l'évolution du critère de Dunn sur les 159 semaines de la période d'étude (Figure 4-29Figure 4-15). Rappelons que plus la valeur du critère est haute, plus le résultat de segmentation est fidèle à la réalité (bonne qualité de segmentation).

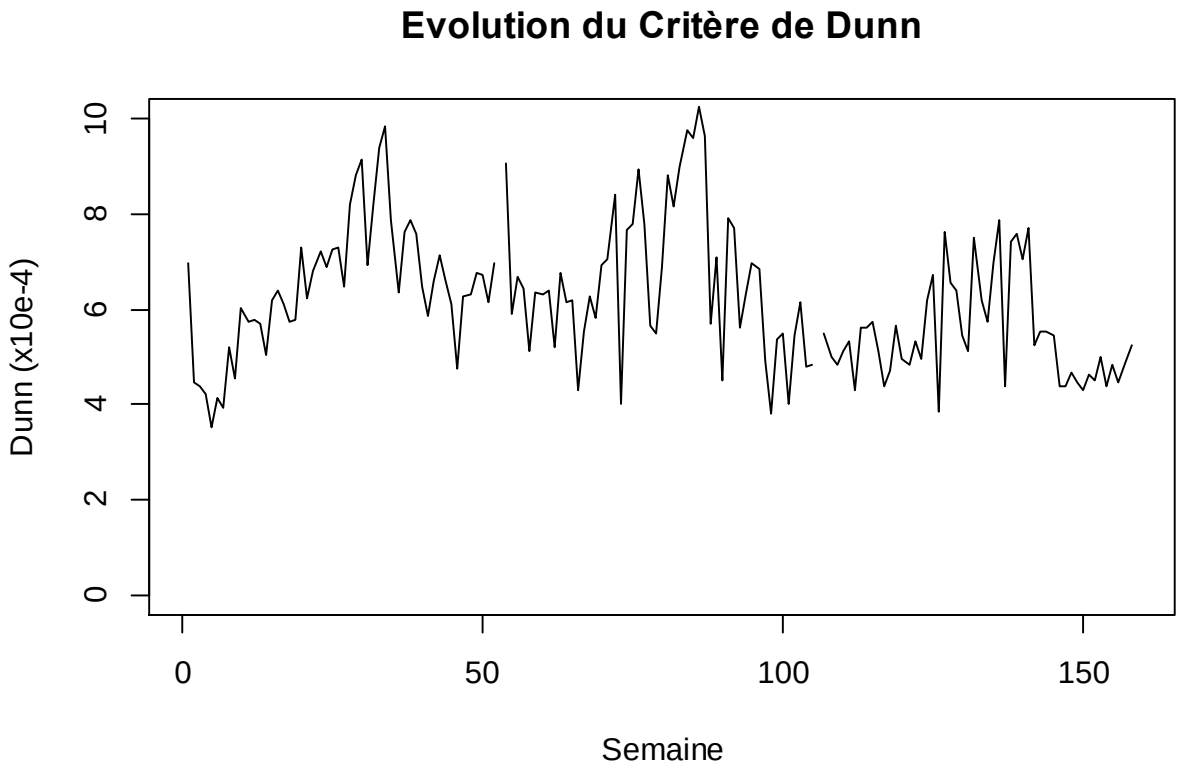


Figure 4-29 : Évolution du critère de Dunn

Fortement influencée par la période de l'année, la valeur de l'indicateur suit une saisonnalité où les maximums se trouvent aux périodes d'été et les minimums en hiver, tandis que la variation globale du nombre de déplacements produit le résultat inverse. On peut conclure que plus il y a d'usagers, plus il y a de patrons de déplacements sensiblement différents de notre approximation par segmentation, et donc, moins bonne est la qualité de cette méthode.

La décomposition de la courbe d'évolution du critère de Dunn pour cette méthode présente une saisonnalité permettant d'augmenter la valeur du critère de 4×10^{-4} en période estivale. On peut confirmer l'influence de la population sur le critère, hypothèse proposée dans l'analyse de la

méthode classique. En effet, le nombre de déplacements Adultes + Desfire (Tableau 4-1) diminue très légèrement entre 2012 et 2013 provoquant une très légère hausse de la tendance, tandis que la forte augmentation de ce nombre de déplacements en 2014 résulte en une décroissance de la tendance et donc une diminution de la qualité de segmentation.

Un des points positifs de la méthode est mis en évidence: il s'agit d'une amélioration conséquente de la qualité de segmentation lorsque l'on traite les semaines à congé. En effet, les chutes de valeurs relevées dans la méthode classique lors des jours fériés sont quasi inexistantes dans les résultats de cette méthode. Et lorsqu'elles apparaissent tout de même, leur effet est estompé.

4.3.5 Analyse de la stabilité de la population

L'indicateur défini selon l'Équation 3.8 est appliqué sur chacune des cartes à puce. Il s'agit d'un critère d'instabilité où plus la valeur est faible plus l'individu en question a un comportement stable, c'est-à-dire un patron de déplacement qui change peu. Cette modification de comportement se traduit par un changement de groupe associé, un individu qui change souvent de groupe possède donc un WSI plus important.

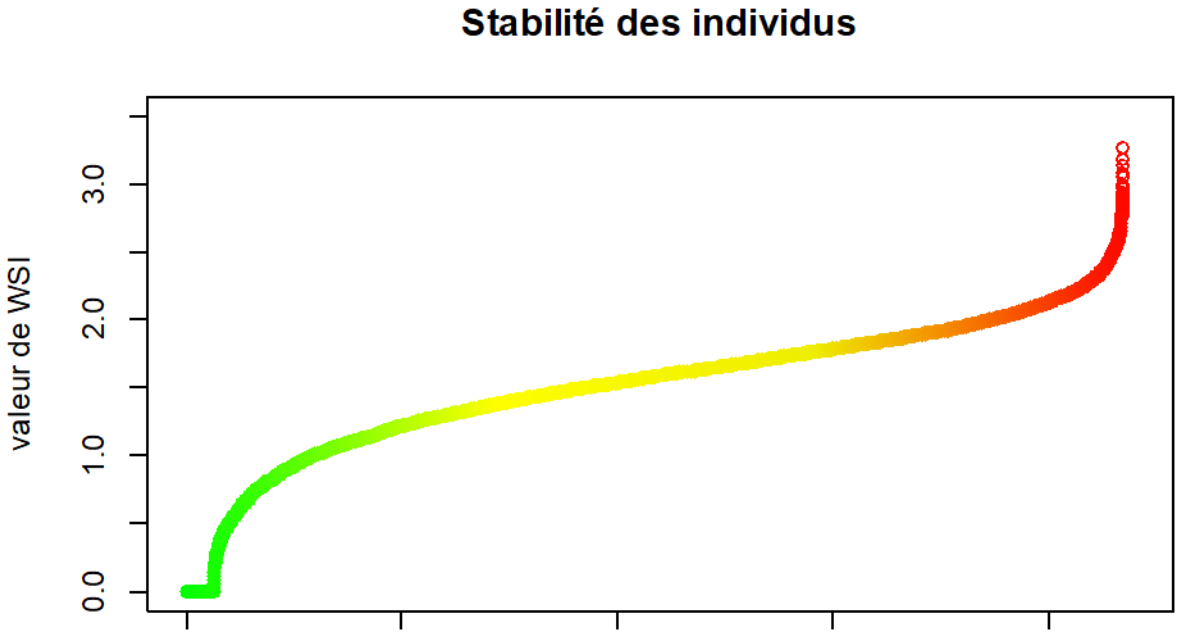


Figure 4-30 : Stabilité des individus par la méthode expérimentale

La Figure 4-30 montre l'indicateur de stabilité associé à chacun des individus de la population ayant réalisé des déplacements sur plus de 4 semaines différentes. Chaque point de couleur représente un usager. En vert, il s'agit de ceux qui ont un comportement très stable et qui changent peu de groupes. En jaune sont représentés ceux dont le comportement est moyennement stable, et en rouge, les comportements très instables puisqu'ils se situent sur une pente de la courbe qui désignent principalement les usagers qui changent constamment de comportement. Cette figure sert principalement de support visuel à l'étude de la distribution des valeurs.

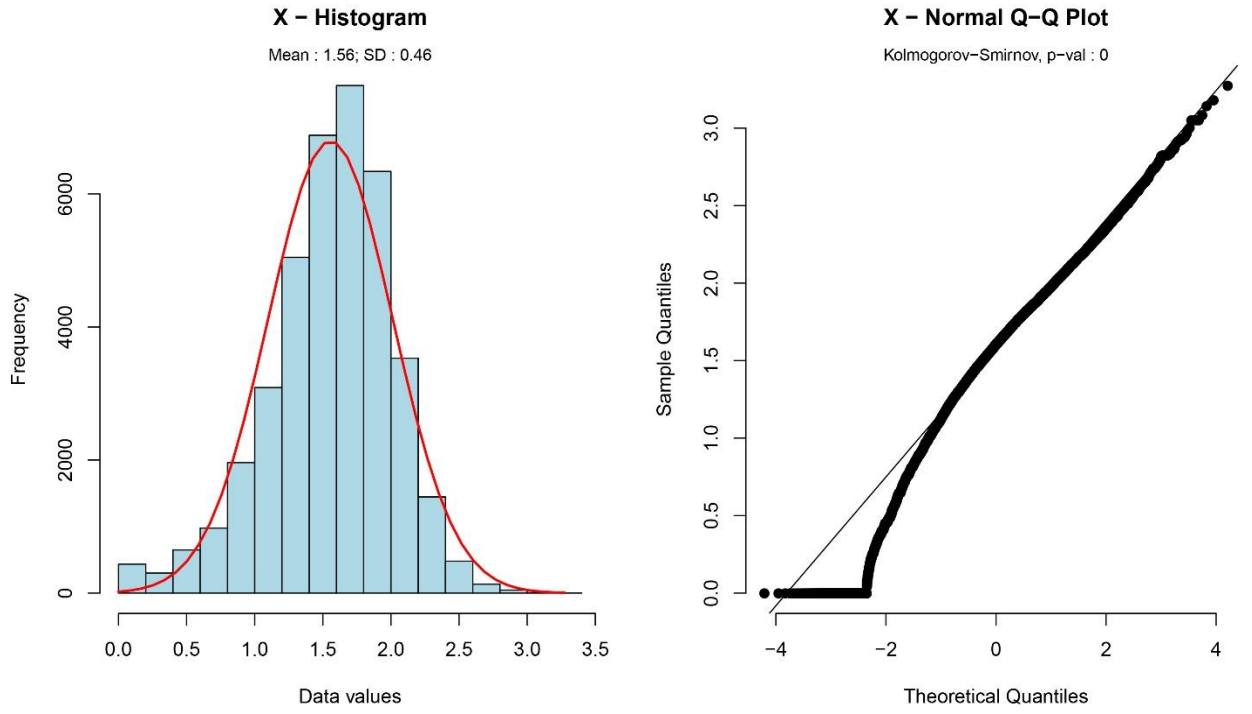


Figure 4-31 : Test Kolmogorov-Smirnov sur la stabilité des individus, méthode expérimentale

La Figure 4-31Figure 4-17 représente l'ensemble des individus ayant réalisé des déplacements sur au moins quatre semaines différentes. Un test de Kolmogorov Smirnov est appliqué pour comparer l'ensemble des WSI_i à une loi normale. Même si la répartition des valeurs semble plus ou moins suivre la loi normale sur l'histogramme, l'hypothèse est rejetée par le test (p-val=0). En effet, la distribution paraît légèrement asymétrique, avec une queue à gauche, ce qui signifie que les valeurs faibles sont plus nombreuses qu'elles ne le sont en distribution normale. Pour comprendre ce résultat, il s'agit d'effectuer l'analyse de la courbe quantile-quantile (Figure 4-31 courbe de droite).

Pour rappel : chaque carte à puce ayant réalisé des déplacements sur plus de quatre semaines est représentée par un point. Sur ce graphique la distribution des WSI est représentée suivant selon une loi normale afin de témoigner de ses défauts. Il s'agit ici de comparer la position des points par rapport à la droite normale.

Pour les quantiles théoriques inférieurs à -1 (valeur réelle : 1,13), les valeurs réelles se situent en dessous de la droite normale signifiant une très forte présence de valeurs WSI faibles par rapport à une distribution normale. On en déduit donc qu'une partie significative de la population a une valeur d'indicateur montrant une excellente stabilité comportementale, tandis que pour les quantiles théoriques supérieurs à -1, les individus se placent sur la droite normale et donc suivent cette loi. On donc peut très facilement caractériser les individus stables des individus moins stables par la simple lecture de cette courbe quantile-quantile.

4.4 Discussion sur les méthodes

Il est à noter que, pour une comparaison efficace des méthodes, l'ensemble des algorithmes sont simulés sur la même machine dotée d'un processeur Intel Core i5 5200U CPU 2,20 GHz avec 8 Go de mémoire vive. La programmation s'est effectuée sur R, utilisant notamment la fonction KMeansrcpp du package ClusterR pour effectuer les segmentations. Ainsi, l'ensemble des temps proposés de calcul dans cette partie sont comparables.

4.4.1 Les choix des paramètres extérieurs

L'application de ces deux méthodes permet une meilleure compréhension de la population que sert la STO, notamment dans les habitudes de trajet des adultes régulier de Gatineau. Afin d'obtenir des résultats évaluables et cohérents, de nombreuses applications ont été simulées avec des paramètres différents. Il s'agit dans cette partie de montrer le chemin qui a permis la sélection des différents paramètres pour les algorithmes.

La Figure 4-32 présente les choix effectués sur les données brutes pour la simulation. Il est à noter que chacune des sorties de l'organigramme a été simulée. Dans un tout premier temps, l'analyse s'est portée sur l'étude des transactions pour l'année 2013 (A). Après avoir effectué des statistiques globales sur ces données, il apparaît que certains usagers valident leur carte beaucoup plus que leurs congénères. En effet, la validation du titre de transport est nécessaire à chaque correspondance

à Gatineau. Un premier problème apparaît ici : un usager qui prend une correspondance se comporte-t-il réellement différemment de celui qui prend un trajet direct ? Supposant une réponse négative, il s'agit donc de s'intéresser aux déplacements, critère qui définit mieux le comportement de l'individu (B). Le choix de traiter uniquement les déplacements permet de réduire la taille de la base de données et donc la mémoire et le temps de calcul. Après simulation, des premiers résultats apparaissent quant à la répartition de la population dans les groupes. Des suppositions sur les variations saisonnières et des fêtes sont émises, mais aucune ne peut être vérifiée si les mêmes périodes ne sont pas comparées annuellement. De plus, il apparaît qu'un changement dans la dénomination des cartes à puce à la fin de l'année 2013 a été mis en place, empêchant tout suivi d'un individu sur un an. Les résultats sont tout de même analysables et cohérents, mais le manque de recul se fait sentir. Il s'agit finalement de traiter les déplacements des usagers sur trois années (D), puisque travailler avec les transactions (C) biaise les résultats. Ces bases de données demandent donc beaucoup plus de mémoire et de temps de calcul, mais fournissent des résultats complets et cohérents.

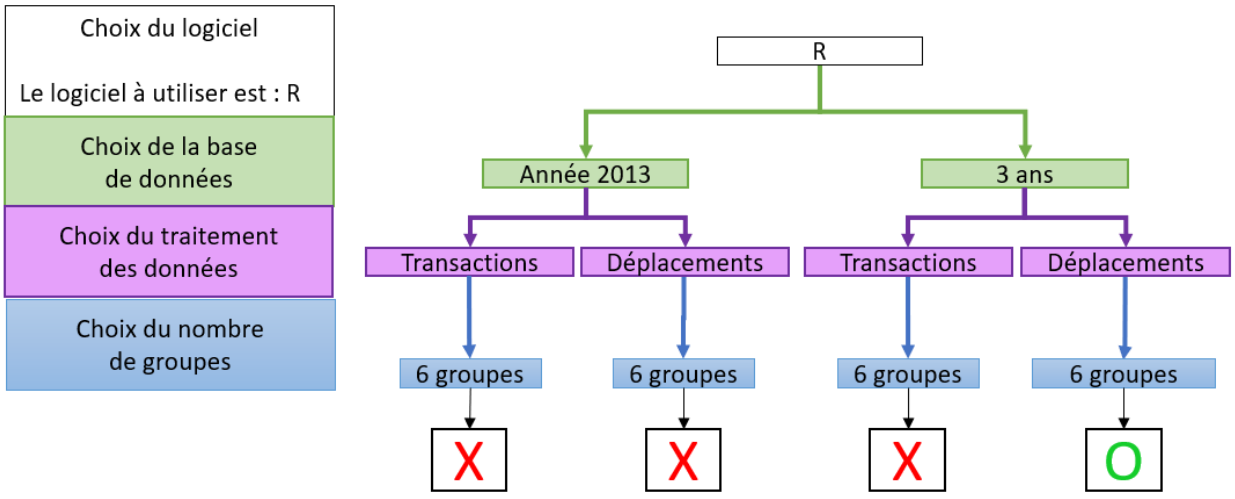


Figure 4-32 : Organigramme des choix d'étude

4.4.2 Comparaison des méthodes

Les deux méthodes analysent les mêmes vecteurs comportements à partir du même algorithme de segmentation. Seuls la logique et les paramètres de l'algorithme sont différents et pourtant, la méthode expérimentale permet de suivre temporellement l'évolution des comportements. Chacune des méthodes est régie par une hypothèse de départ : pour l'un, les comportements des usagers sont fixes, alors que pour l'autre ils évoluent dans le temps. On note que les groupes, résultats de la segmentation à exploiter, sont très similaires dans les deux méthodes. D'après les analyses précédentes, les groupes 1, 2, 3 et 5 représentent à peu près les mêmes populations dans les deux méthodes tandis que le groupe 6 de la méthode expérimentale représente la fusion des groupes 4 et 6 de la méthode classique. Suivre les évolutions de comportements de ces populations apporte une dimension très intéressante dans l'analyse de résultats.

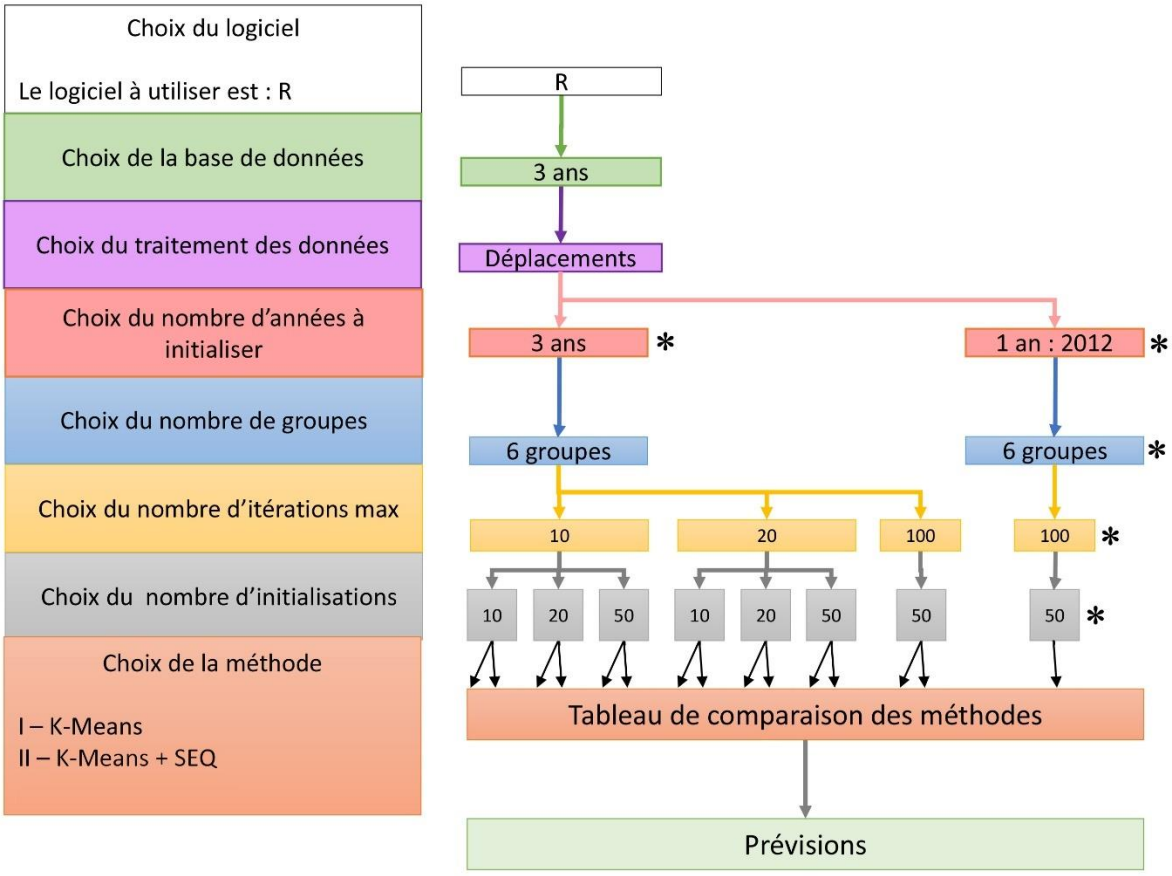


Figure 4-33 : Organigramme des choix des paramètres d'étude

Afin de comparer la viabilité et l'intérêt des résultats des deux méthodes, de nombreuses simulations sont effectuées en modifiant les valeurs des paramètres intrinsèques à la segmentation. La Figure 4-33 montre l'ensemble des combinaisons possibles. Chacune est simulée par les deux méthodes, seules celles présentant un astérisque à leur droite montrent des simulations supplémentaires uniquement traitées par la méthode expérimentale. Cette comparaison générale permet notamment de vérifier l'influence des paramètres sur les résultats de segmentation. Le Tableau 4-7 présente l'ensemble des valeurs des indicateurs moyens ainsi que du temps de calcul de segmentation des deux méthodes en fonction des valeurs des paramètres en jeu.

Tableau 4-7 : Tableau comparatif des résultats

Entrée			Sortie			
Nombre d'itérations	Nombre d'initialisations	Méthode	Temps de calcul (s)	EUC moyen	REL moyen	WSI moyen
10	10	Classique	32.7	1.49730	-1.61E-02	1.4631
		Expérimentale (3 ans appris)	37.9	1.39338	3.91E-18	1.4667
	20	Classique	65.7	1.49730	-1.61E-02	1.4926
		Expérimentale (3 ans appris)	70.5	1.39338	3.91E-18	1.5145
	50	Classique	160.7	1.49002	-1.78E-02	1.5556
		Expérimentale (3 ans appris)	166.1	1.41676	4.60E-18	1.4959
20	10	Classique	61.5	1.49728	-1.61E-02	1.4632
		Expérimentale (3 ans appris)	66.5	1.39338	3.91E-18	1.4667
	20	Classique	121.7	1.49728	-1.61E-02	1.4632
		Expérimentale (3 ans appris)	126.8	1.39338	3.91E-18	1.4667
	50	Classique	301.1	1.48995	-1.77E-02	1.5556
		Expérimentale (3 ans appris)	306.9	1.41676	4.60E-18	1.4178
100	50	Classique	670.8	1.48995	-1.77E-02	1.5741
		Expérimentale (3 ans appris)	678.2	1.41676	4.60E-18	1.5142
100	50	Expérimentale (1 an appris)	221.6	1.41673	3.71E-18	1.4104

Dans un premier temps, les deux méthodes sont simulées successivement avec comme paramètres variables : le nombre d'itérations maximum et le nombre d'initialisations. Ces simulations permettent de vérifier la viabilité des résultats et d'analyser la qualité de la segmentation en fonction de l'évolution des valeurs des paramètres. Comme le montre la littérature, en théorie, plus ces valeurs sont importantes, meilleure est la qualité de segmentation en dépit d'un temps de calcul

logiquement plus long. La comparaison des indicateurs EUC de la méthode classique pour chacune des valeurs des paramètres d'entrée confirme l'affirmation précédente. Ainsi, pour un nombre d'itérations fixe, passer de 10 à 50 initialisations permet une augmentation de la qualité de 0,5% en dépit d'un temps de calcul 5 fois plus long. D'un autre côté, passer de 10 à 100 itérations pour un nombre d'initialisations fixe ne permet une augmentation de qualité que de 7×10^{-3}% pour un temps de calcul 4 fois plus long. Ces améliorations, n'étant que peu significatives, ne sont donc pas rentables à la vue du temps de calcul supplémentaire nécessaire à leurs mises en place. La méthode expérimentale, en revanche développée pour améliorer cette qualité, fournit des résultats plus appréciables. En effet, il apparaît dans un premier temps que l'erreur relative est toujours bien plus faible qu'avec la méthode classique. Mais plus intéressant, la méthode expérimentale permet une amélioration de la qualité (EUC) de 5 à 7% pour un temps de calcul de 5 à 7 secondes, montrant qu'en presque 38 secondes elle fournit une segmentation de bien meilleure qualité que la méthode classique en plus de 11 minutes. On ne peut cependant pas conclure sur l'influence des paramètres d'entrée pour la méthode expérimentale. Le nombre d'itérations et le nombre d'initialisations sont des acteurs directs de la méthode classique, mais seuls ses résultats sont utilisés dans la méthode expérimentale. La qualité de segmentation dépend donc uniquement de la position des groupes classiques. On peut émettre l'hypothèse que, puisque l'indicateur REL définit l'erreur relative, plus sa valeur est haute en méthode classique moins la qualité de la méthode expérimentale sera bonne.

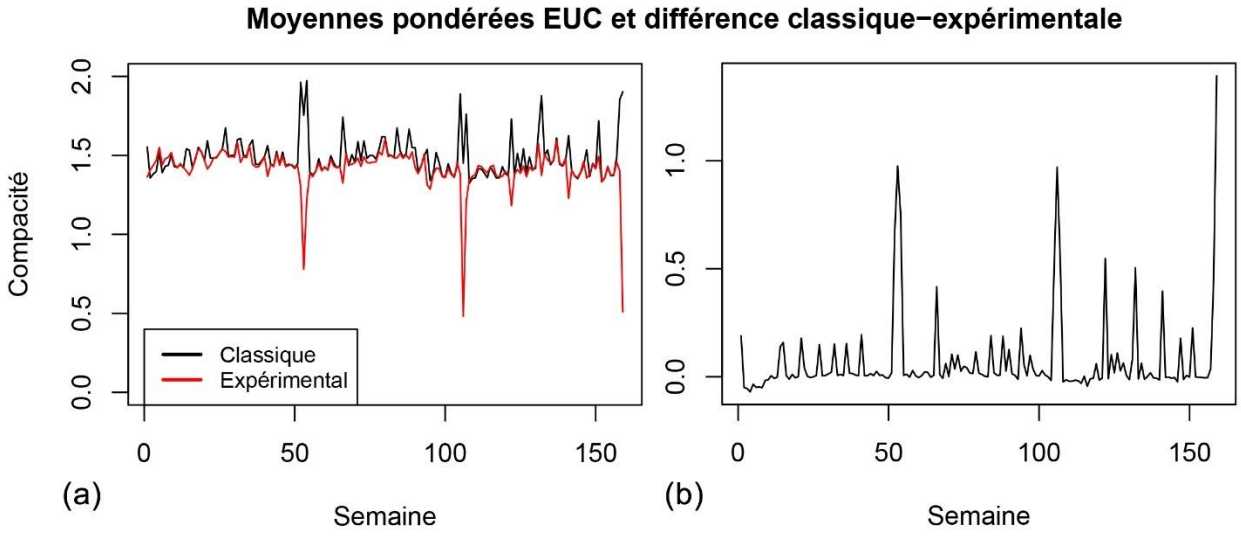


Figure 4-34 : Évolution de la moyenne des EUC des groupes pondérés par leur population (a) et différence des méthodes (b)

Il s'agit maintenant de comprendre comment se traduit cette amélioration en qualité. La Figure 4-34 représente l'évolution de l'indicateur EUC suivant la méthode et leur différence, avec les paramètres nombre d'itérations = 100, nombre d'initialisations = 50 et 3 ans d'apprentissage. Ces courbes montrent finalement que les deux méthodes se valent sur la plupart des semaines, mais que la méthode expérimentale est nettement plus appréciable sur les semaines à congé. En effet, l'analyse des résultats a montré que la méthode classique présentait de grosses difficultés sur ces semaines ainsi que sur les semaines de fin d'années. La courbe de l'évolution de la différence des indicateurs des deux méthodes montre clairement ces améliorations ponctuelles.

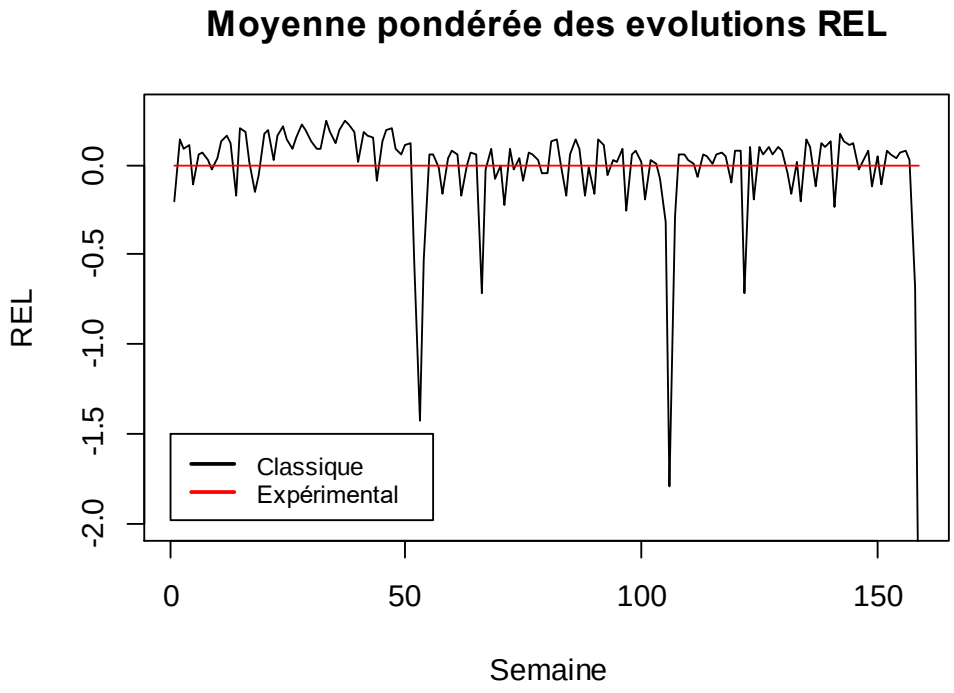


Figure 4-35 : Moyenne de l'indicateur REL pondéré par la population des groupes

La Figure 4-34Figure 4-35 représente l'évolution de l'indicateur REL suivant la méthode, avec les paramètres nombre d'itérations = 100, nombre d'initialisations = 50 et 3 ans d'apprentissage. L'analyse de cette courbe comparative complète la compréhension de l'amélioration en qualité. La méthode expérimentale ne produit que des groupes fixes dans le temps, leur centre n'est donc jamais parfait puisque l'approximation se réalise en mélangeant l'ensemble des semaines. A contrario, la méthode expérimentale est approximée à chaque semaine limitant ainsi les erreurs réelles d'approximation.

Il s'agit ensuite de comprendre comment se traduit l'évolution de l'indicateur de stabilité WSI. Une comparaison directe des moyennes est impossible puisqu'il s'agit de distributions, des tests statistiques sont nécessaires pour cela. On a vu dans les parties précédentes que les distributions des deux méthodes ne suivent pas de loi normale, un test de Wilcoxon est donc préférable à celui de Student.

```
> wilcox.test(stab_classique,stab_experimentale,
+             alternative = "greater", exact =FALSE)

        Wilcoxon rank sum test with continuity correction

data:  stab_classique and stab_experimentale
W = 1022900000, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

Figure 4-36 : Test de Wilcoxon appliqué sur R

La Figure 4-36 présente la comparaison des distributions en stabilité avec les paramètres suivants : 100 itérations maximum, 50 initialisations et 3 ans d'apprentissage. Le résultat montre, comme prévu, une plus grande stabilité dans la méthode expérimentale. Cela signifie simplement que les groupes sont agencés de sorte qu'il y ait moins de déplacements d'usagers entre les groupes.

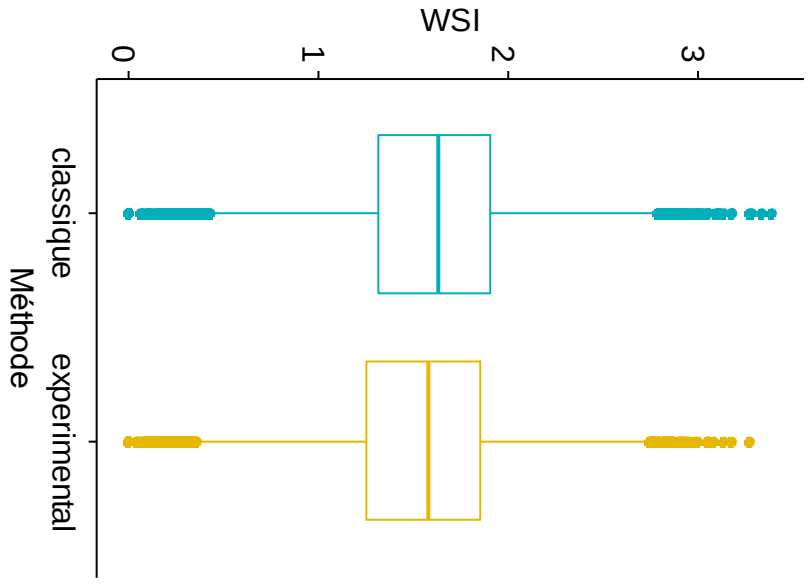


Figure 4-37 : Diagramme boîte comparaison en stabilité

Ce résultat était d'autant plus visible sur la Figure 4-37, puisqu'elle présente une comparaison des distributions sous forme de diagramme en boîte. L'indicateur doit cependant être privilégié pour l'analyse d'une distribution et non d'une comparaison, puisqu'il permet de décrire les comportements stables à l'intérieur de cette même distribution. Il est intéressant de le mettre en lien avec la qualité de segmentation : si une segmentation est de mauvaise qualité, mais fournit une valeur moyenne de WSI faible, cela veut dire que la simulation actuelle propose une approximation erronée de la population et que les mesures prises quant à la variation des comportements des usagers ne sont pas adaptées à la population réelle.

4.5 Prévisions

Une bonne prévision de flux de voyageurs est souvent très appréciée par les planificateurs. En effet, les exploitant peuvent se baser sur les prévisions afin d'anticiper une modification comportementale des usagers due à un événement extérieur qui s'est déjà produit. On peut citer comme exemple, l'arrivée d'un jour férié dans notre cas, ou même les conditions météorologiques difficiles dans certains modèles. Traditionnellement, les études se concentrent sur la prévision de la demande, mais la méthode classique ne fournissant qu'une mauvaise approximation de la population empêche des estimations précises. Il apparaît un avantage de la méthode expérimentale : rendant possible l'évolution des caractéristiques des groupes à travers le temps, elle crée une opportunité pour la prévision de ces évolutions.

Les prévisions sont exécutées à partir d'algorithmes d'approximations par lissage exponentiel définis en 3.6. Il s'agit tout d'abord d'apprendre les centres et populations des 139 premières semaines, puis d'appliquer une prévision à court terme de quatre semaines et une prévision à long terme de vingt semaines. Notons ainsi que, à long terme les semaines estimées vont du 18 août 2014 à la fin de l'année, tandis que le court terme s'arrête au 14 septembre 2014. D'après la méthodologie, deux critères sont mis en place afin d'évaluer les erreurs de prévision. D'un côté, le critère RMSE est appliqué sur l'estimation des centres ainsi que de la population. D'un autre côté, le critère MAPE est calculé uniquement sur la population.

L'Annexe A présente l'ensemble des prévisions à long terme effectuées sur les données issues de la méthode expérimentale. Cette annexe présente ainsi un support visuel comparatif aidant à la compréhension des analyses des indicateurs d'erreurs.

Tableau 4-8 : Comparaison des erreurs des méthodes de prévision sur les populations

		Population court terme 4 semaines				Population court terme 20 semaines			
		Méthode expérimentale		Méthode classique		Méthode expérimentale		Méthode classique	
		MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Holt Winter Additif	Groupe 1	31%	461	75%	790	23%	435	1175%	885
	Groupe 2	23%	1003	114%	1188	31%	633	62%	637
	Groupe 3	9%	168	157%	300	23%	651	48%	278
	Groupe 4	14%	116	25%	618	12%	127	13%	819
	Groupe 5	13%	123	12%	144	10%	116	10%	193
	Groupe 6	10%	282	71%	863	11%	375	50%	923
Holt Winter Multiplicatif	Groupe 1	39%	541	119%	1611	26%	483	60%	968
	Groupe 2	25%	750	75%	1166	46%	1056	51%	1310
	Groupe 3	9%	103	248%	470	17%	524	86%	705
	Groupe 4	16%	123	24%	651	13%	129	15%	808
	Groupe 5	16%	143	10%	129	10%	112	10%	192
	Groupe 6	9%	273	90%	1136	10%	310	70%	1327

Le Tableau 4-8 compile l'ensemble des indicateurs calculés pour l'estimation à court et long terme des populations issues de la méthode classique et expérimentale. Le code couleur utilisé montre en bleu les faibles erreurs et en rouge les fortes erreurs de prévisions. Il apparaît que l'estimation de la population issue de la méthode classique est fortement erronée. En effet, on a vu précédemment l'influence négative des semaines à congé sur l'évolution de la population des groupes, constituant un bruit très difficile à suivre. Les algorithmes de prévision sont certes capables de déterminer la position des semaines à risque, mais sont loin d'estimer la variation réelle. La semaine du 1er septembre (congé au Canada) provoque ainsi des difficultés dans les estimations de courte durée. Moins dépendante aux variations provenant des semaines à congé, la population issue de la méthode expérimentale semble être bien plus facile à prédire. En effet, que ce soit à court ou long terme, les estimations semblent de bonne qualité avec une légère avance pour la méthode Holt-Winters additif. Il est à noter que les variations plus importantes dans le groupe 2 expérimental provoquent une moins bonne fiabilité de prévision sur le long terme.

Tableau 4-9 : Comparaison des erreurs des méthodes de prévision sur les groupes

			Coordonnées de centres							
			Dim.	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.	Moyenne
	MTH	Critère	RMSE							
Groupes court terme : 4 semaines	Holt Winter Additif	Groupe 1	0.04	0.12	0.55	0.02	0.18	0.26	0.04	0.17
		Groupe 2	0.03	0.51	0.30	0.35	0.11	0.29	0.03	0.23
		Groupe 3	0.43	0.50	0.78	0.28	0.43	0.08	0.07	0.37
		Groupe 4	0.34	0.32	0.59	0.25	0.34	0.22	0.10	0.31
		Groupe 5	0.11	0.35	0.25	0.21	0.31	0.28	0.11	0.23
		Groupe 6	0.08	0.29	0.55	0.18	0.22	0.20	0.15	0.24
	Holt Winter Multiplicatif	Groupe 1	0.03	0.13	0.87	0.02	0.17	0.30	0.04	0.21
		Groupe 2	0.03	0.57	0.28	0.35	0.11	0.29	0.03	0.24
		Groupe 3	0.07	0.50	0.76	0.29	0.43	0.10	0.23	0.34
		Groupe 4	0.36	0.39	0.39	0.24	0.33	0.21	0.46	0.34
		Groupe 5	0.19	0.13	0.22	0.19	0.31	0.23	0.12	0.20
		Groupe 6	0.08	0.27	0.84	0.21	0.28	0.24	0.10	0.29
Groupes long terme : 20 semaines	Holt Winter Additif	Groupe 1	0.10	0.10	0.56	0.17	0.42	0.34	0.07	0.25
		Groupe 2	0.09	0.24	0.53	0.42	0.11	0.16	0.15	0.24
		Groupe 3	0.20	0.24	0.57	0.30	0.47	0.13	0.09	0.28
		Groupe 4	0.29	0.24	0.41	0.23	0.26	0.35	0.33	0.30
		Groupe 5	0.23	0.19	0.42	0.39	0.42	0.27	0.24	0.31
		Groupe 6	0.04	0.27	0.47	0.46	0.12	0.12	0.08	0.22
	Holt Winter Multiplicatif	Groupe 1	0.08	0.10	0.88	0.17	0.42	0.32	0.07	0.29
		Groupe 2	0.09	0.32	0.51	0.40	0.10	0.16	0.12	0.24
		Groupe 3	0.07	0.24	0.56	0.30	0.45	0.14	0.10	0.27
		Groupe 4	0.31	0.29	0.50	0.28	0.27	0.34	0.63	0.37
		Groupe 5	0.22	0.51	0.29	0.23	0.42	0.25	0.24	0.31
		Groupe 6	0.04	0.31	0.53	0.45	0.17	0.21	0.05	0.25

Le Tableau 4-9 rassemble l'ensemble des erreurs RMSE calculées sur les évolutions des groupes issus de la méthode expérimentale. Le code couleur utilisé varie du vert au rouge correspondant aux bonnes et mauvaises estimations. Comme prévu, les erreurs sont relativement similaires sur court et long terme. D'après la littérature, l'algorithme Holt-Winters est connu pour être une méthode de prévision peu coûteuse avec une qualité acceptable sur long terme. Sans être parfait, l'algorithme fournit des estimations passables de l'évolution des caractéristiques des groupes. De manière générale, les moins bonnes estimations se font sur le mardi. L'analyse de l'Annexe A montre qu'il s'agit d'un jour où les variations de type bruit sont assez prononcées et donc difficiles à estimer correctement, tandis que sur les fins de semaine, puisque leur volume de déplacement est faible, les variations sont peu importantes permettant ainsi à l'algorithme de produire de bons résultats. En comparant les erreurs des deux horizons de prévision, l'influence du lundi 1er septembre apparaît uniquement dans le court terme. Il est établi que les groupes 2 et 3 sont influencés par ce type de semaine à congé, c'est pourquoi leur indicateur RMSE est très fort. A contrario, ce phénomène disparaît à long terme puisque cette mauvaise estimation est noyée par les bonnes. Il apparaît que l'algorithme Holt-Winters additif est légèrement plus fiable que son opposé. L'Annexe A permet de confirmer cette affirmation. Sur les différentes évolutions, la courbe bleue semble suivre le mieux la courbe noire.

Il faut donc retenir qu'au-delà de proposer une prévision des évolutions des groupes, la méthode expérimentale permet une bien meilleure estimation de la population de ses groupes. De plus, en comparant leurs résultats (indicateurs RMSE et MAPE), l'algorithme de prévision Holt-Winters Additif semble donner de meilleures approximations que l'algorithme Holt-Winters multiplicatif.

Il s'agit ici de la présentation et l'application de deux méthodes de prévisions encore peu utilisées dans le domaine des transports collectif. Montrer les forces et faiblesses de ces méthodes appliquées à notre cas appuient le potentiel de leur utilisation en transport collectif. Néanmoins la pauvre qualité des résultats laisse penser qu'il ne s'agit pas ici de la méthode type à implémenter par l'exploitant des données.

CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS

Dans le cadre de l'ouverture d'un projet en collaboration avec des entreprises exploitantes de transport en commun, l'objectif principal de ce projet était de proposer une méthode universelle d'étude comportementale des usagers à partir de l'exploitation de données issues de cartes à puce. Pour ce faire, l'étude a dû passer par trois sous objectifs définis ci-dessous :

- Mettre en place un algorithme traditionnel et fonctionnel permettant l'analyse comportementale des usagers.
- Développer un algorithme expérimental fonctionnel améliorant la méthode précédente, afin de pouvoir suivre l'évolution des comportements des usagers à travers le temps.
- Proposer une méthode de prévision des évolutions, enrichissant ainsi les connaissances apportées à la planification.

5.1 Synthèse des travaux

Dans un premier temps, une revue de littérature a été réalisée, permettant de montrer les composantes nécessaires à l'élaboration des méthodes et l'état des recherches menées à ce jour. On y retrouve une première étude cherchant à définir le domaine d'applicabilité que représente l'exploitation des cartes à puce. Par cela, de nombreux auteurs montrent l'énorme potentiel de cette technologie notamment dans la compréhension des comportements humains. Définissant une volonté de s'intéresser à l'étude des comportements hebdomadaires à l'échelle individuelle, cette revue présente également les méthodes existantes associées à la résolution de la problématique. Concernant les méthodes de segmentation, les caractéristiques nécessaires au choix du bon algorithme sont clairement définies. Différents processus usuels de fouille de données massives sont ensuite présentés, choisissant ainsi le mieux adapté aux vues de ces précédentes exigences. Dans le but de répondre au premier sous objectif, il a donc été choisi de continuer avec la méthode des k-moyennes, jugé plus efficace et simple à mettre en œuvre. On a introduit l'intérêt et la possibilité d'étudier l'évolution comportementale des individus à travers le temps, idée vers laquelle se tourne le deuxième sous objectif. Prévoir la demande constitue l'aboutissement de l'analyse du planificateur. Ainsi, diverses méthodes de prévisions usuelles en transport sont présentées et comparées dans le but de répondre au troisième sous objectif.

Dans un second temps, la méthodologie a présenté chronologiquement l'ensemble des étapes nécessaires à une analyse comportementale d'usagers, avec l'optique de répondre à ces trois sous objectifs. On y retrouve l'importation et la manipulation des données, deux étapes indispensables dans la transformation des informations « brutes » en des vecteurs exploitables par les algorithmes de segmentation. Deux méthodes de regroupement sont ensuite proposées : le premier processus, plus traditionnel applique une segmentation des k-moyennes sur l'ensemble des vecteurs comportement, permettant ainsi de repérer les habitudes des usagers sur l'ensemble de la durée de la période d'étude. Le second processus part des résultats précédents et applique, de manière incrémentale et chronologique, des segmentations k-moyennes sur chacune des semaines de l'étude. Cette dernière théorie devrait proposer des résultats de comportements évoluant dans le temps. La méthodologie présente ensuite une méthode d'analyse des résultats des segmentations à partir d'indicateurs. On y retrouve des indicateurs de qualité évaluant la fidélité de l'estimation et un indicateur de stabilité montrant la régularité de chaque individu. Finalement, cette partie se conclut par les explications mathématiques des principes de prévision par lissage exponentiel et les indicateurs d'évaluation de ces prévisions.

Finalement, il a été question d'expérimenter ces méthodologies sur un système de transport réel, les données de cartes à puce issues de l'utilisation du réseau exploité par la STO. De manière analogue deux études sont réalisées ici en lien avec les deux méthodes de segmentations proposées :

D'un côté, il s'agissait d'appliquer la méthode classique sur l'ensemble des trois années de transactions fournies par la STO. Les 35,4 millions de transactions sont filtrés et convertis-en 1,3 million de vecteurs comportements par l'algorithme de manipulation de données. Un nombre de six groupes est fixé à partir de la méthode du dendrogramme permettant le traitement de l'ensemble des vecteurs comportements simultanément par le processus de segmentation. On définit donc six habitudes différentes de comportement chez les adultes réguliers, dans lesquels on vient retrouver en grande partie les travailleurs réguliers. L'analyse montre une forte influence des semaines comprenant des jours fériés sur les résultats, générant ainsi un bilan global légèrement biaisé.

D’un autre côté, il s’agissait d’appliquer la méthode expérimentale sur l’ensemble des trois années de transactions fournies par la STO. Les résultats de la méthode classique servent à l’initialisation du processus de segmentation incrémental. On définit six habitudes différentes de comportements qui évoluent au fil du temps. La méthode expérimentale consistant en soit en un échantillonnage temporel optimisé, permet de traiter en moyenne 8000 vecteurs comportements par semaine. Une véritable contribution est mise en lumière puisque d’après la littérature, le processus des k-moyennes est plus efficace sur petites volumétries, c’est pourquoi, en divisant le volume des données par 159, on obtient des résultats de meilleure qualité. L’analyse montre une plus faible influence des semaines à congé et permet l’étude de l’évolution des groupes.

Les segmentations k-moyennes fournissent deux types de résultats : les centres des groupes qui représentent les habitudes moyennes des habitants de ce groupe et les tailles de populations associées. Désignant toutes deux des séries temporelles non-stationnaires avec présence de saisonnalité, on a comparé l’application de deux méthodes de prévision par lissage exponentiel. La forme additive étant la plus fidèle (erreurs de 10 à 30% suivant la population des groupes), il en ressort que la saisonnalité ne semble pas dépendre de la tendance.

Ce projet de recherche a permis de montrer que l’analyse comportementale de moyenne volumétrie (35,4 millions de transactions) peut se faire dans des temps très courts (30 à 40 secondes). En opposition, c’est la création des vecteurs comportements qui est très longue (15 heures de traitement pour l’ensemble des données). Sur ce point, l’aspect incrémental de la méthode expérimentale permet de contourner ce problème puisqu’il ne s’agit que d’interpréter les semaines une à une. La manipulation des données hebdomadaires ne prendrait plus que 6 minutes par semaine, un résultat très faible à la vue de la masse de données engagée.

Pour terminer dans la continuité des travaux présentés lors de la revue de littérature, ce travail démontre le potentiel considérable relatif à l’exploitation des données de cartes à puce et l’intérêt certain pour les sociétés de transports collectifs d’adopter cette technologie.

5.2 Limitations

On a vu dans la revue de littérature que la technologie des cartes à puces présentait des limitations dues à l'incomplétude de ses données. Les travaux d'enrichissement permettent notamment de pallier ce problème. Concernant les contraintes affrontées dans l'élaboration des méthodes et dans les résultats en jeu, il est important de les relever afin d'améliorer les méthodes.

Premièrement, ce travail de maîtrise s'inscrit dans une volonté d'établir une méthode universelle d'analyse. Les processus engagés doivent donc pouvoir s'adapter à tout type de volumétrie. C'est diverses méthodes sont proposées dans le but d'aider à définir un nombre de groupes. Existe-t-il un choix optimal ? Cette problématique prête, depuis longtemps, à discussion et est directement liée à la première limite du processus de segmentation : la nécessité d'intervention de l'utilisateur.

De plus, la méthode traditionnelle travaille ici sur un trop grand nombre de semaines biaisant ses résultats. L'étude cherchait tout d'abord à s'intéresser à l'analyse comportementale sur long terme. Cependant, trop de semaines de vacances sont en jeu faussant ainsi l'étude simultanée de toutes les semaines. Dans les périodes de fêtes de fin d'années par exemple, très peu de déplacements sont à observer et donc les mélanger avec les semaines normales empêche une détection précise des habitudes. La fonction d'apprentissage résout ce problème dans la méthode expérimentale.

D'un point de vue algorithmique, le processus de segmentation utilisé converge généralement vers un optimum local. La méthode usuelle pour améliorer sa qualité est de la simuler un grand nombre de fois à partir de noyaux aléatoires. Cependant, l'algorithme des K-Moyennes doit lire entièrement la base de données à chaque itération. Donc, si l'on travaille avec des données volumineuses comme ici, le temps de calcul devient donc exponentiellement plus long. À l'inverse, le processus expérimental travaille en temps fixe de 5 secondes pour les trois ans de données, puisqu'il n'effectue qu'une seule initialisation et ne traite que des échantillons réduits.

Concernant la méthode expérimentale, il est nécessaire de lui fournir des noyaux de départ optimisés. On a choisi dans cette étude de lui proposer les résultats de la méthode classique afin que l'algorithme cherche des minimums locaux autour de ces noyaux, rallongeant ainsi considérablement le temps de calcul global. Ce choix est justifié puisque les deux méthodes sont par cela comparables. D'autres noyaux auraient très certainement fourni des résultats différents qui n'auraient pas forcement la même interprétabilité.

Finalement, faire travailler le processus expérimental sur très long terme est un bon exercice. Il permet notamment d'évaluer la stabilité de l'algorithme. On a remarqué un échange de groupes sur les premières semaines de l'étude avant de que l'algorithme ne se stabilise de lui-même. La fonction d'apprentissage a été définie de la manière la plus simple possible, mais cette erreur montre qu'elle n'est pas parfaitement adaptée. Sa stabilisation reste tout de même un point positif.

5.3 Recherches futures et perspectives

L'étude proposée cherche à constituer une base solide pour l'élaboration d'une méthode universelle d'analyse efficace. La plupart des méthodes usuelles et applicatives pour chaque type de volumétries y sont présentées. Cependant différents points sont à améliorer :

Tout d'abord, on a remarqué d'énormes disparités dans les comportements des adultes réguliers. On pourrait donc se demander si certains usagers n'ont pas choisi le type tarifaire adapté. Pour témoigner plus clairement de ce phénomène, il s'agit d'appliquer les méthodes décrites dans ce mémoire sur les autres types tarifaires afin de comprendre leurs comportements. Si des similitudes apparaissent entre les groupes de plusieurs types tarifaires, cela signifie que les comportements d'usage de cartes sont proches, et donc que le choix de type tarifaire n'est pas adapté.

Ensuite, les vecteurs comportements, représentant l'utilisation hebdomadaire des usagers, fournissent des résultats peu détaillés. En effet, on ne sait pas s'il s'agit de trajets effectués le matin ou le soir. Un premier axe d'amélioration pourrait se tenir dans la forme de ces vecteurs. En divisant chaque journée en quatre parties, par exemple, on pourrait créer des vecteurs comportements de 28 composantes. Uniquement utilisable avec des fortes volumétries, la segmentation fournirait ainsi des informations supplémentaires sur les habitudes groupales, permettant une meilleure définition des populations engagées. Un second axe d'amélioration serait d'enrichir ces vecteurs d'informations spatiales. En effet, si la granularité des vecteurs comportement est suffisamment forte pour qu'aucun individu n'exécute plus d'un déplacement par tranche de journée, alors il est possible d'ajouter une information spatiale à la place de l'information binaire de déplacement. Par cela, l'algorithme pourrait repérer les habitudes spatio-temporelles des usagers. On doit retenir ici que plus le vecteur comportement sera de grande dimension, plus le processus de segmentation aura besoin de vecteurs pour proposer des résultats non biaisés.

BIBLIOGRAPHIE

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.https://doi.org/10.1109/TAC.1974.1100705

Ali, A., Kim, J., & Lee, S. (2016). Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, *20*(4), 1532-1539. https://doi.org/10.1007/s12205-015-1694-0

Antoniadis, A., Bigot, J., & von Sachs, R. (2009). A Multiscale Approach for Statistical Characterization of Functional Images. *Journal of Computational and Graphical Statistics*, *18*(1), 216-237. https://doi.org/10.1198/jcgs.2009.0013

Assaad, H. E. (2014). *Modélisation et classification dynamique de données temporelles non stationnaires* (phdthesis). Université Paris-Est. Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-01143904/document

Bagchi, M., & White, P. (2004). What role for smart-card data from bus systems? *Municipal Engineer*, *157*, 39-46.

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

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. https://doi.org/10.3141/1817-24

Bentoglio, G., Fayolle, J., & Lemoine, M. (2001). Unité et pluralité du cycle européen. *Revue de l'OFCE, no 78*(3), 9-73. https://doi.org/10.3917/reof.078.0009

Bonnel, P. (2002). *Prévision de la demande de transport* (thesis). Université Lumière - Lyon II. Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-00268919/document

Bougas, C. (2013). Forecasting air passenger traffic flows in Canada : an evaluation of time series models and combination methods. Consulté à l'adresse https://corpus.ulaval.ca/jspui/handle/20.500.11794/24286

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. https://doi.org/10.1016/j.trc.2017.03.021

Calabrese, A., & Paninski, L. (2011). Kalman filter mixture model for spike sorting of non-stationary data. *Journal of Neuroscience Methods*, *196*(1), 159-169. https://doi.org/10.1016/j.jneumeth.2010.12.002

Ceapa, I., Smith, C., & Capra, L. (2012). Avoiding the Crowds: Understanding Tube Station Congestion Patterns from Trip Data. *Proceedings of the ACM SIGKDD International Workshop on Urban Computing* (p. 134–141). New York, NY, USA: ACM. https://doi.org/10.1145/2346496.2346518

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, *5*(2), 249-282. https://doi.org/10.1007/BF01897167

Descoimps, É. (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* (masters). École Polytechnique de Montréal. Consulté à l'adresse https://publications.polymtl.ca/597/

Devillaine, F., Munizaga, M., & Trépanier, M. (2012). Detection of Activities of Public Transport Users by Analyzing Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2276*, 48-55. https://doi.org/10.3141/2276-06

Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, *4*(1), 95-104. https://doi.org/10.1080/01969727408546059

Dzikrullah, F., Setiawan, N. A., & Sulistyo, S. (2016). Implementation of scalable K-Means++ clustering for passengers temporal pattern analysis in public transportation system (BRT Trans Jogja case study). *2016 6th International Annual Engineering Seminar (InAES)* (p. 78-83). https://doi.org/10.1109/INAES.2016.7821911

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

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: Emerging Technologies*, *64*, 1-16. https://doi.org/10.1016/j.trc.2015.12.012

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, *20*(1), 5-10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, *17*(6), 441-458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

Kieu, L. M., Bhaskar, A., & Chung, E. (2014). Transit passenger segmentation using travel regularity mined from Smart Card transactions data. *Faculty of Built Environment and Engineering; Smart Transport Research Centre*. Washington, D.C. Consulté à l'adresse https://eprints.qut.edu.au/66571/

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, 179-191. https://doi.org/10.1016/j.trc.2014.05.012

Lathia, N., Smith, C., Froehlich, J., & Capra, L. (2013). Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, *9*(5), 643-664. https://doi.org/10.1016/j.pmcj.2012.10.007

Lebart, L., Morineau, A., & Fénelon, J.-P. (1982). *Traitement des données statistiques: méthodes et programmes*. Dunod.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489

Lu, H. K. (2007). Network smart card review and analysis. *Computer Networks, 51*, 2234-2248. https://doi.org/10.1016/j.comnet.2007.01.009

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. https://doi.org/10.1016/j.trc.2013.07.010

Mahrsi, M. K. E., 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. https://doi.org/10.1109/TITS.2016.2600515

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting methods and applications*. John wiley & sons.

Marca, D. A., & McGowan, C. L. (1987). *SADT: Structured Analysis and Design Technique*. New York, NY, USA: McGraw-Hill, Inc.

Mcnally, M. G. (2000). The four step model. *Handbook of Transport Modelling* (p. 35–52). Oxford, Pergamon Press.

Morency, C., Trepanier, M., & Agard, B. (2006). Analysing the Variability of Transit Users Behaviour with Smart Card Data. *2006 IEEE Intelligent Transportation Systems Conference* (p. 44-49). Toronto, ON, Canada: IEEE. https://doi.org/10.1109/ITSC.2006.1706716

Morency, Catherine, Trepanier, M., Frappier, A., & Bourdeau, J.-S. (2017). Longitudinal Analysis of Bikesharing Usage in Montreal, Canada. Présenté à Transportation Research Board 96th Annual MeetingTransportation Research Board. Consulté à l'adresse https://trid.trb.org/view/1438718

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. https://doi.org/10.1016/j.trc.2012.01.007

Munizaga, M., Palma, C., & Mora, P. (2010). Public transport OD matrix estimation from smart card payment system data. Présenté à *Proceedings from 12th World Conference on Transport Research, Lisbon, Paper, (No. 2988).*

Ni, M., He, Q., & Gao, J. (2017). Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Transactions on Intelligent Transportation Systems*, *18*(6), 1623-1632. https://doi.org/10.1109/TITS.2016.2611644

O'Connell, R. T., & Koehler, A. B. (2005). *Forecasting, time series, and regression: An applied approach* (Vol. 4). South-Western Pub.

Orfeuil, J.-P. (2001). L'évolution de la mobilité quotidienne: Comprendre les dynamiques, éclairer les controverses. *Recherche - Transports - Sécurité*, *72*, 87.

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

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Seaborn, C., Attanucci, J., & Wilson, N. H. M. (2009). Using Smart Card Fare Payment Data To Analyze Multi-Modal Public Transport Journeys in London. *Wilson via Ann Graham*. Consulté à l'adresse http://dspace.mit.edu/handle/1721.1/69577

Sun, Y., Shi, J., & Schonfeld, P. M. (2016). Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: a case study of Shanghai Metro. *Public Transport*, *8*(3), 341-363. https://doi.org/10.1007/s12469-016-0137-8

Swerling, P. (1958). *A proposed stagewise differential correction procedure for satellite tracking and prediciton*. Rand Corporation.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267-276. https://doi.org/10.1007/BF02289263

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423. https://doi.org/10.1111/1467-9868.00293

Trasarti, R., Pinelli, F., Nanni, M., & Giannotti, F. (2011). Mining Mobility User Profiles for Car Pooling. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 1190–1198). New York, NY, USA: ACM. https://doi.org/10.1145/2020408.2020591

Trépanier, M., & Morency, C. (2010). Assessing transit loyalty with smart card data (p. 11-15). Présenté à 12th World Conference on Transport Research, July.

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. https://doi.org/10.1080/15472450601122256

Ward Jr., J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236-244. https://doi.org/10.1080/01621459.1963.10500845

Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, *6*(3), 324-342. https://doi.org/10.1287/mnsc.6.3.324

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation Research Part C: Emerging Technologies*, *75*, 17-29. https://doi.org/10.1016/j.trc.2016.12.001

ANNEXE A – PRÉVISIONS - MÉTHODE EXPÉRIMENTALE (STO)

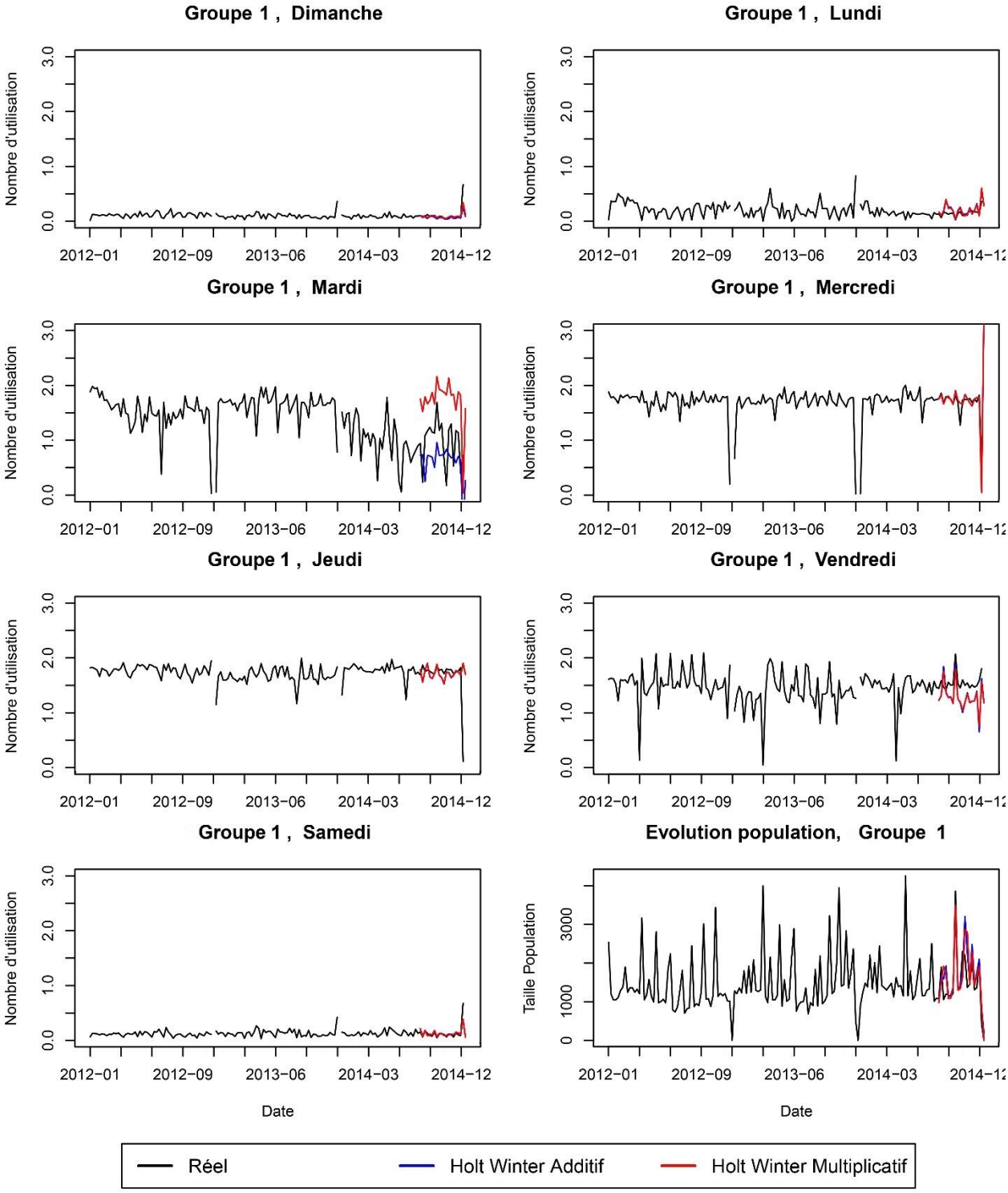


Figure A-1 : Résultats groupe 1 - méthode expérimentale et prévisions

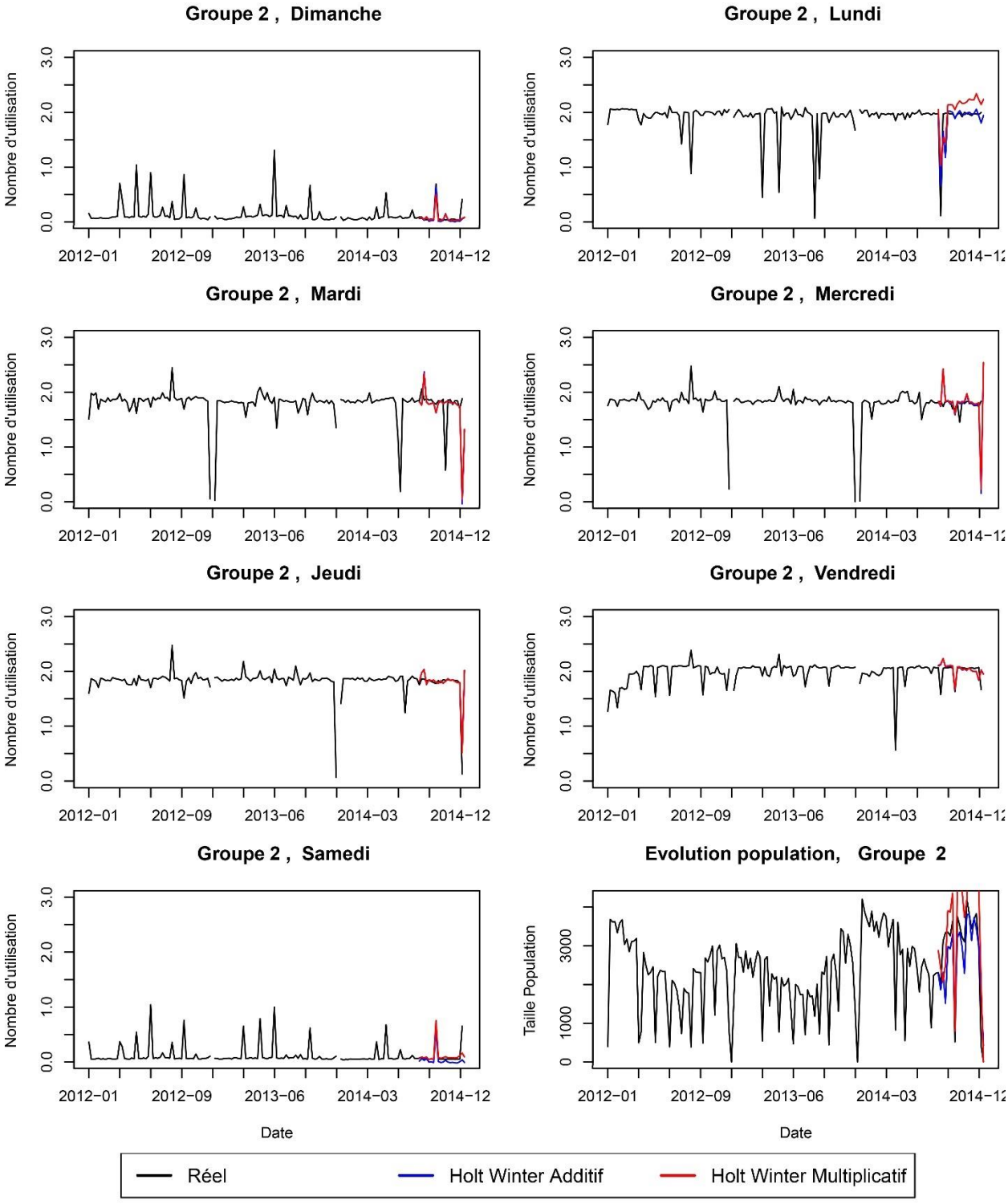


Figure A-2 : Résultats groupe 2 - méthode expérimentale et prévisions

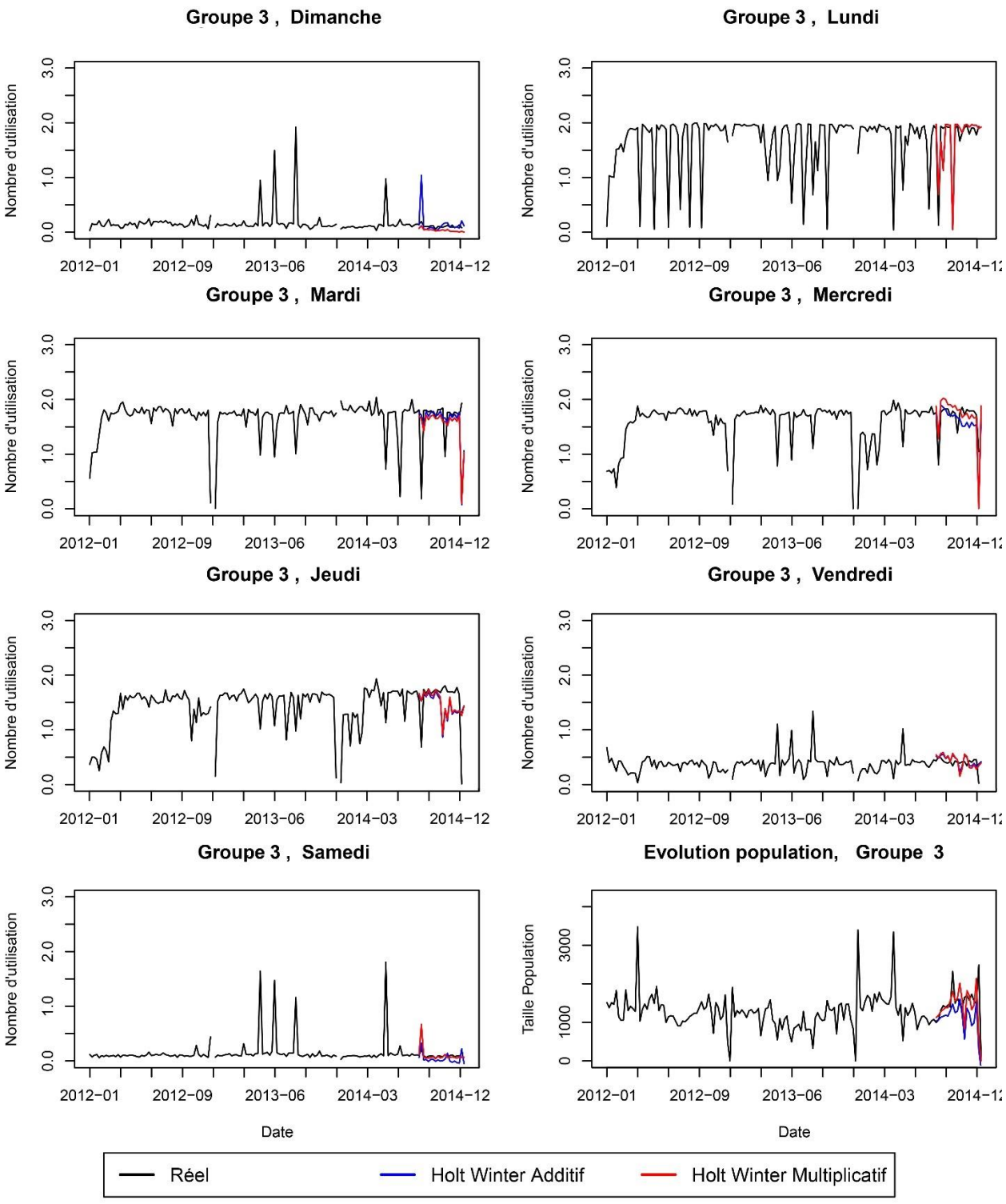


Figure A-3 : Résultats groupe 3 - méthode expérimentale et prévisions

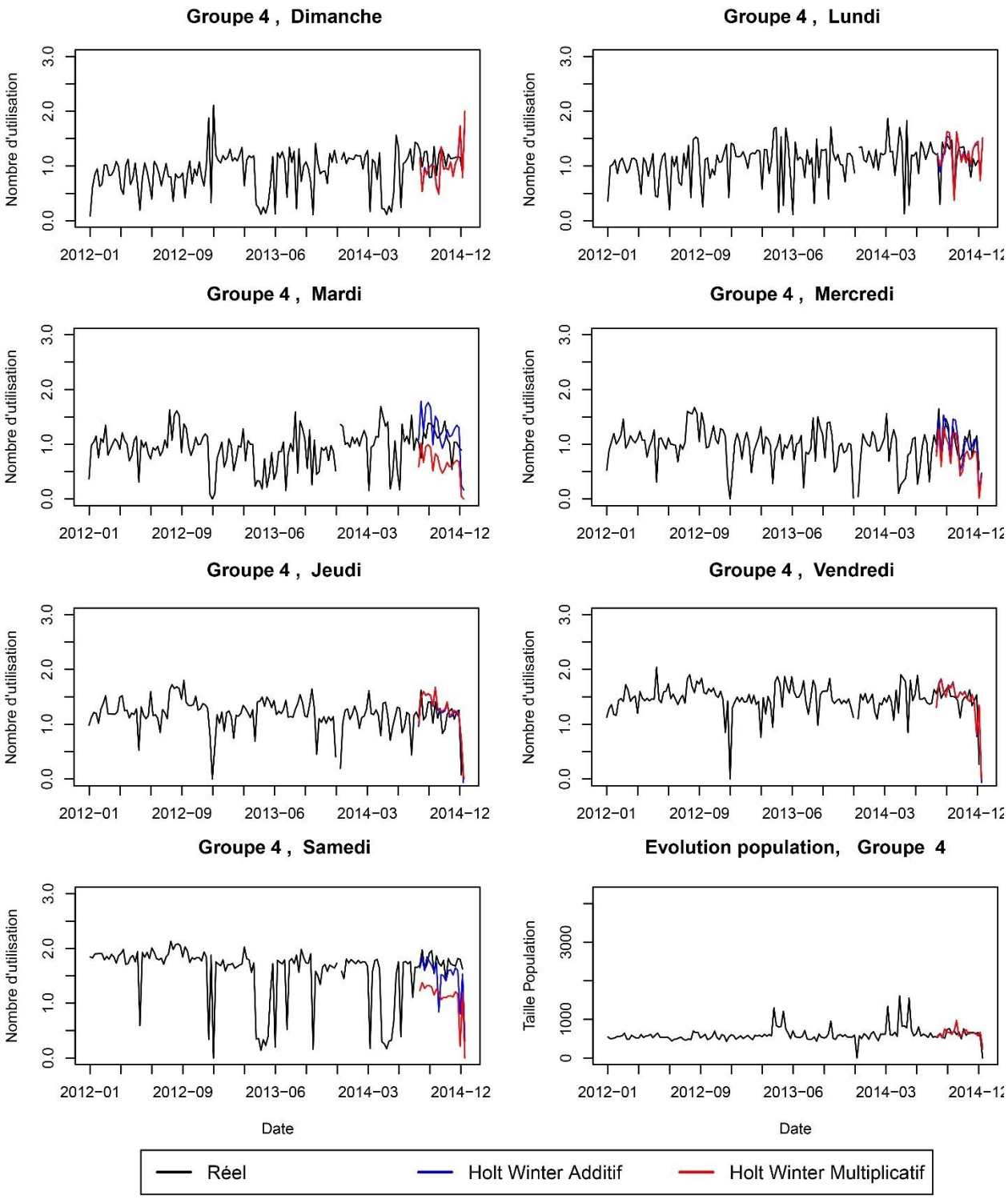


Figure A-4 : Résultats groupe 4 - méthode expérimentale et prévisions

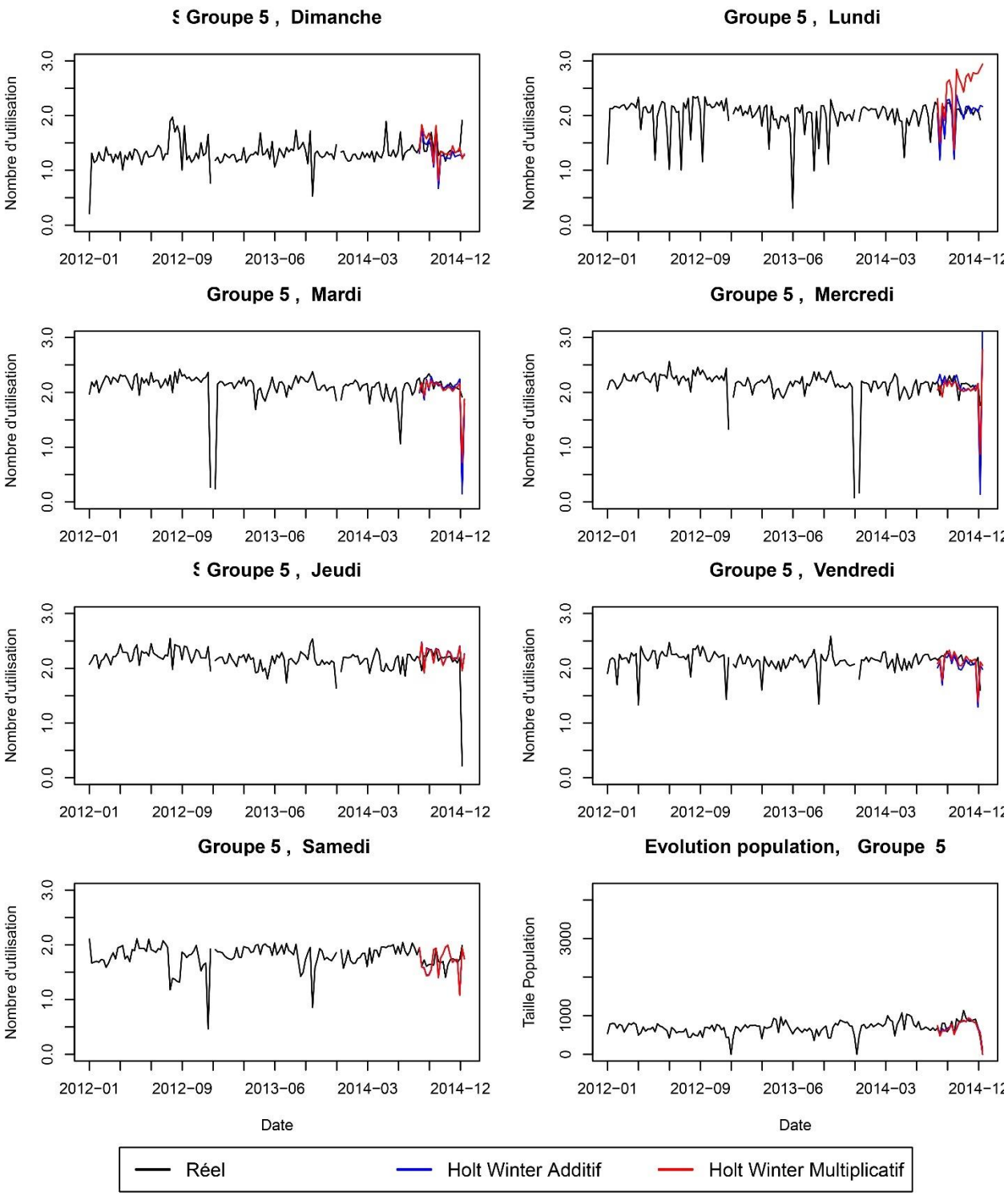


Figure A-5 : Résultats groupe 5 - méthode expérimentale et prévisions

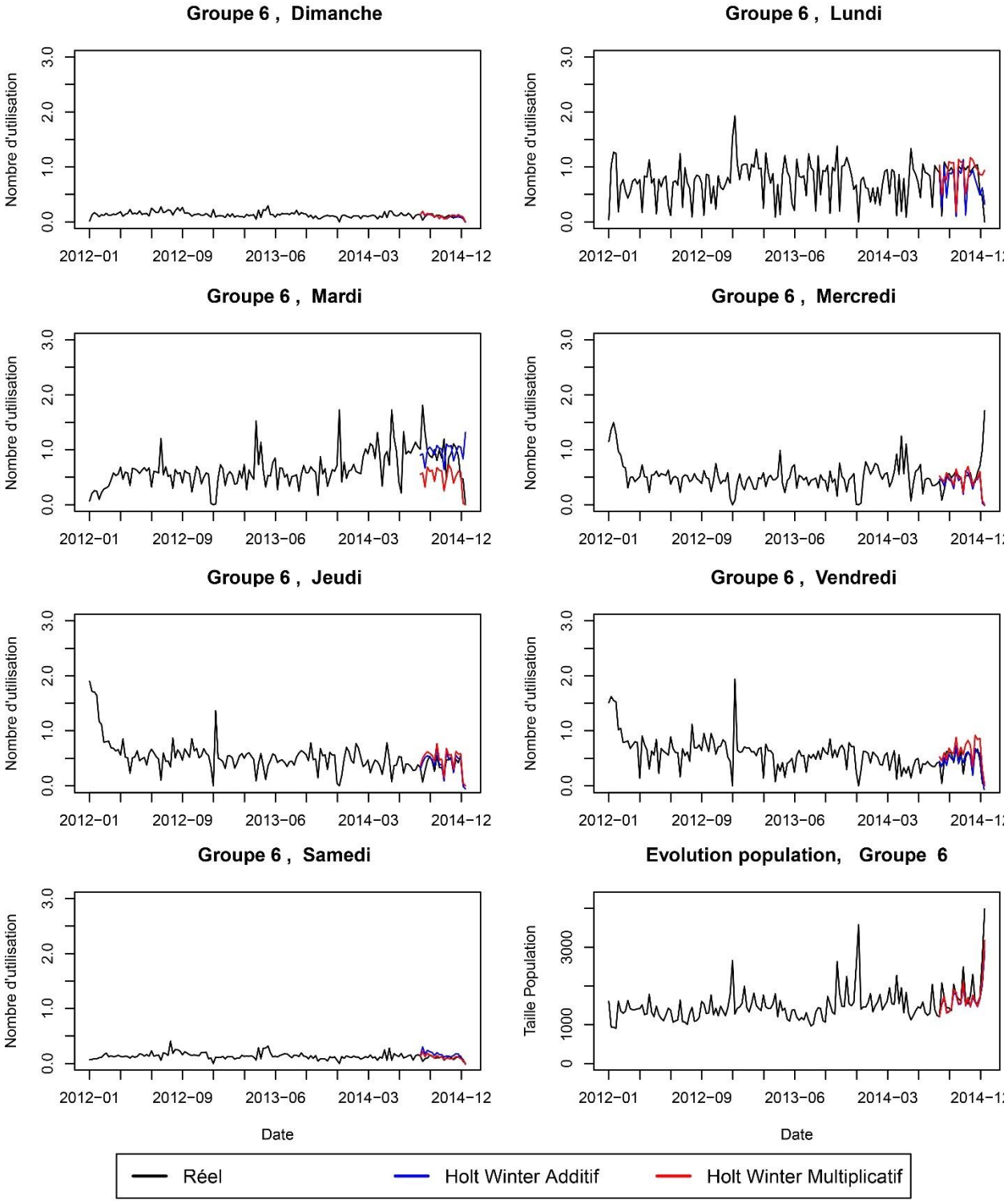


Figure A-6 : Résultats groupe 6 - méthode expérimentale et prévisions