POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Segmentation spatio-temporelle modulable des usagers de cartes à puce
Title:

Auteur: Rémi Decouvelaere
Author:

Date: 2021

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Decouvelaere, R. (2021). Segmentation spatio-temporelle modulable des usagers de cartes à puce [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/6644/
Citation:

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: https://publications.polymtl.ca/6644/
PolyPublie URL:

Directeurs de recherche: Martin Trépanier, & Bruno Agard
Advisors:

Programme: Maîtrise recherche en génie industriel
Program:

Ce fichier a été téléchargé à partir de PolyPublie, le dépôt institutionnel de Polytechnique Montréal
This file has been downloaded from PolyPublie, the institutional repository of Polytechnique Montréal

https://publications.polymtl.ca

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Segmentation spatio-temporelle modulable des usagers de cartes à puce

RÉMI DECOUVELAERE

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Juin 2021

© Rémi Decouvelaere, 2021.

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Segmentation spatio-temporelle modulée des usagers de cartes à puce

présenté par **Rémi DECOUVELAERE**

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

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

Camélia DADOUCHI, présidente

Martin TRÉPANIER, membre et directeur de recherche

Bruno AGARD, membre et codirecteur de recherche

Francesco CIARI, membre

REMERCIEMENTS

J’aimerais tout d’abord remercier mes directeurs de recherche, le professeur Bruno Agard et le professeur Martin Trépanier, du département de mathématiques et de génie industriel de Polytechnique Montréal, pour leur soutien, leur intérêt et leur dévouement, et pour tout ce qu’ils m’ont enseigné au cours de cette maîtrise. Grâce à eux, le projet s’est déroulé de la meilleure des manières, malgré le contexte difficile de la pandémie, et j’en garderai un excellent souvenir.

Merci à l’ensemble du CIRRELT, en particulier à Li He, Zohreh Vaezi, et Laure Siret pour l’aide et le support que vous m’avez apporté.

Merci à Suzanne Guindon et Saroja Chenchooliah pour leur efficacité et leur soutien dans toutes les démarches administratives.

Je tiens également à remercier le Réseau de Transport de la Capitale, et la Société de Transport de l’Outaouais, qui ont fourni les données de cartes à puce que j’ai utilisées, ainsi que le Conseil de Recherches en Sciences Naturelles et en Génie du Canada, et les groupes PROMPT, Thalès, et Cortex pour leurs subventions.

Enfin, merci à ma famille et mes amis pour leur soutien à chaque instant durant ces deux ans de maîtrise.

RÉSUMÉ

L'utilisation des données provenant de systèmes de paiement par cartes à puce permet une compréhension approfondie du comportement des usagers d'un réseau de transport en commun. Que ce soit à des fins de planification ou de stratégie commerciale, il est utile de pouvoir regrouper les usagers en groupes de comportements similaires pour analyser ces comportements. Dans certains cas, le caractère temporel du comportement (heures, nombre et durées des trajets) peut être le plus intéressant, dans d'autres ce sera le caractère spatial (lignes et arrêts empruntés). Dans d'autres cas encore, on peut souhaiter prendre en compte simultanément les caractères spatial et temporel des comportements. À cet effet, dans ce mémoire, nous présentons un outil de segmentation spatio-temporelle qui permet de faire varier l'importance relative du temps et de l'espace dans le critère de regroupement. Le contrôle des poids du temps et de l'espace se fait à l'aide d'un 'paramètre d'équilibre spatio-temporel' qui influe sur la valeur de la métrique utilisée pour comparer deux comportements. Si ce paramètre est proche de 0, la métrique accordera plus d'importance à la composante temporelle des comportements, et s'il est proche de 1, à la composante spatiale.

Dans un premier temps, nous étudions l'influence de chaque paramètre de l'outil de segmentation sur les résultats, mesurée par des indicateurs de performance spatiaux, temporels ou spatio-temporels. Ceci permet d'optimiser certains de ces paramètres (taille d'échantillon, fenêtre temporelle de regroupement), ainsi que de mesurer l'influence du 'paramètre d'équilibre spatio-temporel' sur les natures spatiale et temporelle de la segmentation. Nous observons une amélioration des indicateurs spatiaux lorsque ce paramètre augmente, et une amélioration des indicateurs temporels lorsqu'il diminue. Nous constatons par ailleurs que ces indicateurs ne varient de manière significative que pour un paramètre d'équilibre compris entre 0,2 et 0,8, et qu'ils suivent une tendance à peu près linéaire. Les natures spatiale et temporelle des résultats peuvent donc être contrôlées de manière assez précise à l'aide de ce paramètre.

Dans un second temps, nous utilisons cet outil pour réaliser l'étude du comportement des usagers du Réseau de Transport de la Capitale (Québec, Canada), avec trois valeurs d'équilibre spatio-temporel. Chacune de ces trois études fournit des résultats très différents, mais toujours pertinents. La première présente les comportements temporels les plus répandus à Québec, sans tenir compte

de l'espace, et démontre, par exemple, que les aînés et les étudiants ont des comportements temporels beaucoup plus divers que les adultes, et que les aînés ont des horaires moins tardifs que le reste de la population. La deuxième étude donne une importance à peu près égale au temps et à l'espace. Nous observons alors des couples de *clusters* regroupant des usagers habitant et travaillant à peu près dans la même zone, mais présentant des comportements temporels différents. Des trois études menées, c'est dans celle-ci que les *clusters* diffèrent le plus au niveau de la tranche d'âge de leurs usagers, ce qui démontre l'intérêt de considérer simultanément le temps et l'espace. La troisième étude considère uniquement la nature spatiale du comportement. Les usagers sont alors regroupés en zones géographiques assez précises. Si certaines différences temporelles sont encore observées, c'est que le comportement temporel des individus peut dépendre de la zone où ils habitent. Nous observons encore des couples de *clusters*, mais de nature différente : ceux-ci regroupent des usagers habitant au même endroit, mais se déplaçant dans deux zones différentes. Ainsi, ces trois études fournissent des résultats intéressants, utilisables dans des circonstances différentes. Cette méthode de segmentation spatio-temporelle modulée devrait permettre, dans le futur, une meilleure compréhension du comportement des usagers de cartes à puce, car elle offre, grâce à son paramètre d'équilibre spatio-temporel, plusieurs points de vue aussi divers que pertinents sur le comportement d'une population.

ABSTRACT

The analysis of smart card data offers an in-depth understanding of travel behavior of public transport users. An efficient way to analyze public transport users behavior is to group them into different *clusters* of similar behavior. However, this *cluster*ing process should take into account space and time, because both of these dimensions characterize daily trips. Depending on the expected outcome, we might wish to give more importance to space or to time, or we might wish to balance the two. In this master thesis, we present a spatiotemporal *cluster*ing tool that permits modulation regarding to the importance of space versus time. We control the importance of space and time with a 'space-time balance parameter' which affects the metric used to compare the behavior of two transit users. For a parameter value close to 0, this metric will give more importance to time, and for a value close to 1, to space.

We first test the influence of each parameter of our *cluster*ing tool on the results. Spatial, temporal, and spatio-temporal indicators are used to characterize these results. Based on these tests, we optimize some of the parameters (sample size, DTW window), and we measure the influence of the space-time balance parameter on the spatial and temporal nature of the results. An improvement of temporal indicators when the balance parameter rises, as well as a deterioration of spatial indicators, can be observed. Moreover, these indicators seem to have a linear variation for a parameter value between 0.2 and 0.8. The spatial and temporal nature of the *cluster*ing results can therefore be controlled, with a satisfactory precision.

In a second step, we use this *cluster*ing tool to study the traveling behavior of the Québec transit users, with three different values for the space-time balance parameter. Each of these three studies generates very different, but always interesting, results. The first one considers only the temporal natural of traveling behavior, without any regards for space. It shows, for instance, that the temporal behavior is more diverse for seniors and students, and that seniors tend to have an earlier schedule than the rest of the population. The second study gives approximately the same importance to space and time. We can observe some *clusters* pairs, where both *clusters* have the same origin and destination zones, but a different temporal behavior. Of the three studies, this is the one that creates *clusters* with the most socio-demographical differences, which proves the interest of using time and space simultaneously. The third study considers only the spatial nature of traveling behavior.

The *clusters* created divide Quebec into specific living areas. We also observe some *clusters* pairs, but not of the same kind: these *clusters* regroup users living in the same area but working in two different zones. Therefore, these three studies have generated interesting results, and could all be used, but for different applications. This modulated *cluster*ing method should help, in the future, to better understand the behavior of transit users, because it offers, thanks to its space-time parameter, a variety of diverse and interesting perspectives.

TABLE DES MATIÈRES

REMERCIEMENTS iii

RÉSUMÉ iv

ABSTRACT vi

TABLE DES MATIÈRES viii

LISTE DES TABLEAUX x

LISTE DES FIGURES xi

LISTE DES SIGLES ET ABRÉVIATIONS xiv

LISTE DES ANNEXES xv

CHAPITRE 1 INTRODUCTION 1

CHAPITRE 2 REVUE DE LITTÉRATURE 3

2.1 Utilisation des données de cartes à puce dans les transports 3

2.2 Segmentation des données de cartes à puce 4

2.2.1 Segmentation temporelle 5

2.2.2 Segmentation spatio-temporelle 7

CHAPITRE 3 MÉTHODOLOGIE 12

3.1 Préparation des données 14

3.1.1 Ajustements de l'heure et de la date 15

3.1.2 Création des profils journaliers 16

3.1.3 Élimination des correspondances 17

3.2 Segmentation d'un échantillon 18

3.2.1 Choix de l'échantillon 18

3.2.2 Calcul de la matrice des distances 19

3.2.3 Segmentation hiérarchique de l'échantillon 24

3.2.4 Identification des centroïdes 25
3.3 Affectation des profils restants 25
3.4 Analyse des résultats 26
3.4.1 Outils de visualisation 26
3.4.2 Indicateurs de performance 31
CHAPITRE 4 ÉTUDE ET OPTIMISATION DU MODÈLE CRÉÉ 36
4.1 Mise en œuvre de la méthode 36
4.2 Choix de la taille d'échantillon 39
4.3 Étude de l'influence du paramètre d'équilibre spatio-temporel 42
4.4 Influence du nombre de *clusters* sur l'équilibre spatio-temporel 46
4.5 Choix de la taille de fenêtre 48
CHAPITRE 5 ANALYSE DU COMPORTEMENT SPATIO-TEMPOREL DES USAGERS DE LA RTC 52
5.1 Étude statistique des données 52
5.2 Modèle purement temporel 55
5.2.1 Choix du nombre de *clusters* 55
5.2.2 Étude du comportement temporel des usagers 56
5.3 Modèle spatio-temporel équilibré 61
5.4 Modèle à prédominance spatiale 67
5.5 Conclusion des trois études 73
CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS 75
RÉFÉRENCES 77
ANNEXES 82

LISTE DES TABLEAUX

Tableau 2.1 Segmentation des données de cartes à puce dans la littérature 10

Tableau 3.1 Extrait de la table des transactions 14

Tableau 3.2 Extrait de la table des arrêts 15

Tableau 3.3 Extrait de la table des profils 16

Tableau 3.4 Extrait de la table des résultats de segmentation 26

Tableau A.1 Bibliographie sur l'utilisation des données de cartes à puce 82

LISTE DES FIGURES

Figure 3.1 Méthodologie 12

Figure 3.2 Méthodologie de la segmentation spatio-temporelle modulée 13

Figure 3.3 Algorithme DTW 23

Figure 3.4 Dendrogramme issu de l'algorithme CHA 25

Figure 3.5 Carte des *clusters* les plus fréquents au départ 27

Figure 3.6 Carte de chaleur d'un *cluster* à une heure donnée 28

Figure 3.7 Carte de direction des *clusters* 29

Figure 3.8 Nombre de transactions à chaque heure, dans chaque *cluster* 30

Figure 3.9: Trajectoires espace-temps des *clusters* 30

Figure 3.10 Dispersion des départs dans un *cluster* 4 32

Figure 3.11 Dispersion spatiale du *cluster* k à 4h 33

Figure 3.12 Amplitude de déplacement du centre de gravité du *cluster* k 34

Figure 4.1 Score de silhouette en fonction du nombre de *clusters* 37

Figure 4.2 Carte des *clusters* les plus fréquents au départ 38

Figure 4.3 Nombre de transactions à chaque heure, dans chaque *cluster* 39

Figure 4.4 Temps de calcul total en fonction de la taille d'échantillon 40

Figure 4.5 Variation de R avec la taille d'échantillon 41

Figure 4.6 Dispersion pondérée des départs en fonction de alpha 42

Figure 4.7 Dispersion pondérée des retours en fonction de alpha 43

Figure 4.8 Dispersion spatiale moyenne pondérée en fonction de alpha 43

Figure 4.9 Déplacement moyen pondéré du centre de gravité en fonction de alpha 44

Figure 4.10 Dispersion spatiale moyenne pondérée à 4h, 10h, et 15h 45

Figure 4.11 Dispersion pondérée des départs en fonction de alpha, pour trois nombres de *clusters* 46

Figure 4.12 Dispersion spatiale pondérée en fonction de alpha, pour trois nombres de *clusters* ..47

Figure 4.13 Dispersion pondérée des départs en fonction de alpha, pour trois tailles de fenêtre ..49

Figure 4.14 Dispersion pondérée des retours en fonction de alpha, pour trois tailles de fenêtre...49

Figure 4.15 Dispersion spatiale pondérée en fonction de alpha, pour trois tailles de fenêtre.....50

Figure 5.1 Nombre total de transactions par jour.....52

Figure 5.2 Nombre total de transactions au cours du mois à chaque heure 53

Figure 5.3 Proportion de chaque type d'utilisateur dans la ville de Québec 54

Figure 5.4 Score de Silhouette moyen en fonction du nombre de *clusters* (alpha=0) 55

Figure 5.5 Taille de chaque *cluster* (alpha=0).....56

Figure 5.6 Heures de transaction des usagers dans chaque *cluster* (alpha=0) 57

Figure 5.7 Proportion de chaque type d'usager dans chaque *cluster* (alpha=0).....58

Figure 5.8 Répartition des profils dans chaque *cluster* en fonction du jour (alpha=0).....59

Figure 5.9 Nombre moyen de trajets par profil.....59

Figure 5.10 *Cluster* le plus fréquent au départ de chaque arrêt, avec alpha=0 61

Figure 5.11 Tailles de chaque *cluster* (alpha=0,4).....62

Figure 5.12 Heures de transaction des usagers dans chaque *cluster* (alpha=0,4) 62

Figure 5.13 *Cluster* le plus fréquent au départ de chaque arrêt, avec alpha=0,4 63

Figure 5.15 Carte des directions de chaque *cluster* (alpha=0,4).....64

Figure 5.15 Proportion de chaque type d'usagers dans chaque *cluster* (alpha=0,4) 65

Figure 5.16 Proportion de chaque jour de la première semaine dans chaque *cluster* (alpha=0,4).66

Figure 5.17 Taille de chaque *cluster* 67

Figure 5.18 Heures de transaction des usagers dans chaque *cluster* (alpha=1) 68

Figure 5.19 *Cluster* le plus fréquent au départ de chaque arrêt, avec alpha=1 69

Figure 5.20 Carte des directions de chaque *cluster* (alpha=1) 70

Figure 5.21 Types d'usagers dans chaque *cluster* (alpha=1).....71

Figure 5.22 Proportion de chaque jour de la première semaine dans chaque *cluster* (alpha=1)...72

Figure 5.23 Temps de sortie moyen dans chaque *cluster* (alpha=1).....72

Figure B.1 Carte de chaleur des profils de chaque *cluster* à 4:00.....83

Figure B.2 Cartes de chaleur des profils de chaque *cluster* entre deux pics de transactions 84

Figure C.3 Carte de chaleur des profils de chaque *cluster* à 4:00.....85

Figure C.4 Cartes de chaleur des profils de chaque *cluster* entre deux pics de transaction.....86

LISTE DES SIGLES ET ABRÉVIATIONS

CHA Algorithme de Classification Hiérarchique

DTW Dynamic Time Warping (algorithme de Déformation Temporelle Dynamique)

EM Expectation-maximization Algorithm (algorithme espérance-maximisation)

PCA Principal Component Analysis

RAM Random Access Memory (mémoire vive d'un ordinateur)

RTC Réseau de Transport de la Capitale (ville de Québec)

LISTE DES ANNEXES

Annexe A Bibliographie sur l'utilisation des données de cartes à puce 82

Annexe B Cartes de Chaleur des profils a certaines heures (alpha=0,4) 83

Annexe C Cartes de Chaleur des profils a certaines heures (alpha=1) 85

CHAPITRE 1 INTRODUCTION

Comprendre le comportement des usagers d'un réseau de transport en commun est complexe : chaque individu a son propre comportement et les facteurs qui influent sur celui-ci sont nombreux. Pour améliorer les services d'une compagnie de transport en commun, il ne suffit donc pas d'étudier l'affluence sur chaque ligne et à chaque arrêt, mais il faut suivre le profil de chaque individu au cours de la journée, afin d'avoir un aperçu précis des différents types de comportement. Pour cela, les données provenant des systèmes de paiement par cartes à puce peuvent être utilisées, car elles enregistrent l'heure et le lieu de chaque transaction du propriétaire de la carte, ce qui permet de retracer son profil d'usage du réseau.

Les différents comportements identifiés peuvent alors être séparés en plusieurs groupes, ce qui permet par la suite d'apporter des solutions adaptées à chaque type de comportement. Cependant, comment juger objectivement de la ressemblance entre deux comportements? Par exemple, si un usager A part de la banlieue vers le centre-ville à 7h et rentre à 17h, qu'un autre usager B suit le même trajet, mais en partant à 12h et en rentrant à 14h, et qu'un troisième usager C part du centre-ville vers la banlieue à 7h et rentre à 17h : le comportement de A est-il plus proche de B ou de C? La réponse à ces questions peut dépendre de l'application : dans certains cas on privilégie les caractéristiques temporelles du comportement (à quelle heure l'usager prend-il le bus ou le métro?), dans d'autres les caractéristiques spatiales (quels arrêts l'usager visite-t-il?). Dans d'autres cas encore, on peut souhaiter prendre en compte simultanément le temps et l'espace, en donnant plus ou moins d'importance à l'un des deux. Par exemple, la création d'un nouveau forfait valable uniquement à certaines heures de la journée nécessitera surtout la composante temporelle, alors que la création d'une nouvelle ligne donnera un peu plus d'importance à la composante spatiale.

De nombreuses études traitent du regroupement des comportement des individus en plusieurs *clusters*, mais ces études ne considèrent en général qu'un seul des aspects spatiaux et temporels du comportement, ou bien, lorsqu'elles considèrent les deux, ne permettent pas de contrôler l'influence de chacun de ces aspects sur les résultats. Nous proposons dans ce mémoire un outil de segmentation spatio-temporelle qui permet de faire varier les poids relatifs du temps et de l'espace.

Pour tenir compte du temps et de l'espace simultanément, nous représentons chaque comportement journalier par la liste de ses lieux et nombres de transactions à chaque heure. La métrique que nous utilisons pour comparer deux comportements journaliers s'inspire de la métrique de déformation

temporelle dynamique (*dynamic time warping*, DTW), mais est caractérisée par un 'paramètre d'équilibre spatio-temporel', compris entre 0 et 1, qui donne plus ou moins de poids à la composante spatiale (liste des lieux) ou à la composante temporelle (liste du nombre de transactions) du comportement. C'est à l'aide de ce paramètre que nous contrôlons la nature spatiale ou temporelle des résultats.

Le mémoire s'organise de la façon suivante : après une revue de littérature (**Chapitre 2)** sur l'utilisation des données de cartes à puce et sur les méthodes de segmentation des usagers, nous présenterons en détail le modèle de segmentation spatio-temporelle modulée que nous avons développé (**Chapitre 3**). Ce modèle, inspiré des travaux de Li He (He et al., 2018b), consiste à appliquer un algorithme de classification hiérarchique agglomérative (CHA) à un échantillon, et à affecter ensuite tous les points restants en-dehors de l'échantillon à l'un des *clusters* créés. La spécificité de ce modèle est au niveau de la métrique utilisée pour comparer le comportement de deux individus : la valeur du 'paramètre d'équilibre spatio-temporel' de cette métrique permet de contrôler le poids du temps et de l'espace sur les résultats. Nous étudions ensuite le comportement de ce modèle (**Chapitre 4**), à l'aide d'indicateurs spatiaux et temporels, en faisant varier certains de ces paramètres, et nous chercherons à optimiser ces paramètres. On cherche notamment à optimiser la taille d'échantillon choisie, et la fenêtre temporelle utilisée dans l'algorithme DTW. Nous étudions comment la valeur du paramètre d'équilibre spatio-temporel influe sur les indicateurs spatiaux et temporels. Dans un dernier temps, (**Chapitre 5**) nous présentons trois études détaillées du comportement spatio-temporel des usagers du Réseau de transport de la Capitale (RTC) de la ville de Québec, en utilisant trois valeurs différentes pour le paramètre d'équilibre spatio-temporel de notre modèle. Le but de cette dernière partie est de démontrer qu'en fonction de la valeur de ce paramètre, les usagers du réseau de transport de Québec ne sont pas regroupés de la même façon : ceci souligne la nécessité de contrôler l'influence du temps et de l'espace sur la segmentation des usagers de cartes à puce, et l'importance de choisir une valeur d'équilibre spatio-temporel adéquate à l'application envisagée.

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette revue de littérature présente différents travaux traitant de l'utilisation des données de cartes à puces dans les transports. Nous présentons d'abord de manière générale ce domaine d'étude, puis nous présentons plus en détail la partie de ce domaine d'étude qui touche à la segmentation des usagers.

2.1 Utilisation des données de cartes à puce dans les transports

Les cartes à puce sont utilisées en guise de mode de paiement dans la plupart des réseaux de transport en commun, à cause de leur facilité d'utilisation, de leur fiabilité et de leur durabilité (Lu, 2007), ainsi que pour la quantité et la qualité des données qu'elles permettent de collecter (Blythe, 2004). Il s'agit d'une carte de la taille d'une carte de crédit équipée d'une petite puce de microprocesseur, dotée d'une mémoire lisible par un lecteur de carte. Une grande partie des informations concernant la carte (historique des transactions, renseignements sur l'utilisateur) ne sont cependant pas stockées sur la carte, mais sur un serveur de données (Pelletier et al., 2011). Les données provenant des systèmes de paiement par carte à puces contiennent généralement les informations suivantes :

- l'identifiant de la carte ou du porteur;
- l'heure et le lieu de chaque transaction (c'est-à-dire de chaque validation de la carte);
- la ligne et le sens de la ligne empruntée;
- le type de forfait associé à la carte;
- les informations d'achat (heure et lieu, prix).

Ces données ont été utilisées pour de nombreuses applications, qui peuvent être regroupées en trois catégories, qui se suivent et se complètent : la reconstruction des trajets individuels, l'étude du comportement des usagers, et l'amélioration des services du réseau de transport en commun (cf. Annexe 1).

La reconstruction des trajets a fait l'objet de nombreux travaux (Trépanier et al., 2007; Chu et al., 2008; Seaborn et al., 2009; Devillaine et al., 2012; Kusakabe et al., 2014; He et al., 2015; Alsger et al., 2016; Giraud et al., 2016…). En effet, dans la plupart des réseaux de bus, les usagers ne

doivent valider qu'à la montée dans le bus, si bien que seule l'origine des trajets est connue. Le travail de tous ces auteurs a donc été de reconstituer les destinations de chaque trajet, à l'aide d'algorithmes d'estimation de destination et, parfois, d'autres sources de données (Kusakabe et al., 2014; Giraud et al., 2016…). Une fois les trajets reconstitués, il est désormais possible d'analyser le comportement de chaque individu.

L'étude du comportement des usagers constitue la plus grosse partie de la littérature sur les cartes à puce. De nombreux types d'études ont été menées : étude de la congestion dans les transports en commun (Park et al., 2008; Ceana et al., 2012; Yap et al., 2018), de l'évolution du comportement des usagers (Bagchi et al., 2005; Morency et al., 2007; Trépanier et Morency, 2010), de l'influence de la météo, du type d'usager, ou d'autres facteurs, sur le comportement (Nishiuchi et al., 2013; Arana et al., 2014; Toqué et al., 2019; Zhou et al., 2017), etc… Un domaine d'étude assez important constitue la segmentation des usagers en plusieurs groupes de comportement semblable. Ceci fait l'objet de notre prochaine section.

Enfin, le dernier domaine d'étude sur les cartes à puces explore les possibilités d'amélioration des services des réseaux de transport en utilisant les données de cartes à puces. Il s'agit pour la plupart de revues de littérature sur l'exploration des données de cartes à puces, dans lesquelles les auteurs expliquent comment ces résultats pourraient être utilisés pour l'amélioration des services (Blythe, 2004; Utsunomiya et al., 2006; Lu, 2007; Pelletier et al., 2011; Kurauchi et al., 2017).

2.2 Segmentation des données de cartes à puce

Une façon efficace d'étudier le comportement des usagers d'un réseau de transport est de regrouper ceux-ci en plusieurs *clusters* (groupes) de comportement similaire, afin d'identifier et de comprendre les comportements typiques sur ce réseau. Le comportement d'un usager peut être caractérisé par de nombreux critères, temporels (dates et heures des montées et descentes, durée et nombre de trajets…) ou spatiaux (arrêts et lignes empruntés), et peut être étudié selon plusieurs échelles temporelles (comportement sur un jour, une semaine, ou un mois…). Pour cette raison, les modèles de segmentation des usagers abondent dans la littérature, d'autant plus que les algorithmes et métriques de segmentation sont nombreux. Nous avons regroupé une série d'articles traitant de la segmentation des usagers, selon le type de segmentation (temporelle ou spatio-temporelle), la plage temporelle étudiée, et les données, algorithmes et métriques utilisés (**Tableau 2.1**).

2.2.1 Segmentation temporelle

La segmentation temporelle considère seulement la composante temporelle du comportement d'un utilisateur, c'est-à-dire les heures de ses montées et, lorsque disponibles, descentes. Dans un premier temps, le comportement d'un utilisateur a été étudié sur une semaine (Agard et al., 2006; Morency et al., 2007; Trépanier, 2013) : un comportement est alors représenté par un vecteur de 0 et de 1, renseignant sur l'activité de l'utilisateur durant chaque période de chaque jour de la semaine (chaque jour étant divisé en 4 périodes). Si l'utilisateur a réalisé une transaction durant une de ces périodes, celle-ci est représentée par un 1. Deux comportements peuvent alors aisément être comparés en utilisant la métrique de Hamming, qui renvoie le pourcentage de composantes différentes entre les deux vecteurs représentant ces comportements. Si les deux utilisateurs ont réalisé la plupart de leurs transactions en même temps, la distance entre leurs deux comportements sera donc proche de zéro. Un algorithme de segmentation permet alors de regrouper les comportements en plusieurs *clusters*, de manière à minimiser la distance entre les comportements au sein d'un *cluster*, et à maximiser la distance entre les comportements de deux *clusters* différents.

Cependant, la métrique de Hamming n'est pas très adaptée aux séries temporelles (Agard et al., 2013) : en effet, elle ne considère pas la position relative des éléments des vecteurs temporels. Par exemple si un usager fait une transaction à 6h, il sera aussi éloigné d'un usager faisant une transaction à 7h que d'un usager faisant une transaction à 15h. Pour cette raison, d'autres métriques plus adaptées aux séries temporelles ont été progressivement adoptées. En 2013, Agard et al. proposent de représenter le comportement journalier d'un individu par un point sur un cylindre, dont le rayon est une fonction du nombre de transactions de l'individu au cours de la journée. L'angle formé par le point dépend de l'heure moyenne des transactions (0° pour 0:00 et 180° pour 24:00), et la hauteur du point sur le cylindre est proportionnelle à la dispersion des transactions. La distance entre deux comportements est alors calculée en mesurant la distance euclidienne séparant les deux points représentant ces comportements. De cette façon, deux comportements peuvent être jugés similaires s'ils ont un nombre, une heure moyenne et une dispersion de transactions proches, même si leurs transactions ne sont pas exactement à la même heure. Les paramètres de cette métrique ont été fixés de manière à respecter certaines propriétés, énoncées dans l'article en question. Cette métrique sera reprise plus tard par Ghaemi et al., mais en utilisant un autre algorithme de segmentation (CHA au lieu de k-means, Ghaemi et al., 2017a), puis en appliquant

une méthode de '*biclustering*' spatio-temporelle (Ghaemi et al., 2017b). D'autres métriques temporelles sont aussi utilisées pour la segmentation des usagers : on peut citer l'exemple intéressant de (El-Mahrsi et al., 2014), qui utilise une métrique et un algorithme de reconnaissance de texte pour comparer les vecteurs de comportement temporel hebdomadaire. Il faut également citer les travaux de (He et al., 2018a), qui comparent deux métriques très utilisées pour comparer les séries temporelles : la métrique CCD (Cross-Correlation Distance) et DTW (Dynamic Time Warping). Ces deux métriques ont l'avantage de pouvoir comparer les formes des séries temporelles, même si celles-ci présentent un décalage ou une différence d'échelle (par exemple départ à 8h et retour à 16h au lieu de départ à 7 et retour à 17h). Après avoir testé ces deux métriques sur le réseau de bus de Gatineau, Li He conclut que la métrique CCD est plus adaptée, dans le cas des séries temporelles. Il réutilisera ce modèle en 2019 pour comparer le comportement des usagers dans deux villes très différentes, Gatineau et Santiago. En tentant de regrouper les usagers de ces deux villes en plusieurs *clusters*, il s'apercevra que chaque *cluster* obtenu contient une majorité d'individu d'une seule des deux villes, ce qui prouve que le comportement des usagers est très différent dans ces deux villes.

A partir de 2013, certains auteurs ont commencé à analyser les comportements sur une échelle de temps plus petite qu'une semaine, généralement 24h (Agard et al., 2013; Ghaemi et al., 2017a…). Ceci permet d'avoir un découpage temporel plus précis du comportement des usagers, et tient compte du fait qu'un usager peut avoir un comportement totalement différent selon le jour de la semaine. Certains articles s'intéressent d'ailleurs la question de la variabilité du comportement de l'individu (Agard et al.,2006; Morency et al., 2007; Trépanier, 2013; Ghaemi et al., 2017a…) : celle-ci est mesurée par la variation de la taille de chaque *cluster* de comportement au cours du mois. Elle peut aussi être évaluée en étudiant les changements de *cluster* de chaque individu au cours de la période étudiée (par exemple, le comportement d'un individu pourrait appartenir à un certain *cluster* durant les trois premières semaines, mais à un autre durant la dernière semaine du mois).

Que ce soit pour la segmentation temporelle ou spatio-temporelle, les algorithmes les plus utilisés sont *k-means* et CHA. Si *k-means* présente l'avantage d'être très rapide, il ne permet pas de liberté quant au choix du nombre de *clusters*, celui-ci devant être spécifié à l'avance. CHA, au contraire, en créant des *clusters* emboîtés les uns dans les autres, permet facilement de modifier le nombre

de *clusters*. D'autre part, cet algorithme produit beaucoup plus d'information que *k-means*, car chacun des *clusters* obtenus peut être séparé, de manière récursive, en sous-*clusters*, jusqu'à atteindre des *clusters* de taille 1. Pour résoudre le problème du temps de calcul, dans le cas de l'algorithme CHA, certains auteurs n'appliquent cet algorithme qu'à un échantillon aléatoire, et affectent ensuite les points restants au *cluster* le plus proche (He et al., 2018b). Pour ce qui est du choix du nombre de *clusters*, quel que soit l'algorithme utilisé, on fait souvent appel à l'indicateur Silhouette, qui atteint son maximum lorsque le nombre de *clusters* maximise l'homogénéité au sein des *clusters* et l'hétérogénéité entre *clusters* (Zhao et al., 2017; Faroqi et al., 2019; He et al., 2019b).

Souvent, les données de cartes à puce enregistrent certaines informations sur les usagers (type d'usager, âge, sexe, zone d'habitation…). La plupart des auteurs utilisent ces informations pour comparer les *clusters* (Agard et al., 2006; Morency et al., 2007; Trépanier, 2013, …). Parfois, certaines données socio-démographiques (âge, occupation, revenu, possession d'une voiture…) sont aussi adjointes aux résultats de segmentation temporelle ou spatio-temporelle. Nous observons alors en général des différences socio-démographiques notables entre chaque *cluster* (El-Mahrsi et al., 2014; Langlois et al., 2016).

2.2.2 Segmentation spatio-temporelle

Bien que la segmentation temporelle des usagers puisse déjà offrir de nombreuses informations sur leurs comportements, elle reste cependant assez limitée, car elle néglige l'influence du facteur spatial sur le comportement. Certains auteurs ont donc tenté de prendre en compte ces deux facteurs simultanément. En 2015, Ghaemi et al. proposent de comparer deux usagers en calculant la somme de la distance séparant ces deux usagers à chaque heure. Cependant l'importance donnée au temps est alors minime. Il affine donc son modèle, et 2017, en couplant cette métrique à une autre métrique, temporelle, qui reprend celle proposée par (Agard et al., 2013). L'auteur développe alors un algorithme de segmentation CHA bidirectionnelle, en utilisant ces deux métriques, spatiale et temporelle. Ceci permet de regrouper les usagers selon plusieurs types de comportements spatio-temporels (Ghaemi et al., 2017a). En 2018, He et al. présentent un autre modèle de segmentation spatio-temporelle des usagers. Il représente, lui aussi, chaque usager par la liste de ses positions à chaque heure de la journée (chaque position correspondant à un arrêt de bus), mais il utilise la métrique DTW pour comparer ces listes de positions. Cette métrique, particulièrement adaptée aux

séries temporelles, permet de conserver le caractère temporel du comportement des usagers. A l'aide de l'algorithme CHA, il crée ainsi 10 groupes de comportement spatio-temporels différents. Il visualise ces groupes en représentant pour chacun la trajectoire espace-temps moyenne (outil développé par (Farber et al., 2015)) : cette visualisation prouve que chaque *cluster* représente un comportement spatio-temporel unique. Cependant, si les *clusters* diffèrent beaucoup sur le plan spatial, la plupart d'entre eux ont des caractéristiques spatio-temporelles proches (départ vers 8h et retour vers 16h).

D'autres auteurs ont choisi, plutôt que de regrouper les usagers selon leurs heures de transactions, de les regrouper selon leur 'fidélité' au réseau (Kieu et al., 2014; Ma et al., 2013; Zhao et al., 2017). Dans (Kieu et al., 2014), chaque usager est représenté par le nombre de jours avec un trajet, sur une période de 70 jours, et par le nombre total de trajets sur cette période. La distance entre ces couples de coordonnées est calculée avec la métrique euclidienne, ce qui permet, en utilisant l'algorithme k-means, de répartir les usagers en deux classes : les usagers réguliers et les usagers occasionnels. Cet auteur réalise aussi une autre segmentation selon la régularité du comportement, mais cette fois-ci spatio-temporel, et pas seulement temporel. On calcule pour chaque usager un indicateur de régularité spatiale des trajets et un indicateur de régularité temporelle. Ces couples de coordonnées sont ensuite projetés dans un plan (chaque utilisateur étant représenté par un point), et l'auteur répartit, à vue d'œil, les usagers en 4 groupes, selon la régularité spatiale et temporelle de leur comportement. (Ma et al., 2013), réalise une segmentation spatio-temporelle du même type, en regroupant également les passagers selon la régularité de leur comportement. Un usager est cette fois-ci représenté par quatre coordonnées : le nombre de jours avec au moins un voyage, le nombre d'heures d'embarquement similaires, le nombre de séquences de lignes empruntées similaires, et le nombre de séquences d'arrêts similaires. La distance entre chaque usager est calculée en mesurant la distance euclidienne entre ces vecteurs de 4 coordonnées. Ceci permet de regrouper les usagers en 5 groupes, à l'aide de l'algorithme k-means++. Ces 5 groupes correspondent à 5 niveaux de loyauté au réseau (Very high, High, Medium, Low, Very low). (Zhao et al., 2017) procède aussi à une segmentation spatio-temporelle selon la régularité des usagers, mais il réalise séparément la segmentation spatiale et la segmentation temporelle pour les comparer ensuite. Chaque usager est représenté par un vecteur temporel, de taille 4, donnant la proportion de voyages au cours du mois durant chacune des 4 périodes de la journée; et par un vecteur spatial donnant la

proportion de voyages au cours du mois selon chacun des couples O-D de l'utilisateur, classés par ordre de fréquentation. L'auteur applique l'algorithme k-means aux vecteurs temporels, ainsi qu'aux vecteurs spatiaux, et étudie ensuite les corrélations entre les groupes spatiaux et temporels obtenus.

Le cas de (Faroqi et al., 2019) est assez unique, car ce ne sont pas des comportements journaliers ou hebdomadaires qu'il classifie, mais des trajets. La similarité spatiale entre deux trajets est fonction de leur orientation et de leurs longueurs, et la similarité temporelles dépend de la durée des trajets et de si l'un des deux trajets contient l'autre (départ avant et retour après l'autre). Plusieurs modèles de segmentation spatio-temporelle, qui utilisent chacun l'algorithme CHA, sont alors testés : une segmentation spatiale suivi d'une segmentation temporelle (S-T), une segmentation temporelle suivi d'une segmentation spatiale (T-S), et enfin une segmentation simultanément spatiale et temporelle (ST). Dans le cas de la segmentation S-T (ou dans celui de la segmentation T-S), on regroupe d'abord les trajets en *clusters* spatiaux (ou temporels pour T-S), puis on divise chacun de ces *clusters* en sous-*clusters* temporels (ou spatiaux dans le cas de T-S). Dans le cas de la segmentation ST, on calcule pour chaque couple de trajet le produit de leur similarité spatiale avec leur similarité temporelle. Cette distance permet de regrouper directement les usagers en *clusters* spatio-temporels. L'auteur conclut que la segmentation ST crée des *clusters* plus homogènes individuellement, mais moins hétérogènes deux-à-deux, que pour les segmentations T-S et S-T.

Cependant, la plupart des auteurs ayant mis en œuvre une segmentation de type spatio-temporel, ont dû faire face à un problème : quelle importance relative donner au temps et à l'espace? Par exemple, si un usager A part à 7h d'un arrêt 001 et rentre à 17h d'un arrêt 002, qu'un usager B part à 7h et rentre à 11h des mêmes arrêts, et qu'un usager C part à 7h d'un arrêt 003 et rentre à 17h de l'arrêt 002 : l'usager A est-il plus proche de B ou de C? Dans la plupart des articles en question, les poids du temps et de l'espace ont été fixés de manière assez arbitraire. Il serait pourtant très intéressant de pouvoir contrôler cet équilibre, car en fonction de l'application, on peut vouloir donner plus ou moins d'importance à l'un de ces deux facteurs. Nous proposons donc dans ce mémoire une méthode de segmentation spatio-temporelle dans laquelle l'influence relative du temps et de l'espace peut être contrôlée.

Le tableau 2.1 présente pour chacun des articles cités dans la partie ci-dessus le type de segmentation (spatiale ou spatio-temporelle), le lieu et la durée étudiée, les données utilisées, ainsi que la métrique et l'algorithme de segmentation mis en œuvre. Le nombre des *clusters* est aussi renseigné, et est très varié, entre 2 et 64. Enfin, le critère 'd'étude de variabilité temporelle' indique si l'évolution du comportement des usagers a été étudiée.

Tableau 2.1 Segmentation des données de cartes à puce dans la littérature

Auteurs	**Segmen -tation**		**Lieu**	**Plage temporelle**	**Données utilisées**	**Algorithme**	**Métrique**	***Clusters***	**Etude de variabilité temporelle**
	T	**ST**							
Agard et al.,2006	X		Gatineau Bus	4 mois	Vecteurs d'utilisateur-semaine de taille 20 (5 jours divisés en 4 périodes)	k-means suivi de CHA	Hamming	4	Oui
Morency et al., 2007	X		Gatineau Bus	10 mois	Utilisateur-semaine (7 jours * 4 périodes) Utilisateur-jour (24*1h)	k-means	Hamming	4	Oui
Trépanier, 2013	X		Gatineau Bus	10 mois	Utilisateur-semaine (7 jours * 4 périodes)	k-means	Hamming	4	Oui
Agard et al., 2013	X		Gatineau Bus	1 an	Utilisateurs-jours (24*1h) projetés dans l'espace selon un système de coordonnées cylindriques prenant en compte la fréquence, l'heure moyenne et la dispersion des transactions	k-means	Distance euclidienne entre points projetés	3	Non
El-Mahrsi et al., 2014	X		Rennes: Métro+ bus	28 jours	Profil hebdomadaire de taille 168 (24h * 7 jours)	EM	Unigram (reconnaissance de textes)	16	Non
Kieu et al., 2014	X		Brisbane bus	70 jours	Couple de coordonnées: (nombre de jours avec un trajet, nombre total de trajets)	k-means	Euclidienne	2	Non
		X			Couple de coordonnées: (régularité spatiale des trajets, régularité temporelle des trajets)	A priori	NA	3	
Langlois et al., 2016		X	Londres: Métro + bus	28 jours	Vecteur donnant le numéro de zone de l'individu à chaque heure durant 4 semaines, chaque individu ayant ses propres zones, identifiées par un numéro	PCA	Euclidienne	11	Non
Ghaemi et al., 2017a	X		Gatineau: Bus	1 mois	Utilisateurs-jours (24*1h) projetés dans l'espace selon un système de coordonnées cylindriques prenant en compte la fréquence, l'heure moyenne et la dispersion des transactions (cf. Agard et al.,2013)	CHA	Distance euclidienne entre points projetés	18	Oui
He et al., 2018a	X		Gatineau: Bus	1 mois	Utilisateur-jour (présence de transaction à chaque heure entre 4h du matin et 1h le jour suivant)	CHA	CCD DTW	6 11	Non
Ma et al., 2013		X	Beijing Métro + bus	*Inconnu*	Segmentation par fidélité, mesurée par le nombre de jours avec un voyage, la régularité de l'heure d'embarquement et des lignes et arrêts empruntés	k-means++	Distance euclidienne	5	Non
Ghaemi et al., 2015		X	*NA*	*NA*	Position (coordonnées géographiques) de l'usager à chaque heure	*NA*	Somme des distances entre usagers à chaque heure	*NA*	Non
Zhao et al., 2017		X	Shenzhen: Métro+bus	1 mois	Vecteur temporel donnant la proportion de voyages au cours du mois durant chacune des 4 périodes de la journée Vecteur spatial donnant la proportion de voyages au cours du mois selon chacun des couples O-D de l'utilisateur, classés par ordre de fréquentation	k-means spatial, k-means temporel, Etude des corrélations	Manhattan	4*4	Non
Ghaemi et al., 2017b		X	Gatineau Bus	1 mois	-Position de l'usager à chaque heure -Utilisateurs-jours (24*1h) projetés dans l'espace selon un système de coordonnées cylindriques prenant en compte la fréquence, l'heure moyenne et la dispersion des transactions (cf. Agard et al.,2013)	Bi*cluster*ing CHA	Euclidienne	*NA*	Non
He et al., 2018b		X	Gatineau Bus	1 mois	Profil spatio-temporel journalier (position approximée de l'utilisateur à chaque heure entre 4h du matin et 1h le jour suivant)	CHA	DTW	10	Non

Tableau 2.2 Segmentation des données de cartes à puce dans la littérature (suite et fin)

Faroqi et al., 2019		X	Brisbane Bus + train	4 jours	Segmentation S-T, T-S ou ST des trajets. Similarité spatiale prenant en compte l'orientation et la longueur des trajets; similarité temporelle prenant en compte la durée des trajets et leur imbricabilité 2 à 2	CHA	- Similarité spatiale SS -Similarité temporelle ST -Produit des 2	64, 32	Non
He et al, 2019b	X		Santiago (metro +bus) Gatineau (bus)	1 mois	Utilisateur-jour (présence de transaction à chaque heure entre 5h et 21h)	CHA	CCD	9	Non

CHAPITRE 3 MÉTHODOLOGIE

La méthodologie générale de ce mémoire est divisée en trois étapes (Figure 3.1) :

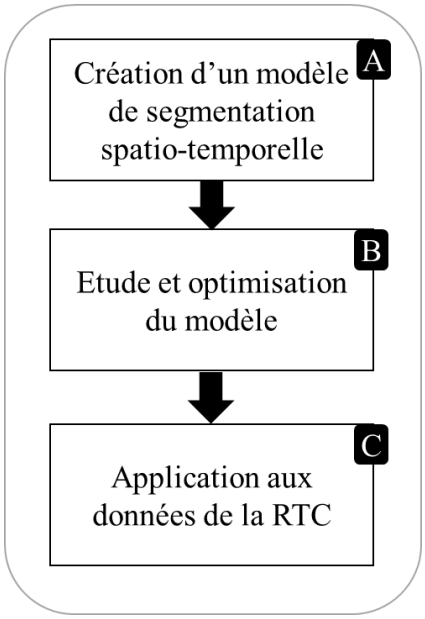


Figure 3.1 Méthodologie

La première étape A correspond à la création d'un modèle de segmentation spatio-temporelle modulée des usagers de cartes à puces. Ce modèle est présenté en détail dans ce chapitre.

La deuxième étape B, présentée dans le chapitre 4, consiste à étudier le comportement du modèle créé durant la première étape, à l'aide d'une série d'indicateurs de performance, et à en optimiser

les différents paramètres. Nous y étudions entre autres l'influence du coefficient d'équilibre spatio-temporel sur la nature spatiale ou temporelle des résultats.

Enfin, dans la dernière étape C, présentée dans le chapitre 5, le modèle est utilisé pour étudier en détail le comportement spatio-temporel des usagers du Réseau de Transport de la Capitale (RTC), sur un mois de données, avec trois rapports espace-temps différents.

Commençons par l'étape A, en présentant notre modèle de segmentation spatio-temporelle modulée :

La Figure 3.2 ci-dessous représente le fonctionnement de notre méthode de segmentation spatio-temporelle modulée, qui permet, à partir des données de cartes à puce d'un groupe d'utilisateurs, de regrouper les comportements journaliers de ces utilisateurs en plusieurs *clusters*, en donnant plus ou moins d'importance au temps ou à l'espace.

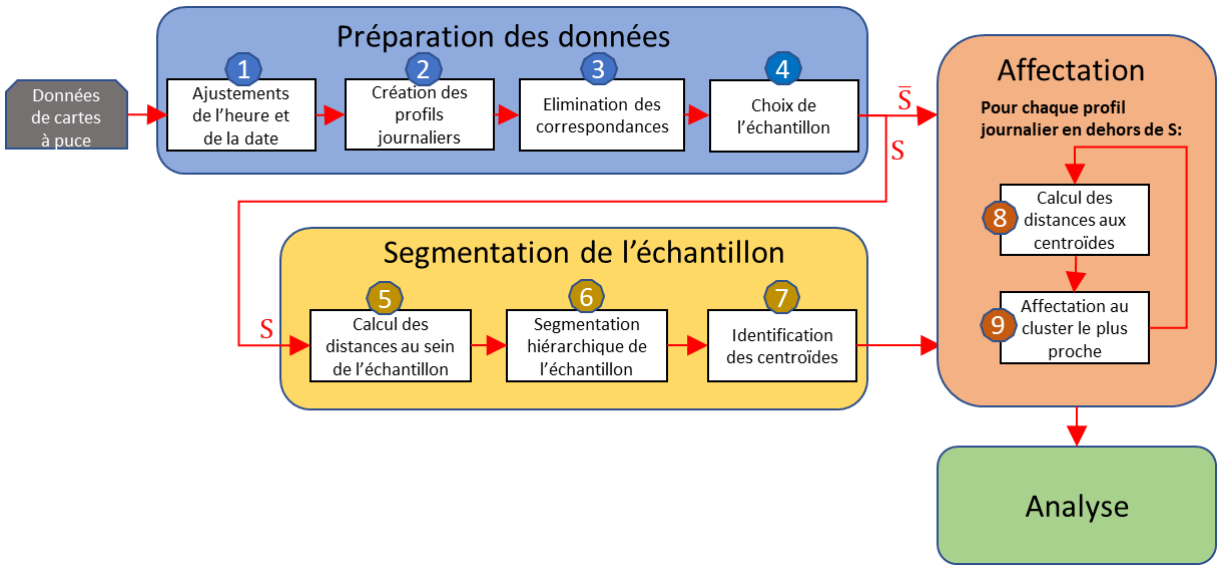


Figure 3.2 Méthodologie de la segmentation spatio-temporelle modulée

Cette méthode est divisible en quatre blocs, chacune décrit dans une sous-partie de ce chapitre :

- Un bloc de préparation des données, dont le but est d'obtenir des profils spatio-temporels journaliers décrivant le comportement d'un utilisateur durant une journée. Et de sélectionner un échantillon de données S.

- Un bloc de segmentation de l'échantillon, qui consiste à regrouper les profils de l'échantillon en plusieurs *clusters*, selon leurs similarités.
- Un troisième bloc consistant à affecter chacun des profils restant à un des *clusters* créés dans la partie précédente.
- Un bloc d'analyse, dont le but est d'observer, à l'aide d'outils de visualisation et d'indicateurs numériques, les différences spatiales et temporelles entre chaque *cluster*.

3.1 Préparation des données

Les données utilisées dans ce modèle sont fournies par le Réseau de Transport de la Capitale (RTC), société de transport publique basée à Québec, au Canada, et qui opère un réseau de 616 bus. Ces données contiennent l'ensemble des transactions de carte à puce des usagers de la RTC durant le mois de décembre 2019. Les voyageurs ne devant valider leur carte qu'à la montée du bus, chaque transaction correspond soit à un départ, soit à une correspondance.

Ces données sont regroupées dans deux tables différentes : une table relative aux transactions, et une table relative à la position des arrêts. Des extraits de ces deux tables sont données dans les tableaux 3.1 et 3.2 :

Tableau 3.1 Extrait de la table des transactions

id_tr	date_opération	code_ligne	direction	stop_o	stop_d	temps_trajet	distance_od	opus_id	code_prod
19641153	07/17/2019 10:15	75	Sud	3157	1790	9	2963	1856714	TLM-E
19641154	07/08/2019 21:20	802	Ouest	2650	1044	5	1535	1929708	TLM-E
19641155	07/12/2019 06:01	82	Sud	3064	1747	34	10304	1785263	TLM-G
19641156	07/11/2019 07:04	803	Ouest	3102	4773	16	4882	1880647	TLM-G

- 'id_tr' est l'identifiant unique de chaque transaction, c'est-à-dire de chaque trajet ou correspondance (puisque les usagers ne valident qu'une seule fois au cours de leur trajet)
- 'date_opération' renseigne sur l'heure et la date de la transaction (c'est-à-dire à quelle heure la carte a été validée)

- ‘code_ligne’ donne le numéro de la ligne de bus empruntée et ‘direction’ le sens de ligne
- ‘stop_o’ correspond à l’arrêt d’origine du trajet, identifié par un numéro. On retrouve les coordonnées de chaque arrêt dans la table des arrêts.
- ‘stop_d’ correspond à l’arrêt de destination du trajet, résultant de l’algorithme d’estimation de destinations. Les trajets dont la destination n’a pas pu être estimée (environ 33%) ont été retirés de la base de données.
- ‘temps_trajet’ donne le temps de trajet en minutes et ‘distance_od’ la distance parcourue en mètres.
- ‘opus_id’ est l’identifiant unique de la carte de transport utilisée. La plupart des usagers (environ 80%) ayant un forfait mensuel ou annuel, et les cartes étant nominatives, on considère que chaque carte représente un utilisateur.
- ‘code_prod’ donne le type de forfait associé à la carte. On regroupera ces forfaits en 3 catégories : étudiants, adultes, et aînés.

La table des arrêts (Tableau 3.2) renvoie les coordonnées de chaque arrêt, selon le système de coordonnées géographiques MTM zone 7.

Tableau 3.2 Extrait de la table des arrêts

ARRET	X_MTM	Y_MTM
1005	249941	5184558
1006	250056	5184637
1007	250153	5184723
1008	250316	5184869
1009	250437	5184966
1010	250543	5184970

3.1.1 Ajustements de l’heure et de la date

Un petit nombre de transactions ont lieu entre minuit et 4h du matin (environ 0,7% de toutes les transactions sur un jour de semaine typique). Presque aucune (moins de 10 par jour) n’ont lieu entre 3h30 et 5h00, car la plupart des lignes sont fermées. Nous avons donc décidé d’utiliser cette période de temps pour séparer une journée de l’autre : au lieu de couper chaque jour à minuit, nous commençons chaque jour à 4h du matin et le finissons à 3h59 le jour suivant (noté 27h59). Nous

supposons ainsi que, pour chaque utilisateur, sa dernière transaction avant 4h est toujours un retour à la maison. De cette façon aucun aller-retour ne peut être coupé en deux, et le domicile d'un utilisateur devrait être correctement identifié.

3.1.2 Création des profils journaliers

Dans l'étape 2, nous créons, pour chaque jour et pour chaque carte, une liste des transactions de cette carte durant ce jour : ceci est surnommé « profil journalier d'un couple carte-jour ». Chaque profil journalier est identifié par une clé formée par l'union de l'ID de la carte ('opus_id' dans le tableau 3.1) avec la date (par exemple 1000006_2019-12-03 pour les transactions du 3 décembre 2019 de la carte 1000006). Nous supposons que chaque carte est associée à un unique utilisateur et que celui-ci utilise toujours la même carte. Ceci est une bonne supposition pour les utilisateurs de transport en commun au Québec : beaucoup d'utilisateurs ont une photo sur leur carte, et la plupart des cartes (environ 80%) sont associées à des forfaits mensuels ou annuels. Le tableau 3.3 donne un extrait de la table alors créée :

Tableau 3.3 Extrait de la table des profils

Daily_profile	Stop_list	Time_list
1000006_2019-12-03	[1457, 1415, 1423]	[453, 1045, 1160]
1000006_2019-12-04	[1455, 1415]	[450, 902]
1000006_2019-12-06	[1455, 1417, 1457]	[449, 776, 1006]
1000013_2019-12-06	[2518, 3457]	[1018, 1314]
1000014_2019-12-02	[1525, 2417]	[471, 1027]

Pour chaque profil journalier ('Daily_profile'), 'Stop_list' donne la liste des lieux de transactions (arrêts visités), et 'Time_list' la liste des heures de transactions (en nombre de minutes). Ainsi par exemple, l'usager possédant la carte 1000006 a réalisé, le 3 décembre 2019, une transaction à l'arrêt 1457 à la 453ème minute de la journée (7h33), une transaction à l'arrêt 1415 à la 1045ème minute (17h25), et une dernière à l'arrêt 1423 à la 1160ème minute (19h20).

Nous choisissons ensuite une certaine période (par exemple, la première semaine de travail de décembre 2019, du lundi 2 décembre au vendredi 6 décembre), et sélectionnons tous les profils journaliers de cette période.

Le transport multimodal peut être non négligeable dans certaines villes : un utilisateur pourrait par exemple aller au travail en bus, puis rentrer chez lui à pied ou en covoiturage. Dans notre cas, seuls les trajets en bus sont détectables; tout autre type de trajet est catégorisé comme un « trajet non détecté ». Certains de nos profils journaliers (environ 30%) ne possèdent qu'une seule transaction, ce qui signifie qu'ils contiennent probablement des trajets non détectés (en faisant l'hypothèse que la première transaction de la journée d'un utilisateur est toujours un départ de chez lui et la dernière un retour chez lui). Tous ces profils, ainsi que tous les profils comportant des trajets dont la destination n'a pas pu être estimée, ont été retirés de la base de données, à cause du manque d'information (difficile d'identifier le domicile de l'utilisateur, ou bien de connaître le temps passé à chaque endroit…). Après cette étape, nous nous retrouvons avec un total d'environ 20 000 profils pour une journée moyenne de semaine.

3.1.3 Élimination des correspondances

Une liste de 24 arrêts de bus, correspondant à l'estimation de la position de l'usager à chaque heure, est créé pour chaque profil journalier. Chaque fois qu'un utilisateur fait une transaction, on lui affecte pour position la destination de ce trajet, jusqu'à sa prochaine transaction. La dernière position de la journée doit être identique à la première (départ et retour chez soi). Pour chaque profil A, nous créons la liste V^A suivante, qui donne à chaque heure de la journée la position de l'utilisateur et le nombre de transactions qu'il a effectué durant cette heure :

$$\forall i \in [\![4,27]\!], \begin{cases} s_i^A : \text{position de l'usager A à l'heure i} \\ n_i^A : \text{nombre de transactions de l'usager A à l'heure i} \\ V_i^A = (s_i^A, n_i^A) \end{cases}$$

Par exemple :

Le 02 décembre 2019, l'utilisateur A a pris le bus à 7h41 à l'arrêt 1362, et est rentré le soir à 17h03 de l'arrêt 1561 :

$$
\begin{array}{lcccccccccc}
 & 4{:}00 & 5{:}00 & 6{:}00 & 7{:}00 & \dots\dots & 16{:}00 & 17{:}00 & \dots\dots & 27{:}00 \\
s_i^A = [& 1362, & 1362, & 1362, & \mathbf{1561}, & \dots\dots, & 1561, & \mathbf{1362}, & \dots\dots, & 1362] \\
n_i^A = [& 0, & 0, & 0, & \mathbf{1}, & \dots\dots, & 0, & \mathbf{1}, & \dots\dots, & 0\ ] \\
V^A = [& (1362,0), & \dots & , & (\mathbf{1561}, \mathbf{1}), & \dots, & (1561,0), & (\mathbf{1362}, \mathbf{1}), & \dots, & (1362,0)]
\end{array}
$$

Cette série spatio-temporelle V^A caractérise le profil journalier d'un utilisateur, et sera utilisée plus tard pour calculer la distance entre deux profils.

Seules l'origine et la destination d'un trajet sont utiles à notre étude. Ainsi, pour chaque profil, toutes les transactions identifiées comme étant des correspondances ont été retirées (étape 3). Les correspondances ont été identifiées de la façon suivante : si deux transactions ont lieu dans un intervalle de moins de 90 minutes et qu'il ne s'agit pas de deux transactions sur la même ligne en sens inverse, alors la deuxième transaction est identifiée comme étant une correspondance.

3.2 Segmentation d'un échantillon

Appliquer notre algorithme de segmentation à l'ensemble des données serait trop chronophage (voir Chapitre 4); nous proposons donc une approche par échantillonnage, en nous inspirant des travaux de (He et al., 2018b).

3.2.1 Choix de l'échantillon

Dans l'étape 4, nous choisissons aléatoirement un échantillon parmi l'ensemble des profils sélectionnés auparavant. Le choix de la taille d'échantillon a fait l'objet d'une étude présentée dans le chapitre 4. Cet échantillon forme un jeu de données S, tandis que le reste des profils sont mis de côté dans un jeu de données $\overline{S}$, et seront plus tard affectés aux *clusters* issus de l'échantillon. Toutes les étapes de la partie 'segmentation d'un échantillon' sont appliquées à S uniquement.

3.2.2 Calcul de la matrice des distances

Nous calculons d'abord la matrice des distances séparant chaque pair de profils journaliers (étape 5). La métrique utilisée pour comparer deux profils est une métrique de type DTW (Dynamic Time Warping).

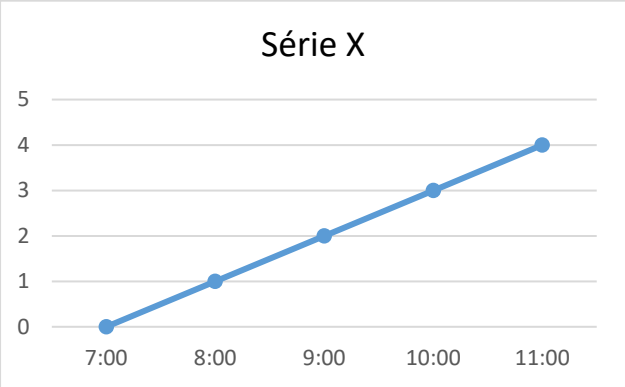
Le principe du DTW est de mesurer la similarité entre deux séries temporelles, en jumelant les points des deux séries d'une manière qui minimise la distance entre deux points jumelés. Les règles suivantes doivent être appliqués pour jumeler les points :

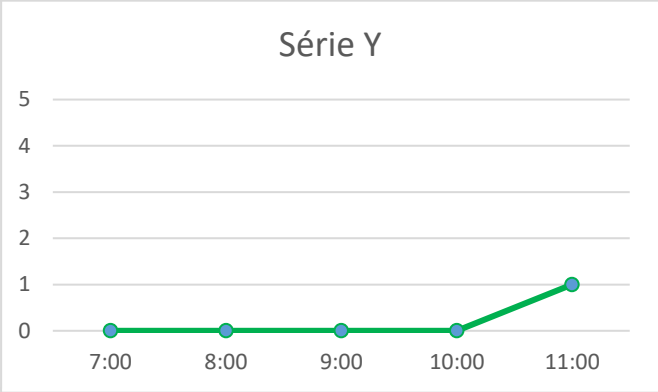
- Le premier point des deux séries doivent être jumelés, ainsi que leur dernier point.
- Les points doivent être jumelés en respectant l'ordre chronologique (un point X de la série A ne peut être jumelé à un point Y de la série B si un point précédant Y dans la série B n'a pas encore été jumelé).
- Un point peut être jumelé à un ou plusieurs points, à condition que ces points ne soient pas déjà jumelés.
- Le paramètre de fenêtre donne le décalage maximal autorisé entre deux points jumelés (par exemple, avec une fenêtre de 2, un point d'une série X à 8h00 peut être jumelé aux points 6h00, 7h00, 8h00, 9h00, ou 10h00 d'une série Y, mais pas à d'autres points).
- La façon dont les points sont jumelés doit minimiser la somme de la distance entre tous les points jumelés.

La distance entre deux points dépend de la métrique choisie (habituellement la métrique Euclidienne).

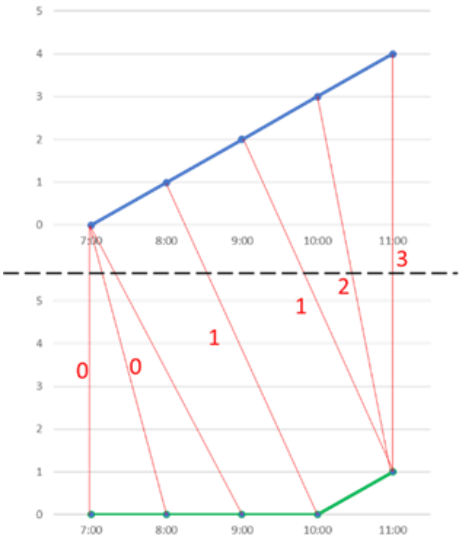
La distance DTW entre deux séries temporelles, renvoyée par l'algorithme DTW, est égal à la somme de toutes les distances entre les points jumelés. Cette distance sera d'autant plus petite que la forme des deux séries est semblable.

Par exemple : si nous prenons les deux séries temporelles : $X = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$ et $Y = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$





Avec une fenêtre de 2h et une métrique Euclidienne, la solution DTW optimale est la suivante :



Le point 7:00 de la série X est jumelé aux points 7:00, 8:00 et 9:00 de la série Y (pas à 10:00 car la fenêtre est de 2h seulement), le point 8:00 de la série X est jumelé au point 10:00 de la série Y, et les points 9:00, 10:00 et 11:00 de la série X sont jumelés au dernier point de la série Y.

La distance DTW entre les séries X et Y est alors la suivante :

$$DTW(X, Y, \text{métrique} = \ euclidienne, \text{fenêtre} = 2)$$
$$= 0 + 0 + 0 + 1 + 1 + 2 + 3 = \mathbf{7}$$

Nous avons décidé d'adapter cette méthode traditionnelle de DTW au cas des séries spatio-temporelles, d'une façon qui permette la modulation de l'influence de l'espace et du temps. Pour un profil journalier A, nous définissons la série VA suivante :

$$\forall i \in [\![4,27]\!], \begin{cases} s_i^A : \text{position de l'usager A à l'heure i} \\ n_i^A : \text{nombre de transactions de l'usagerA à l'heure i} \\ V_i^A = (s_i^A, n_i^A) \end{cases}$$

Nous utilisons la notation : $\begin{cases} V_i^A[1] = s_i^A \\ V_i^A[2] = n_i^A \end{cases}$

Nous définissons les mêmes variables pour un autre profil B :

$$\forall i \in [\![4,27]\!], \begin{cases} s_i^B : \text{position de l'usager B à l'heure i} \\ n_i^B : \text{nombre de transactions de l'usager B à l'heure i} \\ V_i^B = (s_i^B, n_i^B) \end{cases}$$

Par exemple :

L'utilisateur A embarque à 7h41 à l'arrêt 1362, et revient à 17h03 depuis l'arrêt 1561 :

4:00 5:00 6:00 7:00 …… 16:00 17:00 …… 27:00

$$s^A = [1362, 1362, 1362, \mathbf{1561}, \dots\dots, 1561, \mathbf{1362}, \dots\dots, 1362]$$

$$n^A = [\ 0,\ 0,\ 0,\ \mathbf{1},\ \dots\dots,\ 0,\ \mathbf{1},\ \dots\dots,\ 0\ ]$$

$$V^A = [(1362,0), \dots, (\mathbf{1561}, \mathbf{1}), \dots, (1561,0), (\mathbf{1362}, \mathbf{1}), \dots, (1362,0)]$$

L'utilisateur B embarque à 6h12 à l'arrêt 1362 et revient à 16h51 depuis l'arrêt 4111 :

4:00 5:00 6:00 …… 15:00 16:00 …… 27:00

$$s^B = [1362, 1362, \mathbf{4111}, \dots\dots, 4111, \mathbf{1362}, \dots\dots, 1362]$$

$$n^B = [\ 0,\ 0,\ \mathbf{1},\ \dots\dots,\ 0,\ \mathbf{1},\ \dots\dots,\ 0\ ]$$

$$V^B = [(1362,0), \dots, (\mathbf{4111}, \mathbf{1}), \dots, (4111,0), (\mathbf{1362}, \mathbf{1}), \dots, (1362,0)]$$

Nous définissons ensuite $E(s_n, s_m)$ comme la distance Euclidienne entre deux arrêts s_n et s_m :

$$\forall n, m \in [\![1, N]\!], \begin{cases} E(s_n, s_m) = \sqrt{(x_n - x_m)^2 + (y_n - y_m)^2} \\ (x_n, y_n): \textit{coordonnées spatiales de l'arrêt } s_n \\ (x_m, y_m): \textit{coordonnées spatiales de l'arrêt } s_m \end{cases} \quad [\mathbf{1}]$$

N étant le nombre total d'arrêts de bus desservis par la RTC.

A partir de cette distance, nous définissons la métrique spatiale suivante :

$$\forall i,j \in [\![4,27]\!], \begin{cases} S\left(V_i^A, V_j^B\right) = \frac{E\left(s_i^A, s_j^B\right)}{E_{max}} \\ E_{max} = \max\limits_{n,m \in [\![1,N]\!]} E(s_n, s_m) = 35162\ m \end{cases} \qquad [\mathbf{2}]$$

E_{max} étant la distance maximale entre deux arrêts de bus de la RTC.

Nous définissons ensuite la métrique temporelle suivante, qui détermine si les transactions ont été faites à la même heure dans les deux profils :

$$\forall i,j \in [\![4,27]\!], \quad T\left(V_i^A, V_j^B\right) = \left|n_i^A - n_j^B\right| \qquad \mathbf{[3]}$$

Nous utilisons ces deux dernières métriques pour créer la métrique spatio-temporelle suivante :

$$\forall i,j \in [\![4,27]\!], \quad \begin{cases} M\left(V_i^A, V_j^B\right) = \alpha * S\left(V_i^A, V_j^B\right) + (1-\alpha) * T\left(V_i^A, V_j^B\right) \\ \alpha \in [0,1] \end{cases} \qquad [\mathbf{4}]$$

Le choix du paramètre α influe sur les poids des facteurs spatial et temporels de l'équation.

Par exemple, si on souhaite calculer la distance spatio-temporelle $M\left(V_i^A, V_j^B\right)$ entre le profil A à i=5h00 et le profil B à j=7h00 :

- A 5h00, l'utilisateur A est à l'arrêt 1362, et à 7h00 l'utilisateur B est à l'arrêt 4111. Donc :

$$S\left(V_i^A, V_j^B\right) = \frac{E(\mathbf{1362}, \mathbf{4111})}{\boldsymbol{E_{max}}} = \frac{10065\ m}{35162\ m} = 0{,}2862$$

- L'utilisateur A ne fait aucune transaction entre 5h et 5h59 et l'utilisateur B fait une transaction entre 7h et 7h59. Donc :

$$T\left(V_i^A, V_j^B\right) = |0 - \mathbf{1}| = 1$$

Si on choisit $\alpha = 0{,}25$:

$$M\left(V_i^A, V_j^B\right) = \alpha * 0{,}2862 + (1-\alpha) * 1 = 0{,}25 * 0{,}2862 + 0{,}75 * 1 = \mathbf{0{,}82155}$$

La métrique finale D utilisée pour comparer les séries V^A et V^B est une métrique DTW qui utilise $M\left(V_i^A, V_j^B\right)$ comme métrique intermédiaire, avec une fenêtre de 1 :

$$D(V^A, V^B) = DTW\left(V^A, V^B, \text{métrique} = M\left(V_i^A, V_j^B\right), \text{fenêtre} = 2\right) \qquad [\mathbf{5}]$$

Cela signifie que les points des séries V^A et V^B seront jumelés d'une manière qui minimise la distance entre deux points jumelés, cette distance étant calculée avec la métrique $M\left(V_i^A, V_j^B\right)$. Chaque point doit être jumelé avec un ou plusieurs points de l'autre série (en respectant l'ordre

chronologique), et les premiers points de V^A et V^B doivent être jumelés ensemble, ainsi que leurs derniers points. Deux points jumelés ne doivent pas être séparés par plus d'une heure ($|i - j| \leq 1$). En respectant ces règles, l'algorithme DTW trouve une combinaison de points jumelés qui minimise la somme des distances entre chaque paire de points jumelés. La distance $D(V^A, V^B)$ renvoyé est égale à cette somme :

$$D(V^A, V^B) = \sum_{i,j} M\left(V_i^A, V_j^B\right) \qquad [\mathbf{6}]$$

Chaque paire (i,j) symbolisant un couple de points jumelés.

De manière plus rigoureuse, ceci peut être défini ainsi :

$$D(V^A, V^B) = \min_{K \in \mathbb{N}} \sum_{k=1}^{K} M(V_{i_k}^A, V_{j_k}^B) \,/ \begin{cases} \forall k_1, k_2 \in [\![1, K]\!], i_{k_1} \leq i_{k_2} \;\; and \;\; j_{k_1} \leq j_{k_2} \\ i_1 = j_1 \\ i_K = j_K \\ [\![4,27]\!] \subset \{\, i_k \;\; , k \in [\![1, n]\!]\} \\ [\![4,27]\!] \subset \{\, j_k \;\; , k \in [\![1, n]\!]\} \\ \forall k \in [\![1, K]\!], |i_k - j_k| \leq 2 \end{cases} \qquad [\mathbf{7}]$$

Pour reprendre l'exemple précédent, avec une fenêtre de 1, la combinaison optimale de points jumelés serait la suivante :

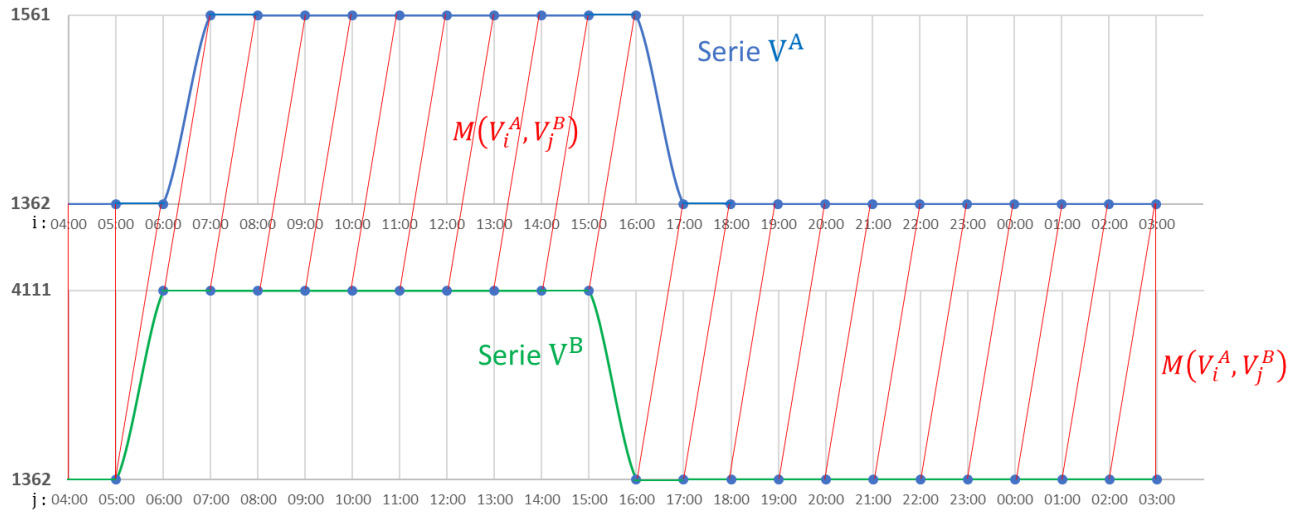


Figure 3.3 Algorithme DTW

Dans ce cas, pour α = 0,25 :

$$D(V^A, V^B) = 3 * 0 + 10 * \left(\alpha * \frac{E(\mathbf{1561, 4111})}{\mathrm{E_{max}}} + (1 - \alpha) * 0\right) + 12 * 0$$
$$= 10 * 0{,}25 * \frac{10357\ m}{35162\ m} = \mathbf{0,7364}$$

3.2.3 Segmentation hiérarchique de l'échantillon

En utilisant la matrice de distances de l'échantillon, il est désormais possible de classifier des profils journaliers de l'échantillon en différent groupes (étape 6), en regroupant les profils les plus proches ensembles. Nous utilisons pour cela l'algorithme CHA (Algorithme de Classification Hiérarchique), avec l'*ultramétrique* médiane ("Statistics toolbox." API Reference Documentation. The MathWorks.). Cet algorithme crée des groupes récursifs s'emboîtant, de sorte que n'importe quel groupe puisse être divisé indéfiniment en sous-groupes jusqu'à ce que des groupes d'un seul profil soient atteints (voir Figure 3.4).

Un nombre spécifique de *clusters* doit être choisi dans cette étape, en fixant une distance maximale acceptable entre deux profils du même groupe. Plusieurs nombres de *clusters* sont testés dans chaque cas : le nombre optimal de *clusters* est alors choisi en étudiant l'évolution du score Silhouette (Rousseeuw, 1987). Pour un point donné, l'indicateur de Silhouette correspond à la différence entre la distance moyenne de ce point aux points extérieurs à son *cluster* et la distance moyenne de ce point aux autres points du même *cluster*. Le score de Silhouette que nous utilisons correspond à la moyenne de cet indicateur sur tous les points d'échantillon. Ce score sera donc d'autant plus élevé que les *clusters* sont éloignés les uns des autres et que les points au sein d'un *cluster* sont rapprochés.

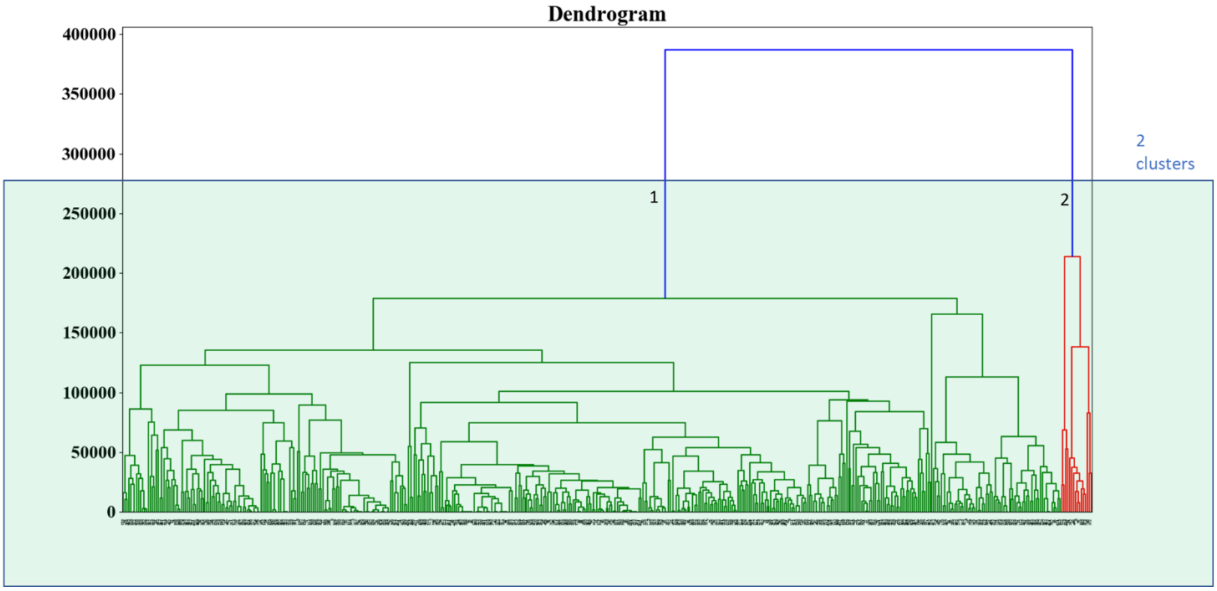


Figure 3.4 Dendrogramme issu de l'algorithme CHA

3.2.4 Identification des centroïdes

Une fois l'échantillon divisé en *clusters*, le centroïde de chaque *cluster* est alors calculé (étape 7). Pour un *cluster* donné, le centroïde est le point qui minimise sa distance moyenne à tous les points du *cluster*. Nous utilisons les centroïdes plutôt que des valeurs moyennes, car ces dernières ne peuvent être calculées avec la métrique utilisée, celle-ci ne respectant pas l'inégalité triangulaire (selon laquelle la distance AB entre deux points A et B est toujours inférieure à la somme des distances AC et CB séparant A et B d'un quelconque point C).

3.3 Affectation des profils restants

À la fin de la segmentation de l'échantillon, chaque profil journalier de l'échantillon a été assigné à un *cluster*. Cependant, tous les profils du jeu de données $\bar{S}$ (de l'étape 4) doivent encore être assignés aux *clusters* de S. Pour identifier le *cluster* le plus proche d'un profil, nous calculons la distance entre ce profil et chacun des centroïdes (étape 8), en utilisant la même métrique que dans l'étape 5. Le *cluster* dont le centroïde est le plus proche est alors choisi, et le profil est affecté à ce *cluster* (étape 9). À la fin de cette partie, chaque profil journalier a été affecté à un *cluster*.

Un extrait de la table alors obtenue est représenté dans le tableau 3.4 ci-dessous :

Tableau 3.4 Extrait de la table des résultats de segmentation

Daily_profile	Stop_time_table	Cluster	code_prod
1838345_2019-12-02	[([249342.0, 5185238.0], 0), ([249342.0, 5185238.0], 0), ([2493	1	TLM-E
1781767_2019-12-02	[([250197.0, 5186033.0], 0), ([250197.0, 5186033.0], 0), ([2501	7	TLM-E
1781770_2019-12-02	[([250147.0, 5189973.0], 0), ([239498.0, 5190836.0], 1), ([2394	1	TLM-G
1797560_2019-12-02	[([247886.0, 5189923.0], 0), ([247886.0, 5189923.0], 0), ([2478	2	TLM-G
1797508_2019-12-02	[([238949.0, 5190316.0], 0), ([238949.0, 5190316.0], 0), ([2389	3	LPUUL
1835962_2019-12-02	[([246777.0, 5182756.0], 0), ([246777.0, 5182756.0], 0), ([2467	2	TLM-G

- ‘Stop_time_table’ est le vecteur V^A décrit dans la partie 3.2.2, donnant la position et le nombre de transactions de l’usager à chaque heure.
- ‘*Cluster*’ est le numéro de *cluster* auquel le profil journalier a été assigné.
- ‘code_prod’ est le type de forfait de la carte de l’utilisateur (voir tableau 3.1).

3.4 Analyse des résultats

Les résultats sont alors analysés. Pour cela, nous avons développé deux types d’outils : des outils de visualisation, qui permettent une analyse qualitative des résultats, et des indicateurs de performance spatiaux, temporels et spatio-temporels, qui permettent une analyse plus quantitative des résultats.

3.4.1 Outils de visualisation

<u>Carte du *cluster* le plus fréquent au départ pour chaque arrêt (Figure 3.5):</u>

Nous nous intéressons ici à l’arrêt de départ de chaque profil, c’est-à-dire à l’arrêt où a lieu la première transaction de la journée pour un profil. Pour chaque arrêt de bus, nous calculons dans chaque *cluster* le nombre d’usagers dont cet arrêt est l’arrêt de départ. Le *cluster* le plus présent à cet arrêt donne alors sa couleur à l'arrêt.

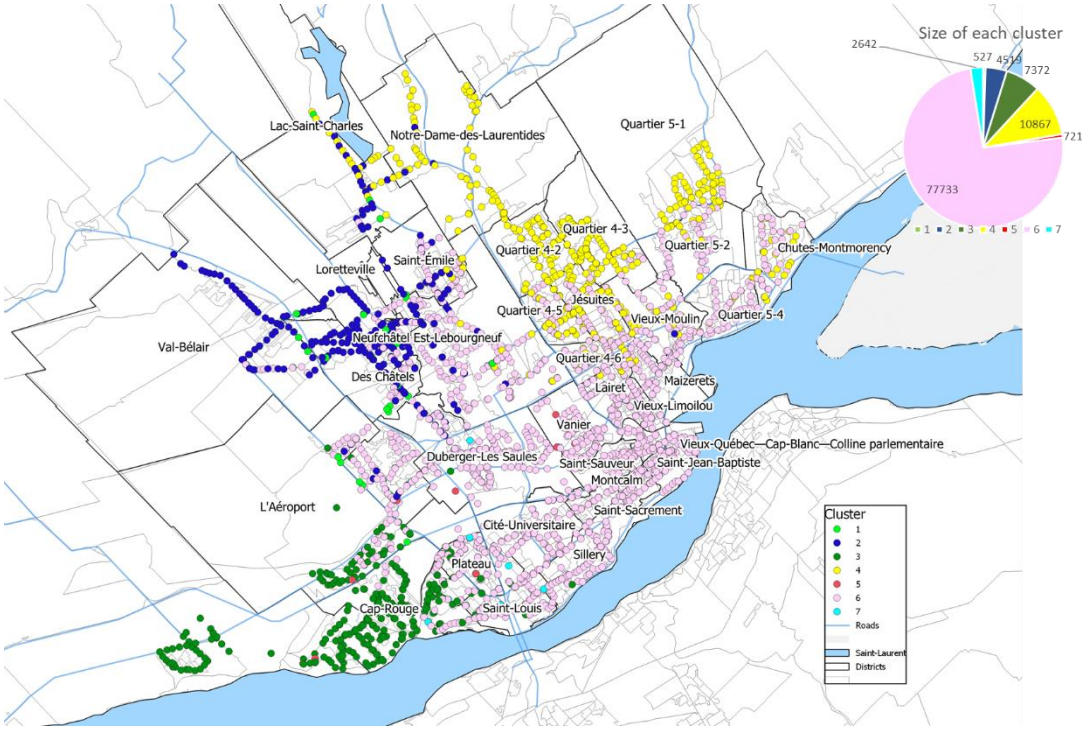


Figure 3.5 Carte des *clusters* les plus fréquents au départ

L'arrêt de départ d'un usager étant en général à proximité de son lieu d'habitation, cette carte permet donc une bonne approximation de la zone d'habitation des usagers présents dans chaque *cluster*.

<u>Carte de la position des usagers à une heure donnée (Figure 3.6) :</u>

Cet outil permet d'étudier le déplacement d'un *cluster* au cours de la journée, et d'identifier les zones d'attraction de ce *cluster*. Il s'agit d'une carte de chaleur donnant la position des usagers d'un *cluster* en particulier à une heure donnée.

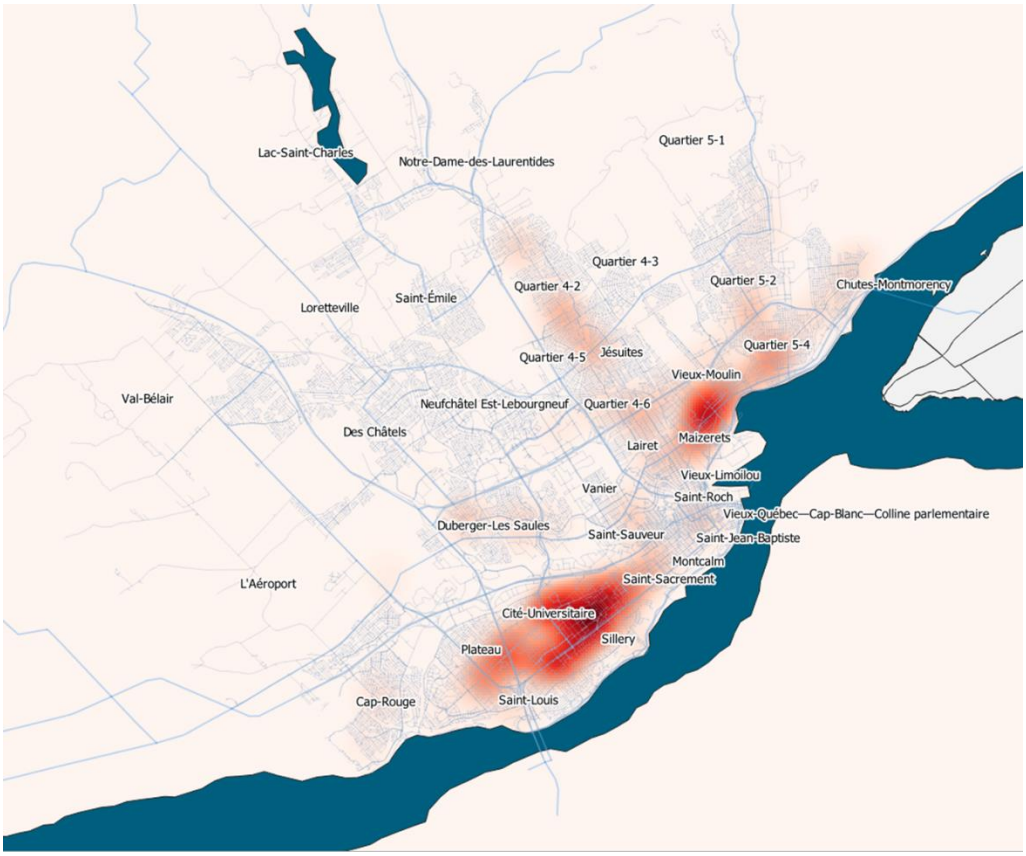


Figure 3.6 Carte de chaleur d'un *cluster* à une heure donnée

<u>Carte des directions de chaque *cluster* (Figure 3.7) :</u>

Nous utilisons ici, pour chaque profil, l’arrêt de départ de l’usager (identifié comme domicile de cet usager) et l’arrêt le plus distant de son domicile, parmi tous les arrêts visités au cours de sa journée. On nomme « vecteur de déplacement d’un profil » le vecteur orienté entre ces deux arrêts (du premier vers le deuxième). Pour chaque *cluster*, nous calculons la moyenne des vecteurs de déplacement de ses profils. Ce vecteur est alors représenté sur la carte par une flèche, dont l’épaisseur est fonction du nombre de profils dans le *cluster*.

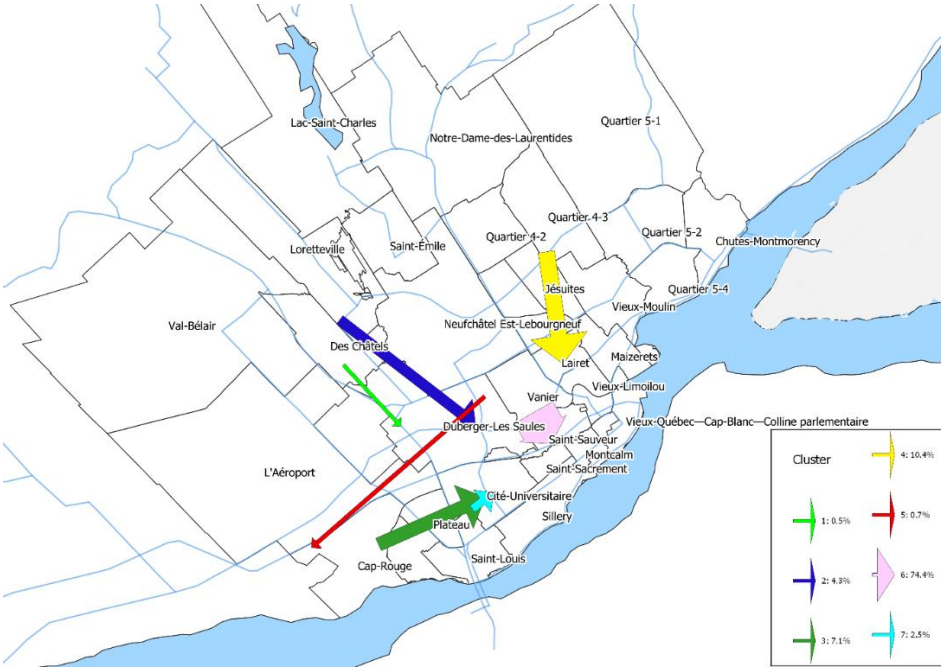


Figure 3.7 Carte de direction des *clusters*

<u>Nombre de transactions d'un *cluster* à chaque heure de la journée (Figure 3.8) :</u>

Cet outil renseigne davantage sur les caractéristiques temporelles des *clusters*. Pour chaque *cluster*, nous représentons le graphique du nombre de transactions à chaque heure de la journée calculé sur toute la durée étudiée (par exemple, pour le point '17h', toutes les transactions entre 17h et 17h59 tous les jours de la période étudiée). Ces graphiques possèdent en général deux pics : un pour les transactions de départ et un pour celles de retour.

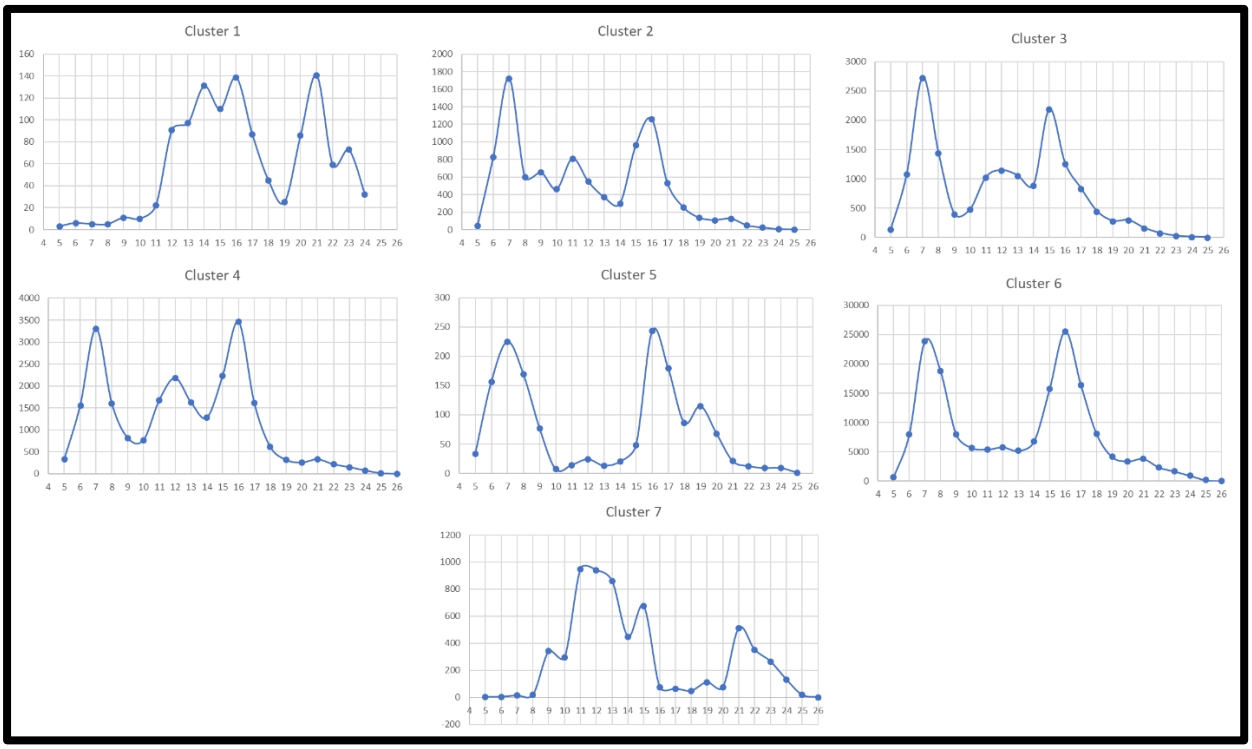


Figure 3.8 Nombre de transactions à chaque heure, dans chaque *cluster*

<u>Trajectoires espace-temps des *clusters* (Figure 3.9) :</u>

Cet outil résume, en trois dimensions, les deux outils précédents. On représente le temps selon l'axe vertical. Chaque courbe donne la position du centre de gravité spatial des usagers d'un *cluster*, au cours de la journée.

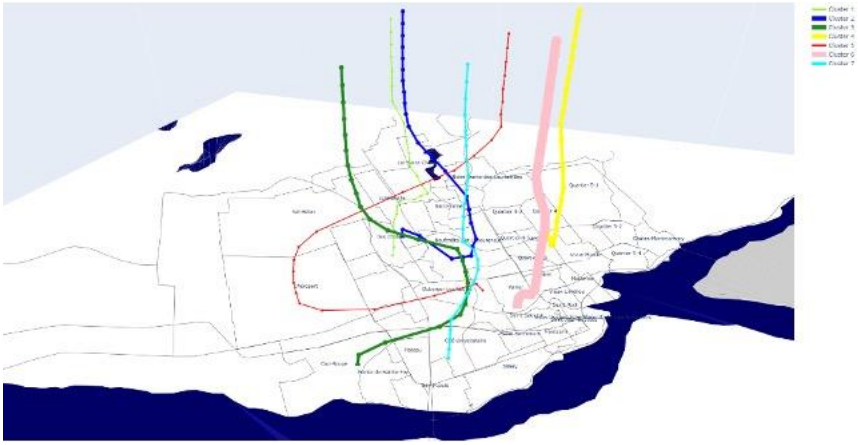


Figure 3.9: Trajectoires espace-temps des *clusters*

3.4.2 Indicateurs de performance

Nous avons développé une série d'indicateurs qui permettent de mesurer et de comparer l'importance des facteurs temporels et spatiaux dans chaque simulation, ou d'évaluer la qualité d'une simulation.

<u>Indicateurs temporels :</u>

Les indicateurs suivants caractérisent un *cluster* en particulier :

- **Heure moyenne de départ dans un *cluster*** (heure moyenne de la première transaction de l'usager)
- **Heure moyenne de retour dans un *cluster*** (heure moyenne de la dernière)
- **Dispersion des départs dans un *cluster*** :

  Cet indicateur s'obtient en mesurant l'écart moyen entre l'heure de départ d'un usager du *cluster* et l'heure moyenne de départ de tous les usagers du *cluster*.

$$\begin{cases} Disp_k = \sum_{i=4h}^{27h} \frac{d_{i,k}}{D_k} * \Delta_{i_k} \\ d_{i,k}: \text{nombre de départs à l'heure i dans le } cluster \text{ k} \\ D_k: \text{nombre total de départs du } cluster \text{ k} \\ \Delta_{i,k}: \text{écart entre l'heure i et l'heure moyenne de départ du } cluster \text{ k} \end{cases} \quad [\mathbf{8}]$$

La figure 3.10 ci-dessous donne une représentation graphique du calcul de cette dispersion des départs sur un *cluster*, numéroté 4.

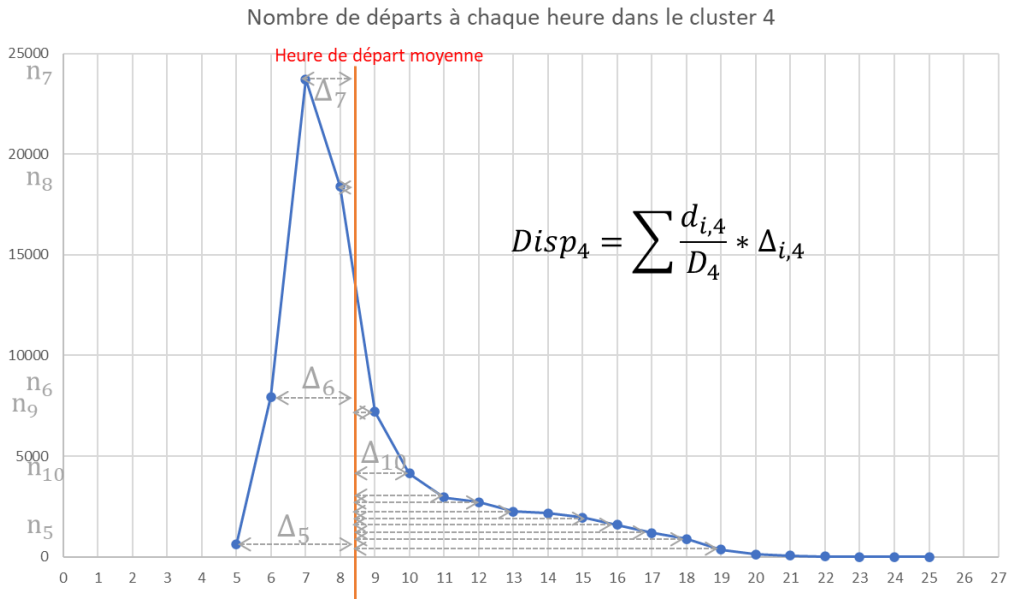


Figure 3.10 Dispersion des départs dans un *cluster* 4

- **Dispersion des retours** (se calcule de la même façon que la dispersion des départs)
- **Temps de sortie moyen du *cluster***: écart entre l'heure de départ moyenne et l'heure de retour moyenne
- **Nombre moyen de transactions du *cluster*** : moyenne du nombre de transactions quotidien de chaque usager du *cluster*

Pour caractériser non pas un *cluster*, mais l'ensemble des *clusters* de la simulation, nous pondérons les indicateurs précédents par la taille de groupe des *clusters* (lorsque ceci est pertinent) :

Par exemple :

- **La dispersion moyenne pondérée des départs** :

$$\begin{cases} Disp = \sum_{k=1}^{K} \frac{n_k}{N} Disp_k \\ n_k : \text{taille du } cluster \text{ k} \\ N : \text{nombre total de profils} \\ Disp_k : \text{Dispersion des départs du } cluster \text{ k} \end{cases} \quad [\mathbf{9}]$$

De la même façon, nous calculons la **dispersion moyenne pondérée des retours**.

Indicateurs spatiaux :

Les indicateurs suivants caractérisent un *cluster* en particulier :

- **Dispersion spatiale des usagers du *cluster*** : distance (euclidienne) moyenne au centre de gravité du *cluster*. Cet indicateur peut être calculé à n'importe quelle heure de la journée. Sa moyenne sur une journée peut également être utilisée.

$$\begin{cases} SDisp_k = \sum_{i=1}^{N_s} \frac{p_{i,k}}{P_k} * \delta_{i,k} \\ p_{i,k} : \text{nombre d'usagers du } cluster \text{ k à l'arrêt i} \\ P_k : \text{Taille du } cluster \text{ k} \\ \delta_{i,k} : \text{distance entre l'arrêt i et le centre de gravité du } cluster \text{ k} \end{cases} \quad [\mathbf{10}]$$

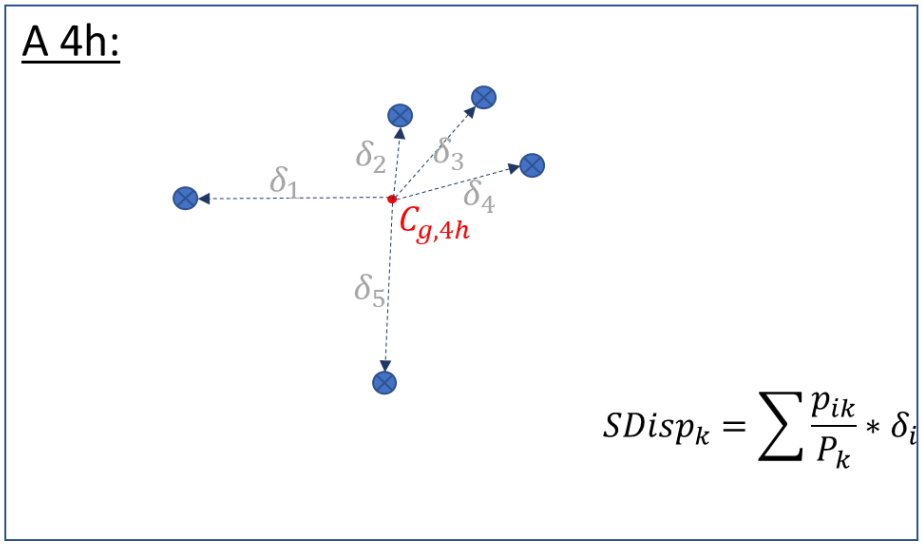


Figure 3.11 Dispersion spatiale du *cluster* k à 4h

- **Amplitude du déplacement du centre de gravité d'un *cluster* :** Distance maximale entre le centre de gravité des usagers du *cluster* à 4h du matin et son centre de gravité à n'importe quelle heure de la journée.

$$\begin{cases} A_k = \max_{4 \le i \le 27} E(Cg_{k,4}, Cg_{k,i}) \\ E : \text{distance euclidienne} \\ Cg_{k,i} : \text{centre de gravité spatiale du } cluster \text{ k à l'heure i} \end{cases} \quad [\mathbf{11}]$$

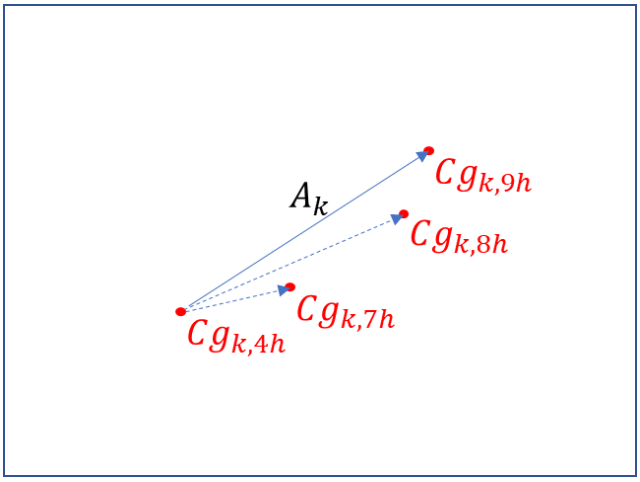


Figure 3.12 Amplitude de déplacement du centre de gravité du *cluster* k

De la même façon que pour les indicateurs temporels, nous pondérons ces indicateurs spatiaux par la taille des *clusters* pour caractériser l'ensemble des *clusters* d'une simulation. Nous utilisons donc la **dispersion spatiale moyenne pondérée** et le **déplacement moyen du centre de gravité**.

<u>Indicateurs spatio-temporels :</u>

La performance d'une simulation se mesure généralement selon deux critères: l'hétérogénéité des différent groupes créés et l'homogénéité au sein de chaque groupe. La première doit être maximisée et la deuxième minimisée. Voici comment nous avons mesuré ces deux critères :

L'hétérogénéité d'un *cluster* est mesurée par la distance moyenne de son centroïde aux centroïdes des autres *clusters*. Pour un *cluster* k, celle-ci est notée H_k :

$$\begin{cases} H_k = \frac{1}{K}\sum_{i=1}^{K} D(m_k, m_i) \\ m_k\text{: centroïde du } cluster \text{ k} \\ m_i\text{: centroïde du } cluster \text{ i} \\ D\text{: métrique spatiotemporelle décrite dans la section 3.2.2} \end{cases} \qquad [\mathbf{12}]$$

L'hétérogénéité globale d'un résultat de segmentation, notée H, correspond à la moyenne de l'hétérogénéité des *clusters* pondérée par la taille de ces *clusters* :

$$\begin{cases} H = \sum_{k=1}^{K} \frac{n_k}{N} H_k \\ n_k\text{: taille du } cluster \text{ k} \\ N\text{: nombre total de profils} \\ H_k\text{: Hétérogénéité du } cluster \text{ k} \end{cases} \qquad [\mathbf{13}]$$

L'homogénéité d'un *cluster* k est mesurée par la distance moyenne de tous ses points à son centroïde. On la note h_k :

$$\begin{cases} h_k = \frac{1}{n_k}\sum_{i=1}^{n_k} D(m_k, e_{i,k}) \\ m_k\text{: centroïde du } cluster \text{ k} \\ e_{i,k}\text{: élément i du } cluster \text{ k} \\ D\text{: métrique spatiotemporelle décrite dans la section 3.2.2} \end{cases} \qquad [\mathbf{14}]$$

L'homogénéité globale d'un résultat de segmentation, notée h, correspond à la moyenne de l'homogénéité des *clusters* pondérée par la taille de ces *clusters* :

$$\begin{cases} h = \sum_{k=1}^{K} \frac{n_k}{N} h_k \\ n_k\text{: taille du } cluster \text{ k} \\ N\text{: nombre total de profils} \\ h_k\text{: Hétérogénéité du } cluster \text{ k} \end{cases} \qquad [\mathbf{15}]$$

Ainsi, H caractérise l'hétérogénéité des *clusters*, que l'on cherche à maximiser, et h leur homogénéité, que l'on cherche à minimiser. Afin d'avoir une mesure globale de la qualité des *clusters*, nous définissons également le rapport R entre hétérogénéité et homogénéité :

$$R = \frac{H}{h} \qquad [\mathbf{16}]$$

CHAPITRE 4 ÉTUDE ET OPTIMISATION DU MODÈLE CRÉÉ

Dans ce chapitre, nous présentons tout d'abord un la mise en œuvre de notre méthode de segmentation sur une journée de données, avec des paramètres fixés de manière arbitraire, afin de vérifier son bon fonctionnement (Partie 4.1). Nous tentons ensuite de déterminer une taille d'échantillon maximisant la qualité des *clusters* (Partie 4.2). Après avoir décidé d'une taille d'échantillon optimale, nous étudions l'influence du paramètre d'équilibre spatio-temporel alpha sur les indicateurs spatiaux et temporels présentés plus tôt (Partie 4.3). Ceci permet d'étudier le contrôle de ce paramètre sur la nature spatiale et temporelle des résultats. Nous vérifions ensuite si les tendances observées restent les mêmes pour un nombre de *clusters* plus petit ou plus grand (Partie 4.4). Enfin, nous étudions dans la partie 4.5 comment la taille de fenêtre DTW influe sur les résultats de la partie 4.3, dans le but de sélectionner une taille de fenêtre optimale.

4.1 Mise en œuvre de la méthode

La méthode a été testé dans un premier temps sur la journée du 2 décembre 2019, avec un échantillon de taille 1000, une fenêtre de 1, et un paramètre d'équilibre spatio-temporel alpha=0,5. L'ordinateur utilisé a un processeur de type Intel® Core™ i7-10510U, de fréquence maximale de 2.30 GHz, et une RAM de 16 Go.

Nous procèdons dans un premier temps à l'étape de préparation des données (étapes 1 à 4). A l'issu de cette étape, nous disposons d'une table de 19297 profils journaliers. Nous calculons ensuite la matrice de distances de l'échantillon (étape 5), et nous appliquons à cet échantillon l'algorithme de segmentation hiérarchique (étape 6). Le calcul de la matrice prend 24 minutes. Le score de Silhouette moyen des *clusters*, en fonction du nombre de *clusters*, est ensuite calculé (Figure 4.1).

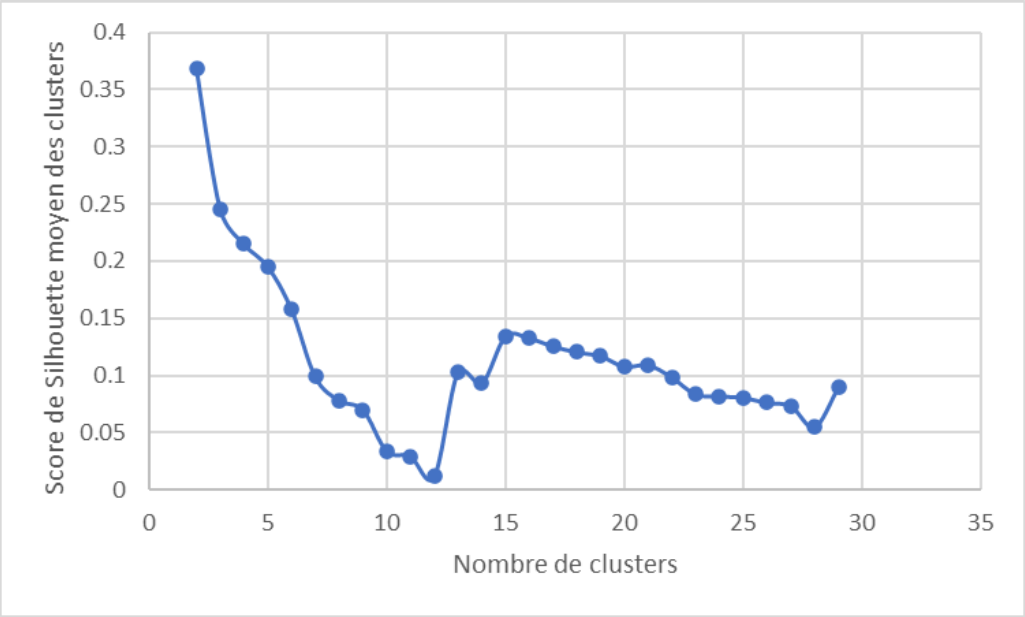


Figure 4.1 Score de silhouette en fonction du nombre de *clusters*

Nous constatons que celui-ci atteint un maximum local pour 15 *clusters*. Nous divisons donc l'échantillon en 15 *clusters*, et nous calculons les centroïdes de ces *clusters* (étape 7).

Nous affectons ensuite tous les profils restants à l'un des *clusters* (étapes 8 et 9). Cette opération prend 13 minutes. En ajoutant à ceci le au temps de calcul de la matrice (24 minutes) et les autres temps de calcul (algorithme CHA, création des indicateurs et des graphiques), de moins de 2 minutes, le temps de calcul total est d'environ 39 minutes.

Voici les résultats obtenus (Figures 4.2 et 4.3):

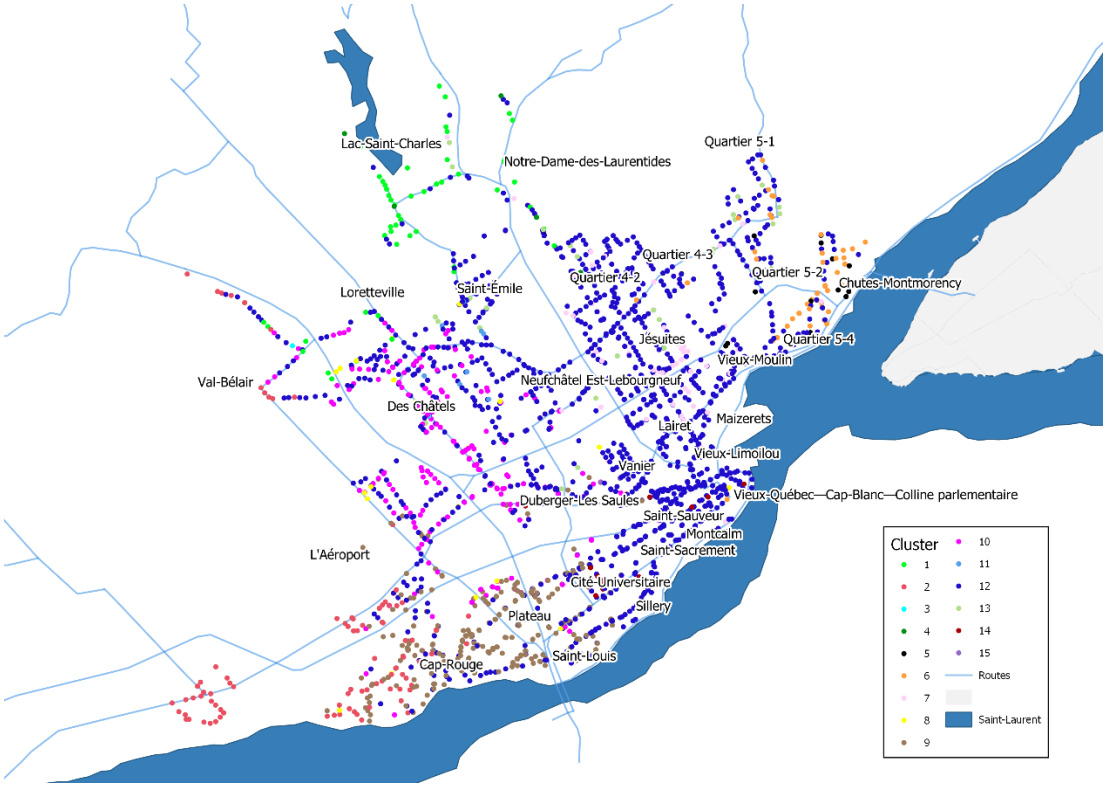


Figure 4.2 Carte des *clusters* les plus fréquents au départ

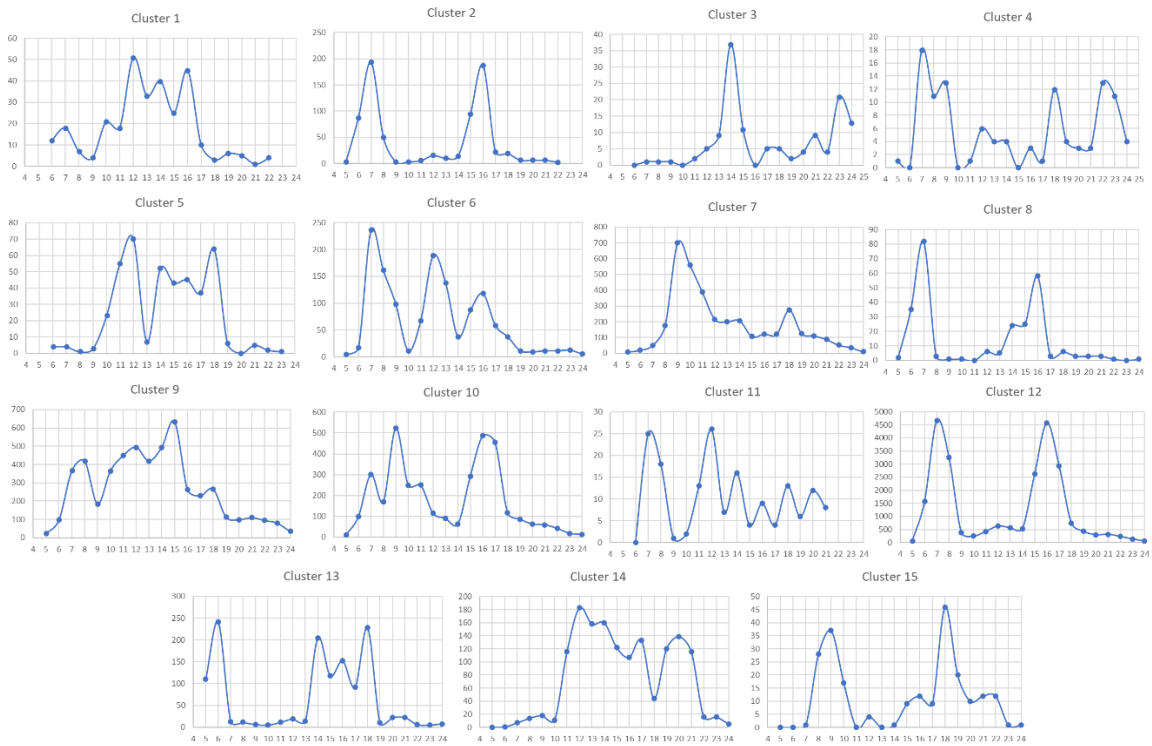


Figure 4.3 Nombre de transactions à chaque heure, dans chaque *cluster*

Nous constatons que les *clusters* créés diffèrent tant sur le plan temporel (Figure 4.3) que sur le plan spatial (Figure 4.2) : notre modèle permet bien de réaliser une segmentation spatio-temporelle des profils journaliers. La section 4.3 (étude de l'influence de alpha) permet d'étudier si l'équilibre entre l'influence du temps et de l'espace sur cette segmentation est véritablement contrôlable.

4.2 Choix de la taille d'échantillon

Dans un premier temps, nous avons étudié l'influence de la taille d'échantillon sur les résultats, dans le but de trouver, pour un certain nombre de profils total, une taille d'échantillon qui optimiserait les résultats tout en respectant un temps de calcul raisonnable.

Nous avons donc réalisé une série de tests, en conservant toujours une fenêtre de 1, un nombre de *clusters* de 10, et les mêmes profils journaliers, appartenant à la journée du 2 décembre 2019 (soit 19297 profils); et en faisant varier la taille d'échantillon. Pour chaque taille d'échantillon, nous avons réalisé 4 tests et avons calculé la moyenne de ces 4 tests, afin de tenir compte du caractère aléatoire des échantillons.

A partir de cette moyenne, nous avons constaté que le temps de calcul suivait une fonction de type polynomiale, variant avec le carré de la taille d'échantillon, et avec un terme constant de 11 minutes, correspondant au temps d'allocation. Cette fonction est représentée sur la figure 4.4.

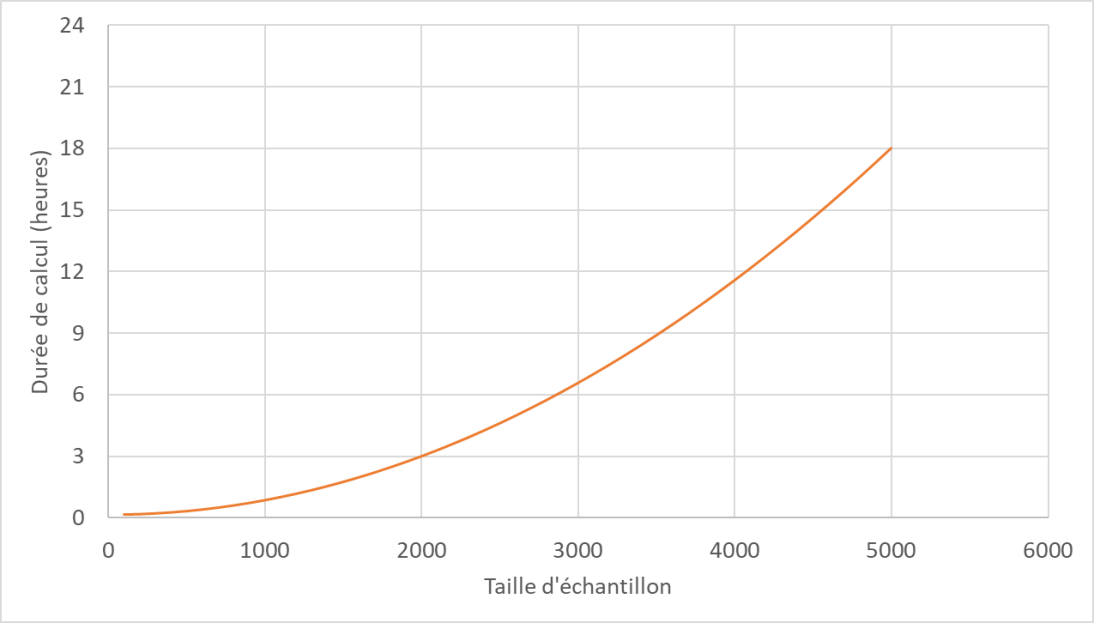


Figure 4.4 Temps de calcul total en fonction de la taille d'échantillon

Au-delà d'une taille d'échantillon de 2000, le temps de calcul dépasse 3h et augmente de plus en plus pour atteindre 18h pour une taille d'échantillon de 5000. Nous choisissons de ne dépasser une taille d'échantillon de 3000 pour ne pas dépasser 6h de simulation.

Pour mesurer la qualité des résultats nous utilisons le rapport R entre l'hétérogénéité et l'homogénéité moyenne des *clusters*, défini dans la section 3.4.2. La variation de ce rapport avec la taille d'échantillon est représenté sur la figure 4.5.

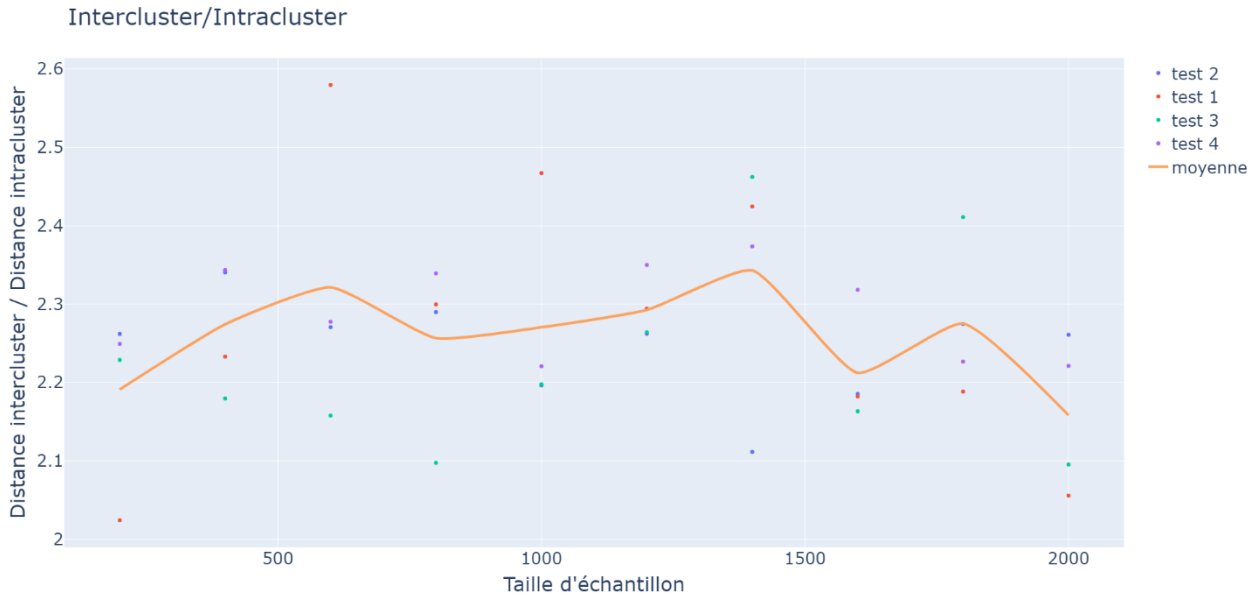


Figure 4.5 Variation de R avec la taille d'échantillon

Si on constate une certaine hausse de R entre 0 et 600, ce rapport semble stagner jusque 1500, puis décliner légèrement jusque 2000. Par ailleurs une variabilité importante entre les tests 1 à 4 est observée, quelle que soit la taille d'échantillon. Cette variabilité excède l'amplitude des variations de R : nous ne pouvons donc pas tirer de résultats très précis de cette analyse.

Il est à noter cependant que les centroïdes utilisés pour calculer ces deux indicateurs correspondent aux centroïdes des *clusters* avant la phase d'allocation. En effet, calculer les centroïdes réels des *clusters*, après la phase d'allocation, reviendrait à calculer 10 matrices de distance d'environ 2000 profils (19297/10), ce qui, d'après la figure 4.4, reviendrait à un temps de calcul d'environ 30h pour chaque test. Il est fort possible cependant que l'utilisation des centroïdes réels apporteraient des résultats plus faciles à interpréter.

Cette étude pourrait également être poursuivie en utilisant d'autres indicateurs de performance de segmentation, tels que les indices de Dunn ou de Davies-Bouldin, plus utilisés dans la littérature.

Nous nous contenterons dans la suite d'une taille d'échantillon supérieure à 500 et inférieure à 3000 pour limiter le temps de calcul. Nous utiliserons une taille d'échantillon de 1000 pour la segmentation sur une journée (environ 20 000 profils) et une taille d'échantillon de 3000 pour la segmentation sur un mois (environ 300 000 profils). Ceci revient à prendre une taille d'échantillon

comprise entre 1% et 5% du nombre total de profils, ce qui s'approche des tailles d'échantillon utilisés par Li He dans ces études de segmentation spatio-temporelle (He et al., 2018b).

4.3 Étude de l'influence du paramètre d'équilibre spatio-temporel

Pour mieux comprendre l'influence du paramètre d'équilibre spatio-temporel α sur les résultats, nous avons étudié l'évolution de certains indicateurs spatiaux et temporels lorsque α varie. Pour chaque valeur de alpha, nous avons réalisé trois tests (test 1, test 2, et test 3) avec des échantillons aléatoires différents. Le nombre de *clusters* est toujours de 10, et le paramètre fenêtre de l'algorithme est fixé à 1. Nous traçons la dispersion pondérée des départs en fonction de alpha pour chacun des tests, ainsi que sa moyenne (Figure 4.6). De même nous traçons la dispersion pondérée des retours (Figure 4.7), la dispersion spatiale pondérée (Figure 4.8) et le déplacement moyen pondéré du centre de gravité (Figure 4.9)

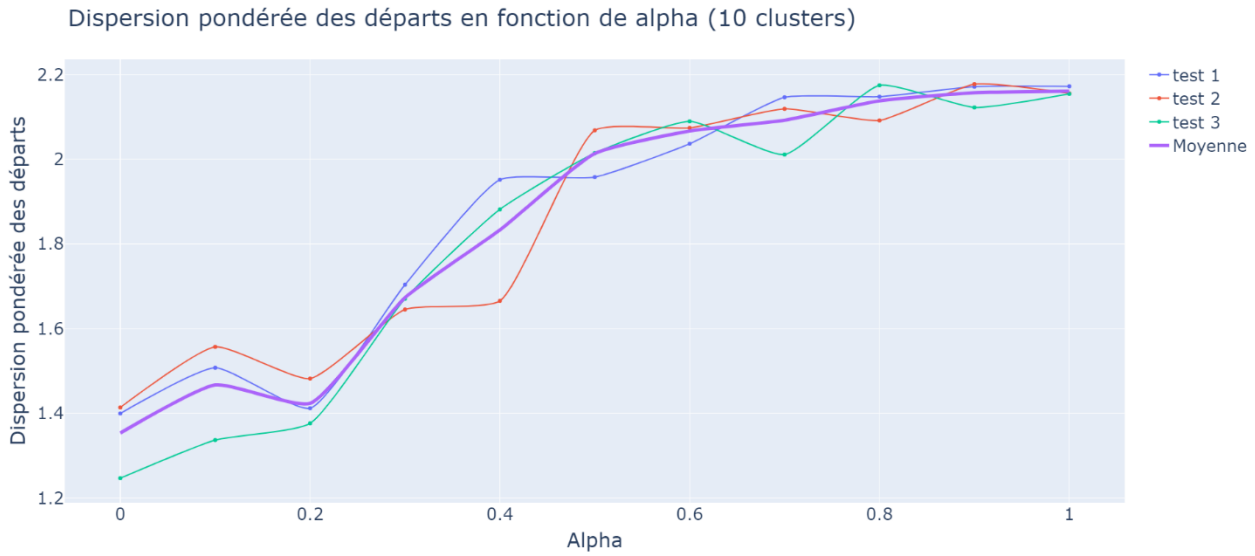


Figure 4.6 Dispersion pondérée des départs en fonction de alpha

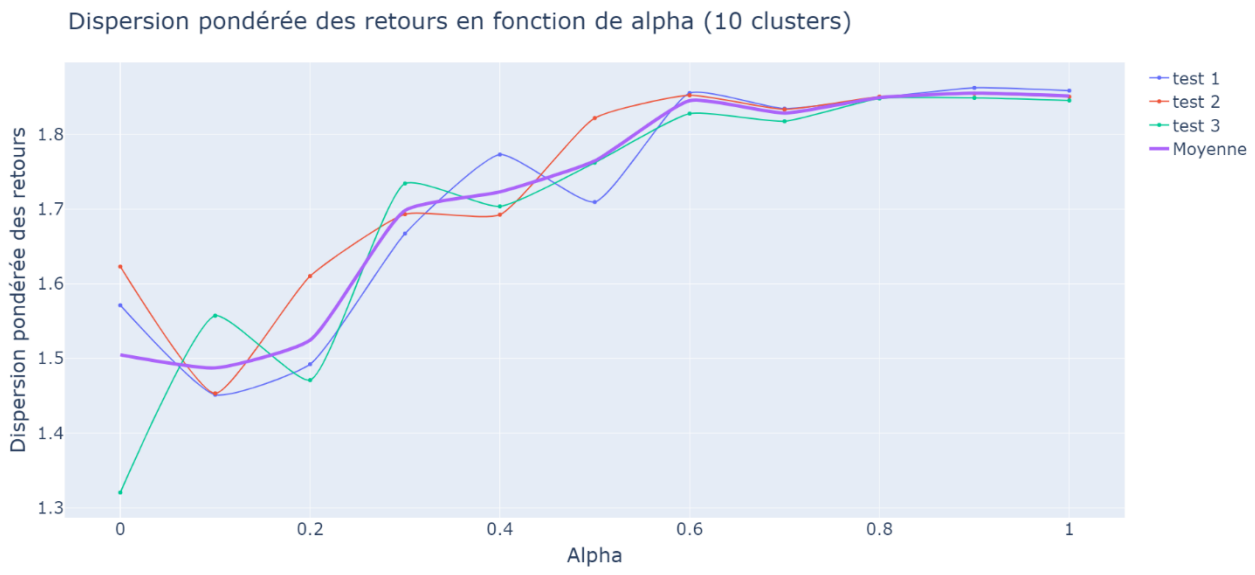


Figure 4.7 Dispersion pondérée des retours en fonction de alpha

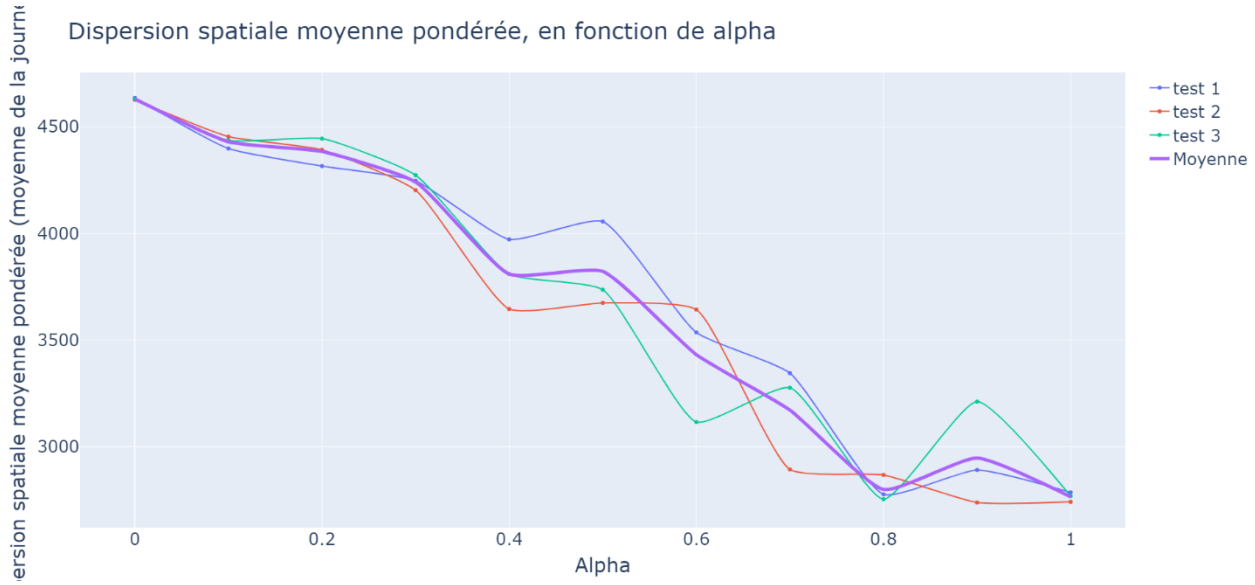


Figure 4.8 Dispersion spatiale moyenne pondérée en fonction de alpha

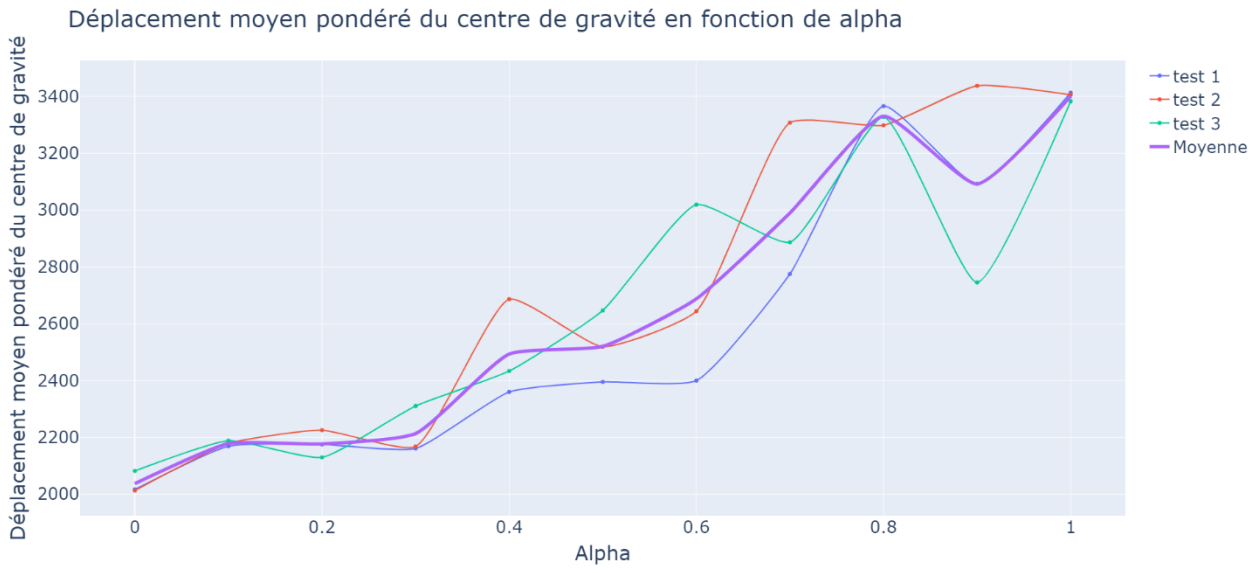


Figure 4.9 Déplacement moyen pondéré du centre de gravité en fonction de alpha

Nous observons plusieurs résultats intéressants : tout d'abord, lorsque alpha augmente, la dispersion temporelle des départs (Figure 4.6) et des retours (Figure 4.7) augmente, et la dispersion spatiale (Figure 4.8) diminue, ce qui démontre que alpha a véritablement une influence sur la nature spatiale ou temporelle des résultats. Nous constatons également une augmentation du déplacement moyen du centre de gravité des *clusters* (Figure 4.9) : plus le modèle est fort spatialement, plus le lieu de domicile et le lieu de travail de chaque *clusters* seront différenciés.

De manière plus détaillée, la dispersion temporelle semble converger vers une valeur maximale lorsque alpha se rapproche de 1 (Figure 4.6). Cette valeur maximale, à peu près 2.17, c'est-à-dire 2h10, correspond à l'écart moyen entre la première transaction de deux usagers quelconques. Ceci prouve que pour alpha=1, les usagers ne sont plus différenciés temporellement. Nous observons également que la dispersion temporelle stagne entre alpha=0 et alpha=0,2, comme si en-deçà de la valeur seuil alpha=0,2, les performances temporelles ne pouvaient pas être améliorées. La dispersion spatiale présente un phénomène symétrique : au-delà de la valeur seuil 0,8, elle stagne. Ces deux valeurs-seuils semblent donc jouer un rôle important dans le modèle : ce sont les valeurs entre lesquelles l'équilibre spatio-temporel peut être modifié.

Si l'on souhaite trouver une valeur de alpha pour laquelle l'influence du temps et parfaitement égale à celle de l'espace, on ne peut se contenter de alpha=0,5. En effet, si la dispersion spatiale

semble varier de manière assez continue, la dispersion temporelle connaît une hausse importante entre alpha=0,2 et alpha=0,6 puis converge vers son maximum. Alpha=0,4 correspond donc à un meilleur compromis entre espace et temps.

Nous nous intéressons maintenant à la dispersion spatiale moyenne à une heure donnée (et non à sa moyenne sur une journée). (Figure 4.10)

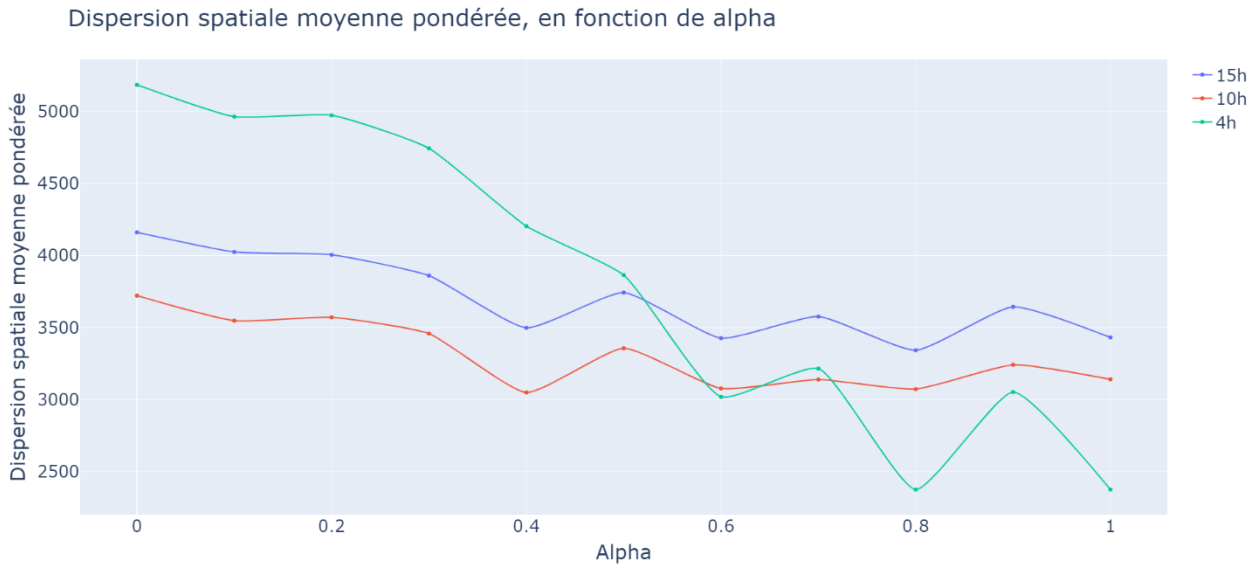


Figure 4.10 Dispersion spatiale moyenne pondérée à 4h, 10h, et 15h

On s'aperçoit que celle-ci diffère en fonction de l'heure ainsi qu'en fonction de alpha. Pour la dispersion spatiale moyenne à 4h, qui symbolise la dispersion des lieux d'habitation des membres du *cluster*, nous observons une variation avec alpha plus importante que pour la dispersion spatiale à 10h ou à 15h. Pour les valeurs de alpha proches de 0 (modèle fort temporellement), la dispersion spatiale est supérieure à 4h que le reste de la journée, alors que pour des valeurs de alpha proches de 1, elle est minimale à 4h. Ceci s'explique de la façon suivante :

- Pour alpha proche de 0, les profils ne sont presque pas différenciés spatialement. Or, la population de la ville de Québec, comme dans la plupart des villes, a tendance à passer la journée au centre de la ville et à habiter plus en périphérie. Ainsi, à 4h du matin, les usagers, encore chez eux, seront dispersés tout autour de la périphérie de Québec. Les *clusters* n'étant pas différenciés spatialement, la dispersion spatiale d'un *cluster* sera alors très

importante; tandis qu'à 10h ou 15h, les usagers s'étant déplacés vers le centre-ville, leur dispersion spatiale est plus petite, quel que soit le *cluster*.

- Pour alpha proche de 1, les *clusters* sont spatialement très différents. La plupart des habitants de Québec passant la journée vers le centre de la ville, les *clusters* ne sont pas très différenciés sur la zone de travail, mais ils sont très différenciés quant à la zone d'habitation. Ainsi à 4h du matin, tous les membres d'un *cluster* se trouve dans une zone restreinte propre à ce *cluster*, ce qui explique leur faible dispersion spatiale; tandis qu'à 10h ou 16h, une partie importante se trouve vers le centre-ville mais une partie non négligeable est encore au lieu de domicile, ce qui fait hausser la dispersion spatiale du *cluster*.

4.4 Influence du nombre de *clusters* sur l'équilibre spatio-temporel

Étudions désormais comment ces résultats varient si nous modifions le nombre de *clusters*. Pour chaque nombre de *cluster* (5, 10, et 15 *clusters*), et pour chaque valeur de alpha, on réalise trois tests avec trois échantillons différents et on conserve la moyenne de ces trois tests (Figure 4.11 et Figure 4.12).

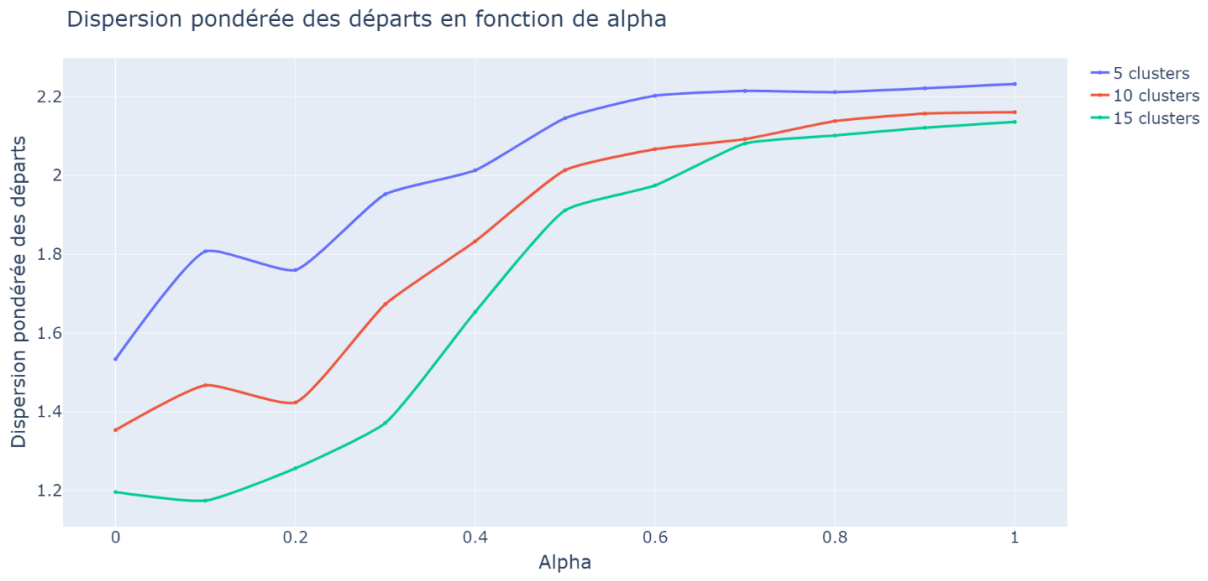


Figure 4.11 Dispersion pondérée des départs en fonction de alpha, pour trois nombres de *clusters*

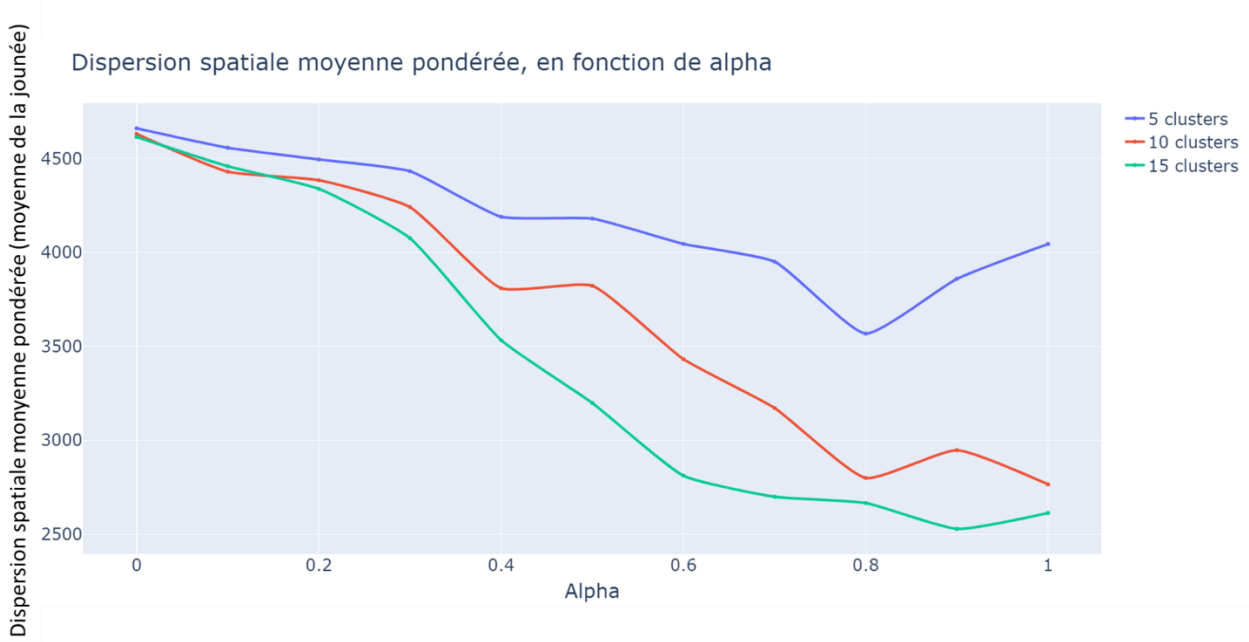


Figure 4.12 Dispersion spatiale pondérée en fonction de alpha, pour trois nombres de *clusters*

On s'aperçoit tout d'abord que les courbes ont à peu près la même forme quelle que soit le nombre de *clusters*. D'autre part, plus le nombre de *clusters* est petit, plus la dispersion moyenne pondérée, spatiale ou temporelle, est importante. Ce résultat est tout à fait logique : diminuer le nombre de *clusters* a pour effet de créer des *clusters* plus grands et donc d'augmenter la dispersion moyenne d'un *cluster*. Cependant, nous observons que cette différence de dispersion est plus ou moins accentuée en fonction de la valeur de alpha. Pour la dispersion temporelle (dispersion des départs), nous observons une différence importante entre chaque courbe pour les faibles valeurs de alpha, et une différence minime pour les valeurs de alpha proches de 1. Pour la dispersion spatiale, on a le phénomène inverse : l'écart est maximal pour les valeurs de alpha proches de 1, mais devient presque nul lorsque alpha se rapproche de zéro. Ce résultat peut être expliqué de façon suivante :

- Pour des valeurs de alpha proches de 0, le modèle est très fort temporellement; chaque *cluster* présente des caractéristiques temporelles bien différentes. Cependant, lorsque alpha se rapproche de 1, les *clusters* ne sont presque plus différenciés sur le plan temporel. On pourrait dire que « l'entropie » temporelle des *clusters* est alors maximale, puisque les usagers ne sont plus ordonnés selon des critères temporaux. Ainsi, la dispersion temporelle dans chaque *cluster* est à peu près la même : il s'agit de la valeur maximale (environ 2h10)

identifiée dans la partie précédente, et qui correspond à l'écart moyen entre la première transaction de deux usagers quelconques. Ainsi, puisque chaque *cluster* a la même dispersion temporelle, augmenter le nombre de *clusters* n'a presque plus d'effet sur la dispersion temporelle moyenne.

- Nous retrouvons le phénomène inverse pour la dispersion spatiale. Lorsque alpha se rapproche de 0, les *clusters* ne sont presque plus différenciés spatialement et ils atteignent leur « entropie spatiale » maximale : leur dispersion spatiale converge vers une valeur maximale (environ 4600 m), qui correspond à la distance moyenne d'un usager au centre de gravité de tous les usagers à une heure quelconque.

Par ailleurs, nous retrouvons les valeurs seuils observés dans la partie précédente : alpha=0,2 et alpha=0,8. En dehors de l'intervalle formé par ces deux valeurs, il n'y a pas de nette amélioration de la dispersion spatiale ou temporelle. Nous retrouvons également que alpha=0,4 correspond à peu près à un équilibre entre l'espace et le temps.

4.5 Choix de la taille de fenêtre

Dans un dernier temps, nous étudions comment la taille de fenêtre temporelle utilisée dans le DTW influe sur l'équilibre spatio-temporel. Pour trois tailles de fenêtre différentes (0, 1 et 2), nous étudions la variation des indicateurs spatiaux et temporels lorsque alpha augmente (en fixant le nombre de *clusters* à 10). Trois tests sont effectués pour chaque taille de fenêtre, et chaque valeur de alpha : nous représentons la moyenne de ces trois tests, pour chacune des trois tailles de fenêtre (*window*) (Figures 4.13, 4.14 et 4.15).

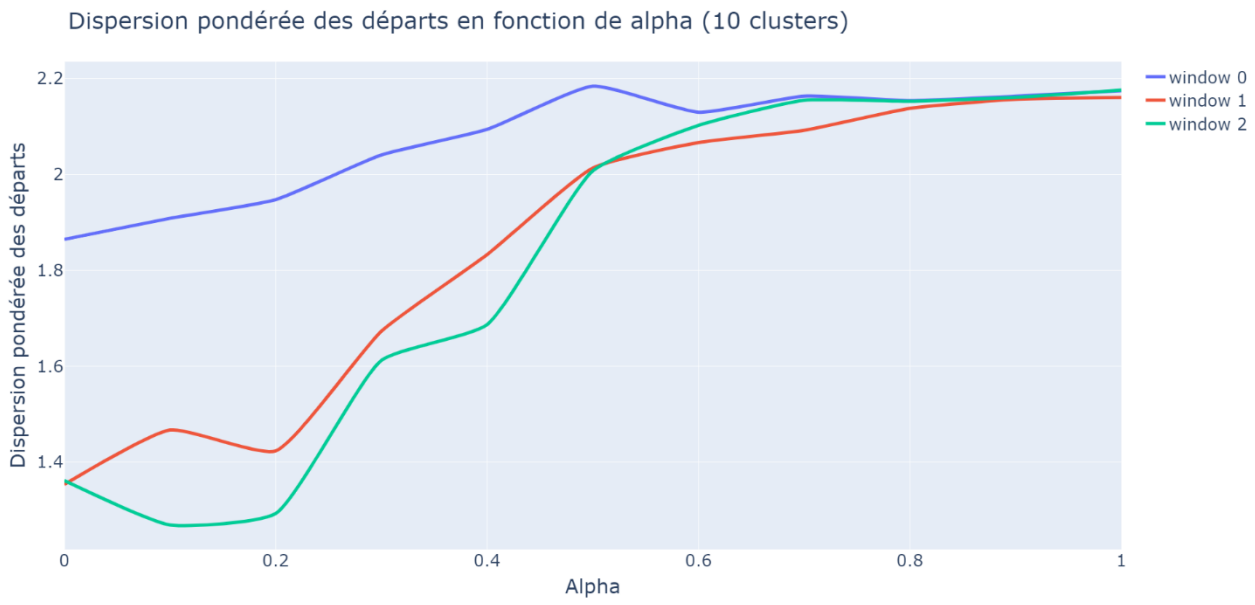


Figure 4.13 Dispersion pondérée des départs en fonction de alpha, pour trois tailles de fenêtre

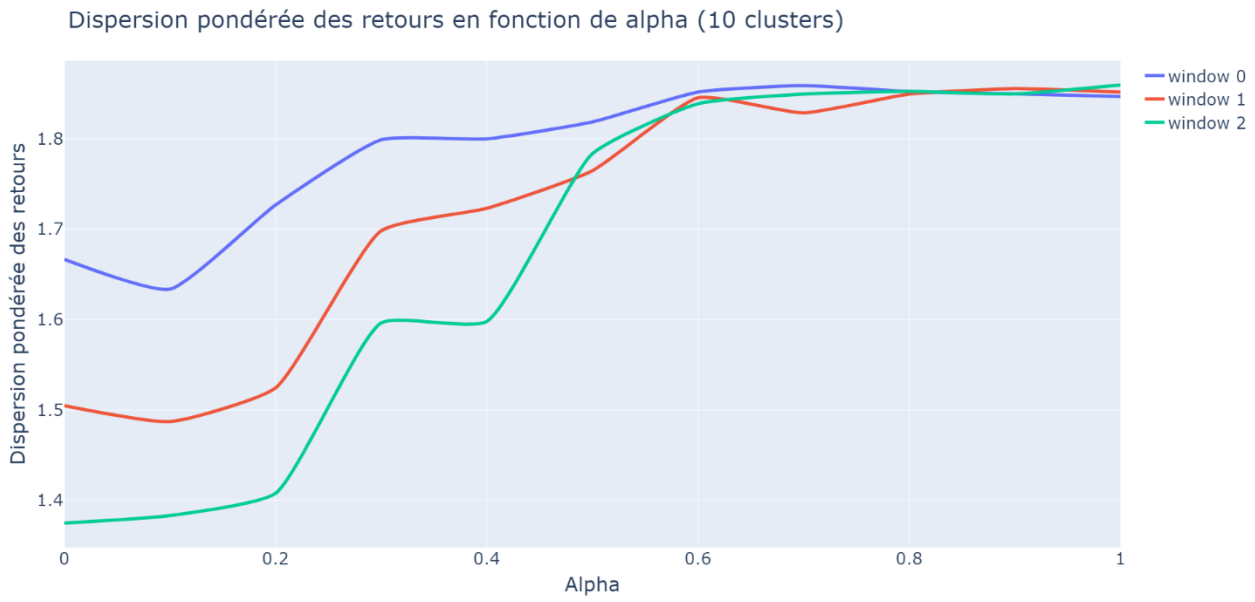


Figure 4.14 Dispersion pondérée des retours en fonction de alpha, pour trois tailles de fenêtre

Figure 4.15 Dispersion spatiale pondérée en fonction de alpha, pour trois tailles de fenêtre

Nous observons que la dispersion temporelle (des départs ou des retours) diminue lorsque la fenêtre augmente, en particulier pour les valeurs de alpha proches de zéro. Augmenter la fenêtre améliore donc les performances de l'algorithme DTW, lorsqu'il s'agit de regrouper les comportements selon leur similarité temporelle. Pour les valeurs de alpha proches de 1, les *clusters* atteignent leur « entropie temporelle » maximale, et la valeur de la fenêtre n'a plus d'importance.

Cependant, pour la dispersion spatiale, nous observons un résultat très intéressant : quelle que soit la valeur de alpha, la dispersion spatiale atteint son minimum pour une fenêtre de 1. D'autre part, nous observons que la différence entre la dispersion temporelle des retours pour une fenêtre de 1 et une fenêtre de 2 est très faible (ce résultat n'est cependant pas valable pour la dispersion des retours). Il semble donc qu'une fenêtre de 1 soit un excellent compromis entre espace et temps, d'autant plus que l'augmentation de la taille de fenêtre a une forte influence sur le temps de calcul.

Pour conclure cette étude du comportement de notre modèle spatio-temporel, voici les principaux résultats que nous avons pu observer. Le paramètre alpha a une influence majeure sur les caractéristiques spatiales et temporelles des *clusters* : son augmentation fait diminuer la dispersion spatiale et augmenter la dispersion temporelle. Ainsi, en faisant varier alpha de 0 à 1, on passe progressivement d'un modèle temporel à un modèle spatial. Nous avons également identifié deux

valeurs seuils de alpha (alpha=0,2 et alpha=0,8) entre lesquelles il est possible de faire varier les caractéristiques spatiales et temporelles des *clusters*. En-dehors de cet intervalle, aucune diminution de la dispersion spatiale ou de la dispersion temporelle n'est observée. Par ailleurs, nous avons identifié, à titre indicatif, une valeur de alpha (alpha=0,4) qui correspondrait à peu près à l'équilibre spatio-temporel, c'est-à-dire à l'équilibre entre l'influence du temps et de l'espace sur la segmentation. Il est à noter cependant que cette valeur d'équilibre est propre à la ville de Québec et à la période étudiée. Sur une ville avec une utilisation différente du réseau, ou simplement de taille différente, cette valeur serait probablement différente. Enfin, une étude de l'influence de la taille de la fenêtre nous a permis de conclure qu'une fenêtre de taille 1 est un excellent compromis entre l'optimisation des performances spatiales et des performances temporelles des *clusters*. Nous utilisons donc cette taille de fenêtre dans l'étude suivante.

CHAPITRE 5 ANALYSE DU COMPORTEMENT SPATIO-TEMPOREL DES USAGERS DE LA RTC

Dans cette partie, nous étudions, avec trois valeurs différentes de alpha, le comportement des usagers sur un mois de données (soit 361365 profils journaliers). Après une étude statistique générale des données, nous présentons ces trois études, qui utilisent respectivement une valeur de paramètre d'équilibre spatio-temporel alpha de 0; 0,4 et 1. Chaque *cluster* est étudié de manière exhaustive, en mettant l'accent ici sur le type de titres des usagers.

5.1 Étude statistique des données

Nous étudions la période du lundi 02 décembre au dimanche 29 décembre 2019. La figure 5.1 montre le nombre total de transactions par jour.

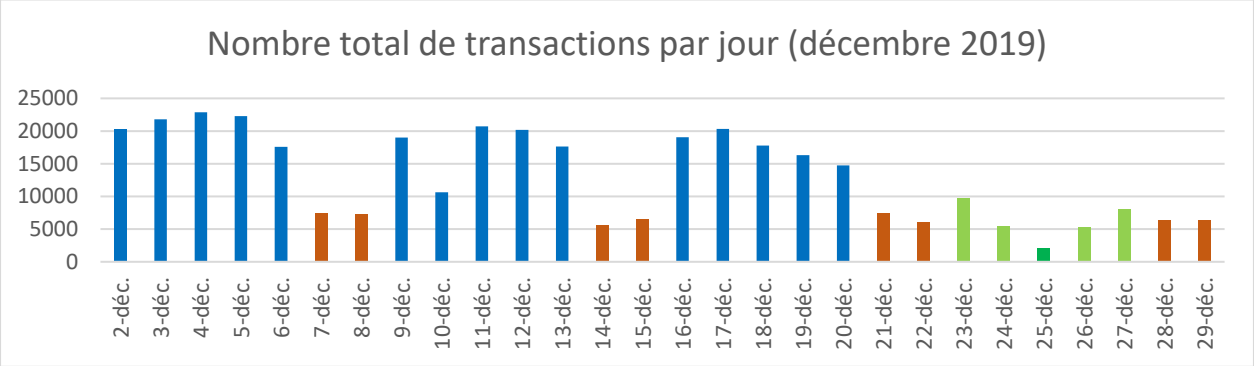


Figure 5.1 Nombre total de transactions par jour

Nous remarquons durant les trois premières semaines une fréquentation du réseau beaucoup moins accrue le week-end que le reste de la semaine (avec environ 65% de transactions en moins durant un jour de week-end que durant un jour de début de semaine, lors de la première semaine). Nous remarquons aussi une certaine baisse de fréquentation du réseau en semaine à l'approche de Noël, très visible sur le 3ème début de semaine (16, 17, 18, 19 et 20 décembre). La 4ème semaine de décembre, marquant le début des vacances de Noël, connait une fréquentation à peu près égale à celle d'un jour de week-end, avec un minimum de fréquentation le 25 décembre. Nous observons enfin une baisse importante de la fréquentation le mardi 10 décembre, pour des raisons inconnues.

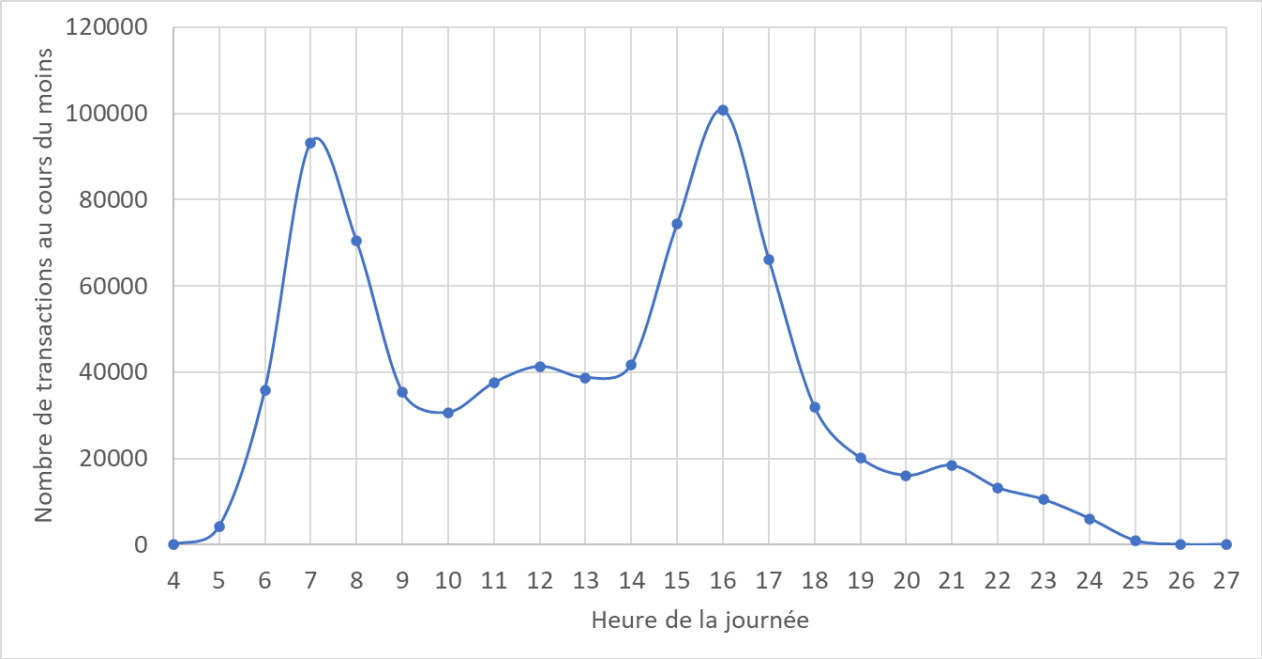


Figure 5.2 Nombre total de transactions au cours du mois à chaque heure

La Figure 5.2 représente le nombre total de toutes les transactions effectuées au cours du mois, à chaque heure de la journée. Nous constatons deux pics de transaction très importants et de tailles à peu près égale vers 7h du matin et 16 du matin. Ces deux pics correspondent aux allers et retours des travailleurs à horaires de travail réguliers. Nous constatons également un plateau entre 11h et 13h, qui pourrait correspondre aux déplacements vers des restaurants ou autres à l'heure du repas de midi. Enfin, un petit plateau entre 19h et 23h pourrait être dû aux sorties nocturnes le week-end ou en semaine. Il est à noter cependant que, Québec étant peu desservi par le réseau de bus la nuit (la plupart des lignes ferment vers 1h du matin), les usagers de la ville de Québec préconisent la voiture ou d'autres modes de transport pour leurs sortie nocturnes.

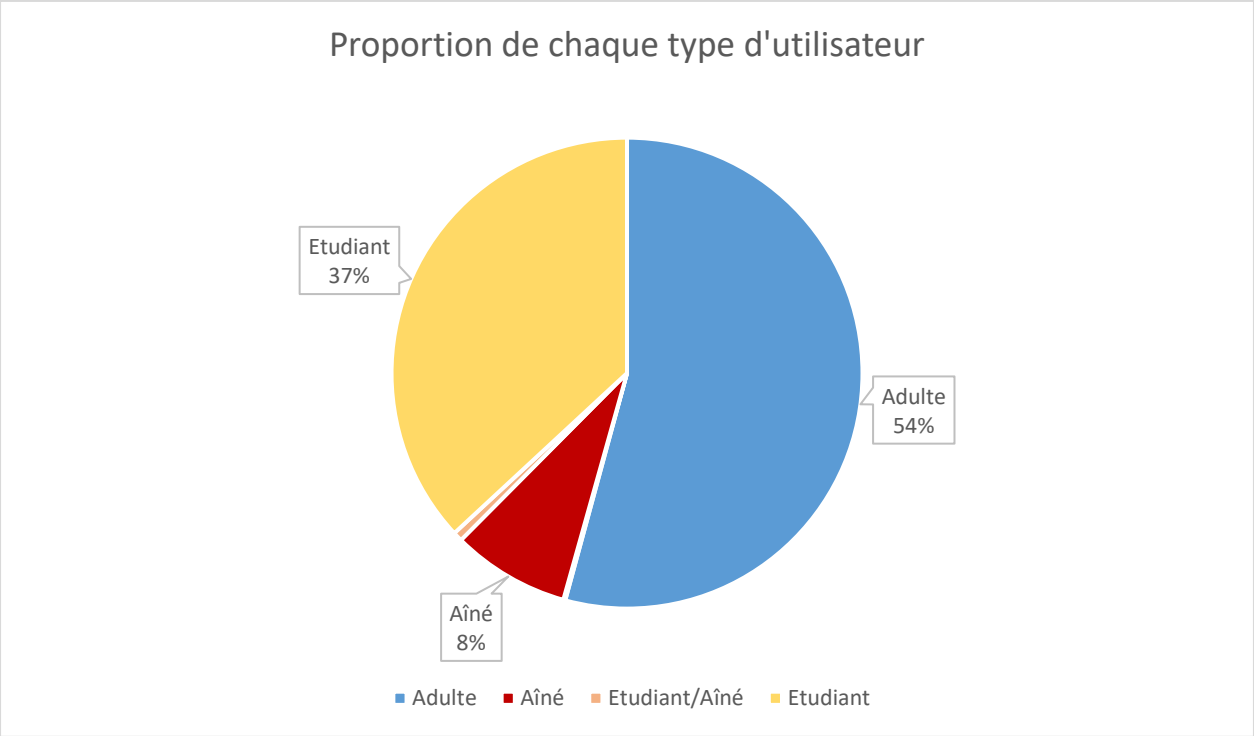


Figure 5.3 Proportion de chaque type d'utilisateur dans la ville de Québec

On dispose aussi pour chaque transaction du type de titre utilisé par l'usager. Nous avons regroupé ceux-ci en trois catégories correspondant au type d'utilisateur ayant accès à ces titres : adulte, étudiant et aîné. Nous avons représenté sur la Figure 5.3 la proportion de chaque type d'utilisateur pour tous les profils journaliers de la période étudiée. Nous constatons une proportion importante d'étudiants (37%), et une proportion non négligeable d'aînés (8%).

5.2 Modèle purement temporel

Dans un premier temps, nous étudions le comportement des usagers de la RTC avec un modèle purement temporel (alpha=0). La distance entre les usagers n'est donc pas prise en compte ici, et les usagers sont classifiés à l'aide de leurs heures de transactions uniquement. Les *clusters* obtenus sont identifiés ici par un 0 (pour alpha=0) suivi de leur numéro de *cluster* (par exemple le *cluster* 0.3 pour le troisième *cluster*).

5.2.1 Choix du nombre de *clusters*

Pour choisir un nombre de *clusters* optimal, nous utilisons l'indicateur de Silhouette, que nous calculons pour chaque nombre de *clusters* (Figure 5.4).

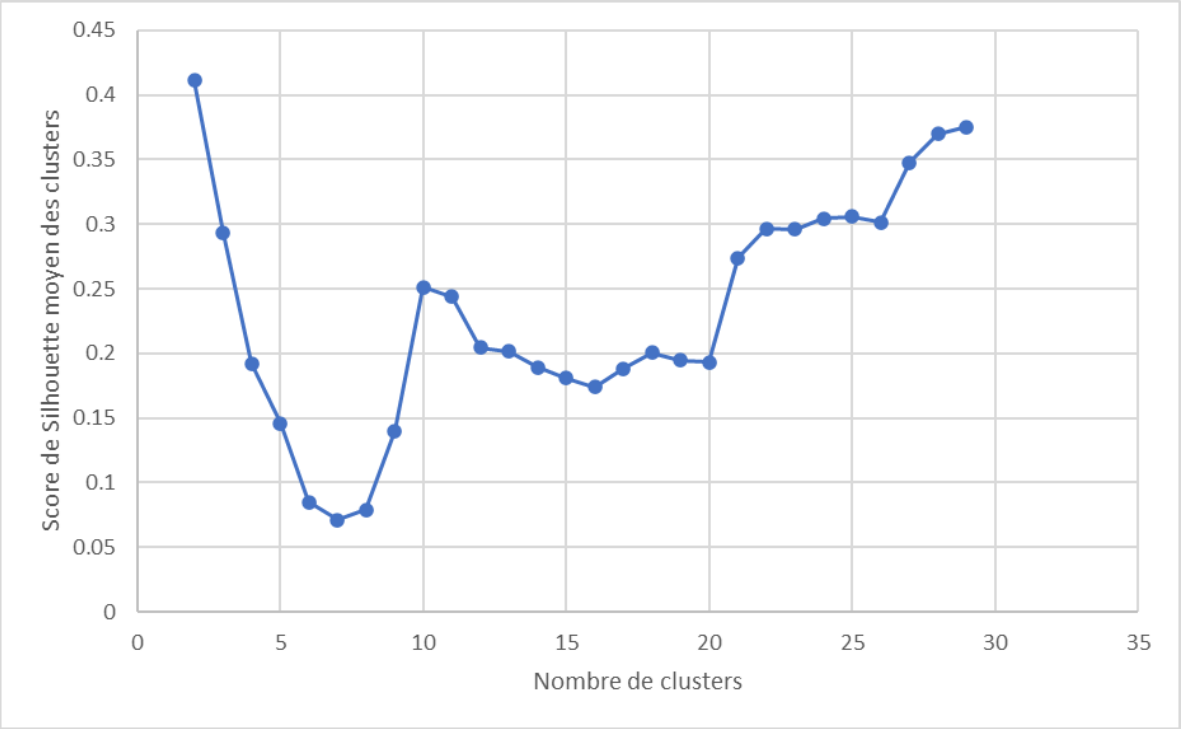


Figure 5.4 Score de Silhouette moyen en fonction du nombre de *clusters* (alpha=0)

Nous observons un maximum local pour 10 *clusters*. En-dessous de 4 *clusters* l'analyse ne serait pas assez détaillée, et au-delà de 20, elle serait trop fastidieuse. Nous choisissons donc de travailler avec 10 *clusters*. On conservera cette valeur de 10 *clusters* pour les études suivantes, afin de pouvoir comparer plus facilement les différents modèles spatio-temporels.

Voici donc les résultats de notre simulation, avec alpha=0, une fenêtre de 1, une taille d'échantillon de 3000, et 10 *clusters*, sur la période du 2 au 29 décembre 2019.

5.2.2 Étude du comportement temporel des usagers

Les résultats de la simulation renvoient des *clusters* assez équilibrés (Figure 5.5), à l'exception du *cluster* 0,5, qui est particulièrement petit, et du *cluster* 0.10 qui contient environ 30% des profils.

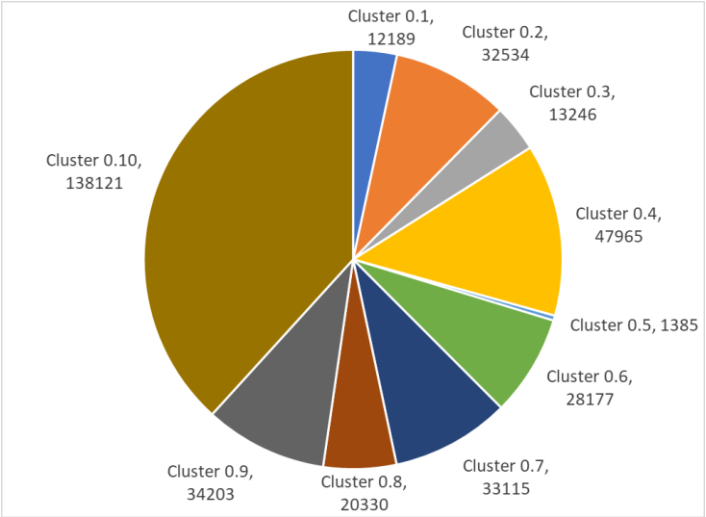


Figure 5.5 Taille de chaque *cluster* (alpha=0)

Nous étudions dans un premier temps le nombre de transactions à chaque heure dans chaque *cluster* (Figure 5.6). Ces profils temporels sont très divers, et peuvent être caractérisés par le nombre et la position de leurs pics de transactions.

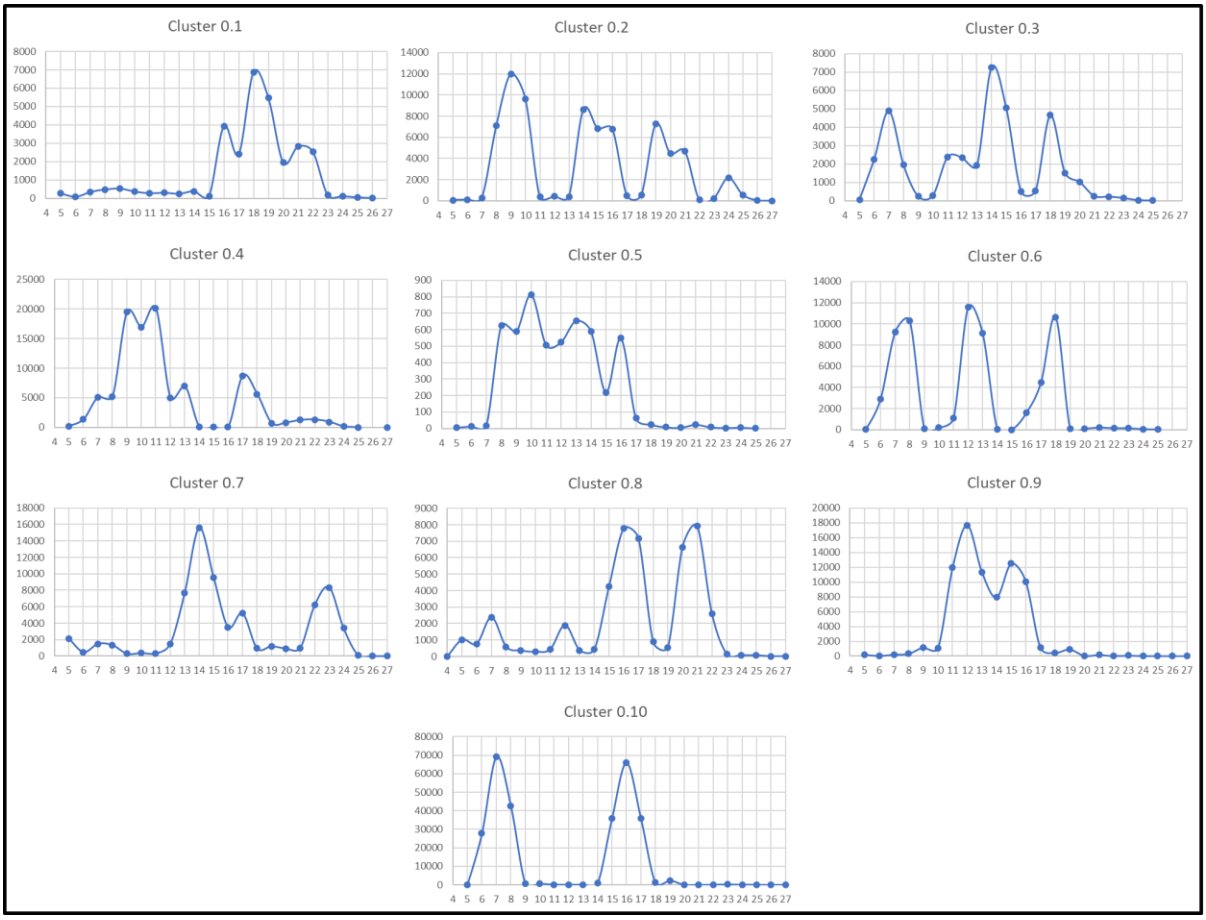


Figure 5.6 Heures de transaction des usagers dans chaque *cluster* (alpha=0)

Le *cluster* 0.10, contenant environ 30% des profils, correspond au comportement temporel le plus répandu à Québec : un départ vers 7h du matin et un retour vers 16h. Il s'agit de la journée de travail typique d'un utilisateur de Québec. Nous constatons, par ailleurs, en regardant la Figure 5.7, que ce *cluster* contient une très forte proportion d'adultes (73%) et relativement peu d'élèves ou d'aînés par rapport aux autres *clusters*. Enfin, selon la Figure 5.8, qui représente la répartition des profils dans chaque *cluster* au cours de la première semaine, la majorité des profils du *cluster* 0.10 sont observés en semaine (et non le week-end). Le *cluster* 0.10 est donc celui des travailleurs de semaine aux horaires réguliers.

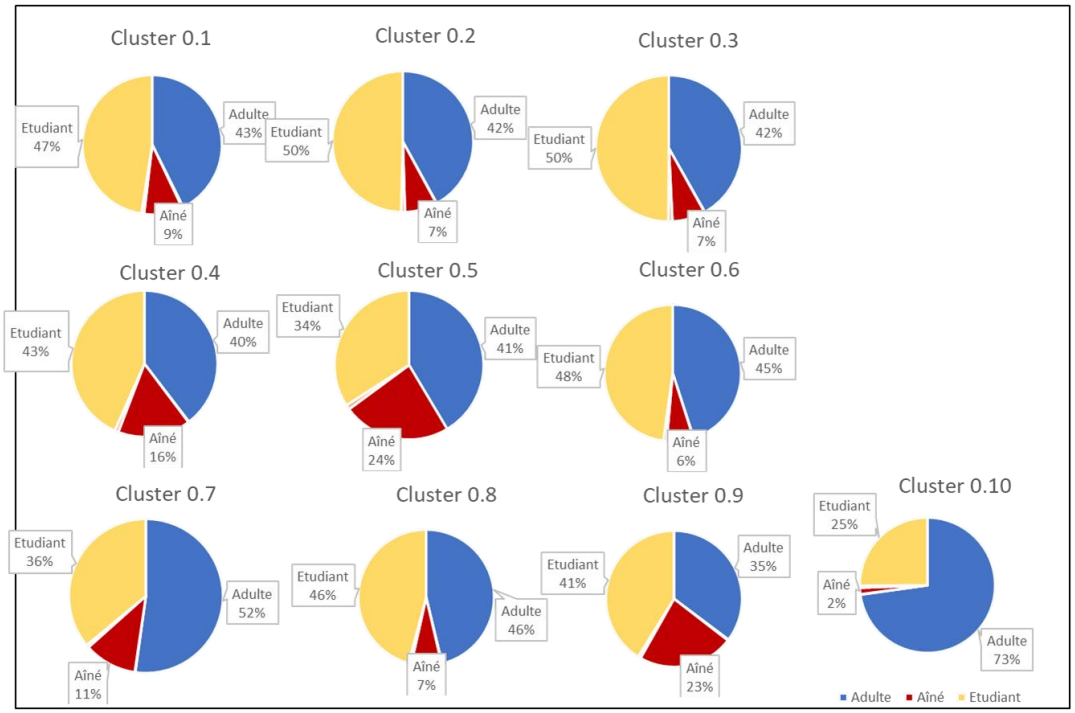


Figure 5.7 Proportion de chaque type d'usager dans chaque *cluster* (alpha=0)

Le *cluster* 0.7 contient aussi une proportion assez importante d'adultes (52%), et son profil temporel moyen contient aussi deux pics de transactions, mais ceux-ci sont beaucoup plus tardifs que pour le *cluster* 0.10. Nous observons un pic de départ aux alentours de 14h, prolongé jusque 17h, et un pic de retour entre 22h et 24h. A l'inverse du *cluster* 0.10, ce *cluster* apparait surtout les jours de week-end (Figure 5.8). On peut donc supposer que le *cluster* 0.7 représente les sorties du week-end (restaurants, fêtes ou autres…), et qu'une partie importante des adultes faisant partie du *cluster* 0.10 en semaine se retrouvent dans le *cluster* 0.7 le week-end.

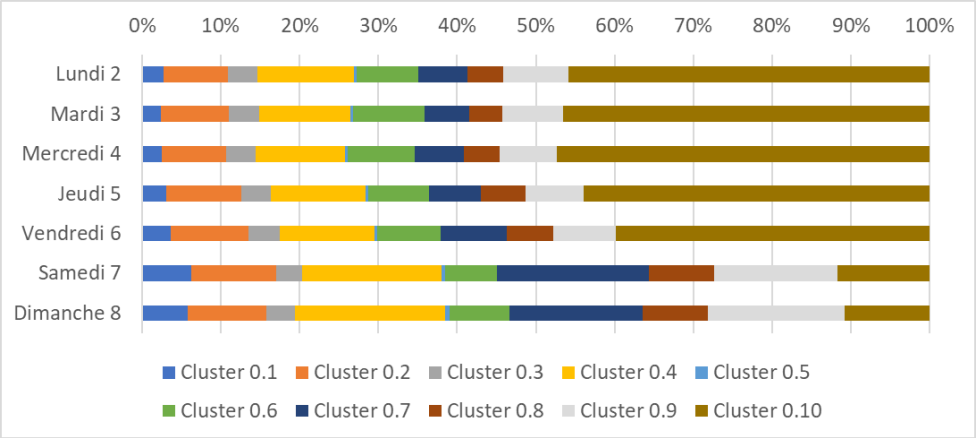


Figure 5.8 Répartition des profils dans chaque *cluster* en fonction du jour (alpha=0)

Nous retrouvons aussi pour le *cluster* 0.9 une augmentation significative le week-end (Figure 5.8). Son profil temporel correspond à 2 pics de transactions assez rapprochés, l'un vers midi, l'autre vers 16h. Compte tenu de l'heure des transactions et de la durée sur place, ce comportement correspond vraisemblablement à une sortie au restaurant, ou dans un autre endroit dans le but de déjeuner. Les sorties au restaurant étant plus fréquentes le week-end, cela expliquerait la hausse soudaine de la taille du *cluster* les jours de week-end. Par ailleurs, nous remarquons que la proportion d'aînés dans le *cluster* 0.9 est particulièrement élevée (23%) par rapport à la normale (8%). Les sorties du midi au restaurant seraient donc assez fréquentes pour les aînés.

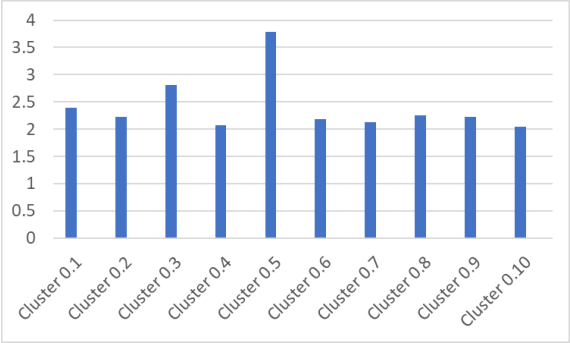


Figure 5.9 Nombre moyen de trajets par profil

Un autre *cluster* contenant une proportion élevée d'aînés (24%) est le petit *cluster* 0.5. La particularité de ce *cluster* est le nombre particulièrement élevé de trajets par profils (presque 4 contre un peu plus de 2 pour les autres *clusters*) (Figure 5.9 ci-contre). Nous constatons 4 pics de transactions vers 8h, 10h, 14h et 16h (Figure 5.6). Les raisons de ce nombre si important de trajets

sont obscures, mais ce groupe étant très petit (avec moins de 50 profils par jour), il n'est pas vraiment représentatif du comportement de la population québécoise.

Le *cluster* 0.4, de taille beaucoup plus importante, et ayant aussi une proportion importante d'ainés (16%), il est plus à même d'expliquer le comportement des aînés à Québec. D'après la Figure 5.9, ses profils contiennent en général 2 trajets. D'après la Figure 5.6, la plupart de leurs transactions ont lieu le matin entre 7h et 14h, ou bien l'après-midi vers 17h ou 18h. Tout comme pour les *clusters* 0.5 et 0.9, les pics de transactions ne sont pas très marqués, ce qui prouve que ces groupes, constitués d'une proportion importante d'ainés, ont un comportement moins régulier que le groupe 0.10, et ne sont pas soumis à des horaires fixes comme les usagers du groupe 0.10. De manière générale, on peut constater que les *clusters* avec une proportion importante d'aînés ont des horaires moins tardifs que le reste de la population, et ont très peu de transactions après 18h.

Pour ce qui est des étudiants, ils sont représentés par les *clusters* 0.1, 0.2, 0.3 et 0.6, où ils atteignent quasiment 50% des profils (contre 37% en temps normal). Les *clusters* 0.2, 0.3 et 0.6 sont très semblables, car ils présentent tous trois pics de transactions bien marqués (Figure 5.6), alors que le nombre moyen de leurs trajets est proche de 2 (Figure 5.9). Ceci peut être expliqué par le fait qu'en fonction de leur emploi du temps, les étudiants peuvent aller en cours soit le matin, soit l'après-midi, soit toute la journée, mais que le début et la fin des cours le matin et l'après-midi se fait toujours à la même heure (d'où la netteté des pics). La différence entre ces trois *clusters* se fait donc par la position des trois pics : en fonction de l'heure de ces pics, on pourrait identifier le type d'école et le niveau scolaire des élèves de ces *clusters*.

Le *cluster* 0.8 possède également une forte proportion d'étudiants (46%), mais présente uniquement 2 pics de transactions vers 16h et 22h (Figure 5.6). Il s'agit probablement des étudiants suivant des cours du soir. Enfin, le *cluster* 0.1, avec 47% d'étudiants, et des horaires de transactions moins fixes, entre 16h et 22h pourrait représenter les sorties d'étudiants le soir, en semaine ou le week-end.

Enfin, si nous observons les caractéristiques spatiales des *clusters*, nous constatons que celles-ci ne dépendent absolument pas du *cluster*. En effet, d'après la carte représentant le *cluster* le plus fréquent au départ de chaque arrêt (Figure 5.10), nous retrouvons autant de profils de chaque *cluster* quelle que soit la zone de Québec : les *clusters* ne sont donc pas différenciés spatialement.

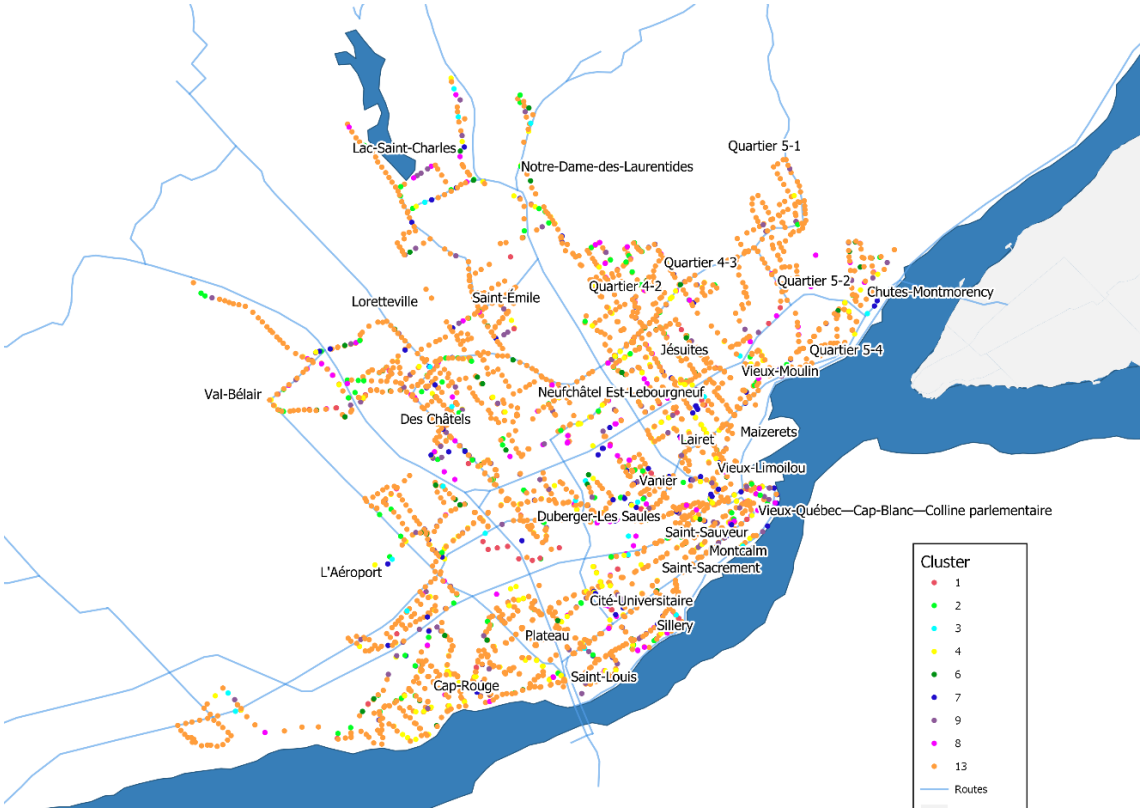


Figure 5.10 *Cluster* le plus fréquent au départ de chaque arrêt, avec alpha=0

Nous avons donc réussi à mettre en valeur, à l'aide de ce modèle, plusieurs comportements temporels typiques des usagers de la RTC, et nous avons constaté que certains de ces comportements sont spécifiques à certaines tranches d'âge d'usagers, ainsi qu'à certains jours de la semaine. Si ce modèle permet d'identifier des comportements temporels variés, il ne permet cependant pas de différencier la population par rapport au facteur spatial de son comportement.

5.3 Modèle spatio-temporel équilibré

Nous lançons une nouvelle simulation en prenant cette fois-ci alpha=0,4 comme paramètre spatio-temporel. D'après le chapitre 4, cette valeur devrait correspondre à peu près à l'équilibre entre l'influence du temps et de l'espace sur la classification. Nous utilisons à nouveau une fenêtre de 1, une taille d'échantillon de 3000, et un nombre de *clusters* de 10. Les *clusters* obtenus sont identifiés ici par un 4 (pour alpha=0,4) suivi de leur numéro de *cluster* (par exemple le *cluster* 4.3 pour le troisième *cluster*).

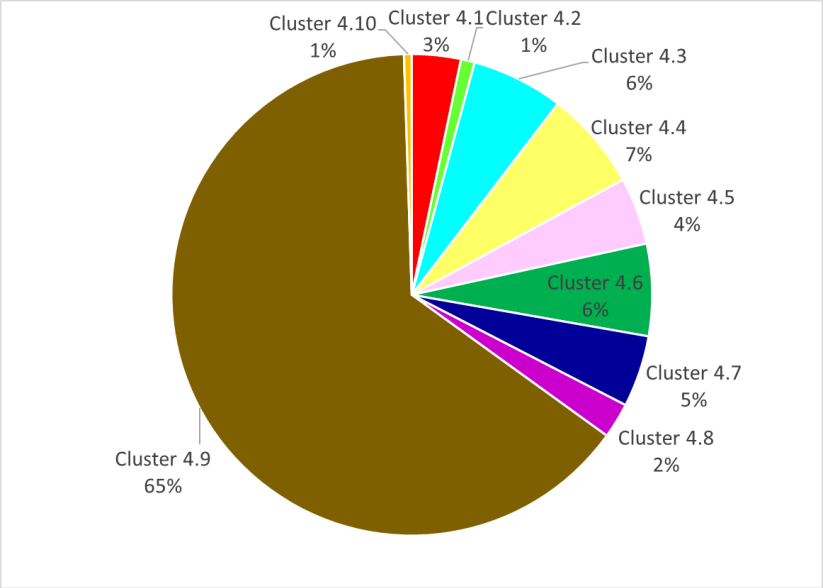


Figure 5.11 Tailles de chaque *cluster* (alpha=0,4)

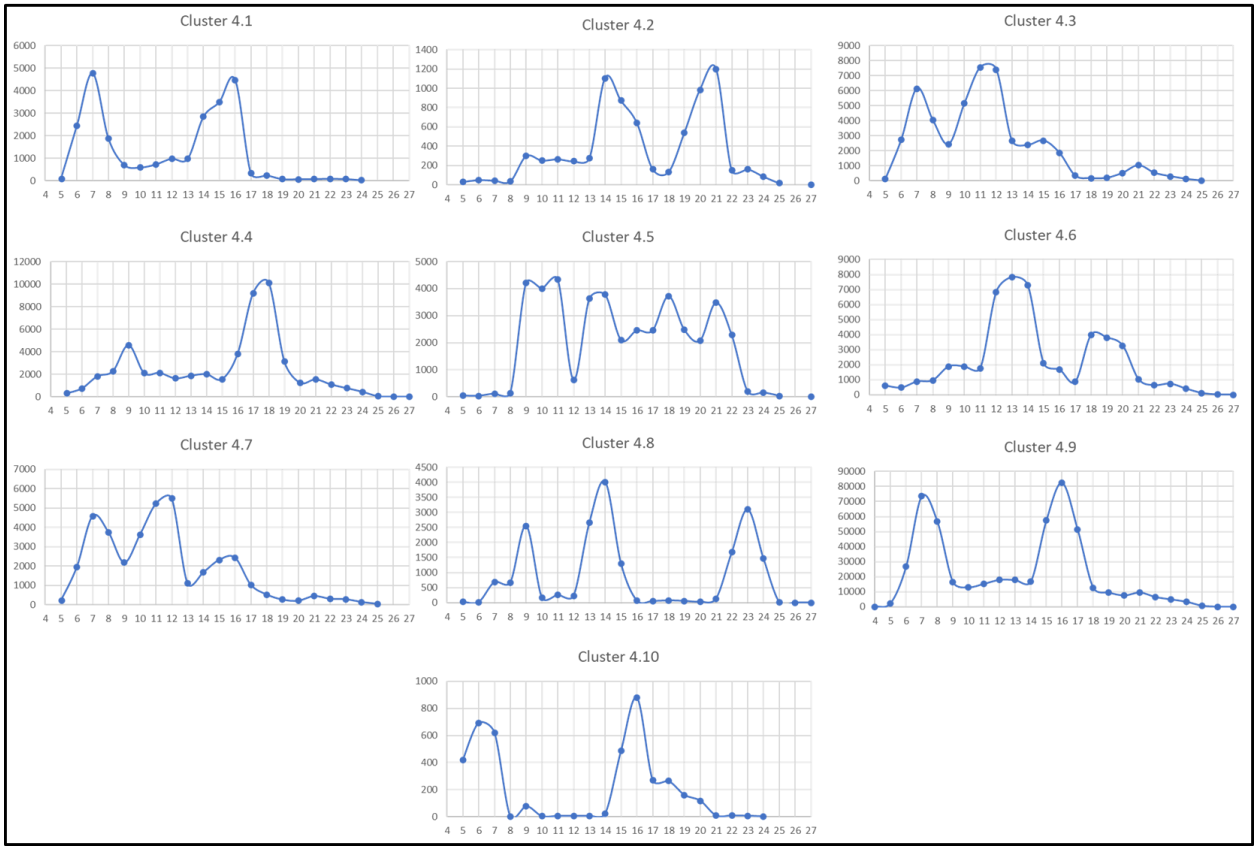


Figure 5.12 Heures de transaction des usagers dans chaque *cluster* (alpha=0,4)

Nous constatons tout d'abord que les *clusters* sont beaucoup moins équilibrés que dans la partie précédente (Figure 5.11). Le *cluster* 4.9 contient environ 65% des profils tandis que les autres *clusters* n'en contiennent qu'entre 1 et 7%.

Étudions plus en détail ce *cluster* principal. Avec deux pics de transactions autour de 7h et 16h et un nombre assez important de transactions entre ces deux pics (Figure 5.12), son profil est proche du profil temporel moyen de toutes les transactions à Québec (Figure 5.2). D'après les Figures 5.13 et 5.14, les usagers de ce *clusters* habitent et travaillent dans la zone centrale de Québec. La forte densité de population du centre-ville explique la taille si importante de ce *cluster*. Pour résumer, il s'agit du *cluster* des usagers habitant le centre-ville, ou à proximité. Ce *cluster* contient une variété assez importante de comportements spatio-temporels. Il pourrait être intéressant de le diviser en plusieurs sous-*clusters*.

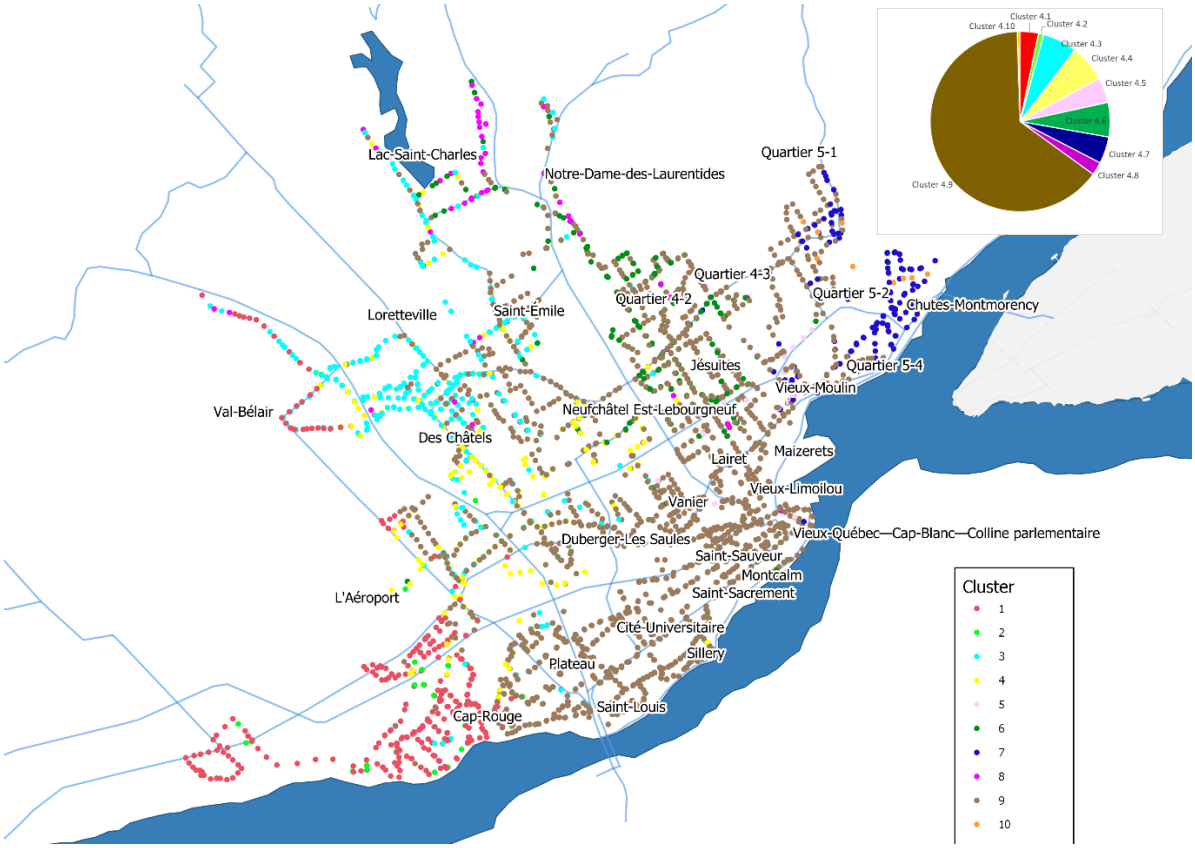


Figure 5.13 *Cluster* le plus fréquent au départ de chaque arrêt, avec alpha=0,4

Les *clusters* 4.1 et 4.2 regroupent les habitants de la zone de Cap-Rouge, à l'extrémité ouest de Québec (Figure 5.13), mais n'ont pas le même comportement temporel. Celui du *cluster* 4.1 est

assez habituel, avec un départ vers 7h et un retour vers 16h, tandis que celui du *cluster* 4.2 est un peu plus tardif, avec un pic autour de 14h et un deuxième autour de 21h (assez semblable au *cluster* 0.8 de la partie précédente). D'autre part, d'après la figure 5.14, les usagers du *cluster* 4.1 ont tendance à se déplacer vers le centre-ville, alors que ceux du *cluster* 4.2 se déplacent plutôt vers la cité universitaire (voir Annexe B). Ceci est en accord avec la Figure 5.15, qui donne un pourcentage d'étudiants beaucoup plus élevé dans le *cluster* 4.2 que dans le *cluster* 4.1. Nous constatons par ailleurs, que le comportement du *cluster* 4.1 apparait surtout les jours de semaines, alors que celui du *cluster* 4.2 apparait également le week-end (Figure 5.16). Pour récapituler, les *clusters* 4.1 et 4.2 regroupent les habitants de l'extrémité ouest, avec les travailleurs (ou étudiants) aux horaires réguliers dans le *cluster* 4.1 et les étudiants du soir dans le *cluster* 4.2.

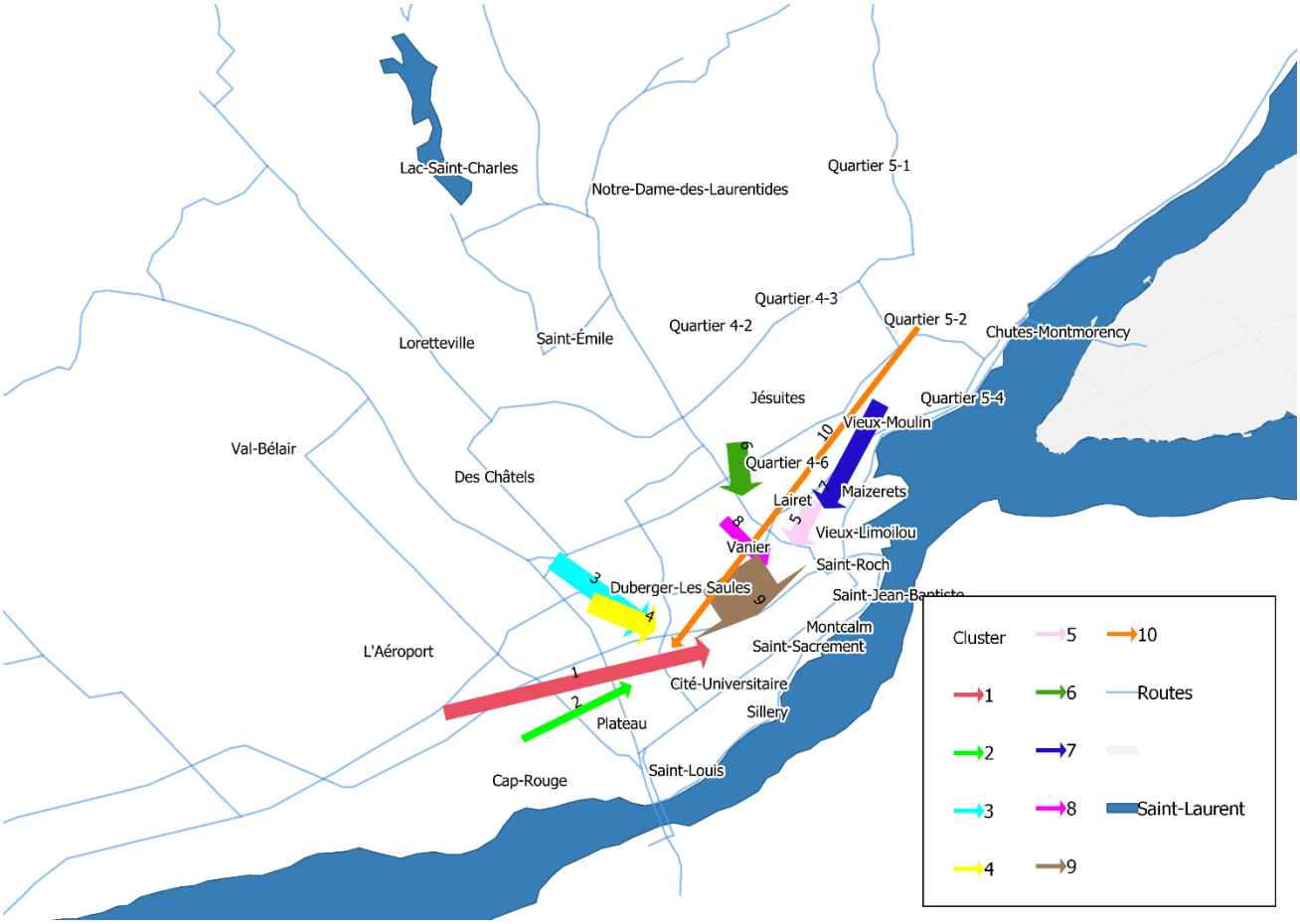


Figure 5.14 Carte des directions de chaque *cluster* (alpha=0,4)

Les *clusters* 4.3 et 4.4 présentent la même forme de dualité : leurs usagers habitent dans les deux cas à proximité du quartier Des Châtels (Figure 5.13) et se déplacent pour la plupart vers la cité-universitaire (Figure 5.14). Il s'agit en effet de *clusters* à fort pourcentage d'étudiants (Figure 5.15), avec plus de 51% d'étudiants (contre 37% en temps normal). De même que pour les *clusters* 4.1 et 4.2, la différence entre les *clusters* 4.3 et 4.4 est principalement temporelle. Tandis que la plupart des transactions du *cluster* 4.3 ont lieu le matin ou en début d'après-midi, avec 2 pics vers 7h et 11h30, celles du *cluster* 4.4 ont lieu toute la journée avec un pic de retour vers 18h.

Nous retrouvons également une certaine dualité entre les *clusters* 4.7 et 4.10 : tous deux correspondent aux usagers habitant l'extrémité est de Québec, près des chutes Montmorency (Figure 5.13). La différence entre ces deux *clusters* se fait au niveau du lieu de travail (Figure 5.14) : les usagers du *cluster* 4.7 travaillent pour la plupart au centre-ville, tandis que ceux du *cluster* 4.10 travaillent du côté de la Cité Universitaire (voir Annexe B). Il ne s'agit cependant pas d'étudiants, mais bien de travailleurs, comme en témoigne le pourcentage très élevé d'adultes (72%) (Figure 5.15).

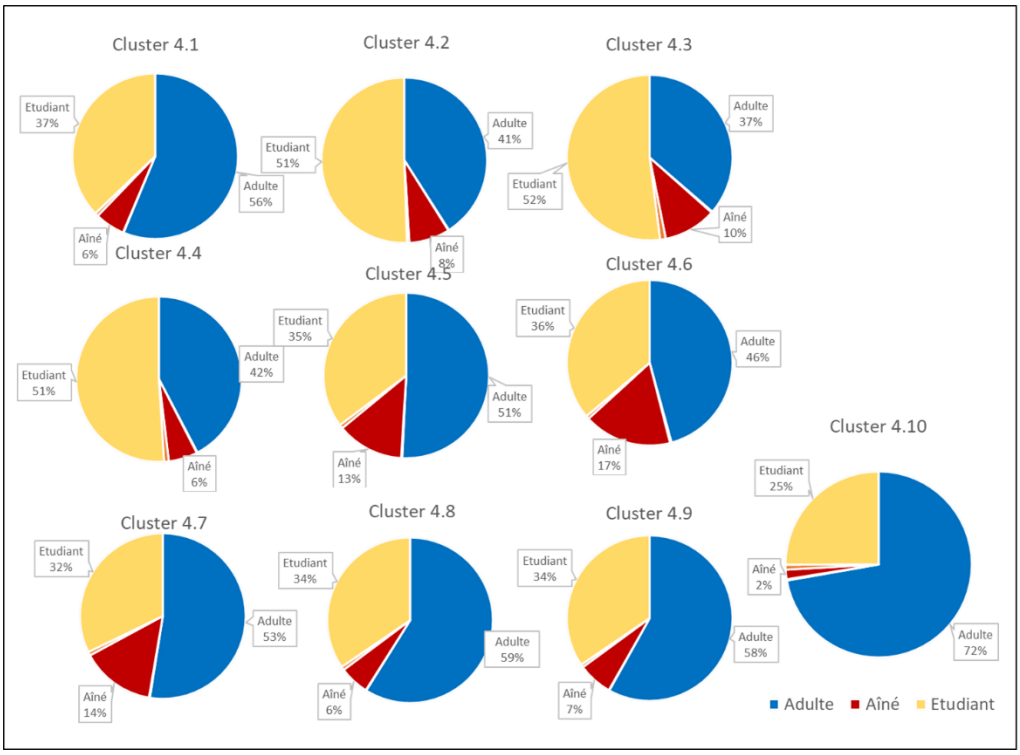


Figure 5.15 Proportion de chaque type d'usagers dans chaque *cluster* (alpha=0,4)

De la même façon, les usagers des *clusters* 4.6 et 4.8 habitent à peu près la même zone, comprise le long de la ligne entre le Vieux-Québec et le lac Saint-Charles (voir Annexe B), mais diffèrent sur le plan temporel. Le *cluster* 4.6, composé d'une forte proportion d'ainés, a des horaires plus floues et moins tardives que le *cluster* 4.8. (Figure 5.12)

Enfin, le *cluster* 4.5, caractérisé par un grand nombre de transactions, tout au long de la journée, (Figure 5.12) est un *cluster* d'usagers habitant et se déplaçant dans le centre-ville, pour des raisons probablement récréatives, à cause du nombre et du manque de régularité de leur comportement.

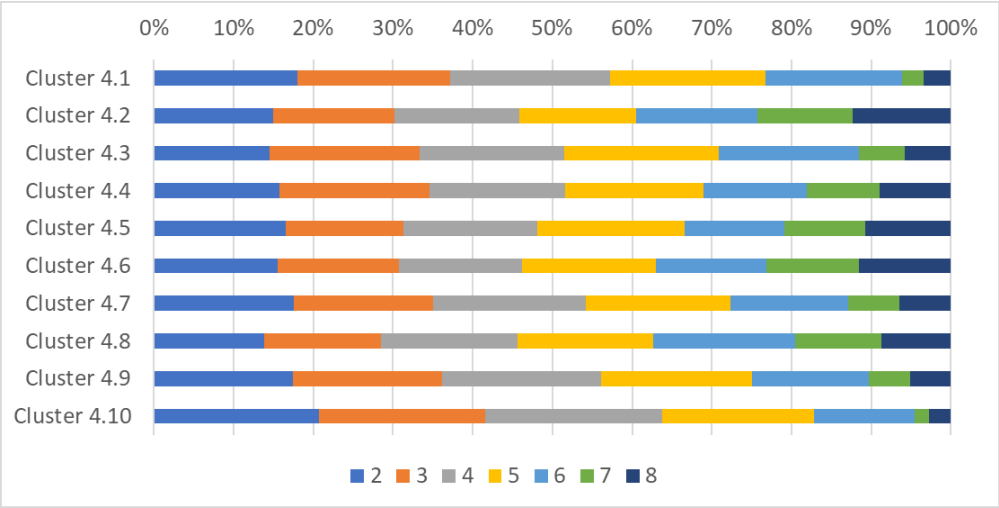


Figure 5.16 Proportion de chaque jour de la première semaine dans chaque *cluster* (alpha=0,4)

Ainsi, pour conclure, cette étude a permis de regrouper les profils en plusieurs groupes de comportement spatio-temporels. Certains de ces groupes peuvent être regroupés en paires, possédant à peu près le même comportement spatial (même zone de domicile et éventuellement même zone de travail) mais ayant des comportements temporels différents. Le *cluster* 4.9 regroupe tous les profils qui ne ressortent ni pas leur comportement temporel (ayant un comportement proche du comportement moyen), ni par leur comportement spatial (habitant et travaillant à proximité du centre-ville). Il serait intéressant de diviser ce *cluster* en sous-*clusters* plus petits pour mieux comprendre le comportement des usagers du centre-ville. Des trois études menées dans ce chapitre, celle-ci est celle où la différence de comportement entre les différentes tranches d'âge d'usagers apparait le plus (Figure 5.15) : ceci démontre l'intérêt de considérer simultanément le temps et l'espace dans la segmentation du comportement des usagers.

5.4 Modèle à prédominance spatiale

De même que pour les deux études précédentes, nous prenons de nouveau une taille d'échantillon de 3000, une fenêtre de 1 et 10 *clusters*, mais cette fois-ci avec un paramètre alpha=1. Les *clusters* obtenus sont identifiés ici par un 1 (pour alpha=1) suivi de leur numéro de *cluster* (par exemple le *cluster* 1.3 pour le troisième *cluster*). Ceci correspond à un modèle fortement spatial : l'heure des transactions n'a plus d'incidence sur la segmentation, seule compte la séquence d'arrêts visités.

Tout comme pour l'étude précédente, nous observons un *cluster* principal (le *cluster* 1.9), contenant une grande partie des profils (Figure 5.17), dont les usagers habitent le centre-ville (figure 5.19) et dont le profil temporel (Figure 5.18) correspond au profil temporel moyen d'un usager de Québec (Figure 5.2). Les *clusters* sont encore moins équilibrés que dans la partie précédente, avec quelques *clusters* de taille moyenne (*clusters* 1.5 et 1.8), quelques petits *clusters* avec 1 à 3% des profils, et d'autres *clusters* minuscules avec moins de 0% de la population (*clusters* 1.4, 1.6 et 1.7).

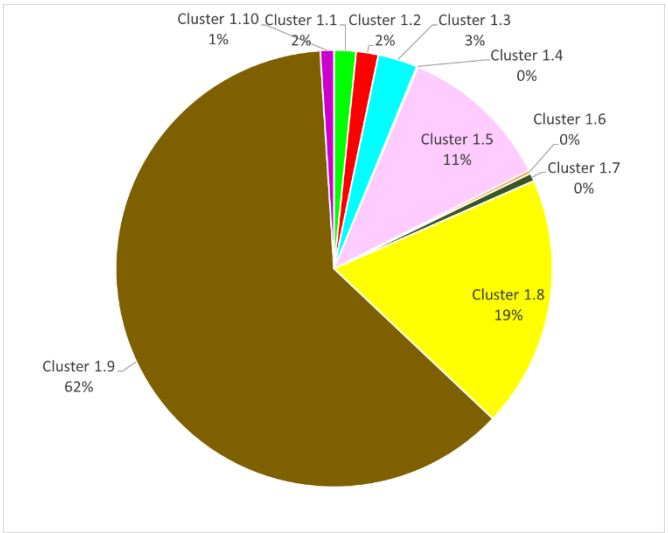


Figure 5.17 Taille de chaque *cluster*

On peut constater sur la Figure 5.18 que les *clusters* présentent beaucoup moins de variété sur le plan temporel que dans les deux études précédentes. La plupart des *clusters* ont le même profil : un pic de transaction à 7h et un deuxième à 16h (ce qui correspond à la tendance moyenne identifiée dans la partie 1, Figure 5.2). Cependant, certains *clusters* présentent tout de même des profils temporels assez différents : le *cluster* 1.1 possède un troisième pic de transaction vers 12h, en plus des pics habituels vers 7h et 16h; le *cluster* 1.4 présente quant à lui un troisième pic vers 21h; et

les *clusters* 1.6 et 1.10 présentent des profils à part assez semblables, avec un petit pic vers 7 ou 8h, une hausse progressive du nombre de transactions jusqu'à 16h, puis une baisse progressive, avec une petite remontée vers 21h. Nous constatons donc que le rapport espace-temps de notre modèle n'est pas parfaitement symétrique : en effet, si pour alpha=0 les *clusters* ne présentent aucunes différences spatiales, pour alpha=1 certains *clusters* diffèrent sur le plan temporel. Cependant, comme on le verra par la suite, ceci s'explique par le fait que les habitants de Québec n'ont pas le même comportement temporel en fonction de leur zone d'habitation, ou plus exactement, de leur couple origine-destination.

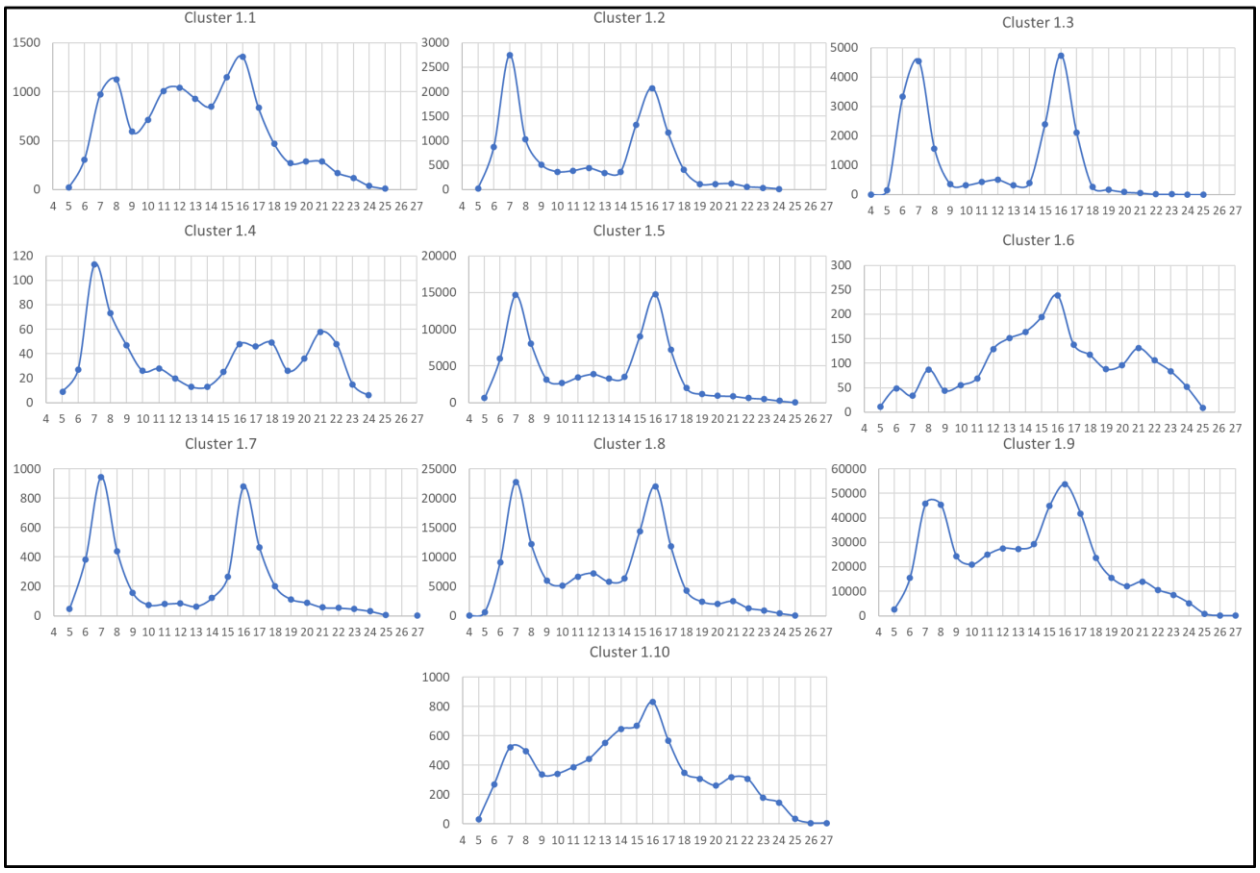


Figure 5.18 Heures de transaction des usagers dans chaque *cluster* (alpha=1)

Néanmoins, c'est surtout sur le plan spatial que les *clusters* diffèrent dans cette étude. La Figure 5.19 (ou pour plus de détail l'annexe C) prouve que chaque *cluster* a une zone d'habitation bien précise, qui lui est propre. Les seuls cas où les zones d'habitation de 2 *clusters* coïncident correspondent aux cas où l'origine de leurs trajets est dans la même zone mais la destination dans

une zone complètement différente pour les deux *clusters*. Ainsi chaque *cluster* présente des caractéristiques spatiales uniques. Par exemple, les usagers du *cluster* 1.1 et 1.2 habitent dans les deux cas la zone de Cap-Rouge, mais tandis que les usagers de *cluster* 1.2 se déplacent vers la Cité Universitaire ou vers le centre-ville, ceux du *cluster* 1.1 restent dans la zone de Cap-Rouge (Figure 5.19 et Annexe C). Ces *clusters* diffèrent aussi sur le plan temporel : le *cluster* 1.2, avec les horaires de travail habituels (Figure 5.18), correspond à un comportement de travail ordinaire; d'ailleurs ce comportement apparaît majoritairement les jours de semaine (Figure 5.22). Le *cluster* 1.1, quant à lui, est soumis à des horaires moins réguliers, et a plus tendance à sortir seulement une demi-journée, d'où le troisième pic de transactions vers 12h (Figure 5.18). Ceci est confirmé par la Figure 5.23, selon laquelle le temps de sortie moyen est bien inférieur dans le *cluster* 1.1 que dans le *cluster* 1.2 (300 minutes par rapport à 500).

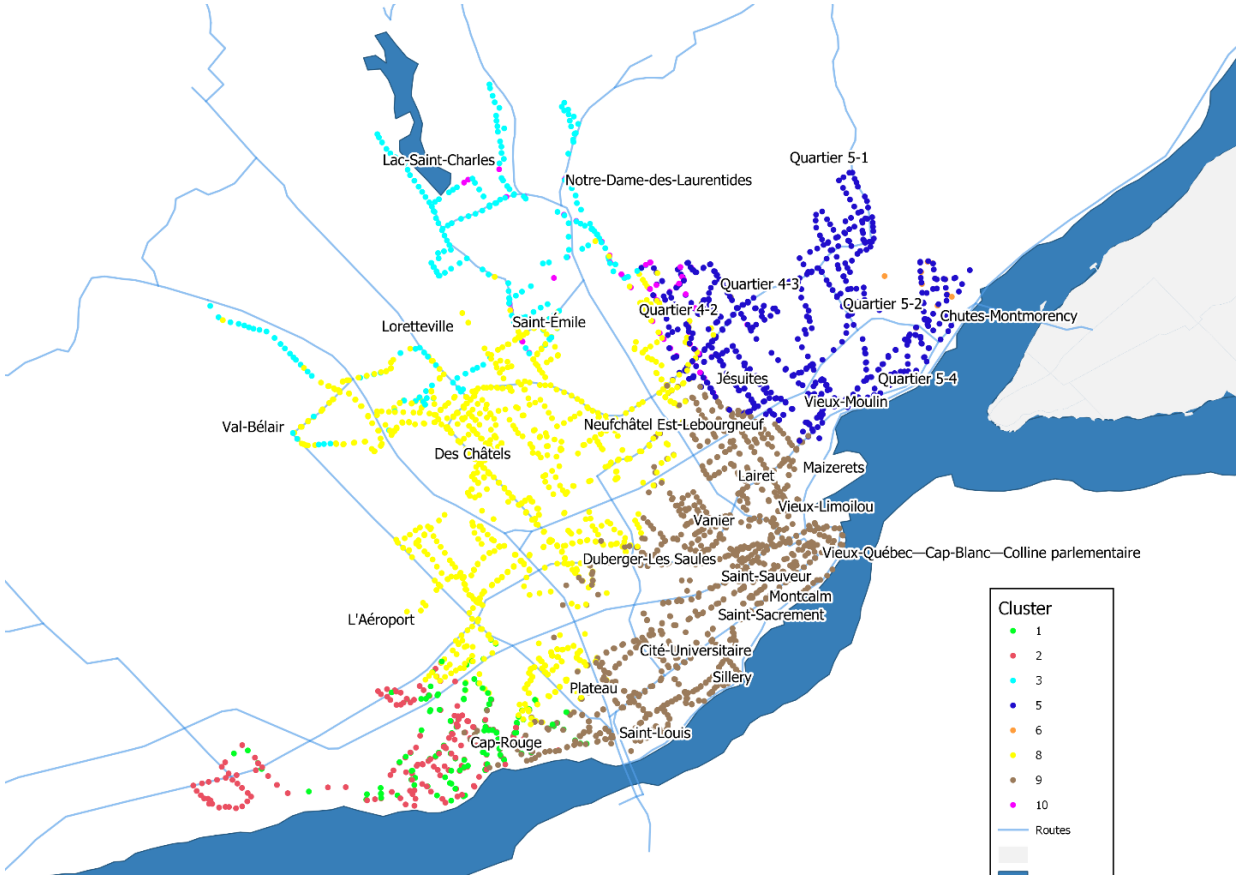


Figure 5.19 *Cluster* le plus fréquent au départ de chaque arrêt, avec alpha=1

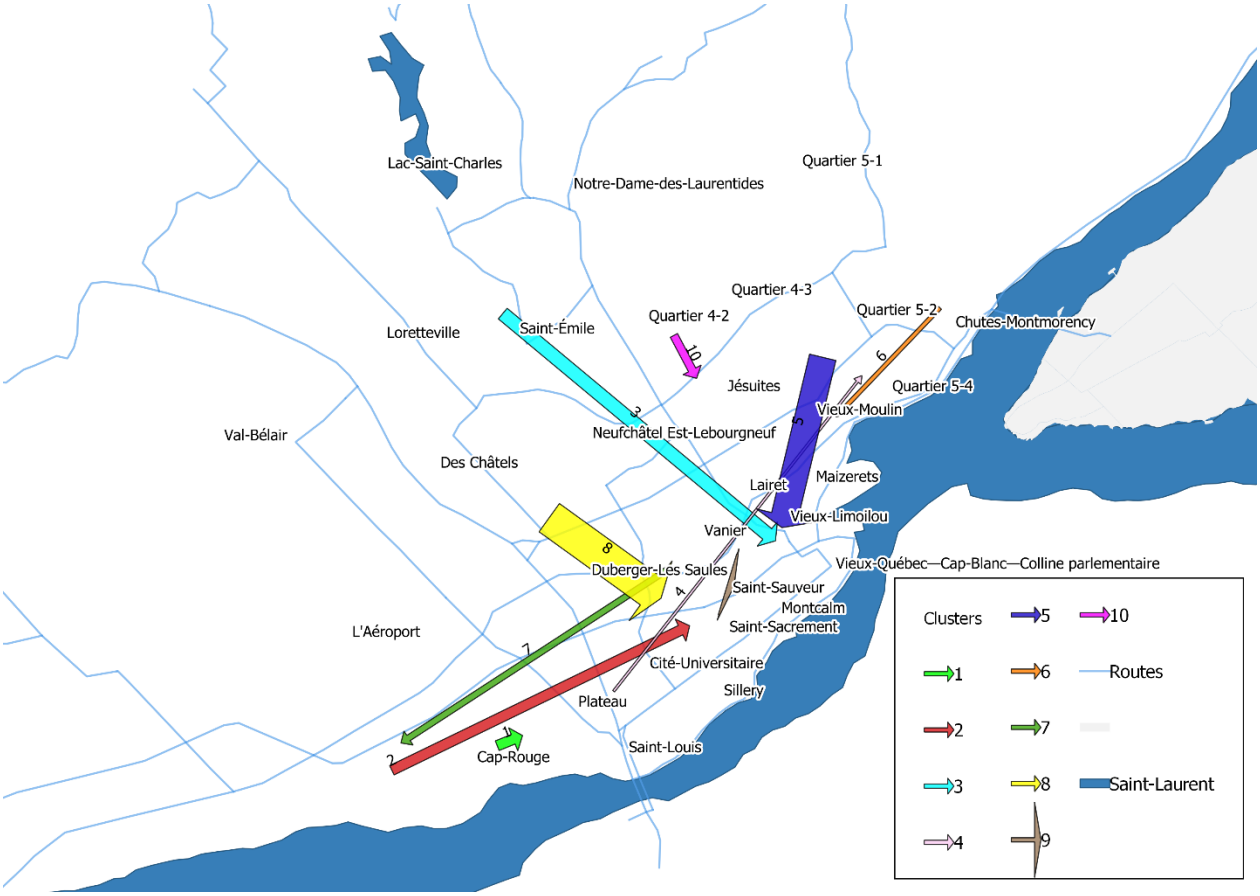


Figure 5.20 Carte des directions de chaque *cluster* (alpha=1)

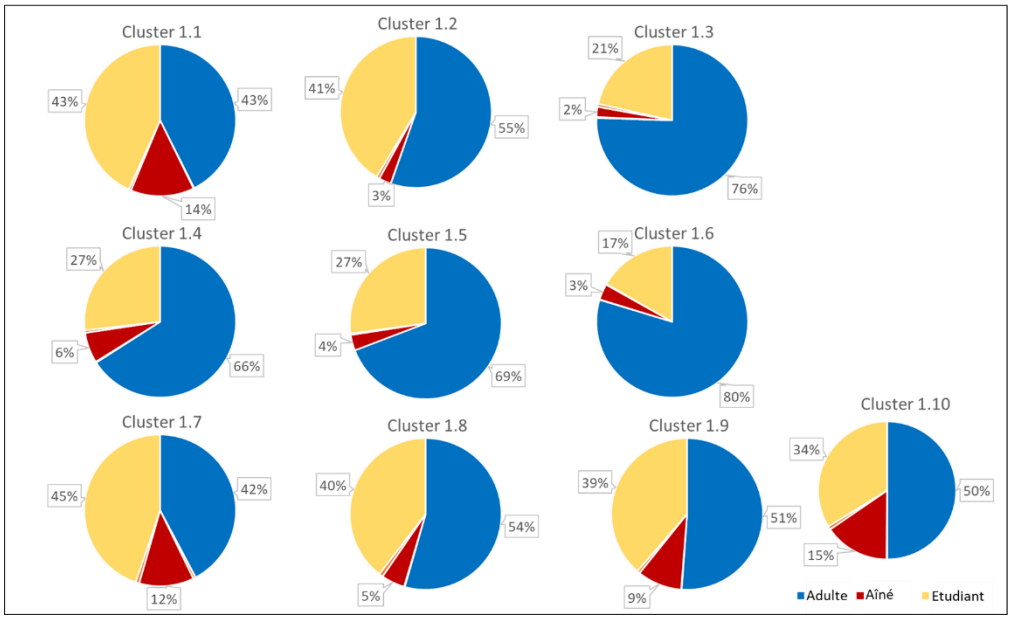


Figure 5.21 Types d'usagers dans chaque *cluster* (alpha=1)

De même, si les usagers des *clusters* 1.3 et 1.10 habitent dans les deux cas au Nord de Québec, du côté de Notre-Dame-des-Laurentides, ceux du *cluster* 1.3 vont travailler dans le centre-ville, tandis que ceux du *cluster* 1.10 restent dans une zone plus proche de leur lieu d'habitation (Figure 5.19 et Annexe C). Ces deux *clusters* diffèrent également sur le plan temporel, avec un comportement de travailleur habituel pour le *cluster* 1.3 et un comportement beaucoup plus libre pour le *cluster* 1.10 (selon la forme de la courbe des transactions Figure 5.18) avec un faible temps de sortie moyen (Figure 5.23). Nous retrouvons aussi que le comportement du *cluster* 1.3 apparait presque uniquement en semaine tandis que celui du *cluster* 1.10 ne dépend pas du jour. Nous constatons aussi une proportion d'adultes beaucoup plus importante dans le *cluster* 1.3 (Figure 5.21). Ainsi, le *cluster* 1.3 serait celui des travailleurs de Québec nord, tandis que le *cluster* 1.10 serait le reste de la population de cette zone (non-travailleurs, ainés et étudiants).

Nous retrouvons exactement le même couple de *clusters* à l'extrémité est de Québec (près des chutes Montmorency), avec les *clusters* 1.5 et 1.6 : le *cluster* 1.5 travaillant dans le centre-ville (Figure 5.19 et annexe C) selon les horaires de travail habituels (Figure 5.18), et le *cluster* 1.6, non soumis à des horaires fixes, restant à proximité du domicile, et sortant pendant une durée moyenne plus réduite que le *cluster* 1.5 (Figure 5.23)

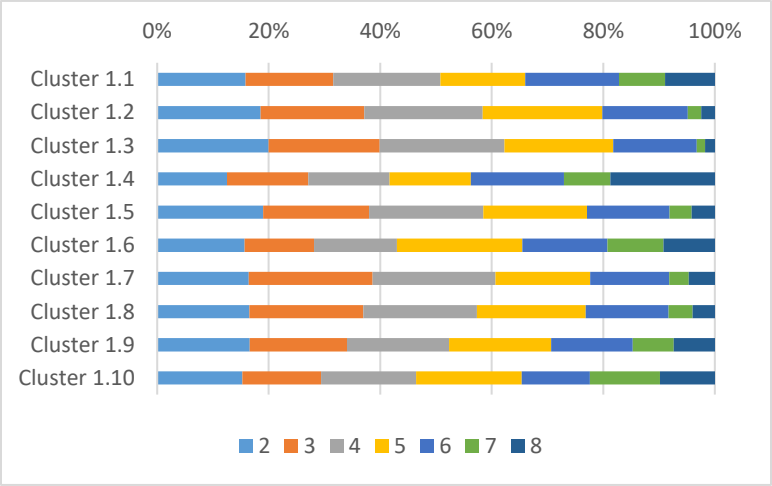


Figure 5.22 Proportion de chaque jour de la première semaine dans chaque *cluster* (alpha=1)

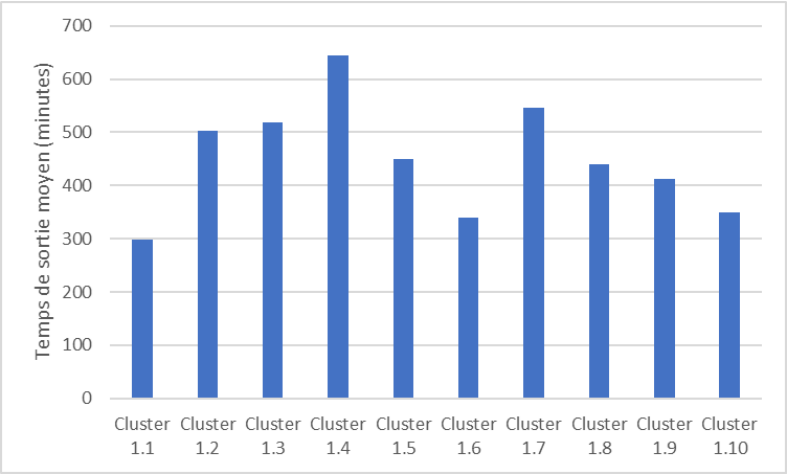


Figure 5.23 Temps de sortie moyen dans chaque *cluster* (alpha=1)

Les *clusters* 1.4 et 1.7 présentent tous deux un comportement spatial vraiment à part, caractérisé par une longue distance de parcours, et par des destinations inhabituelles, à l'extérieur de la ville. Les usagers du *cluster* 1.4 habitent près de Cap-Rouge et vont travailler à l'autre bout de la ville près des chutes Montmorency. Ceux du *cluster* 1.7 regroupent les usagers se rendant à l'extrémité ouest de Québec, près du village de Saint-Augustin de Desmaures. Avec moins de 1% des profils, ces deux *clusters* sont vraiment marginaux, et leur comportement est très peu répandu à Québec. Si ces *clusters* ont été formés, c'est que leur distance DTW aux autres *clusters* est très élevée. Ainsi, ces *clusters* ne seraient probablement pas être pris en compte lors d'une étude de comportement des usagers.

Enfin, le *cluster* 1.8, très semblable au *cluster* 1.9, représente un groupe important d'usagers travaillant dans la même zone que le *cluster* 1.9 (centre-ville et, dans une moindre mesure, Cité Universitaire), selon les mêmes horaires de travail, mais habitant un peu plus en périphérie, au Nord-Ouest du centre-ville. Ils diffèrent du *cluster* 1.9 par leur temps de trajet légèrement plus long.

Pour résumer, chacun des *clusters* de cette étude présente un comportement spatial très différent, à la fois au niveau de l'origine de leurs trajets et de leurs destinations. Certains des *clusters* en périphérie peuvent être regroupés en paires : un des deux *clusters* regroupant les usagers se déplaçant vers le centre-ville ou la Cité Universitaire, et un deuxième regroupant les usagers restant à proximité de leur domicile. Deux des *clusters* (1.4 et 1.7) correspondent à des comportements spatiaux vraiment à part, caractérisés par des très longues distances de parcours et des destinations à la périphérie de la ville : ces *clusters* représentant une très faible partie des profils, ils peuvent être négligés. Les deux *clusters* centraux (1.8 et 1.9) regroupent le reste des habitants, et pourraient éventuellement être divisés en d'autres sous-*clusters* pour un découpage plus précis. Si certains *clusters* minoritaires présentent des comportements temporels différents, c'est parce que les usagers empruntant les couples origine-destination de ces *clusters* n'ont pas des horaires habituels. Cependant, c'est bien le facteur spatial du comportement, caractérisé par les séquences d'arrêts visités, qui prédomine dans la détermination des *clusters*. On peut donc dire que ce modèle est à influence purement spatiale.

5.5 Conclusion des trois études

Nous avons étudié le comportement des usagers de la RTC avec trois modèles de segmentation différents, correspondant à trois valeurs différentes pour le paramètre alpha : 0; 0,4 et 1. Le premier modèle (alpha=0) renvoie une segmentation purement temporelle des usagers, et permet d'identifier des comportements temporels très variés, et représentent chacun une part non négligeable de la population. Ce modèle permet d'avoir un bon aperçu des différents comportements temporels de la population. Le deuxième modèle (alpha=0,4) correspond à peu près à l'équilibre entre l'influence du temps et de l'espace. Les *clusters* créés diffèrent à la fois temporellement et spatialement. On distingue plusieurs 'paires' de *clusters*, identifiant des groupes d'usagers habitant au même endroit mais ayant un comportement temporel différent. Le dernier

modèle (alpha=1) maximise l'influence de l'espace sur les résultats. Les usagers sont répartis en plusieurs zones géographiques, plus ou moins éloignées du centre-ville. A cause de la différence de densité démographique entre le centre de Québec et sa périphérie, les *clusters* éloignés du centre-ville sont beaucoup plus petits. Comme dans la deuxième étude, nous observons des paires de *clusters*, ayant la même zone d'habitation mais différant par leur zone de déplacement. Nous observons également quelques différences temporelles entre les *clusters*, dues au fait que le comportement temporel d'un usager n'est pas le même en fonction de son couple origine-destination. Dans chacun des modèles, nous constatons la présence d'un *cluster* beaucoup plus grand (en particulier pour alpha=0,4 et alpha=1), qui regroupe tous les profils ne ressortant des autres profils ni par leur comportement temporel ni par leur comportement spatial : ces gros *clusters* pourraient facilement être scindés en plusieurs sous-*clusters*, en utilisant la segmentation hiérarchique issue de l'algorithme CHA. Dans les trois modèles également, nous observons des différences socio-démographiques (représentées par la tranche d'âge de la population) entre les *clusters*, mais celles-ci sont le plus marquées pour alpha=0,4, ce qui démontre l'intérêt de considérer simultanément le temps et l'espace dans un modèle de segmentation.

CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS

Nous sommes donc parvenus, au cours de nos travaux, à créer un modèle de segmentation spatio-temporelle dans lequel l'influence de l'espace et du temps peuvent être contrôlées de manière précise. Selon les résultats de notre première expérience (chapitre 4), la variation du paramètre d'équilibrage spatio-temporel alpha engendre une variation continue et à peu près linéaire des caractéristiques spatiales et temporelles des *clusters* : lors de l'augmentation de alpha, la dispersion temporelle augmente progressivement, tandis que la dispersion spatiale diminue progressivement. Ceci prouve que le paramètre alpha permet un grand contrôle sur l'équilibre espace-temps des résultats. Notre deuxième expérience (chapitre 5) consistait à réaliser l'étude du comportement des usagers de la RTC avec trois valeurs différentes pour le paramètre d'équilibrage espace-temps alpha. Ces trois études ont fourni des résultats très différents, mais toujours pertinents. Selon la valeur de alpha, les différences entre *clusters* étaient de nature temporelle uniquement (alpha=0), spatiale et temporelle simultanément (alpha=0,4), ou principalement spatiales (alpha=1). Dans chacun des trois cas, les usagers sont répartis de manière très différente, mais nous observons toujours des différences sociodémographiques entre chaque *cluster*, en particulier pour alpha=0,4, ce qui prouve l'intérêt de considérer simultanément le temps et l'espace.

Notre travail devrait permettre de nombreuses avancées dans l'étude du comportement des usagers des réseaux de transport en commun : en effet, en fonction de l'application, le rapport souhaité entre l'influence de l'espace et le temps peut varier. Pour la première fois, à notre connaissance, nous disposons d'un modèle qui permet de contrôler cet équilibre. De plus, cet équilibre est très facilement modifiable, à l'aide d'un paramètre uniquement, et permet donc d'obtenir une grande diversité de points de vue sur un même réseau, comme en témoignent les résultats du chapitre 5. Cette facilité d'utilisation devrait permettre aux utilisateurs de l'outil de trouver rapidement une valeur d'équilibre spatio-temporel qui convienne à leur application. Par exemple, si l'on souhaite créer une nouvelle ligne de bus, on peut augmenter progressivement alpha jusqu'à voir apparaître des tendances spatiales spécifiques pour chaque cluster, tout en veillant à ne pas augmenter trop alpha afin de conserver plusieurs clusters temporels au sein de la même zone spatiale : de cette façon, on pourrait identifier plusieurs lignes de bus potentielles, ainsi que les horaires d'opération optimales de ces lignes.

Cependant, ce modèle reste très perfectible. Le choix de l'échantillon doit être amélioré : il faudrait mener une étude d'optimisation de la taille d'échantillon, en utilisant les indices de Dunn, de Davies-Bouldin, ou autres. Nous observons en effet une grande variabilité des résultats en fonction de l'échantillon choisi. Peut-être qu'une plus grande taille d'échantillon permettrait de réduire cette variabilité? La taille d'échantillon reste cependant limitée par le temps de calcul, proportionnel à son carré. Une autre option serait de sélectionner l'échantillon de manière non aléatoire, en réalisant par exemple une première segmentation uniquement spatiale des profils, avec un algorithme *k-means*. Par ailleurs, notre métrique spatiale pourrait être améliorée en utilisant les temps de parcours entre deux arrêts, plutôt que la distance géographique entre ces arrêts. Ceci serait plus adéquat avec la réalité. Il faut aussi noter que les profils utilisés dans notre segmentation ne représentent qu'une partie des usagers de la RTC (environ 70%) : en effet, les profils contenant un seul trajet, ou bien des trajets dont la destination n'a pas pu être estimée, ont été retirés de la base de données.

Il serait aussi intéressant d'étudier plus en profondeur comment les résultats varient, de manière qualitative, lorsque l'on fait varier alpha continument. On pourrait par exemple fixer un échantillon précis, et faire tourner notre modèle sur cet échantillon avec 20 valeurs de alpha comprises entre 0 et 1, en prenant toujours le même nombre de *clusters* : on observerait alors comment chaque *cluster* apparait, évolue ou disparait lorsque alpha varie. Enfin, l'évolution du comportement des usagers serait un sujet intéressant à analyser avec notre modèle. On pourrait par exemple suivre un usager au cours du mois, et étudier à quel *cluster* il est assigné chaque jour, pour évaluer la régularité de son comportement. Ainsi, certains *clusters* apparaitraient comme des *clusters* de comportement plus 'occasionnels' tandis que d'autres seraient des *clusters* de comportement 'réguliers'. Les usagers eux-mêmes pourraient être divisés en plusieurs *clusters* en fonction de la régularité de leur comportement.

RÉFÉRENCES

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

Agard, B., Partovi Nia, V., & Trépanier, M. (2013). *Assessing public transport travel behaviour from smart card data with advanced data mining techniques.* Paper presented at the World Conference on Transport Research.

Alsger, A., Assemi, B., Mesbah, M., & Ferreira, L. (2016). Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C: Emerging Technologies, 68,* 490-506.

Alliance activities : Applications : Transportation. (2021, March 02). Retrieved April 05, 2021, from https://www.securetechalliance.org/applications-transportation/

Arana, P., Cabezudo, S., & Peñalba, M. (2014). Influence of weather conditions on transit ridership: A statistical study using data from Smartcards. *Transportation research part A: policy and practice, 59,* 1-12.

Attoh-Okine, N., & Shen, L. (n.d.). Security issues of emerging smart cards fare collection application in mass transit. *Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS. A Ride into the Future*. doi:10.1109/vnis.1995.518887

Bagchi, M., & White, P. (2005). The potential of public transport smart card data. *Transport Policy, 12*(5), 464-474.

Blythe, P. T. (2004). Improving public transport ticketing through smart cards. *Municipal Engineer, 157*(1), 47-54. doi:10.1680/muen.157.1.47.36592

Ceapa, I., Smith, C., & Capra, L. (2012, August). Avoiding the crowds: understanding tube station congestion patterns from trip data. In *Proceedings of the ACM SIGKDD international workshop on urban computing* (pp. 134-141). ACM.

Cheung, F. (2006). Implementation of nationwide public transport smart card in the Netherlands. *Transportation Research Record: Journal of the Transportation Research Board, 1971*(1), 127-132. doi:10.1177/0361198106197100115

Chu, K.A., & Chapleau, R. (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (2063), 63-72.

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.

El Mahrsi, Mohamed & Côme, Etienne & Baro, Johanna & Oukhellou, Latifa. (2014). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. *The 3rd International Workshop on Urban Computing (UrbComp 2014)*

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of transport geography*, *49*, 26-38.

Faroqi, H., Mesbah, M. and Kim, J., 2019. Comparing Sequential with Combined Spatiotemporal *Cluster*ing of Passenger Trips in the Public Transit Network Using Smart Card Data. *Mathematical Problems in Engineering*, 2019.

Ghaemi, M. S., Agard, B., Nia, V. P., & Trépanier, M. (2015). Challenges in Spatial-Temporal Data Analysis Targeting Public Transport. *IFAC-PapersOnLine*, *48*(3), 442-447.

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

Ghaemi, M. S. (2017b). *Forestogram: biclustering visualization framework with applications in public transport and bioinformatics.* École Polytechnique de Montréal. Thèse de doctorat.

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

He, L., Trépanier, M., Hickman, M., & Nassir, N. (2015). Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems (No. CIRRELT-2015-52). *CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et*

le transport= Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation.

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

He, L., Agard, B., & Trépanier, M. (2018b). Space-time classification of public transit smart card users' activity locations from smart card data. *Conference on Advanced Systems in Public Transport and TransitData 2018*, paper 62.

He, L., Trépanier, M., Agard B., Munizaga M., Bustos B. (2019b). Comparing transit user behavior of two cities using smart card data. *Annual Meeting of the Transportation Research Board*, Washington, DC. No. 19-05564.

Kieu, L. M., Bhaskar, A., & Chung, E. (2014). Transit passenger segmentation using travel regularity mined from Smart Card transactions data. In *Transportation Research Board 93rd Annual Meeting*, 12-16 January 2014, Washington, D.C.

Kurauchi, F., & Schmöcker, J. D. (Eds.). (2017). Public Transport Planning with Smart Card Data. CRC Press.

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.

Langlois, G. 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.

Lu, H. K. (2007). Network smart card review and analysis. *Computer Networks*, *51*(9), 2234-2248.

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.

Meadowcroft, P. (2005). Hong Kong raises the bar in smart card innovation. *Card Technology Today*, *17*(1), 12-13.

Morency, C., Trepanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, *14*(3), 193-203.

Nishiuchi, H., King, J., & Todoroki, T. (2013). Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data. *International Journal of Intelligent Transportation Systems Research*, *11*(1), 1-10.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids *cluster*ing. *Expert systems with applications*, *36*(2), 3336-3341.

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

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.

Seaborn, C., Wilson, N.H., Attanucci, J., 2009. Using Smart Card Fare Payment Data to Analyze Multi-Modal Public Transport Journeys (London, UK). 88th. *Annual Meeting of the Transportation Research Board, Washington*, 16 p.

Spurr, T., Chapleau, R., & Piché, D. (2014). *Discovery and Partial Correction of Travel Survey Bias Using Subway Smart Card Transactions.* Paper presented at the Transportation Research Board 93rd Annual Meeting. 2405, 56-67.

Statistics and machine learning toolbox. (n.d.). Retrieved April 06, 2021, from https://www.mathworks.com/help/stats/index.html

Trépanier, M., Barj, S., Dufour, C., & Poilpré, R. (2004). Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain. *Congrès de l'Association des transports du Canada*. Québec. 10–14.

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.

Trépanier, M., & Morency, C. (2010). *Assessing transit loyalty with smart card data.* Paper presented at the 12th World Conference on Transport Research, Lisbon, Portugal.

Trépanier, Morency, C., Agard, B., Descoimps, E., & Marcotte, J. (2012). *Using smart card data to assess the impacts of weather on public transport user behavior.* Paper presented at the Conference on Advanced Systems for Public Transit-CASPT12, Santiago, Chile.

Trépanier, M. (2013). L'exploitation des données de cartes à puce à des fins de planification des transports collectifs urbains. *Recherche Transports Sécurité, 2012*(02), 139-152.

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential uses of transit smart card registration and transaction data to improve transit planning. *Transportation Research Record: Journal of the Transportation Research Board, 1971*(1), 118-126.

Yap, M., Cats, O., & van Arem, B. (2018). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science*, 1-20.

Zhao, J., Qu, Q., Zhang, F., Xu, C., & Liu, S. (2017). Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems, Nov. 2017*, vol. 18, no. 11, 3135-3146

ANNEXE A BIBLIOGRAPHIE SUR L'UTILISATION DES DONNÉES DE CARTES À PUCE

Tableau A.1 Bibliographie sur l'utilisation des données de cartes à puce

Auteurs	Titre de l'article
Reconstruction des trajets	
Estimation des destinations	
Trépanier et al., 2007	Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System
Chu et al., 2008	Enriching archived smart card transaction data for transit demand modeling
Seaborn et al., 2009	Using Smart Card Fare Payment Data To Analyze Multi-Modal Public Transport Journeys in London
Devillaine et al., 2012	Detection of public transport user activities through the analysis of smartcard data
Kusakabe et al., 2014	Behavioural data mining of transit smart card data: A data fusion approach
He et al., 2015	Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems
Alsger et al., 2016	Validating and improving public transport origin–destination estimation algorithm using smart card fare data
Giraud et al., 2016	Data fusion of APC, Smart Card and GTFS to Visualize Public Transit Use
Etude du comportement des usagers	
Segmentation des usagers	
Cf. tableau suivant	
Etude de la congestion	
Park et al., 2008	Use of smart card data to define public transit use in Seoul, South Korea
Ceapa et al., 2012	Avoiding the crowds: understanding tube station congestion patterns from trip data
Yap et al., 2018	Crowding valuation in urban tram and bus transportation based on smart card data
Mise en relation avec d'autres types de données	
Trépanier, 2012	L'exploitation des données de cartes à puce à des fins de planification des transports collectifs urbains
Spurr et al., 2014	Use of Subway Smart Card Transactions for the Discovery and Partial Correction of Travel Survey Bias
Evolution du comportement des usagers	
Bagchi et al., 2005	The potential of public transport smart card data
Morency et al., 2007	Measuring transit use variability with smart-card data
Trépanier et Morency, 2010	Assessing transit loyalty with smart card data
Trépanier et al., 2012	Are transit users loyal? Revelations from a hazard model based on smart card data
Kusakabe et al., 2014	Behavioural data mining of transit smart card data: A data fusion approach
Langlois et al., 2016	Inferring patterns in the multi-week activity sequences of public transport users
Influence de la météo	
Arana et al., 2014	Influence of weather conditions on transit ridership: A statistical study using data from Smartcards
Zhou et al., 2017	Impacts of weather on public transport ridership: Results from mining data from different sources
Influence du type d'usager	
Trépanier, 2013	L'exploitation des données de cartes à puce à des fins de planification des transports collectifs urbains
Nishiuchi et al., 2013	Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data
El Mahrsi et al., 2014	Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France
Amélioration des services	
Etude des impacts positifs de l'utilisation des cartes à puces	
Attoh-Okine et al., 1995	Security issues of emerging smart card fare collection application in mass transit
Meadowcorft, 2005	Hong Kong raises the bar in smart card innovation.
Cheung, 2006	Implementation of Nationwide Public Transport Smart Card in The Netherlands: Cost-Benefit Analysis
Smart Card Alliance	https://www.securetechalliance.org/applications-transportation/
Stratégies d'amélioration des réseaux de transports	
Trépanier et al., 2004	Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain
Blythe, 2004	Improving public transport ticketing through smart cards
Utsunomiya et al., 2006	Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning
Lu., 2007	Network smart card review and analysis
Pelletier et al., 2011	Smart card data use in public transit: A literature review.
Kurauchi et al., 2017	Public Transport Planning with Smart Card Data

ANNEXE B CARTES DE CHALEUR DES PROFILS A CERTAINES HEURES (ALPHA=0,4)

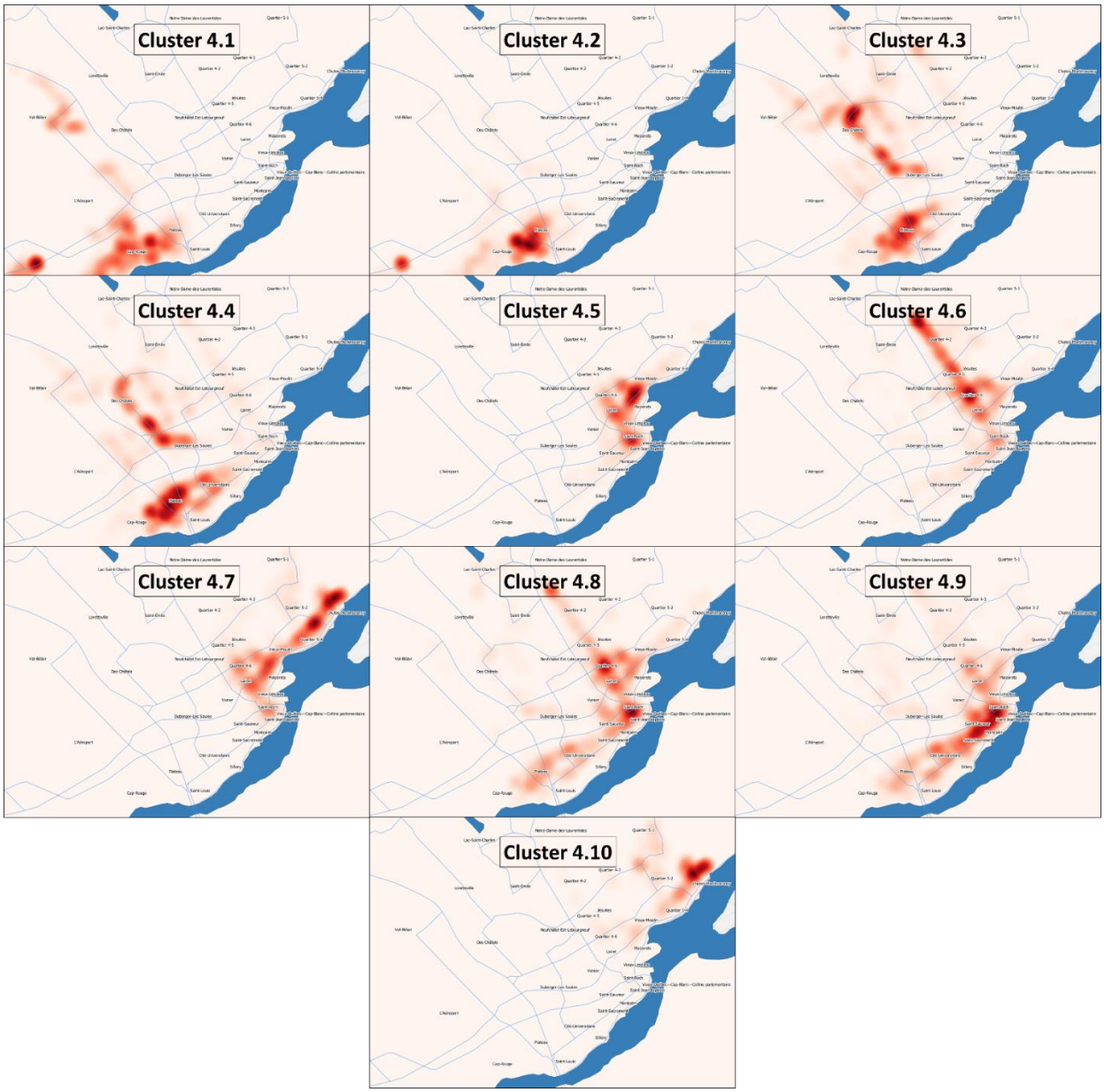


Figure B.1 Carte de chaleur des profils de chaque *cluster* à 4:00

La première figure peut être assimilée aux zones d'habitation des usagers de chaque *cluster*, tandis que la deuxième correspondrait plus tôt à leurs zones de travail.

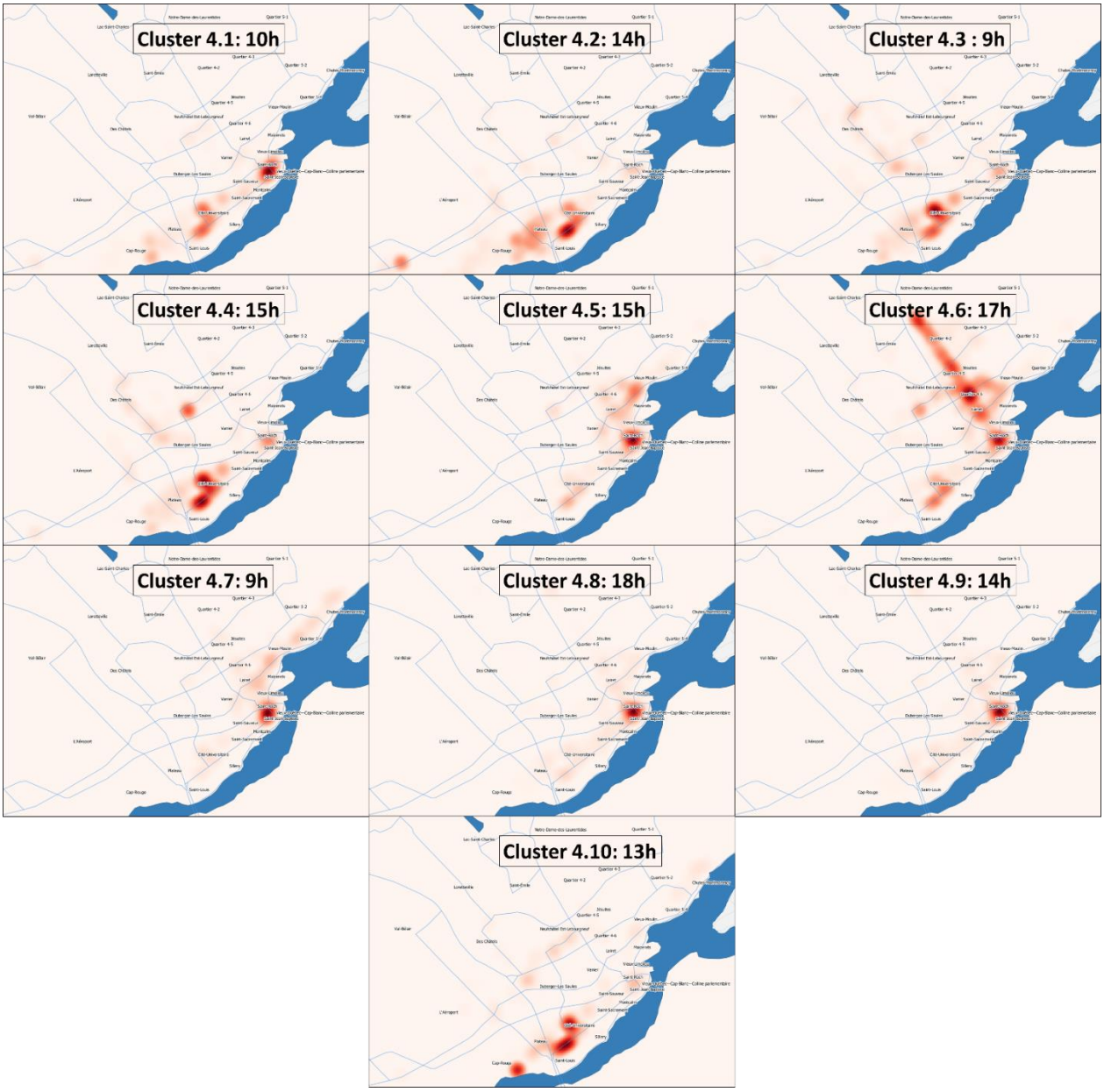


Figure B.2 Cartes de chaleur des profils de chaque *cluster* entre deux pics de transactions

ANNEXE C CARTES DE CHALEUR DES PROFILS A CERTAINES HEURES (ALPHA=1)

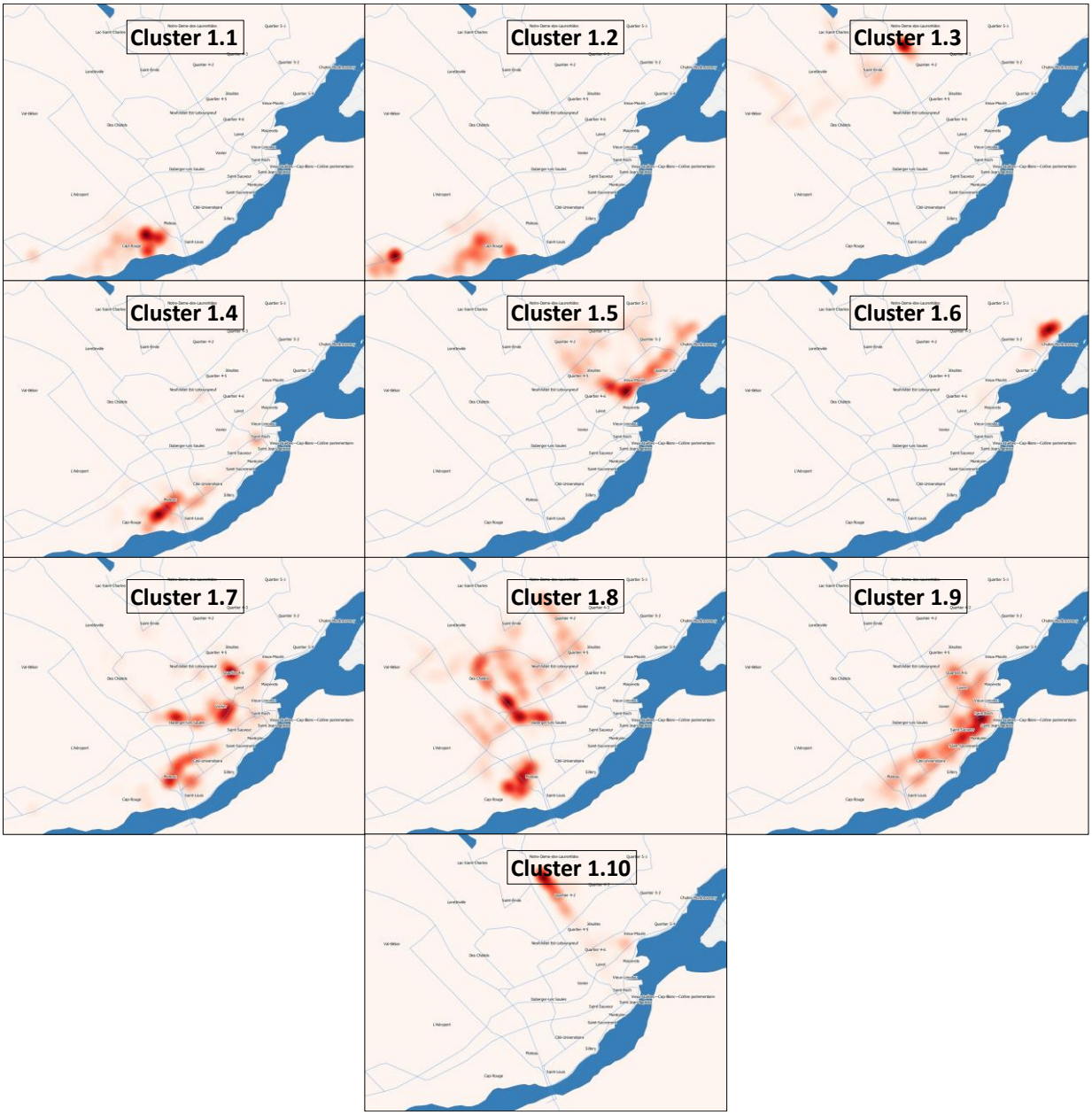


Figure C.3 Carte de chaleur des profils de chaque *cluster* à 4:00

La première figure peut être assimilée aux zones d'habitation des usagers de chaque *cluster*, tandis que la deuxième correspondrait plus tôt à leurs zones de travail.

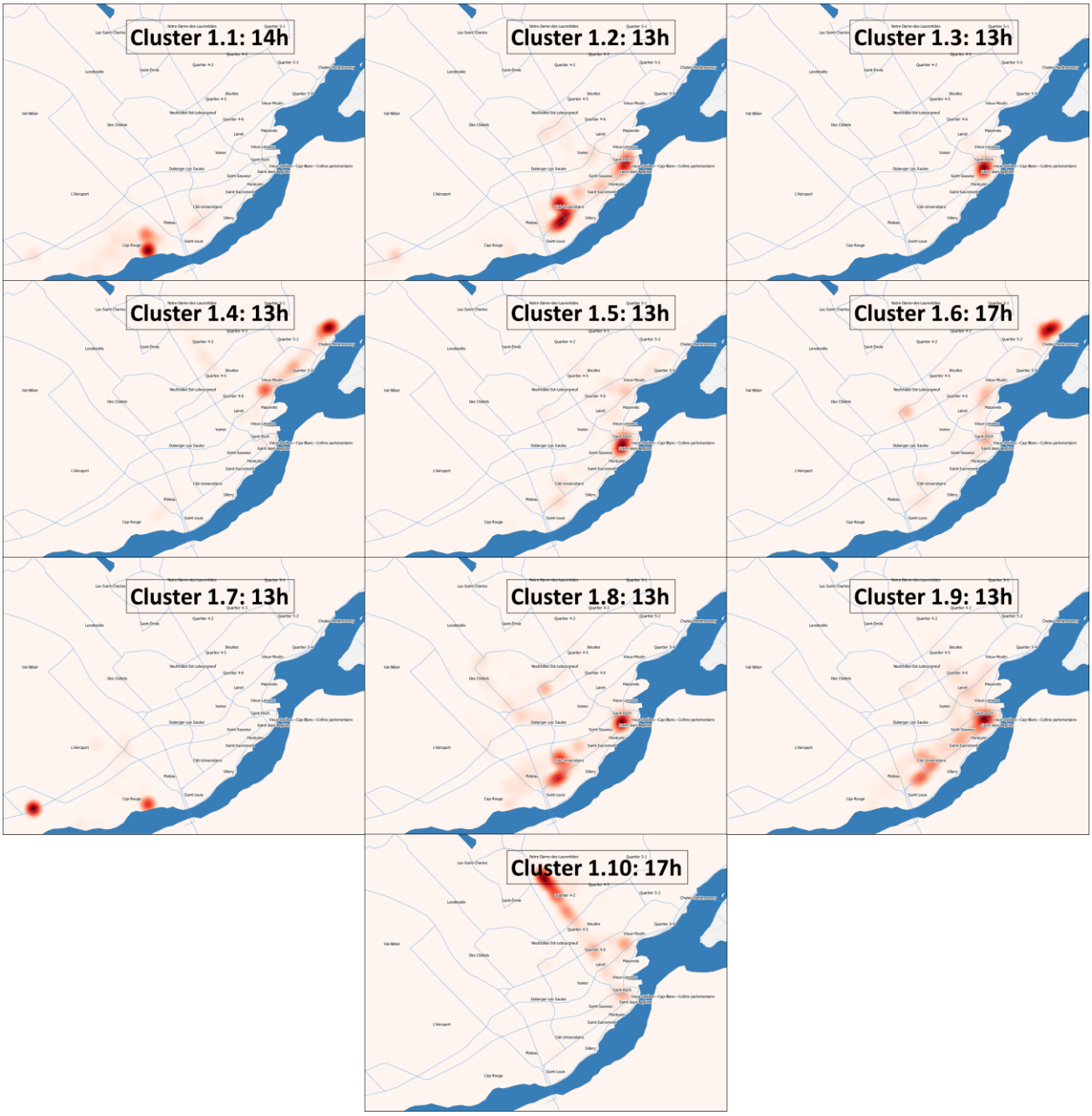


Figure C.4 Cartes de chaleur des profils de chaque *cluster* entre deux pics de transaction