POLYPUBLIE
Polytechnique Montréal

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Titre: Title:	Modélisation longitudinale de l'usage d'un réseau de transport collectif
Auteur: Author:	Antonin Darrou
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Darrou, A. (2021). Modélisation longitudinale de l'usage d'un réseau de transport collectif [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10007/

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10007/
Directeurs de recherche: Advisors:	Catherine Morency, & Catherine Trépanier
Programme: Program:	Génie civil

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

Modélisation longitudinale de l'usage d'un réseau de transport collectif

ANTONIN DARROU

Département des génies civil, géologique et des mines

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

Génie Civil

Septembre 2021

© Antonin Darrou, 2021.

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé:

Modélisation longitudinale de l'usage d'un réseau de transport collectif

Présenté par **Antonin DARROU**

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

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

Francesco CIARI, président

Catherine MORENCY, membre et directrice de recherche

Martin TRÉPANIER, membre et codirecteur de recherche

Atylla BRITTO, membre

REMERCIEMENTS

Je tiens tout d'abord à remercier ma directrice Catherine MORENCY pour m'avoir introduit au domaine des transports. Ses compétences et son dynamisme m'ont permis de mener à bien ce mémoire et d'acquérir des connaissances précieuses. Je souhaite également remercier mon codirecteur Martin TREPANIER. Son expertise et ses conseils m'ont permis d'avancer dans ce projet. Travailler avec eux a été une expérience épanouissante et enrichissante.

Je souhaite également remercier l'École Polytechnique Montréal et l'École Nationale Supérieure d'Arts et Métiers pour m'avoir permis d'effectuer un double diplôme et d'avoir ainsi eu l'opportunité de réaliser cette maitrise.

Je remercie également toutes les personnes ayant contribué à la réalisation de ce travail. Notamment, je remercie Attyla BRITO pour avoir été le contact principal au sein de la STO et dont la connaissance du réseau de transport a été une aide précieuse dans l'exploitation des résultats. Je remercie également Alexis VIALLARD pour m'avoir fourni et aidé dans l'exploitation des données de cartes à puce. Je remercie finalement Hubert VERREAULT pour m'avoir introduit aux modèles faisant l'usage de typologies.

RÉSUMÉ

La compréhension des évolutions de la fréquentation d'un réseau de transport en commun est une caractéristique importante dans la gestion d'une société de transport collectif. Avoir une prévision de l'achalandage futur sur un réseau de transport permet d'anticiper les changements et ainsi d'adapter l'offre à la demande. Suite à des augmentations de l'achalandage à sein du réseau de la Société de transport de l'Outaouais (STO - Gatineau, Québec, CA), le besoin pouvoir prévoir l'évolution de l'achalandage futur a été formulé. L'objectif de ce mémoire est de comprendre et d'estimer les évolutions de la fréquentation du réseau de bus à moyen terme. Sans un modèle de prévision de l'achalandage, il est difficile d'identifier et principalement de quantifier l'effet des facteurs influençant la demande. Aucune décision stratégique ne peut donc être prise avec certitude. La problématique majeure concerne l'anticipation d'acquisition de ressources matérielles et humaines. Un modèle de prévision longitudinale de la fréquentation du réseau de transport est donc mis en place dans ce mémoire.

Le développement d'un modèle de prévision longitudinale adapté à la zone d'étude est déterminé par la comparaison de plusieurs modèles afin de déterminer le plus pertinent. Le mémoire s'organise autour de plusieurs étapes. La revue de la littérature identifie un panel de modèles potentiels. L'analyse rétrospective de la demande et d'indicateurs d'évolution de la région de Gatineau confirme les modèles de prévision à calibrer ainsi que les facteurs explicatifs à utiliser. Ces modèles sont mis en place par la suite au fils de trois niveaux d'agrégation différents : le réseau global, les secteurs municipaux et un niveau d'agrégation précis déterminé spécifiquement pour remplir les objectifs du projet. La performance des modèles de prévision est comparée afin de déterminer le plus pertinent pour le réseau en question.

La revue de la littérature est la première étape de tout travail de recherche, car elle permet d'acquérir des connaissances supplémentaires sur le sujet d'étude, mais principalement de situer les travaux dans le contexte pour identifier les contributions possibles. Cette partie aboutie sur l'identification de modèles potentiellement pertinents pour remplir les objectifs de ce mémoire. Les modèles utilisés sont principalement constitués de modèles autorégressifs, de régressions linéaires sur séries temporelles et de modèles de génération de taux de montants. Elle identifie également un manque de travaux dans la littérature visant à comparer les performances de plusieurs modèles sur un même territoire d'étude. La méthodologie est justifiée par cette lacune et a pour but

de comparer différentes structures de modèles afin d'identifier les forces et faiblesses de chacun. Cette comparaison aboutit à l'identification du modèle le plus pertinent afin de remplir les objectifs fixés.

Un travail préliminaire est effectué avant toute expérimentation. Pour permettre de capturer des modifications de la fréquentation du réseau de bus à l'échelle locale malgré la constante évolution du réseau, le besoin de créer des zones d'agrégation de petite taille a été constaté. L'analyse au niveau des arrêts n'était pas possible à cause de la modification des arrêts au fil des années empêchant d'avoir une continuité temporelle de cet élément. Un maillage hexagonal pertinent pour l'étude est développé afin d'agréger toutes les données utilisées. Les prévisions à ce niveau d'agrégation permettent d'obtenir une image de l'utilisation du réseau de bus à l'échelle locale.

L'enjeu étant de déterminer un modèle adapté au territoire de la STO, une analyse rétrospective de la demande et de divers indicateurs permettant de décrire la région d'étude est effectuée. Les principales sources de données sont constituées des données de carte à puce du réseau de la STO, du recensement et de l'enquête Origine-Destination (OD) sur la région de Gatineau. Malgré l'identification d'un manque de données longitudinales, les indicateurs présents dans ces bases de données ont permis d'identifier des facteurs d'influence de la demande. Les données de permis de construire ainsi que celles concernant les stationnements incitatifs ont été identifiées comme les données les plus utiles pour le développement de modèles de prévision. L'analyse rétrospective met en valeurs l'influence de multiples facteurs explicatifs comme le taux d'emploi, mais le manque de données dans le temps empêche leur utilisation dans des modèles longitudinaux.

Les parties d'expérimentation sont donc les étapes qui suivent. Des modèles longitudinaux de prévision de l'achalandage sont calibrés à trois échelles d'agrégation : le réseau global, les secteurs municipaux et les zones d'agrégation précises déterminées spécifiquement pour répondre aux enjeux de l'étude. Cette structure permet d'identifier quels paramètres permettent d'expliquer la demande aux différentes échelles. Au niveau du réseau global, les variables socio-économiques sont pertinentes pour obtenir une estimation de la tendance décrite par la fréquentation du réseau. À l'échelle des zones hexagonales d'agrégation, la demande est plus sensible à des facteurs locaux. Le nombre de stationnements incitatifs ou l'évolution du nombre de logements dans chaque zone permettent de capturer les évolutions à l'échelle locale. Mais, les variables socio-économiques ne le permettent pas à l'échelle d'un quartier.

Pour chacune des unités d'agrégation des données, plusieurs modèles sont développés en suivant la logique suivante : commencer par un modèle simple et le complexifier en prenant compte des erreurs constatées sur les modèles précédents. Cette méthode permet d'améliorer la performance au cours des chapitres en étudiant les résidus des modèles. Le modèle autorégressif ARIMA est pour tous les niveaux d'agrégation la référence. Utiliser uniquement l'historique de la demande pour prévoir la demande future permet de produire des prévisions en supposant qu'aucune perturbation majeure ne vient influencer la fréquentation du réseau. L'étude des résidus permet d'identifier des instants où la prévision diverge et ainsi d'identifier un paramètre explicatif permettant de prendre en compte ce changement de tendance. Ces expérimentations aboutissent sur deux modèles de prévision adaptés au réseau de la STO. Ces modèles concernent tous deux une agrégation des données selon les zones hexagonales fines. Un modèle de régression linéaire sur séries temporelles, prenant en compte l'historique de la demande, l'évolution du nombre de logements et des données sur les stationnements incitatifs, permet de bien capter l'évolution du territoire par le biais du peu de données longitudinales disponibles. Un autre modèle prometteur est issu d'une segmentation de la population basée sur des critères de tranches d'âge. Il permet de générer des taux de montants sur le réseau de bus pour par la suite utiliser des données projetées de composition de la population pour effectuer des prévisions.

Le travail se termine sur l'énonciation d'enjeux pour de futurs travaux et de recommandations sur l'utilisation des modèles calibrés sur le réseau de la STO.

ABSTRACT

Understanding the evolution of ridership on a transit system is an important feature in the management of a public transit company. Having a forecast of future ridership on a transit system allows to anticipate changes and thus to adapt the level of supply to the demand. The Société de transport de l'Outaouais (STO - Gatineau, Quebec, CA) has expressed the need to forecast ridership on its network. The objective of this paper is to understand and estimate the evolution of the bus network ridership in the medium term. Without a ridership forecasting model, it is difficult to identify and primarily quantify the effect of factors influencing demand. Therefore, no strategic decision can be made with certainty. The main problem concerns the anticipation of the acquisition of material and human resources. A longitudinal forecasting model of the transport network's use is therefore implemented in this research.

The development of a longitudinal forecasting model suited for the study area is determined by comparing several models to determine the most relevant one. The study is organized around several steps. The literature review identifies a panel of potential models. The retrospective analysis of transit demand and indicators of the Gatineau region confirms which forecasting models are to be estimated and the explanatory factors to be used. These models are then implemented at three different levels of aggregation: the overall network, the municipal sectors and a specific level of aggregation determined to meet the project objectives. The performances of the forecasting models are then compared to determine the most appropriate for the network in question.

The literature review is the first step of any research work because it allows acquiring additional knowledge on the subject of study but mainly to assess previous works and to identify possible contributions. This part leads to the identification of potentially relevant models to fulfill the objectives. The models used are mainly autoregressive models, linear regressions on time series and segmentation of effects. It also identifies a lack of work in the literature aiming at comparing the performance of several models on the same study area. The methodology is justified by this gap and aims to compare different model structures in order to identify the strengths and weaknesses of each. This comparison leads to the identification of the most relevant model to meet the objectives.

Preliminary work is done before any experimentation. In order to capture changes in bus ridership at the local level on an everchanging bus network, the need to create small aggregation areas was

identified. The study at the stop level is not possible due to the constant modification of the stops creating a lack of temporal continuity of this element. A hexagonal mesh is developed to aggregate all data used and to eliminate certain issues of the disaggregated data. The forecasts at this level of aggregation provide a picture of the bus network usage at the local level.

In order to determine a model that is adapted to the STO's territory, a retrospective analysis of the demand and various indicators that describe the region is conducted. The main sources of data are the STO's smart card data, the census and the OD study of the Gatineau region. Although a lack of longitudinal data was identified, the indicators extracted from those databases allowed for the identification of demand influencing factors. Building permit data and park-and-ride parking lots were identified as the most useful data for developing forecasting models. The retrospective analysis highlights the influence of multiple explanatory factors such as the employment rate, but the lack of data over time prevents their use in longitudinal models.

The experimental parts are therefore the following steps. Longitudinal ridership forecasting models are calibrated at three aggregation scales: the overall network, municipal sectors, and specific aggregation areas determined to address the study issues. This structure allows us to identify which parameters explain demand at the different scales. At the level of the overall network, socio-economic variables are relevant to obtaining an estimate of the trend described by network ridership. At the scale of the hexagonal aggregation zones, demand is more sensitive to local factors. The number of park-and-ride lots or the evolution of the number of dwellings in each zone allow capturing the evolution at the local scale. But socio-economic variables do not allow this at the neighborhood level.

For each of the data aggregation units, several models are developed according to the following reasoning: start with a simple model and make it more complex by taking into account the errors found in the previous models. This method allows improving the performance during the chapters by studying the residuals of the models. The ARIMA autoregressive model is the reference for all levels of aggregation. Using only historical demand to forecast future ridership allows producing forecasts assuming that no major disruptions affect network ridership. The study of the residuals allows us to identify the moments where the forecast diverges and thus to identify an explanatory parameter allowing taking into account this change in trend. These experiments lead to two forecasting models adapted to the STO network. These models both concern an aggregation of the

data according to the fine hexagonal zones. A linear regression model on time series, taking into account the history of demand, the evolution of the number of dwellings and data on park-and-ride lots, allows capturing the evolution of the territory through the few longitudinal data available. Another promising model is derived from a segmentation of the population based on age group criteria. It allows the generation of bus ridership rates and the subsequent use of projected population composition data for forecasting.

The work concludes with a statement of recommendations for future work and guidelines for the use of calibrated models on the bus network of the STO.

TABLE DES MATIÈRES

REMERCIEMENTS III

RÉSUMÉ IV

ABSTRACT VII

TABLE DES MATIÈRES X

LISTE DES TABLEAUX XIV

LISTE DES FIGURES XVI

LISTE DES SIGLES ET ABRÉVIATIONS XXIII

LISTE DES ANNEXES XXIV

CHAPITRE 1 INTRODUCTION 1

1.1 Problématiques et objectifs 2

1.2 Structure du mémoire 3

CHAPITRE 2 REVUE DE LA LITTÉRATURE 7

2.1 Modèles autorégressifs 7

2.2 Modèle de prévision direct de la demande 9

2.2.1 Caractérisation du modèle 9

2.2.2 Zone d'influence 11

2.3 Modèles d'élasticité linéaire 12

2.4 Autres types de modèles 14

2.4.1 Modèles d'activités 14

2.4.2 Modèles âge-période-cohorte 15

2.4.3 Modèles multiniveaux 16

2.4.4 Modèles de choix modal 16

2.4.5 Autres modèles 17

2.5 Synthèses des paramètres 17
2.5.1 Les variables explicatives 17
2.5.2 Synthèse des modèles et de leurs paramètres 19
CHAPITRE 3 MÉTHODOLOGIE 21
3.1 Description des données 22
3.1.1 Mise en contexte 22
3.1.2 Données utilisées 23
3.2 Méthodologie des modèles 26
3.2.1 ARIMA 27
3.2.2 Régression linéaire multiple 28
3.2.3 Regression linéaire sur séries temporelles 28
3.2.4 Régression linéaire avec correction de l'erreur ARIMA 30
3.2.5 Réseau de neurones autorégressif 30
3.2.6 Modèles multiniveaux 32
3.2.7 Modèle de génération de taux de montants 33
3.3 Travaux préliminaires sur les données 34
3.3.1 Variables des permis de construire 35
3.3.2 Création de zones d'agrégation 46
CHAPITRE 4 ANALYSE RÉTROSPECTIVE 64
4.1 Analyse rétrospective du territoire global 64
4.1.1 Analyse de corrélation des variables du rapport 65
4.1.2 Synthèse de l'évolution des variables caractérisant le territoire global 77
4.1.3 Étude du taux d'utilisation des TC 79
4.2 Analyse rétrospective : secteurs municipaux 83

4.2.1 Demande TC 84

4.2.2 Démographie et offre TC 86

4.3 Analyse rétrospective : agrégation au niveau des arrêts 87

4.3.1 Demande TC 88

4.3.2 Permis de construire résidentiels 92

4.3.3 Variables démographiques 93

4.4 Analyse rétrospective : agrégation en zones hexagonales 96

4.4.1 Demande TC 96

4.4.2 Logements 98

4.4.3 Analyse des parcs-o-bus 99

4.4.4 Étude des données OD 102

CHAPITRE 5 PRÉVISION DE LA DEMANDE : RÉSEAU GLOBAL 108

5.1 Modèle autorégressif 109

5.1.1 Méthodologie de création d'un modèle de séries temporelles 110

5.1.2 Détermination de l'horizon de prévision d'un modèle ARIMA 120

5.2 Modèle de réseau de neurones autorégressif 124

5.3 Modélisation d'une année par 3 mois 126

5.3.1 Prévision 128

5.4 Introduction de la population comme variable explicative 129

5.5 Modèle économétrique 133

5.5.1 Variables 133

5.5.2 Modèle 135

CHAPITRE 6 PRÉVOIR LA DEMANDE AU NIVEAU DES SECTEURS 140

6.1 Modèle ARIMA par secteur 141

6.2 Modèles multiniveaux 145
6.2.1 Le modèle « idéal » 145
6.2.2 Modèles multiniveaux 146
CHAPITRE 7 PRÉVISION SUR LES ZONES D'AGRÉGATION PRÉCISES 160
7.1 Modèles précédemment mis en place 161
7.1.1 Modèles autorégressifs 161
7.1.2 Intégration de la variable logement 166
7.1.3 Modèle sur 3 mois 167
7.2 Modèle de régression sur série temporelle 170
7.2.1 Stationnarité et indépendance 171
7.2.2 Détermination du retard à utiliser 173
7.3 Création d'une typologie de zones 178
7.3.1 Typologie basée sur la demande 178
7.3.2 Typologie basée sur le nombre de logements 182
7.3.3 Typologie basée sur les logements et les parcs-o-bus 186
7.4 Modèle spatial 188
7.5 Typologie de génération de déplacements 193
CHAPITRE 8 CONCLUSION (ET RECOMMANDATIONS) 198
8.1 Synthèse de la recherche 198
8.2 Contributions 200
8.3 Limites 201
8.4 Perspectives 203
RÉFÉRENCES 206
ANNEXES 209

LISTE DES TABLEAUX

Tableau 2-1 Synthèse des classes de modèles de la littérature 20

Tableau 3-1 Tableau traduit de United States Census Bureau (https://www.census.gov/construction/nrc/lengthoftime.html) 45

Tableau 3-2 Moyenne du délai séparant l'obtention du permis et la fin des travaux 45

Tableau 3-3 Résultats étude de variance 53

Tableau 4-1 Proposition de variables pour l'analyse de corrélation croisée 65

Tableau 4-2 Proposition de variables pour l'analyse de corrélation croisée 66

Tableau 5-1 Performances du modèle ARIMA 115

Tableau 5-2 Comparaison de l'effet de l'ensemble de données d'entraînement pour un modèle ARIMA 119

Tableau 5-3 Schématisation de la méthode de détermination de l'horizon de prévision 121

Tableau 5-4 Indicateurs pour la validation des modèles en fonction de leur horizon 122

Tableau 5-5 Comparaison des modèles NNAR et ARIMA 125

Tableau 5-6 Sélection des mois les plus achalandés pour la modélisation 127

Tableau 5-7 Résultats de la comparaison de l'année type d'automne avec l'année complète 129

Tableau 5-8 Comparaison de la finesse des données et de la population de référence 131

Tableau 5-9 Performance des modèles de régression 138

Tableau 6-1 Résultats des erreurs pour les modèles ARIMA par secteur 142

Tableau 6-2 Indicateurs de performance du modèle multiniveau 154

Tableau 6-3 Classes du modèle de classes latentes 159

Tableau 7-1 Échantillon des tests réalisés sur les séries différenciées d'ordre 2 173

Tableau 7-2 Logements selon la typologie basée sur la demande 181

Tableau 7-3 Composition de la typologie de logements 182

Tableau 7-4 Résumé de la significativité des variables pour chaque modèle 192

Tableau 7-5 Légende tableau 7-4 192

Tableau 7-6 Centroïdes typologie 1 194

Tableau 7-7 Centroïdes typologie 2 194

Tableau 7-8 Taux de génération de montants moyens pour les deux typologies 196

Tableau 8-1 Synthèse de la performance des modèles 200

Tableau 0-1 Analyse de corrélation méthode de la courbe de tendance 210

Tableau 0-2 Analyse de corrélation méthode interpolation linéaire 211

Tableau 0-3 Analyse de corrélation valeurs réelles 212

LISTE DES FIGURES

Figure 1-1 Structure du mémoire 4

Figure 3-1 Carte réseau STO 23

Figure 3-2 Schéma d'un réseau de neurones 31

Figure 3-3 Test d'autocorrélation de la demande basé sur une unité annuelle 37

Figure 3-4 Test d'autocorrélation sur les données de permis de construire basé sur une unité annuelle 38

Figure 3-5 Étude du retard de la variable de permis de construire 39

Figure 3-6 Test d'autocorrélation sur les données de permis de construire basé sur une unité annuelle (1 à 4 logements par bâtiment) 40

Figure 3-7 st d'autocorrélation sur les données de permis de construire basé sur une unité annuelle (5 et plus logements par bâtiment) 40

Figure 3-8 Étude de corrélation pour les bâtiments de 1 à 4 logements 41

Figure 3-9 Étude de corrélation pour les bâtiments de 5 logements et plus 41

Figure 3-10 Étude de corrélation entre la demande et différents retards du cumul des permis des logements autorisés 42

Figure 3-11 Pourcentage de création et suppression d'arrêts 47

Figure 3-12 Nouveaux arrêts sans les arrêts déplacés 48

Figure 3-13 Schéma de l'implantation des nouveaux permis de construire 48

Figure 3-14 Moyenne du LISA moyen par taille de maillage 55

Figure 3-15 Maximum des LISA moyen par taille de maillage 56

Figure 3-16 Moyenne du 1er quartile du LISA par taille de maillage 57

Figure 3-17 Maximum du 1er quartile du LISA par taille de maillage 57

Figure 3-18 Valeurs du LISA moyen interpolées en fonction de la position du maillage 59

Figure 3-19 Exemple d'un maillage peu pertinent 60

Figure 3-20 LISA du maillage le plus pertinent pour les années 2016-2019 61

Figure 3-21 LISA des incréments de validations en pointe AM pour le mois de novembre 2019 agrégé selon les îlots de diffusion 62

Figure 3-22 LISA des incréments de validations en pointe AM pour le mois de novembre 2019 agrégé selon les aires de diffusion 63

Figure 4-1 Disponibilité des données dans le temps 67

Figure 4-2 Véhicule km pour les TC 69

Figure 4-3 Validations carte à puce 69

Figure 4-4 Passage arrêt TC 69

Figure 4-5 Véhicules par permis 69

Figure 4-6 Nombre de déplacements total pour motif travail 70

Figure 4-7 Durée moyen de l'activité travail pour les travailleurs 70

Figure 4-8 Nombre de personnes âgées 70

Figure 4-9 Nombre de personnes 70

Figure 4-10 Part modale TC 70

Figure 4-11 Nombre de travailleurs 70

Figure 4-12 Nombre de véhicules 71

Figure 4-13 Longueur du réseau routier 71

Figure 4-14 Pourcentage de logements unifamiliaux 71

Figure 4-15 Densité d'unités de logements 71

Figure 4-16 Nombre d'unités de logements 71

Figure 4-17 Prix du litre d'essence en cent 71

Figure 4-18 Durée moyenne de l'activité travail pour tous les déplacements travail 72

Figure 4-19 Schéma des méthodes de projection 73

Figure 4-20 Évolution des variables associées à l'accroissement de la population 75

Figure 4-21 Évolution de la part modale TC et des validations de CAP 76

Figure 4-22 Évolution des validations de CAP par personne et des personnes âgées 76

Figure 4-23 Taux de croissance annuelle moyen des variables du rapport 79

Figure 4-24 Évolution du taux moyen de fréquentation des TC 80

Figure 4-25 Distribution temporelle des validations de CAP du réseau d'Ottawa 81

Figure 4-26 Distribution spatiale des validations de CAP du réseau d'Ottawa 82

Figure 4-27 Taux de déplacement TC par segment de population 83

Figure 4-28 Montant par secteurs en pointe AM 84

Figure 4-29 Montants par secteur en pointe PM 85

Figure 4-30 Pourcentage de validations de CAP en pointe du matin 85

Figure 4-31 Densité de population par secteur 86

Figure 4-32 Indicateur de l'offre en TC par secteur 86

Figure 4-33 Taux de génération de validations de CAP par secteur 87

Figure 4-34 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2013 88

Figure 4-35 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2014 89

Figure 4-36 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2015 89

Figure 4-38 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2016 90

Figure 4-37 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2017 90

Figure 4-39 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2019 91

Figure 4-40 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2018 91

Figure 4-41 Permis de construire autorisés entre 2009 et 2019 93

Figure 4-42 Taux d'activité du recensement de 2016 94

Figure 4-43 Densité de population en 2016 (recensement) 95

Figure 4-44 Moyenne des montants en pointe AM pour les mois de novembre des années 2012-2019 97

Figure 4-45 Nombre de logements par zone pour novembre 2012 98

Figure 4-46 Nombre de logements par zone pour novembre 2019 98

Figure 4-47 Positionnement des parcs-o-bus en 2019 99

Figure 4-48 Capacité de parcs-o-bus par zone hexagonales en 2019 100

Figure 4-49 Utilisation des parcs-o-bus en 2019 selon les relevés de plaques 101

Figure 4-50 Rayon d'attraction moyen des parcs-o-bus en 2019 101

Figure 4-51 Étude des 100 trajets référencés les plus empruntés pour un jour ouvrable moyen de novembre 2019 103

Figure 4-52 Matrice OD d'un jour ouvrable moyen de novembre 2019 104

Figure 4-53 Matrice OD d'un jour ouvrable moyen de novembre 2018 104

Figure 4-54 Matrice OD de la pointe AM d'un jour ouvrable moyen de novembre 2019 105

Figure 4-55 Matrice OD de la pointe AM d'un jour ouvrable moyen de novembre 2018 106

Figure 4-56 Matrice OD de la pointe PM d'un jour ouvrable moyen de novembre 2019 106

Figure 5-1 Décomposition additive des validations CAP mensuelles 112

Figure 5-2 Analyse d'autocorrélation des validations de CAP 112

Figure 5-3 Résultats prévision ARIMA saisonnier 114

Figure 5-4 Prévision validations CAP sur 1 an du modèle ARIMA 115

Figure 5-5 Résidus du modèle ARIMA des validations de CAP 116

Figure 5-6 Prévision des CAP pour pointe AM d'un jour ouvré moyen mensuel à 5 ans..... 117

Figure 5-7 Prévision des CAP pour pointe AM d'un jour ouvré moyen mensuel à 5ans calibré sans régime transitoire 118

Figure 5-8 Modèle ARIMA validations CAP mensuelles 123

Figure 5-9 Modèle ARIMA à un horizon de 2 ans 124

Figure 5-10 Validation des modèles NNAR manuel et ARIMA 126

Figure 5-11 Prévision ARIMA pointe AM moyenne d'un jour ouvré 128

Figure 5-12 Évolution de la population de la RMR O-G (rive Québec) 130

Figure 5-13 Régressions avec correction ARIMA pour les validations de CAP en pointe d'un jour ouvrable moyen..... 131

Figure 5-14 Étude des résidus du modèle de régression avec correction de l'erreur ARIMA pour les validations moyennes d'un jour ouvré 133

Figure 5-15 Évolution des variables socio-économiques 134

Figure 5-16 Corrélogramme des variables trouvées pour la RMR O-G (rive QC) 135

Figure 5-17 Prévision régression linéaire 137

Figure 5-18 Prévision régression linéaire sur séries temporelles..... 137

Figure 6-1 Secteurs municipaux de Gatineau 140

Figure 6-2 Modèle ARIMA pour CAP d'un jour ouvré moyen par mois pour le centre-ville d'Ottawa 143

Figure 6-3 Modèle ARIMA pour CAP d'un jour ouvré moyen par mois pour Hull..... 143

Figure 6-4 Comparaison du modèle combiné par secteur 144

Figure 6-5 Schéma d'un modèle idéal..... 145

Figure 6-6 Résultat du modèle de moyenne inconditionnelle dans R 148

Figure 6-7 Résultats du modèle inconditionnel de croissance..... 150

Figure 6-8 Résultats du modèle classique..... 151

Figure 6-9 Résultats du modèle multiniveau 153

Figure 6-10 Modélisation multiniveaux sur le secteur Aylmer 154

Figure 6-11 Modélisation multiniveaux sur le secteur BMA 155

Figure 6-12 Modélisation multiniveaux sur le secteur Gatineau 155

Figure 6-13 Modélisation multiniveaux sur le secteur Hull 156

Figure 7-1 Pourcentage d'erreur quadratique moyen pour le modèle sur les jours ouvrés 161

Figure 7-2 Pourcentage d'erreur quadratique moyen pour le modèle sur les pointes AM 162

Figure 7-3 Erreur quadratique moyenne pour le modèle sur les jours ouvrés 162

Figure 7-4 Erreur quadratique moyenne pour le modèle sur les pointes AM 163

Figure 7-5 Erreurs du modèle ARIMA, jour ouvrable moyen de novembre 2017 164

Figure 7-6 Erreurs du modèle ARIMA, jour ouvrable moyen de novembre 2018 165

Figure 7-7 Erreurs du modèle ARIMA, jour ouvrable moyen de novembre 2019 165

Figure 7-8 Résidus des mois de novembre pour le modèle de régression linéaire avec correction de l'erreur ARIMA 167

Figure 7-9 Comparaison des deux modèles pour novembre 2017 168

Figure 7-10 Comparaison des deux modèles pour novembre 2018 168

Figure 7-11 Comparaison des deux modèles pour novembre 2019 169

Figure 7-12 Distributions des erreurs pour le modèle sur 3 mois en fonction des années 170

Figure 7-13 BICs du modèle pour un retard de 12mois 174

Figure 7-14 BICs du modèle pour un retard de 12 et 24 mois 174

Figure 7-15 BICs du modèle pour un retard de 1 et 12 mois 175

Figure 7-16 Résultat du modèle de régression sur séries temporelles 176

Figure 7-17 Exemple des résidus d'une zone 177

Figure 7-18 Exemple de prévision sur une zone 177

Figure 7-19 Réparation spatiale des types de la typologie basées sur la demande 179

Figure 7-20 Évolution de la fréquentation TC en pointe AM pour chaque segment de la typologie 180

Figure 7-21 Répartition spatiale des segments de la typologie..... 181

Figure 7-22 Évolution temporelle du taux de génération de montants moyen pour chaque type 183

Figure 7-23 Évolution temporelle de l'écart-type du taux de génération de montants moyen pour chaque type 183

Figure 7-24 Erreurs projection typologie de logements 2019 184

Figure 7-25 Erreurs projection typologie de logements 2018 185

Figure 7-26 Erreurs projection typologie de logements 2017 185

Figure 7-27 Erreurs du modèle pour une projection en novembre 2017 187

Figure 7-28 Erreurs du modèle pour une projection en novembre 2018 187

Figure 7-29 Erreurs du modèle pour une projection en novembre 2019 188

Figure 7-30 Coefficient de corrélation linéaire des variables explicatives considérées (Pearson) 190

Figure 7-31 Distribution du taux de montants par habitant 191

Figure 7-32 Distribution des taux de montants pour la typologie 1 195

Figure 7-33 Distribution des taux de montants pour la typologie 2 195

Figure 7-34 Erreurs de la modélisation pour la typologie 1 196

LISTE DES SIGLES ET ABRÉVIATIONS

AM Ante Meridiem (matin)

AP Auto-Passager

ARIMA Auto Regressive Integrated Moving Average

BMA Buckingham Masson-Angers

DRM Direct Ridership Model

GARCH Generalized Auto Regressive Conditional Heteroskedasticity

GTFS General Transit Feed Specification

LISA Local Indicator of Spatial Autocorrelation

OD Origine-destination

PM Post Meridiem (après-midi)

RMR Région métropolitaine de recensement

SLR Système Léger sur Rail

SRB Système de Bus Rapide

STO Société des Transport de L'Outaouais

TC Transport en commun

LISTE DES ANNEXES

Annexe A Démonstration du taux de croissance moyen 209

Annexe B Tableaux analyse de corrélation 210

Annexe C Fiches analyses descriptives des secteurs 213

CHAPITRE 1 INTRODUCTION

Les transports en commun sont aujourd'hui considérés comme inhérents à toute ville développée. Les réseaux de transport collectifs sont présents dans le monde entier sous différentes formes. Ils sont pour beaucoup un moyen de transport essentiel. Les transports publics considérés dans beaucoup de pays comme acquis nécessitent une gestion rigoureuse pour s'adapter aux besoins des usagers. Face à une clientèle en constante évolution, adapter l'offre à la demande est une tâche à ne pas négliger pour maintenir la performance d'un réseau de transport. L'augmentation constante de la population ou les changements d'habitudes de mobilité mettent une pression sur le transport collectif. Anticiper ces changements d'achalandage est une préoccupation majeure des entreprises de transport public. Que ce soit à des fins de planification stratégique ou dans l'optique de gérer les acquisitions de ressources, être capable d'anticiper les évolutions de la demande est primordial.

Un grand nombre de chercheurs se sont penchés sur cette question : comment anticiper la demande sur un réseau de TC. Prévoir la demande est un terme vaste qui comprend plusieurs types de prévision. Cela peut concerner de la prévision spatiale en anticipant par exemple le nombre d'usagers pouvant être captés par l'implantation d'un nouvel arrêt. Dans ce cas, des modèles spatiaux sont déployés pour tenter d'identifier la demande latente. D'autres modèles ont pour but d'avoir une image de la demande dans le temps pour anticiper les évolutions de la fréquentation. Prévoir la demande à un certain horizon nécessite l'utilisation de données longitudinales. Ces modèles se basent sur un principe simple : étudier les variations passées de l'achalandage en lien avec des paramètres explicatifs pour par la suite projeter les dynamiques observées afin de produire des résultats. Derrière un principe simple en apparence, un grand nombre de modèles variant en complexité existent, mais aucun ne peut répondre à tous les cas de figure. La littérature a montré des performances très hétérogènes pour même modèle, mais sur plusieurs réseaux de transport en commun. Adapter un modèle aux caractéristiques d'un réseau est l'enjeu que se voit confier une société de transport en commun voulant anticiper l'évolution de la fréquentation de son réseau.

La Société de Transport de l'Outaouais (STO), entreprise de transport en commun de la ville de Gatineau au Québec, ne fait pas exception et souhaite anticiper la fréquentation de son réseau de bus. Trouver le modèle qui correspond à ce cas de figure est donc nécessaire.

1.1 Problématiques et objectifs

Les évolutions de la fréquentation d'un réseau sont inhérentes à la vie d'une société de transport. L'enjeu est de savoir comment gérer ces fluctuations pour adapter au mieux l'offre à la demande. Une augmentation de la fréquentation est souvent un bon signe pour une société de transport en commun. Cela peut signifier une évolution des habitudes de mobilité des usagers ou encore que le réseau devient de plus en plus performant. Malgré, une signification positive d'une telle évolution, les enjeux liés à un accroissement de l'achalandage sont multiples. Ne pas être capable d'anticiper ces variations peut être néfaste pour la société de transport : offre non adéquate ou encore surpopulation dans le réseau.

Les données issues des comptages de cartes à puce montrent une augmentation majeure de la demande en transport : presque 15% d'augmentation entre 2016 et 2018. Mais il n'était pas possible de déterminer précisément l'ampleur de cette hausse de l'achalandage ni d'en identifier les causes et de la quantifier. Sans information sur les facteurs ayant déclenché cette augmentation de la fréquentation, il est très difficile de déterminer si cette tendance est susceptible de se maintenir ou non et pendant combien de temps. Face à une telle modification de la demande, un réseau de transport peut rapidement être surchargé causant des problèmes au niveau du confort, de la sécurité et de la satisfaction de la clientèle.

Des mesures ont été mises en place pour augmenter l'offre de transport au sein de la STO. L'ajout de service a permis d'absorber une partie de cette hausse. Mais une telle modification se fait toujours dans la limite des ressources matérielles et humaines disponibles. Sans prévision de la future demande, il est compliqué de prendre des décisions sur l'acquisition de nouvelles ressources. Qu'il s'agisse d'acheter de nouveaux bus ou de recruter de nouvelles personnes, ces processus prennent du temps et nécessitent d'être dimensionnés. Compte tenu des défis auxquels la STO fait face en regard de la transformation de la demande, l'estimation du poids des facteurs endogènes et exogènes influençant la demande est primordiale pour pouvoir offrir un service répondant pleinement aux besoins des usagers. Le dimensionnement de cette offre passe principalement par l'anticipation des acquisitions de ressources matérielles et humaines.

L'objectif de ce projet est d'élaborer un modèle permettant de prévoir la fréquentation du réseau de bus de la STO. Identifier un ou plusieurs modèles de prévision permettant d'estimer les variations de l'achalandage à un horizon de 5 ans est l'enjeu principal de ce projet. Il doit prendre

en compte l'évolution des différents indicateurs, comme ceux socioéconomiques, démographiques, d'occupation du territoire et toute autre donnée jugée pertinente à l'explication de la demande, pour fournir des prévisions.

Ces prévisions auront pour but d'assister la STO dans deux tâches majeures :

- Adapter au mieux l'offre à la demande
- Planifier la gestion et l'acquisition des ressources

Ce mémoire de maîtrise tente de répondre à ces attentes par l'acquisition d'une connaissance sur les facteurs influençant les variations de la demande et le développement d'un modèle permettant d'anticiper ces variations. L'important est également de pouvoir quantifier et situer dans le temps et dans l'espace les variations de la fréquentation du réseau de transport en commun.

1.2 Structure du mémoire

Pour simplifier la compréhension de la structure du mémoire, la figure ci-dessous illustre le déroulement des étapes principales. La figure est explicitée et commentée par la suite.

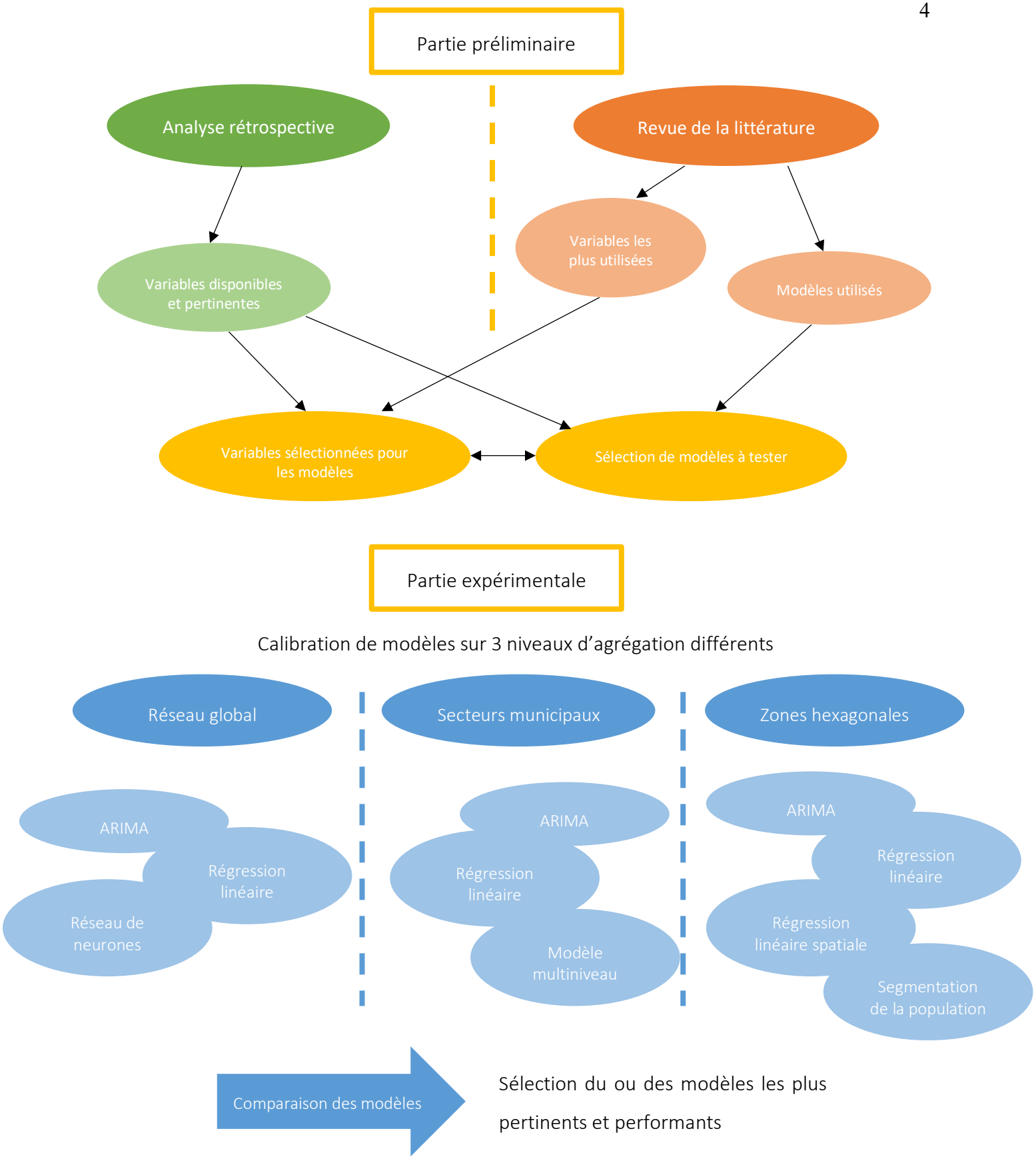


Figure 1-1 Structure du mémoire

La figure précédente décrit la structure générale du mémoire. L'enjeu est de réaliser deux étapes préliminaires afin de déterminer les modèles et les variables à tester dans les modèles de prévision. La revue de la littérature et l'analyse rétrospectives constituent ces deux étapes.

La revue de la littérature a comme but de dresser un catalogue des modèles les plus communément utilisés dans le milieu des transports ainsi que les variables explicatives utilisées. D'autres domaines sont également étudiés afin de multiplier les approches différentes pour effectuer un modèle de prévision. Ce catalogue de modèles et de variables est par la suite utilisé pour orienter la suite du projet. Les modèles les plus pertinents pour le territoire de Gatineau sont sélectionnés ainsi que les variables qu'ils impliquent. Ces modèles seront testés dans la phase expérimentale du projet.

Ayant une connaissance du domaine dans lequel s'inscrit le projet, il est nécessaire de prendre par la suite en compte les caractéristiques du territoire d'étude. La revue de la littérature montrera que les modèles à employer dépendent fortement de la zone considérée. L'analyse rétrospective est donc la première étape permettant d'adapter l'approche de modélisation à la ville de Gatineau. Cette partie a pour but de déterminer les dynamiques qui régissent l'évolution de la demande dans la région. Les grandes transformations subies par la région sont étudiées et leurs impacts sur les habitudes d'usage des TC sont constatés. L'objectif est d'avoir une vue globale de la région et de déterminer quels sont les facteurs qui impactent la demande. Les interactions entre ces facteurs explicatifs sont également étudiées afin de juger s'ils peuvent par la suite être intégrés à des modèles prévisionnels.

En combinant les résultats de la revue de la littérature et de l'analyse rétrospective, il est possible de déterminer quels modèles peuvent potentiellement être pertinents pour la ville de Gatineau ainsi que les variables explicatives qui leur sont associées. Ces modèles sont testés par la suite.

Les chapitres qui suivent concernent l'expérimentation des modèles de prévision de la demande. L'organisation suit la logique suivante, un chapitre concerne chaque niveau d'agrégation : réseau global de la STO, secteurs municipaux et zones d'agrégation hexagonales. Chacun des trois chapitres d'expérimentation de modèles de prévision de la demande sera constitué du calage de différents modèles. Les modèles gagneront en complexité au cours du chapitre pour permettre à un modèle donné de s'affranchir au maximum des erreurs des modèles qui le précèdent. Cela peut passer par l'ajout de variables ou encore du changement de la structure du modèle utilisé. Chaque

modèle sera comparé aux autres dans l'optique de déterminer le plus pertinent pour l'étude en cours.

Finalement, la conclusion déterminera le ou les modèles les plus adaptés pour répondre aux besoins du projet. Des pistes d'amélioration seront également données pour permettre d'améliorer d'autant plus la performance du modèle.

CHAPITRE 2 REVUE DE LA LITTÉRATURE

Ce chapitre présente une synthèse d'informations recueillies dans des publications scientifiques en lien avec le sujet de ce mémoire. La revue de la littérature permet de situer ce projet de mémoire parmi les travaux existants. Ainsi, il est possible d'orienter le projet pour avoir un apport nouveau au domaine des transports. Les connaissances acquises dans ce chapitre sont mises à profit dans l'ensemble du mémoire.

La documentation concernant les modèles de prévision de la demande en transport est abondante. Une multitude de modèles sont décrits au fil des articles scientifiques allant de modèles simples à des modèles très complexes. Mais peu d'articles permettent réellement de comparer ces différents modèles. Dans la mesure où l'objectif de ce mémoire est de déterminer un modèle de prévision pertinent pour le réseau de la STO, il faut être capable de comparer des modèles.

Les modèles décrits dans la littérature ne sont pas comparables, car ils concernent souvent des zones géographiques différentes ou sont estimés sur des périodes différentes. En consultant la littérature de ce domaine, il n'est donc pas possible de connaître de manière certaine quel modèle est le plus approprié pour l'étude présente. D'autant plus que des modèles très complexes sont développés sans justifier le gain en performance par rapport à des modèles « simples ».

La revue de la littérature s'organise autour de l'étude de plusieurs types de modèles présents dans la littérature et pouvant être pertinents pour le territoire d'étude. À savoir que les modèles du type traditionnel à quatre étapes (Bonnel, 2004) ne sont pas considérés dans cette étude, car ils sont jugés trop complexes pour le cas présent. Cette méthode permet la modélisation de la demande de l'ensemble des modes de transport présents sur un territoire. Ce modèle n'est pas adapté aux enjeux du projet. Des approches modernes et plus simples sont privilégiées. Pour chaque type de modèle, les variables dépendantes et indépendantes sont étudiées ainsi que les principales caractéristiques du modèle.

2.1 Modèles autorégressifs

Les modèles autorégressifs sont une catégorie de modèle prenant uniquement comme variables explicatives l'historique de la variable dépendante. Dans le cas présent, un tel modèle utiliserait l'historique de la demande pour prévoir la demande future. Ces modèles sont généralement utilisés dans le cas ou très peu de données sont disponibles.

Le modèle ARIMA (*Auto Regressive Integrated Moving Average*) (Ding, Duan, Zhang, Wu, & Yu., 2017) est le modèle autorégressif le plus répandu dans la prévision de la demande en transport et est également considéré comme le plus performant devant des modèles de type GARCH (*Generalized AutoRegressive Conditional Heteroskedasticity*). Il est généralement utilisé pour de la prévision à court terme. Dans le cas de Ding (2017), il est appliqué aux données de carte à puce pour prévoir l'évolution de la demande au cours d'une journée dans une station. Les intervalles de prévision sont par conséquent courts et permettent donc une image précise de la fréquentation des réseaux. La force de ce type de modèle est sa simplicité, sa versatilité et le fait qu'il produit des résultats souvent très précis à court terme. Il est versatile dans la mesure où il peut être appliqué à la prévision sur presque tous les intervalles de temps possibles. Le modèle peut être journalier ou annuel et, tant que la tendance se maintient ou qu'aucun évènement majeur ne vient perturber le réseau de transport, il offrira des résultats précis sur un certain nombre de pas de temps.

L'inconvénient d'un tel modèle est que cela ne tient pas compte des perturbations extérieures. Un modèle ARIMA a souvent tendance à dériver au cours du temps ou à être entièrement biaisé en cas de perturbations majeures. C'est pour cette raison que des modèles autorégressifs prenant en compte des évènements extérieurs ont été développés (Abadi, Rajabioun, & Ioannou, 2015). L'objectif est de communiquer au modèle autorégressif qu'un évènement perturbateur est arrivé à un moment précis et lui permettre de se calibrer de nouveau. Dans son article, Abadi (2015) applique cette méthode aux données de comptage de véhicules. Il est donc capable de prévoir le nombre de véhicules empruntant un axe sur un horizon de 30 minutes tout en prenant en compte des évènements perturbateurs comme des accidents, des fermetures de routes ou encore des évènements publics. Les résultats affichent de bonnes précisions, 2% à 5 minutes et 12% à 30 minutes.

L'inconvénient constitué par la non-prise en compte des perturbations extérieures peut s'avérer être une force. Les résidus d'un tel modèle constituent donc une image de ces perturbations qui ne s'inscrivent pas dans une tendance générale. Leur étude peut donc permettre l'identification de facteurs modifiant la tendance de la variable dépendante.

2.2 Modèle de prévision direct de la demande

2.2.1 Caractérisation du modèle

Les modèles de prévision directe de la demande sont les plus abondants dans la récente littérature. Ils permettent la prévision de la demande aux stations d'un réseau de transport en se basant sur les caractéristiques de l'environnement des stations (population, emplois, etc.), de l'occupation des sols et de l'offre de transport. Toutes ces variables sont calculées sur une zone décrivant le voisinage de la station de transport en question.

Une des principales caractéristiques de ce type de modèle est de pouvoir différencier le nombre de montants ou descendants aux multiples stations sur un réseau de transport. L'utilisation des données de points d'intérêts (Sun, Wang, Yao, Rong, & Ma, 2016) (Cardozo, García-Palomares, & Gutiérrez, 2012) permet par exemple de différencier les stations et ainsi, grâce aux données de cartes à puce, permettre de prévoir le nombre de montants sur une journée pour les stations hors de l'échantillon de calibration. Zhang & Wang, 2014, font appels à l'interpolation de krigeage pour déterminer les montants à une nouvelle station de transport en commun. L'interpolation se base sur la distance réseau. Mais un tel modèle qui permet de faire de la prévision spatiale n'est pas pertinent pour l'étude présente.

La plupart des modèles de prévision directe de la demande permettent toutefois de considérer cet aspect de projection temporelle. Frei & Mahmassani, 2013 introduisent un indicateur de la période de la journée pour obtenir le nombre de montants aux stations par tranche de 30 minutes. Le modèle utilisé est un modèle de régression logarithmique sur des variables telles que le nombre de stations de métro à moins de 100m, le nombre de personnes de moins de 18ans, etc. L'utilisation d'un tel modèle permet de produire des prévisions pertinentes à l'horizon d'une journée sur les 11000 arrêts de bus qui compose le réseau de Chicago. Durning & Townsend, 2015 réalisent un modèle similaire en ajoutant un indice d'utilisation du territoire. Il permet d'intégrer plusieurs variables qui pourraient être initialement corrélées.

L'horizon d'une journée ne correspond pas non plus aux exigences du projet. Mucci & Erhardt, 2018 étudient la possibilité d'utiliser ce type de modèle à des fins de prévision à moyen terme. L'étude est réalisée sur le SLR et le SRB de San Francisco. Leur objectif est de prévoir les montants et descendants journaliers moyens en 2009 et 2016, en calibrant le modèle sur des années

antérieures. Dans ce cas les variables qui pouvaient dans les modèles évoqués précédemment qualifier les différences spatiales prennent ici en compte également l'évolution temporelle. Ces variables sont constituées de donnée telles que le recensement américain, le coût de stationnement, la fréquentation des réseaux voisins ou encore les caractéristiques des lignes. Ces variables sont combinées dans une régression linéaire multiple pour produire les résultats de montants et de descendants. Les résultats montrent une précision similaire pour 2009 et 2016 avec des valeurs qu'ils qualifient d'encourageantes. Il est tout de même stipulé que ce modèle n'est pas pertinent en cas de changements majeurs du réseau.

Les modèles de prévision directe de la demande exposés jusqu'à maintenant sont structurés autour de régressions linéaires pouvant faire appel à des transformations diverses. Ce type de modèle qualifie uniquement le fait de prévoir le nombre de montants ou de descendants aux stations d'un réseau en prenant en compte les caractéristiques d'une zone d'influence déterminée autour de la station en question. Ainsi d'autres structures sont présentes dans la littérature et notamment en utilisant des algorithmes issus de l'intelligence artificielle tels que les arbres de décision (Li, Yue Liu, Gao, & Liu, 2016) ou encore des réseaux de neurones (Baek & Sohn, 2017). Ces modèles sont généralement basés uniquement sur le traitement de données et ne prennent que rarement en compte les limitations qui pourraient être dictées par le domaine des transports. De plus, les modèles d'intelligence artificielle rendent souvent difficile l'étude du poids des différentes variables explicatives sur la variable dépendante.

Le modèle réalisé par Ding, 2016, s'appuie sur une structure d'arbre de décision à gradient boosté. Ce modèle permet de s'affranchir de cette problématique rendant la plupart des modèles d'intelligence artificielle similaires à des « boîtes noires ». Cette structure innovante permet de quantifier et classer les poids relatifs des variables explicatives contrairement aux autres modèles faisant appel à l'intelligence artificielle. Cette information est primordiale dans un processus de prévision et également pour permettre d'orienter des décisions stratégiques sur le développement d'un réseau de transport. Ce modèle développé sur le réseau de transport en commun de Beijing a la particularité d'utiliser des données sur l'historique de la demande comme variables explicatives. C'est une caractéristique qui n'avait pas été constatée sur les autres modèles de prévision directe de la demande. La prévision est à court terme et prévoit le nombre de montants aux stations de métro sur des intervalles de 15 minutes en fonction des montants à t-1 et t-2, les descendants aux stations de bus à proximité à t-1 et t-2 ainsi que des variables temporelles pour décrire la plage de

temps qui est prédite. Cette approche produit des résultats plus précis que tous les autres modèles utilisant une structure basée sur l'intelligence artificielle. À savoir que ces résultats sont à prendre avec du recul, car ces comparaisons ne sont pas faites sur des modèles utilisant les mêmes données expérimentales.

2.2.2 Zone d'influence

Une discussion revient dans une multitude d'articles scientifiques sur les modèles de prévision directe de la demande : c'est la définition de la zone d'influence au niveau de chaque arrêt. La prévision du nombre de montants à un arrêt se base sur des variables calculées dans un voisinage défini de l'arrêt en question. La définition de cette zone dite d'influence est une caractéristique majeure dans création d'un tel modèle. Bien que les résultats qui sont exposés dans cette partie concernent les modèles de prévision directe de la demande, ils sont pertinents pour d'autres modèles dans la mesure où la caractéristique observée est la zone influençant significativement la demande à un arrêt. Même dans le cas de modèles faisant l'usage d'agrégation spatiale, il est possible d'en déduire des zones tampons qui peuvent faire varier la demande d'une zone voisine.

Plusieurs manières de qualifier cette zone d'influence sont étudiées dans la littérature. La première, et la plus simple, est l'utilisation d'une zone circulaire déterminée par un rayon à vol d'oiseau autour de chaque station du réseau de TC. C'est également la méthode la plus répandue. Il est souvent constaté qu'une valeur ne peut pas être prise pour tous les cas de figure et notamment cette valeur de rayon varie en fonction des modes de transport étudiés. Pour un même mode de transport, Li & Yang 2019 montrent que le rayon de la zone d'influence peut varier entre 600 et 1200 mètres pour un même mode, ici le métro, mais pour des villes différentes. Des résultats sont disponibles en grand nombre sur des réseaux de trains avec en moyenne des rayons de 500 mètres (He, Zhao, & Tsui, 2020). Pour les réseaux de bus, moins de valeurs sont disponibles, mais Pulugurtha & Agurla, 2012 en étudiant le réseau de bus de Charlotte en Caroline du Nord, déterminent que pour les zones d'influence à vol d'oiseau le rayon le plus pertinent est de 0.25 mile soit environ 400 mètres. L'intérêt majeur de cet article est qu'il compare cette structure de zone d'influence à une autre : la pondération des variables en fonction de la distance à l'arrêt.

Caractériser les variables explicatives par une pondération décroissante en fonction de la distance permet de tenir compte de l'effet non constant dans une zone d'influence. Prenons l'exemple de la population, la personne à forte proximité de l'arrêt est plus susceptible d'utiliser le réseau que les

personnes plus éloignées. Le poids de la population fortement rapprochée de l'arrêt est donc plus important que celui de la population éloignée. Une telle pondération tente donc de tenir compte de ce phénomène. Pulugurtha & Agurla, 2012 déterminent que la pondération décroissante en $1/D^2$ où D représente la distance à l'arrêt, fournit de meilleurs résultats que la zone d'influence définie par un rayon à vol d'oiseau et sans pondération. D'autres pondérations de ce type sont également utilisées pour effectuer des prévisions. C'est le cas de Sun, 2016, qui définit trois zones autour des stations. Une première qui est appelée la zone piétonne caractérisée en tant que zone où les usagers se rendent à la station à pied. Une deuxième zone appelée zone réseau définie par l'accès aux stations à vélo ou via le réseau de bus. La dernière zone nommée zone potentielle qui est définie par les limites des polygones de Thiessen autour de chaque station. L'usage de cette géométrie permet de signifier que la station est la plus proche pour tous les habitants de cette zone, mais que l'accessibilité est limitée. C'est trois zones sont donc pondérées en accord avec leur fonction pour calculer les variables explicatives du modèle. Cette hiérarchie très imagée permet donc d'adapter différemment la pondération de chaque variable en fonction de leur relation avec les différentes zones. Cette méthode apporte des résultats satisfaisants et permet de ne pas introduire un biais lié à une géométrie fixe comme les zones d'influence circulaires.

Comme il a été mentionné au début de cette partie, ces résultats sont également intéressants pour la création d'autres modèles, car ils permettent de quantifier l'effet du voisinage d'une station de transport en commun sur la demande en son sein.

2.3 Modèles d'élasticité linéaire

Les modèles d'élasticité sont basés sur le principe que la demande évolue d'une certaine manière à des variations de variables explicatives. Les modèles linéaires utilisent l'évolution linéaire de la demande en réponse à l'évolution d'un paramètre et il est possible de quantifier cette élasticité. Les modèles de prévision directe de la demande pourraient être assimilés à des modèles d'élasticité dans la mesure où ils utilisent cette notion d'élasticité pour prédire à l'échelle de l'arrêt. Mais ils sont structurés d'une manière qui leur est propre et c'est pour cette raison que le choix de les étudier dans une partie indépendante a été fait.

Les modèles d'élasticité permettent de produire des prévisions de la demande à plusieurs niveaux d'agrégation différents en fonction des besoins auxquels le modèle souhaite répondre. Prévoir la

demande sur l'intégralité du réseau est chose commune avec ce type de modèle (Ko, Kim, & Etezady, 2019). Les variables démographiques et économiques sont prépondérantes lors de la prévision du nombre total de montants sur la globalité du réseau de TC. Ces paramètres varient relativement lentement au cours du temps, il n'est donc pas possible de réaliser des modèles journaliers uniquement en utilisant ces données. Chiang (2011) utilise le budget d'exploitation, le prix des titres de transport, le prix de l'essence et un indicateur de pauvreté pour réaliser des prévisions du nombre total de montants sur le réseau de transport par mois. Ces variables économiques ayant un effet à moyen terme sur la demande, étudier l'élasticité de la demande par rapport à ces paramètres à une échelle de temps inférieure n'est pas pertinent. C'est pour cette raison qu'il est important d'adapter les variables explicatives au niveau d'agrégation temporelle des données de demande. Un modèle effectuant de la prévision journalière (Ko, Kim, & Etezady, 2019) privilégiera des variables plus précises comme l'offre à travers des données de fréquence ou encore de taille de flotte en service.

La prévision de la demande totale sur le réseau peut être réalisée par le calcul immédiat de la demande finale, mais il existe des modèles permettant de détailler ce calcul. La segmentation des usagers peut par exemple être un moyen d'obtenir des valeurs de la demande en TC pour différents groupes d'individus. Outwater, et al. (2003) mettent au point un modèle permettant de tenir compte de l'évolution différente des habitudes de mobilité des usagers d'un réseau de traversiers. La construction du modèle commence par une segmentation de la population afin de lier des segments à des habitudes de mobilité. Les segments de population sont projetés pour connaître leur composition à un certain horizon temporel. L'hypothèse qui est faite par la suite et qui permet l'obtention de résultat en termes de nombre d'usagers du réseau de traversiers est la suivante : l'évolution du choix modal de chaque segment de la population suit un comportement linéaire. Ainsi, un modèle de régression linéaire au sein de chaque segment est établi pour étudier cette évolution en fonction de paramètres explicatifs. Le modèle se base donc sur l'élasticité de la demande vis-à-vis de plusieurs paramètres au sein de chaque segment et non pas sur la globalité des usagers comme les modèles étudiés précédemment. Cette approche permet de capturer les différents comportements et de les isoler pour les projeter. Il existe également d'autres modèles fournissant des résultats pertinents et qui supposent la constance des habitudes de mobilité de chaque segment. Ainsi, uniquement projeter les segments de la population permet l'obtention de résultats concernant la demande de transport.

Prévoir les paires OD sur un réseau de transport est une donnée très utile dans un grand nombre de projets. En effet, connaître les origines et les destinations permet d'avoir une image précise de la charge sur les différents liens du réseau. Cette information est importante quand il s'agit de réaliser de la planification d'offre. Oort (2015) a mis au point un modèle permettant d'actualiser à court terme les matrices origines-destinations (OD). Ces matrices sont générées grâce aux données de carte à puce et le modèle se base sur l'élasticité du coût généralisé sur la demande sur chaque paire OD. La valeur de l'élasticité est trouvée dans la littérature. En dressant une matrice des coûts généralisés pour les origines et les destinations, il est possible de réaliser des scénarios d'évolution de ces coûts et ainsi d'actualiser la matrice OD par conséquent. Umlauf, Galicia, & Cheu, 2015 effectuent un travail similaire sur les matrices OD en générant des taux de croissance qu'ils appliquent à chaque paire. Ces taux de croissance sont obtenus par une modélisation des caractéristiques des arrêts.

Le principe d'élasticité est souvent utilisé dans les modèles de prévision, car il permet immédiatement de quantifier l'effet d'un scénario de projection sur la demande. La plupart des modèles font appel à ce principe, mais ceux présents dans cette partie se basent presque uniquement dessus. D'autres caractéristiques de ces modèles sont à étudier dans la suite du projet comme l'utilisation de variables démographiques et économiques pour la prévision mensuelle globale et le principe de segmentation de la population selon les habitudes de mobilité.

2.4 Autres types de modèles

2.4.1 Modèles d'activités

Les modèles d'activités dérivent du modèle séquentiel classique. Ils se basent sur des étapes similaires que le modèle à quatre étapes : génération, distribution, choix modal et affectation (Bonnel, 2004). Le modèle d'activités vise à créer des chaînes d'activités pour la population afin d'orienter les résultats des trois premières étapes du modèle séquentiel classique. Cette approche tente de modéliser au mieux les habitudes de mobilité d'une population en tentant de répondre à la question : pour quelle raison sont réalisés les déplacements. En trouvant la cause de ces déplacements, il est plus simple de les quantifier et de connaître les origines et les destinations de ces derniers. Un tel modèle vise à modéliser l'ensemble des mouvements d'une population, quel

que soit le mode de transport. Ce type de modèle est peut-être trop complexe lorsque le but est de modéliser uniquement le réseau de TC, mais il s'agit ici de s'inspirer d'idées utilisées dans ces modèles.

La génération de patrons d'activités pour la population est la base de ce type de modèle. Généralement, la création de ces patrons d'activités est faite grâce à l'étude d'une enquête OD ou à une enquête spécifique sur les activités. Il est par exemple possible par la suite de lier des chaînes d'activités avec des caractéristiques socio-économiques (Hafezi, Liu, & Millward, 2019). Grâce à ces enquêtes, un panel de plusieurs séquences d'activités spécifiques est identifié. Il est par la suite possible d'associer chaque patron d'activité à des caractéristiques socio-économiques. Ainsi en connaissant les caractéristiques des individus d'une population, il est possible d'estimer le patron d'activité qui sera réalisé pendant une journée type.

D'autres méthodes de création de patrons d'activités sont constatées dans la littérature. Notamment certaines faisant l'usage de variables comportementales. Le modèle FEATHERS (Linh, et al., 2019) par exemple dresse des chaînes d'activités en décomposant de manière simple et grâce à des choix heuristiques le raisonnement permettant d'aboutir à cette chaîne d'activité. Le modèle ALBATROSS (Arentze, Hofman, van Mourik, & Timmermans, 2000) fait également usage de choix heuristiques pour simuler des programmes d'activités, mais à l'avantage sur le précédent de permettre aux motifs d'activités d'évoluer et d'être modifié par des changements d'habitude ou de l'environnement d'un individu.

Ces motifs d'activités permettent par la suite d'associer des déplacements aux individus d'une population et de les affecter sur les différents réseaux de transport. Il est donc possible d'obtenir une image de l'utilisation de tous les modes de transport et de manière assez précise. Ces modèles sont performants sur des horizons à long terme, mais sont très complexes et demandent souvent la création d'une enquête spécifique. Beaucoup de données sont nécessaires à la calibration d'un tel modèle et cela n'est possible que sur un nombre limité de territoires.

2.4.2 Modèles âge-période-cohorte

Un modèle âge-période-cohorte (Krakutovski, 2012) se base sur l'étude séparée de l'influence de plusieurs paramètres. L'influence de l'âge, de la période et de la cohorte sur le comportement de mobilité est quantifiée indépendamment. Le nombre de trajet en transport en commun n'est pas le

même au fils de l'âge. Une relation linéaire est donc établie entre l'âge et la fréquentation d'un réseau de transport public. Cela est effectué également pour la période : comment une certaine période vient modifier les habitudes de mobilité d'une population entière. Puis, l'influence de la cohorte est également déterminée. En combinant ces relations linéaires, il est possible de déterminer le nombre de déplacements générés par un groupe de personnes du même âge, à la même période et appartenant à une même cohorte. En utilisant des prévisions de la composition de la population, il est possible d'obtenir une projection de la fréquentation d'un réseau de transport en commun. Les résultats observables dans la littérature suggèrent que ces modèles permettent de prédire à long terme. Cette capacité serait expliquée par l'inertie forte des relations existantes entre ces paramètres.

2.4.3 Modèles multiniveaux

Les modèles multiniveaux (Wegener, 1982) ont une structure de linéaires permettant de prendre en compte l'effet de plusieurs variables ayant une organisation hiérarchique. Sur un réseau de transport, si l'objectif est de déterminer la demande au niveau de différents secteurs, cela signifie qu'il est possible considérer des variables explicatives ayant une influence sur l'ensemble du territoire et d'autres sur les secteurs indépendamment. La structure est linéaire et comporte plusieurs niveaux. Un niveau permet par exemple d'intégrer les variables ayant une influence globale et un autre niveau permet de relier les facteurs explicatifs au niveau de secteur à la demande en leur sein. Les facteurs d'influence des niveau supérieurs modifient les coefficients des relations linaires des niveaux inférieurs. Cette approche adopte une structure proche de la réalité et permet de limiter l'effets de certains enjeux comme la disponibilité de données à des niveaux d'agrégation différents.

2.4.4 Modèles de choix modal

Ces modèles ne sont pas les plus pertinents pour ce projet car ils nécessitent souvent de prédire la demande pour l'ensemble des modes d'un territoire. Des modèles de choix modales (Idris, Habib, & Shalaby, 2015) permettent par exemple de décomposer les choix des usagers en termes de choix modal en fonction de certaines caractéristiques comme le motif de déplacement ou des facteurs socio-économiques. Cette décomposition permet par la suite d'obtenir la fréquentation de chaque mode de transport d'une région.

Cheng, et al., 2020, mettent en place un modèle de choix modal en utilisant une structure d'intelligence artificielle via un modèle de régression logistique aléatoire. Le modèle en question a pour but de prévoir le mode utilisé pour un trajet d'un certain ménage. Il permet de prendre en compte les caractéristiques du ménage, celles des individus, l'environnement du logement et le motif et le temps du déplacement.

2.4.5 Autres modèles

La partie qui suit a pour but de passer en revue des modèles de prévision de la demande existant dans la littérature mais dont leurs structures sont particulières et il n'est pas possible de les attribuer à une catégorie.

Lawson, Muro, & Krans, 2019 ont créé un modèle combinant plusieurs méthodes existantes mais qui jusqu'à présent n'avaient pas été utilisées simultanément. Cette méthode a pour but de prévoir l'occupation d'un réseau. Elle génère des déplacements vers le travail grâce à des régressions linéaires à partir des données du recensement. Ces déplacements sont organisés dans une matrice OD qui est corrigée par un taux de correction calculé grâce à une enquête OD. Les déplacements sont ensuite repartis sur le réseau grâce à des données GTFS (General Transit Feed Specification). Le modèle permet donc de combiner plusieurs sources de données différentes pour effectuer des prévisions.

D'autres comme Cervero R., 2006, ont développé des modèles de post-traitement qui permettent d'affiner les résultats d'un modèle existant pour tenir compte de nouveaux paramètres. R. Cervero met en place des modèles d'élasticité pour corriger les résultats des modèles séquentiels classiques. Il est donc possible de tenir compte de paramètres comme l'accès aux stations.

2.5 Synthèses des paramètres

2.5.1 Les variables explicatives

La plupart des modèles recensés font appel à des variables explicatives. Trois grandes catégories de variables sont souvent prises en compte : l'aménagement du territoire, des données socio-économiques et l'offre en transport en commun. Ces données sont récoltées à travers des recensements, des données ouvertes des villes, des sociétés de transport, des cartographies, des

GTFS, ou encore de diverses enquêtes menées au sein de la population. Parmi ces catégories, des variables plus concrètes apparaissent plus fréquemment :

- Aménagement du territoire : ce sont principalement des données sur des points d'intérêts. La variable peut effectuer un comptage de chaque unité de ces lieux ou encore déterminer la surface couverte par chaque catégorie.
    - Lieux de résidence
    - Lieux de travail
    - Lieux d'éducation
    - Lieux de loisir/magasinage
    - Place de stationnement
- Données socio-économiques : cette catégorie regroupe des informations sur les personnes d'une certaine population.
    - Population (segmentée ou non par âge)
    - Possession automobile
    - Taux d'activité / de chômage
- Offre en transport : ces données sont souvent fournies par les sociétés de transport en commun (TC)
    - Densité des réseaux TC et routier
    - Accessibilité : nombre de logements dans un rayon défini autour des stations de TC
    - Fréquence des lignes
    - Transférabilité vers d'autres modes

Les modèles étudiés possèdent souvent une déclinaison de ces variables, en ajoutent, en suppriment, font varier les résolutions spatiales, temporelles ou en utilisent des proxys. La liste dressée ci-dessus n'est pas exhaustive, mais elle permet de se rendre compte du type de variable à envisager lorsqu'il s'agit de mettre en place un modèle de prévision de la demande.

2.5.2 Synthèse des modèles et de leurs paramètres

Tous les modèles mentionnés dans la revue de la littérature sont résumés dans le tableau récapitulatif ci-dessous. La liste n'est pas exhaustive et ce tableau résume les caractéristiques des modèles vus dans ce chapitre. Sur la première partie du tableau, il est possible d'avoir une vue globale des variables explicatives qui peuvent être utilisées par les classes de modèles. Toutes ces variables ne sont pas systématiquement toutes utilisées, elles représentent les types de paramètres explicatifs qu'il est possible d'inclure pour effectuer de la prévision. La seconde partie décrit le niveau d'agrégation spatiale sur lequel la demande est prédite. La précision commence aux montants et descendants à chaque station du réseau de TC et se termine à la prévision du nombre global de trajets effectués sur l'ensemble du réseau. Certains modèles permettent également de prévoir la charge sur tous les liens de tous les modes de transport, une colonne est dédiée à cette caractéristique. Enfin, il est possible de connaître, quand cela est mentionné dans la littérature, l'horizon de prévision des classes de modèles. À noter qu'il s'agit de l'horizon de prévision le plus long constaté dans la littérature et ayant donné des résultats pertinents.

Tableau 2-1 Synthèse des classes de modèles de la littérature

Modèles		Modèles autoregressifs	DRM	Modèle d'élasticité	Modèle d'activité	Modèle de choix modale	Age-periode-cohort
Variables explicatives	Utilisation du territoire		•	•			
	Caractéristiques socio-économiques		•	•	•	•	•
	Structure du réseau TC		•	•			
	Accès autres modes		•	•			
	Variables comportementales /activités				•	•	
	Historique de la demande	•					
Variable d'outcome	Charge sur tous les réseaux				•	• (part)	
	Demande sur le réseau de TC			•			•
	Demande sur des secteurs d'agrégation	•		•			
	M/D aux stations		•				
Horizon de plannification	Court terme	•					
	Moyen terme		•	•			
	Long terme				•		•
	Non spécifié					•	

CHAPITRE 3 MÉTHODOLOGIE

Le chapitre précédent a révélé une littérature abondante sur les modèles de prévision. Prévoir la demande future sur un réseau de transport en commun est une thématique sur laquelle une multitude de chercheurs ont travaillé. Des modèles très divers existent pour répondre à cette question omniprésente dans la gestion d'un réseau de transport : comment la demande va-t-elle évoluer ?

Les manières de répondre à cette question sont nombreuses. La littérature a montré qu'un seul modèle ne pouvait convenir à tous les cas de figure. La prévision à court terme est la plus répandue et fait appel à des modèles spécifiques ainsi que des types de variables explicatives précis. Les besoins exprimés pour la réalisation d'un tel modèle sont très différents de ceux pour la réalisation d'un modèle de prévision à long terme. Souvent un modèle court terme vient répondre à des questions de gestion du trafic en temps réel ou à l'information des usagers. Il est rarement considéré pour la prise de décisions stratégiques ou à des fins de planification.

La littérature manque d'explication sur la sélection d'un modèle de prévision de la demande. Le cas présent appelle un modèle pouvant effectuer des projections à minima à moyen terme. Pour une telle situation, une multitude de modèles existent encore et permettent de répondre à différentes problématiques. Ces modèles ont donc pour objectif d'être testés dans la suite de ce projet. Mais d'autres paramètres restent à être déterminés afin d'affiner la sélection des modèles à expérimenter. Il faut déterminer la résolution spatiale et temporelle de la variable dépendante ainsi que sa définition exacte. En effet, la demande peut être définie par plusieurs variables ou proxys différents. Une étude des données disponibles permettra donc de déterminer ce paramètre ainsi que de trouver quelles sont les variables explicatives potentielles.

La littérature a montré que les variables explicatives disponibles limitent souvent le type de modèle qu'il est possible d'utiliser. L'approche de modélisation est très différente pour un réseau riche en données diverses pouvant expliquer les variations de la demande que pour un territoire dont seule la fréquentation des transports en commun est connue.

Ce chapitre vise donc à présenter le territoire d'étude ainsi que les données disponibles pour ensuite décrire la méthodologie générale. La méthodologie générale inclut également un catalogue de l'ensemble des modèles utilisés dans le projet. Chaque modèle est explicité et la méthode utilisée pour les mettre en place est décrite. Cette partie est présente afin de permettre une meilleure

compréhension des modèles utilisés ainsi que d'en permettre l'utilisation dans de futurs travaux. Une première étape de traitement préliminaire des données est également présente dans ce chapitre afin de faciliter la compréhension des chapitres expérimentaux qui suivent.

3.1 Description des données

3.1.1 Mise en contexte

La Société de transport de l'Outaouais (STO) est l'opérateur de transport en commun de la région de Gatineau. Gatineau située dans la province de Québec constitue la quatrième ville québécoise la plus peuplée. Sa position géographique au bord de la rivière des Outaouais et sur la rive opposée d'Ottawa, lui confère une importance majeure dans cette région administrative. Ottawa en tant que capitale du Canada, est le berceau des principales administrations canadiennes. Le gouvernement fédéral représente ainsi le principal employeur de la région. D'après le journal Le Droit, plus de 123 000 habitants d'Ottawa et de Gatineau sont employés par le gouvernement fédéral. Ce nombre est associé à une augmentation majeure du nombre de Gatinois étant fonctionnaires fédéraux: une hausse de 20.6% a été constatée entre 2005 et 2015 uniquement pour les habitants de la rive de Gatineau.

Le réseau de la STO est constitué de lignes de bus régulières et d'une ligne de bus à haut niveau de service (BRT). L'ensemble du réseau s'étend sur les villes d'Aylmer, Hull, Gatineau, Ottawa, Buckingham et Masson-Angers. Une carte du réseau est disponible sur la Figure 3-1.

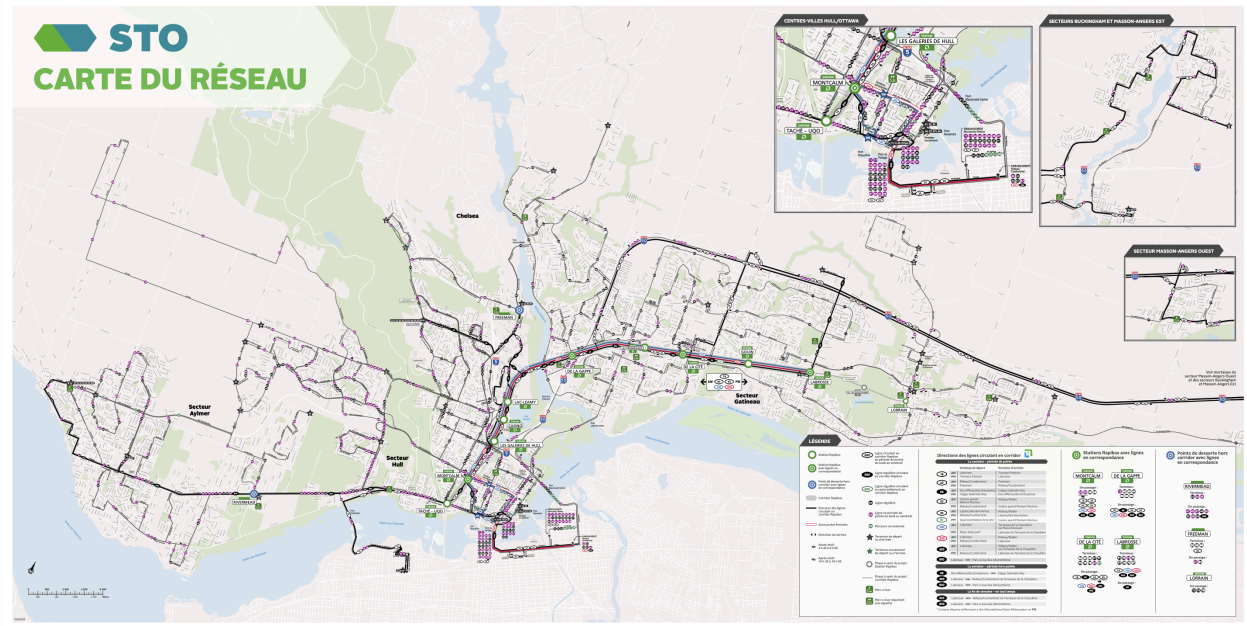


Figure 3-1 Carte réseau STO

(http://www.sto.ca/fileadmin/user_upload/communications/PDF/Carte_reseau/Carte_reseau.pdf)

La STO est l'un des deux principaux opérateurs de transport en commun de la région Ottawa-Gatineau. OC Transpo est l'autre société majeure de transport. Elle assure les transports en commun dans la ville d'Ottawa. Les deux sociétés fonctionnent en symbiose pour assurer l'interconnectivité des deux rives.

Le demande sur le réseau de transport de la STO a connu en 2016 en forte croissance. L'accroissement de la population dans la zone, le taux de chômage bas, la congestion routière grandissante et les développements immobiliers de plus en plus nombreux sont autant de facteurs pouvant expliquer une telle croissance. La facilité de transit en transport en commun entre Ottawa et Gatineau permettant d'accueillir les lieux de résidence d'une partie des fonctionnaires directement sur la rive de Gatineau peut également constituer un moteur de l'accroissement de la demande.

3.1.2 Données utilisées

Les données utilisées sont mentionnées dans cette partie, mais le chapitre traitant de l'analyse rétrospective vient par la suite détailler ce panel de données. Il a notamment pour but d'étudier la dynamique des données disponibles et leurs interactions avec la demande.

3.1.2.1 Données de demande en transport

La demande, comme il a été mentionné précédemment, peut être quantifiée par plusieurs variables différentes. Les données de cartes de puce de l'ensemble des usagers du réseau de la STO ont été mises à disposition. Les variables matérialisant la demande sont donc issues de cette base de données de carte à puce. Cette donnée est très versatile, car brute, elle regroupe un grand nombre d'informations.

Les usagers du réseau de la STO doivent se procurer un titre de transport afin d'accéder au réseau de bus. L'acquisition de ce titre, qu'il soit un trajet simple ou un abonnement, se fait sur un support de carte à puce, principalement sur une carte appelée Multi. Pour accéder à un bus, l'usager est dans l'obligation de valider son titre de transport sur des bornes dédiées.

La versatilité de cette donnée réside dans les informations récoltées lors de la validation du titre de transport. Parmi les informations récoltées par ce système, il y a :

- Le numéro d'identification de la carte (qui est anonymisé pour des raisons de protection de données personnelles)
- Le type de forfait utilisé
- La date et l'heure de la validation
- L'arrêt correspondant à la montée de l'usager
- L'identifiant du bus
- La direction
- Le type de la validation : savoir si c'est l'origine du trajet ou une correspondance

D'autres données sont disponibles, mais ces informations correspondent à celles qui permettent la fabrication des bases de données utilisées dans la suite du projet.

Ces informations permettent de construire plusieurs variables utiles pour effectuer de la prévision. Il est par exemple possible d'extraire la moyenne mensuelle de montants sur le réseau de la STO pour un jour ouvrable. Cette donnée transformée donne une indication de la fréquentation du réseau et peut être calculée à divers niveaux d'agrégation spatiale. La dimension brute de la donnée permet d'avoir la possibilité d'aller jusqu'à une agrégation par arrêt.

Avoir des informations sur le type de forfait utilisé sur la transaction permet également de pouvoir segmenter les usagers du réseau de TC. Il est possible de savoir si l'usager est un étudiant, un adulte ou encore une personne âgée. Les modèles de segmentation de la population vus dans la revue de littérature peuvent profiter d'une telle information. À noter que l'identifiant de l'usager ne peut pas être utilisé pour identifier ses caractéristiques, car il est anonyme.

Dans les expérimentations, des déclinaisons de ces données de cartes à puce sont utilisées. La variable précise est explicitée avant toute expérience lorsque cela est nécessaire.

3.1.2.2 Variables explicatives potentielles

Des modèles de prévision de la demande ont pour but d'être calibrés dans les chapitres expérimentaux. Dans cette mesure, il y a un besoin de données pouvant expliquer les variations de la demande en transport. Toutes les variables permettant d'avoir une représentation quantifiée des transformations subies par le territoire d'étude sont bénéfiques.

La revue de la littérature a permis de connaître les données utilisées dans la création de modèles de prévision. Cette information est donc utilisée pour récolter des variables explicatives similaires dans les bases de données du territoire d'étude. Les principales sources de données pour le territoire de Gatineau sont :

- ISQ : l'institut statistique du Québec
- Statistique Canada : les données issues des recensements y sont disponibles.
- La ville de Gatineau : le site internet de la ville de Gatineau regroupe une multitude de statistiques et d'informations sur les évolutions du territoire.

Ces bases de données regorgent d'informations sur la démographie, l'occupation des sols, l'économie et autres. Autant de données qui sont pertinentes à inclure dans un modèle de prévision de la demande. Néanmoins, la plupart des statistiques ne tiennent pas d'historique accessible au grand publique. Les données ne sont donc pas continues dans le temps. Intégrer de telles données dans un modèle de prévisions de la demande est impossible, car leur évolution temporelle n'est pas connue et ainsi leur effet sur la demande ne peut pas être pris en compte.

Ces bases de données permettent la consultation des deux grandes enquêtes réalisées à intervalle régulier dans la région:

- Le recensement de la population : il a pour but de brosser un portrait statistique du pays. Il est réalisé à l'échelle du Canada et conduit tous les cinq ans. C'est une source d'information importante, car il est possible d'en extraire des données précises sur la population ou encore les emplois.
- L'enquête origine-destination : c'est une enquête descriptive réalisée à grande échelle visant à brosser un portrait des habitudes de mobilité des personnes vivant dans la zone d'étude. Elle permet de mesurer, à l'aide de divers indicateurs, la mobilité et son évolution ainsi qu'analyser le comportement et les choix de modes de transport des usagers. Cette enquête est également réalisée à intervalles réguliers d'environ cinq ans.

Ces enquêtes ne possèdent pas non plus un caractère continu dans le temps. Pour effectuer de la prévision, un intervalle entre chaque mesure de cinq ans est trop long. Cela ne permet pas d'être mise en parallèle des évolutions de la demande en transport. Des données actualisées annuellement ou mensuellement sont disponibles pour certaines données telles que la population ou le nombre d'emplois. Des données actualisées sont pertinentes pour effectuer des analyses rétrospectives, mais il est souvent difficile de les intégrer dans un modèle de prévision.

Cette partie a permis d'identifier les bases de données desquelles sont extraites les variables utilisées dans ce projet ainsi que de les introduire. Une description rigoureuse de ces données est effectuée dans le chapitre d'analyse rétrospective. Ce chapitre a pour but d'étudier l'évolution de toutes données récoltées et de les mettre en parallèle entre elles afin d'identifier des dynamiques régissant leur comportement dans le temps. Les données ne sont qu'introduites dans la partie actuelle afin d'éviter les redondances.

3.2 Méthodologie des modèles

Les données mentionnées dans la partie précédente sont pour la plupart mises à contribution des objectifs de prévision au sein de modèles prédictifs. Cette partie a pour objectif de recenser l'ensemble des modèles calibrés dans le projet afin de présenter la méthodologie mise en place pour produire des prévisions. Ce catalogue recense uniquement les modèles utilisés dans le projet; certains modèles évoqués précédemment n'apparaissent donc pas dans cette partie. Le choix de ne pas les inclure est justifié dans la suite du mémoire. De plus, les méthodes sont présentées dans

cette partie, mais certains détails de la méthodologie sont uniquement présentés dans les parties expérimentales qui suivent. En effet, certains choix méthodologiques découlent de résultats expérimentaux intermédiaires. À des fins de concision, ils ne sont pas explicités une fois les résultats intermédiaires obtenus.

3.2.1 ARIMA

Le modèle ARIMA (*autoregressive integrated moving average*) est une approche permettant la prévision de séries temporelles en utilisant l'aspect autorégressif de la série. Elle utilise notamment une combinaison linéaire des valeurs passées. Une méthode autorégressive est appliquée aux valeurs passées de validations de CAP et une méthode de moyenne flottante (moving average) est appliquée aux résidus. Pour ce type de modèle, la série doit être stationnaire et c'est pour cela que préalablement, la série temporelle peut être transformée en utilisant, dans le cas d'ARIMA, la différentiation (transformation de la série en la variation entre deux valeurs consécutives). Une telle transformation appliquée à un certain degré (d) a pour effet de rendre la série stationnaire. C'est un des paramètres qui doivent être déterminés dans la création d'un tel modèle. Il faut également trouver le paramètre du terme autorégressif (p) et celui de la moyenne flottante (q). Une méthode simple ARIMA est définie par ces trois paramètres : ARIMA (p, d, q). Dans certains cas, il faut également rajouter un terme qui définit l'aspect saisonnier et qui est défini de la même manière, mais appliqué à la variation saisonnière.

Dans ce projet, le modèle ARIMA est calibré sur le logiciel RStudio, un logiciel de statistique et de gestion de données. Afin de pouvoir calibrer ce modèle, il est nécessaire d'utiliser le package Forecast, dont la documentation est disponible en ligne. Cet outil permet l'utilisation de l'ensemble des fonctions indispensables à la mise en place du modèle. Premièrement, il faut déterminer les paramètres explicités précédemment. Cela peut être réalisé de deux manières : automatiquement et manuellement. La méthode manuelle se base sur la détermination du degré de différenciation afin de stabiliser la série temporelle. Puis sur l'expérimentation de plusieurs combinaisons de paramètres concernant le terme autorégressif et celui de moyenne flottante, afin d'obtenir le meilleur critère d'information d'Akaike. La méthode automatique disponible via la fonction auto.arima dans RStudio, se base sur la détermination de ces paramètres par l'intermédiaire de l'algorithme Hyndman-Khandakar. Les prévisions du modèle ARIMA sont obtenues en utilisant

la fonction forecast. La documentation de l'ensemble de ces fonctions est disponible sur le site RDocumentation.

3.2.2 Régression linéaire multiple

Le principe de prévision d'une régression linéaire repose sur la combinaison linéaire de plusieurs variables indépendantes. Le modèle peut être explicité de la manière suivante :

$$y_t = \beta_0 + \beta_1 * x_{1,t} + \beta_2 * x_{2,t} + \cdots + \epsilon_t$$

Où y_t est la variable dépendante à l'instant t avec t un entier naturel allant de 0 à T,

$(x_1, x_2, \dots, x_n)$ sont les n variables indépendantes avec n un entier naturel,

Et ϵ_t l'erreur du modèle à l'instant t.

Les hypothèses pour le modèle sont les suivantes :

- Pour tout t, les ϵ_t sont indépendant les uns des autres, leur espérance est nulle et ils ont une distribution normale.
- Les $(x_1, x_2, \dots, x_n)$ sont indépendants et de distribution normale.
- Les paramètres $\beta_0, \beta_1, \beta_2, \dots, \beta_i$ sont supposés constants.

Les paramètres du modèle $\beta_0, \beta_1, \beta_2, \dots, \beta_i$ sont déterminés afin de réduire au maximum l'erreur tout en respectant les hypothèses. Différentes méthodes peuvent être utilisées, la plus courante étant la méthode des moindres de carrées qui sera utilisée dans le mémoire. Le but est principalement de réduire le terme d'erreur :

$$E = \sum_{t=0}^{T} \epsilon_t^2$$

Le modèle est calibré dans le logiciel RStudio. Le package utilisé est Forecast et les fonctions permettant de calibrer et de projeter le modèle sont *tslm()* et *forecast()*.

3.2.3 Regression linéaire sur séries temporelles

Un modèle de régression linéaire sur séries temporelles est similaire à un modèle de régression linéaire où la variable dépendante peut également être une variable explicative si un retard est

intégré à cette variable. Ce modèle permet d'utiliser le caractère autorégressif de la variable dépendante.

Le modèle s'explicite de la manière suivante :

$$
\begin{aligned}
y_t = {} & \beta_0 + \beta_1 * y_{t-1} + \beta_2 * y_{t-2} + \cdots + \beta_p * y_{t-p} \\
& + \delta_{11} * x_{1,t-1} + \delta_{12} * x_{1,t-2} + \cdots + \delta_{1q} * x_{1,t-q} \\
& + \ldots \\
& + \delta_{k1} * x_{k,t-1} + \delta_{k2} * x_{k,t-2} + \cdots + \delta_{kq} * x_{k,t-q} \\
& + u_t
\end{aligned}
$$

Où Y est la variable dépendante et p représente le nombre de retards pris en compte,

(X_1, X_2, …, X_k) sont k variables explicatives dont q valeurs du retard sont prises en compte,

δ représente des coefficients réels calculés pour calibrer le modèle,

u_t représente l'erreur sur la prévision de Y_t.

Pour que ce modèle soit pertinent, les hypothèses suivantes doivent être vérifiées :

- Le terme d'erreur u doit posséder une moyenne nulle.
- (Y, X_1, X_2, …, X_k) doivent être stationnaires
- (Y_t , $X_{1,t}$, $X_{2,t}$, …, $X_{k,t}$) et (Y_{t-j} , $X_{1,t-j}$, $X_{2,t-j}$, …, $X_{k,t-j}$) deviennent indépendants lorsque j est grand.
- La probabilité d'obtenir des valeurs absurdes est faible
- Il n'existe aucune multicolinéarité parfaite.

Le modèle est également mis en place en utilisant l'outil RStudio. Le package utilisé est Forecast et les fonctions permettant de calibrer et de projeter le modèle sont *tslm()* et *forecast()*.

3.2.4 Régression linéaire avec correction de l'erreur ARIMA

Ce modèle se différencie d'une régression linéaire simple par l'intégration du caractère autorégressif en plus des variables explicatives. Le modèle consiste en un modèle de régression linéaire classique avec les mêmes hypothèses dans lequel on autorise l'erreur à être auto-corrélée. À partir des valeurs de cette erreur, un modèle ARIMA est créé. Pour la prévision, le modèle s'appuiera donc sur la régression linéaire et le modèle ARIMA pour tenter de corriger l'erreur. La structure est la suivante :

$$y = \beta_0 + \beta_1 * x + \eta + \epsilon$$

Où, y est la variable dépendante,

β_0 et β_1 sont les scalaires de la régression linéaire classique,

x une variable explicative (à savoir qu'il peut y en avoir plusieurs),

η le terme d'erreur issu de la correction par un modèle ARIMA,

Et ϵ l'erreur réelle.

Ce modèle est également calibré sur le logiciel RStudio. Il utilise les mêmes fonctions que le modèle ARIMA en ajoutant un terme nommé *xreg,* dans les fonction *arima()* (ou *auto.arima()*) et *forecast()*, qui correspond à la variable explicative.

3.2.5 Réseau de neurones autorégressif

Le réseau de neurones autorégressif est une méthode d'intelligence artificielle se basant sur une structure similaire au cerveau humain. Il permet de prendre en compte les dynamiques non-linéaires de la variable étudiée. Le caractère autorégressif indique que la variable dépendante est uniquement projetée en utilisant les variations passées de cette dernière.

Un tel modèle peut être pensé comme un réseau de « neurones » organisés sur différentes couches ou strates dont la première est constituée des entrées (les prédicteurs) et la dernière, la sortie (la prévision). L'entrée est dans le cas présent uniquement représentée par les valeurs passées de la variable dépendante et la sortie est une valeur future de la variable dépendante.

Pour chaque couche qui est constituée de nœuds placés en parallèle, il y a une entrée et une sortie. La sortie d'une couche est l'entrée de la suivante. Chaque nœud appartenant à une couche prend en entrée une combinaison linéaire des nœuds de la couche précédente.

Schématiquement, le modèle peut être représenté de la manière suivante :

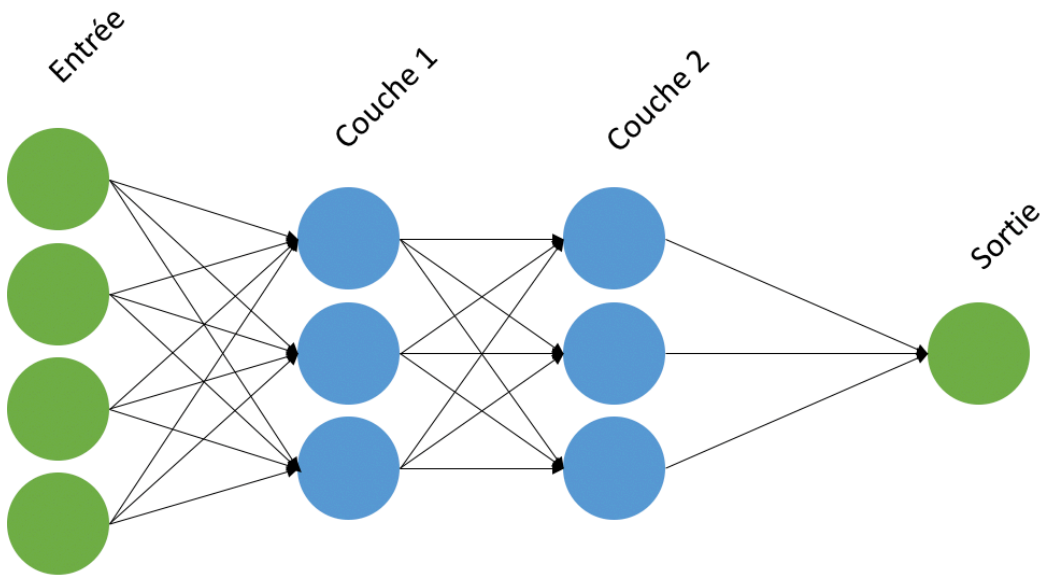


Figure 3-2 Schéma d'un réseau de neurones

Le schéma représente arbitrairement un réseau de neurones avec deux couches composées chacune de 4 nœuds. Pour créer un modèle, il faut déterminer le nombre de couches et de nœuds. Il Faut également déterminer le nombre d'entrées qui correspond au nombre d'incréments de retard de la variable dépendante qui sera utilisé pour effectuer de la prévision. Si dans la fonction utilisée sur RStudio ces paramètres ne sont pas spécifiés, elles sont sélectionnées automatiquement.

Le modèle est calibré dans RStudio. Le package utilisé est Forecast et les fonctions utilisées pour produire une prévision sont *nnetar()* et *forecast()*.

3.2.6 Modèles multiniveaux

Un modèle multiniveau est une méthode de modélisation linéaire permettant de considérer l'influence de variables indépendantes à plusieurs échelles. En prenant l'exemple d'un territoire divisé en de multiples zones, il est possible de mieux comprendre la structure et les avantages de ce modèle.

Soit un territoire subdivisé en I zones (I un entier naturel), chaque zone est nommée par un entier i allant de 1 à I. L'objectif est de modéliser une variable dépendante dont l'évolution est temporelle et dont la valeur est différente pour chaque zone i. La dépendance temporelle est notée par l'indice t. Ainsi, à un instant t dans la zone i, la variable dépendante s'écrit y_{ti}. Les variables explicatives disponibles sont constituées de deux groupes :

- Un premier (niveau 1) regroupant n variables dépendantes du temps et influençant l'ensemble du territoire. Sa résolution géographique n'est pas disponible au niveau de zones i indépendantes. Ces variables sont désignées par $(X_{1t}, X_{2t}, \dots, X_{nt})$.
- Le deuxième groupe (niveau 2) rassemble des variables ne possédant pas de dépendance temporelle, mais une valeur est disponible pour chaque zone i. Ce sont des variables permettant de différencier les zones entre elles. Ces variables sont désignées par $(x_{1i}, x_{2i}, \dots, x_{ki})$.

L'avantage de ce modèle réside dans cette capacité à séparer l'influence des variables explicatives selon leur influence géographique et de les considérer dans un même modèle.

La mise en équation du modèle est donc de la forme suivante :

- Niveau 1 :

$$y_{ti} = \beta_{0i} + X_{1t} * \beta_{1i} + X_{2t} * \beta_{2i} + \cdots + X_{nt} * \beta_{ni} + \epsilon_{ti}$$

- Niveau 2 :

$$\beta_{0i} = \beta_{00} + \beta_{01} * x_{1i} + \beta_{02} * x_{2i} + \cdots + \beta_{0k} * x_{ki} + \alpha_{0i}$$

$$\beta_{1i} = \beta_{10} + \beta_{11} * x_{1i} + \beta_{12} * x_{2i} + \cdots + \beta_{1k} * x_{ki} + \alpha_{1i}$$

$$\beta_{2i} = \beta_{20} + \beta_{21} * x_{1i} + \beta_{22} * x_{2i} + \cdots + \beta_{2k} * x_{ki} + \alpha_{2i}$$

...

$$\beta_{ni} = \beta_{n0} + \beta_{n1} * x_{1i} + \beta_{n2} * x_{2i} + \cdots + \beta_{nk} * x_{ki} + \alpha_{ni}$$

Avec,

y_{ti} représente la variable dépendante évaluée à l'instant t sur la zone i,

L'ensemble des βdouble indicé représente une constante qui sera calculée par le modèle,

ϵ_{ti} est un terme d'erreur qui suit une loi normale $\mathcal{N}(0, \sigma_{ti}^2)$,

α_{ni} est un terme d'erreur qui suit une loi normale $\mathcal{N}(0, \sigma_{ni}^2)$,

Les résultats concernant les modèles multiniveaux suivants sont obtenus à l'aide du logiciel R Studio et plus particulièrement avec l'utilisation de package *nlme*.

3.2.7 Modèle de génération de taux de montants

Le modèle de génération de taux de montants se base sur la création de typologies. Cette typologie s'applique aux différentes zones du territoire sur lequel est présent un système de transport en commun. L'hypothèse réalisée pour permettre d'obtenir une projection de la fréquentation des transports en commun est qu'au sein d'un segment de cette typologie, les comportements de mobilité sont similaires. La similarité est modélisée par un taux de génération de montants constant pour chaque type. Ce taux peut être établi à la fois en termes de personnes ou de logements.

Pour calibrer ce modèle, quatre étapes sont nécessaires :

- Réaliser une typologie de zones : En étudiant quels paramètres influençent les habitudes de mobilité, il est possible de sélectionner des paramètres qui sont déterminants dans la différenciation des zones au regard du taux de génération de montants. Par la suite, il faut également établir les bornes ou centroïdes de la typologie. Des fonctions comme *kmeans* du package Stats dans RStudio peuvent être utiles.
- Associer chaque zone à un type : La typologie est fixe, c'est-à-dire qu'une même zone peut changer de type au fils des années. À chaque incrément temporel, une zone doit être associée à un certain type.

- Calculer le taux de génération de montants pour chaque type : Plusieurs méthodes sont possibles pour déterminer le taux de génération de montants pour chaque type. Il est possible de sélectionner pour tous les incréments temporels, pour toutes les zones d'un même type les taux de génération de montants et d'effectuer une moyenne. Ce taux sera par la suite appliqué à toutes les zones du type pour effectuer une projection. Il faut aussi prendre en compte de quoi dépend le taux. Généralement, il peut correspondre à un taux de génération de montants par personne ou par logement. Cela doit également être cohérent avec la typologie choisie.
- Projeter la fréquentation : Les caractéristiques des zones sont projetées afin de pouvoir projeter le type de chaque zone et également l'élément déterminant du taux (personne ou logement). Avec ces informations, il est donc possible de multiplier le taux du type correspondant par le nombre de personnes ou le nombre de logements afin d'obtenir un nombre de montants au sein de la zone.

Cette méthode a l'avantage de ne pas nécessiter beaucoup de variables explicatives pour pouvoir être calibrée. Elle peut également être enrichie par la création de plusieurs typologies différentes résultant en des nombres de montants différents pour une même zone qui seront par la suite adaptés pour obtenir un nombre unique.

3.3 Travaux préliminaires sur les données

Dans le déroulement du projet, deux étapes importantes ont été réalisées préalablement aux expérimentations afin d'être en mesure de conduire la calibration de certains modèles. Ces travaux sont décrits dans ce chapitre méthodologique, car ils sont indispensables à la compréhension des futures expérimentations ainsi qu'à la justification de la structure du projet. La première étape est la création d'une variable permettant de connaître l'évolution du nombre de logements sur le territoire. La deuxième étape consiste à créer des zones d'agrégation pertinentes pour mettre en place des modèles de prévision.

3.3.1 Variables des permis de construire

Une donnée très importante à ce projet est le registre des permis de construire tenu par la ville de Gatineau. Cette donnée est importante du fait de sa précision à la fois spatiale et temporelle. Le registre est précisément daté et localisé. Il est donc possible d'avoir cette composante temporelle manquante a beaucoup d'autres variables disponibles.

La littérature a montré l'influence non négligeable du nombre de logements sur la demande. Mais les permis de construire ne constituent pas une image directe du nombre de logements. L'objectif de ce travail préliminaire est d'utiliser les données de permis de construire pour actualiser les nombres de logements dans la zone, disponible aux cinq ans, grâce aux données de recensements. La partie qui suit a pour but de détailler les étapes ayant mené à l'actualisation des données de logements.

La variable étudiée dans cette partie reflète le nombre de permis de construire attribués sur le territoire de la RMR de Gatineau chaque mois. Mais l'effet de cette variable sur la demande n'est vraisemblablement pas immédiat. En effet, la déclaration d'un permis de construire ne signifie pas l'implantation immédiate d'un ou plusieurs nouveaux logements ni de la prise de possession du logement par les propriétaires. L'enjeu de cette partie est de déterminer une durée moyenne entre l'acceptation du permis de construire et l'effet sur la demande en transport en commun.

Connaître ce retard d'effet a pour objectif de pouvoir intégrer cette variable dans les modèles de prévision. Par l'intermédiaire du nombre de logements autorisés par chaque permis de construire, l'objectif est d'actualiser les données du nombre de logements fournies par le recensement de 2016. Le recensement 2016 permet d'obtenir ces données agrégées à l'échelle des îlots de diffusion. Un îlot de diffusion est la plus petite unité géographique sur laquelle les données démographiques issues du recensement sont disponibles. Ces zones sont généralement définies par le réseau routier et comptent en règle générale au moins 15 habitants. En secteur dense, ces zones sont très fines et permettent d'avoir des données sur les populations et les logements très précises. L'objectif est donc d'actualiser les données de logements pour les îlots de diffusion afin de garder la précision géographique de la variable du nombre de logements fournie par le recensement. Il sera par la suite possible d'obtenir des approximations pour les années qui suivent. L'actualisation a pour but d'avoir des données mensuelles sur l'évolution du nombre de logements.

Deux méthodes sont utilisées pour tenter de déterminer le retard de l'effet sur la demande et ainsi construire la variable du nombre de logements : une étude de la corrélation entre la demande et les permis de construire et une revue de la littérature à ce sujet.

3.3.1.1 Étude de la corrélation entre les permis et la demande

Pour comprendre ce phénomène, la première étape consiste à étudier la corrélation entre la demande et l'attribution des permis de construire. Plusieurs études de corrélation peuvent être réalisées avec de multiples retards de la variable (simple décalage temporel) pour tenter d'avoir une idée de la valeur de ce délai.

Cette étude est réalisée sur le nombre de logements autorisés mensuellement par permis de construire sur la totalité du territoire d'étude : la RMR Ottawa-Gatineau (rive Québec). Ce nombre est représenté par le cumul de logements autorisés depuis le début de l'ensemble de données. La demande est matérialisée par le nombre de montants mensuels sur le réseau de la STO pendant une pointe du matin moyenne d'un jour ouvrable.

Mais tout d'abord, pour comprendre les résultats suivants de manière pertinente, il faut réaliser une étude d'autocorrélation sur les variables utilisées. Les figures suivantes mettent en évidence si la variable est autocorrélée ou non. La ligne bleue en pointillé est la limite à partir de laquelle la variable est considérée comme significativement autocorrélée. Le coefficient de corrélation utilisé est celui de Pearson prenant en compte uniquement la corrélation linéaire.

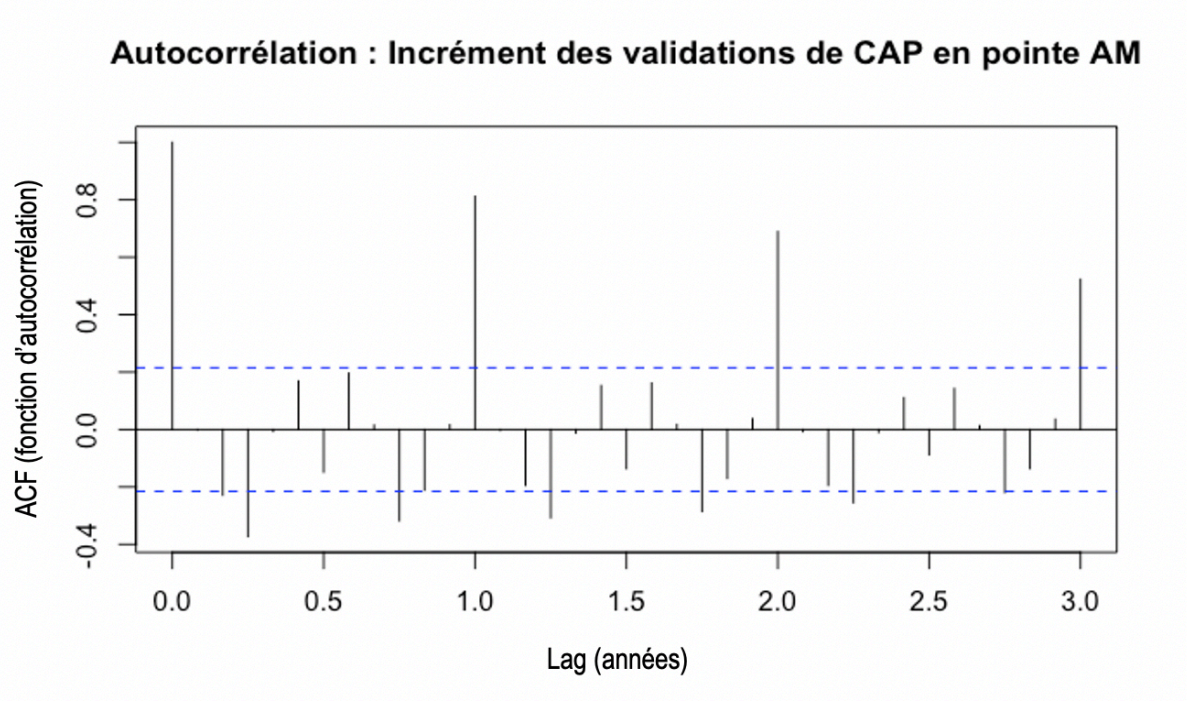


Figure 3-3 Test d'autocorrélation de la demande basé sur une unité annuelle

Les validations de CAP font preuve d'une autocorrélation annuelle ce qui est lié à la nature de la variable possédant un caractère saisonnier prononcé. Cette caractéristique peut s'avérer fournir des résultats sur le retard de la variable modulo un an. Par la suite, une étude tente de s'affranchir de ce problème.

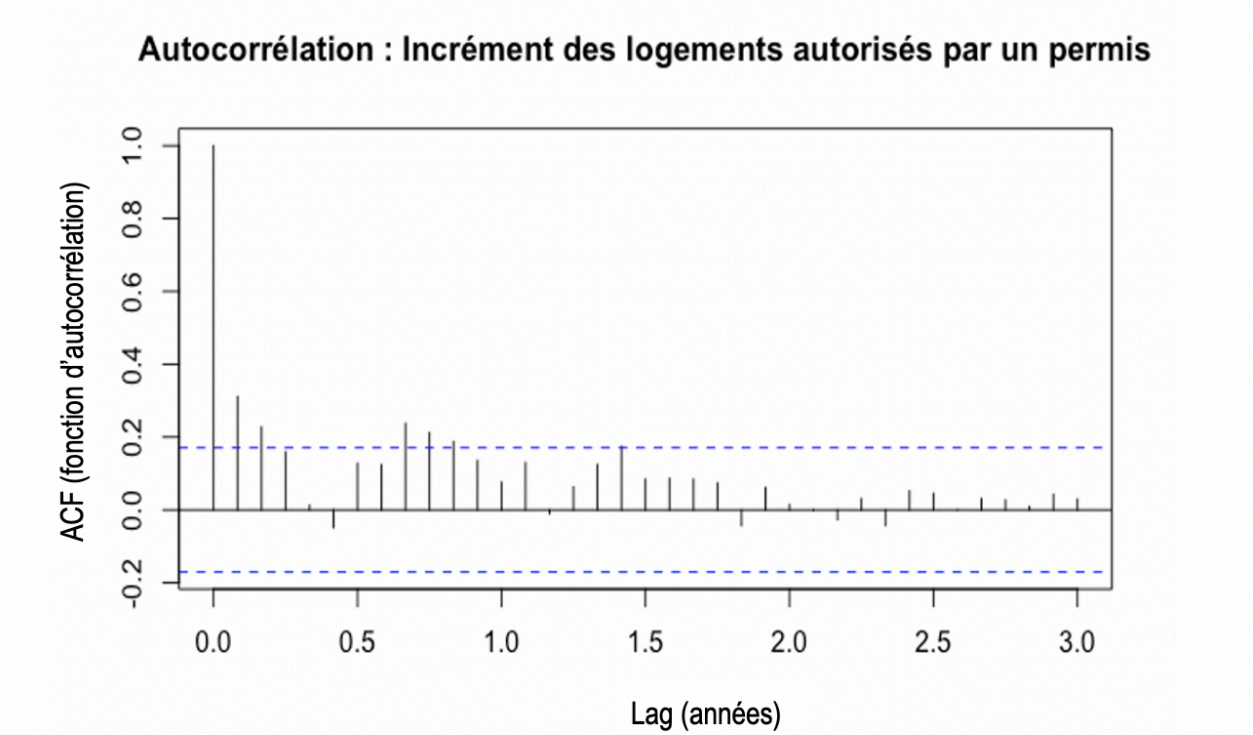


Figure 3-4 Test d'autocorrélation sur les données de permis de construire basé sur une unité annuelle

L'attribution des permis de construire semble être corrélée à avec une fréquence de 7 mois, mais ce phénomène est rapidement atténué dans le temps. L'expérimentation consiste à étudier le coefficient de corrélation entre les deux variables en décalant la variable des permis de construire d'un mois à chaque fois. Le délai de la variable est étudié avec un incrément d'un mois allant de 0 à 3 ans. Sur la Figure 3-5, il est possible de constater l'évolution du coefficient de corrélation en fonction de ce décalage.

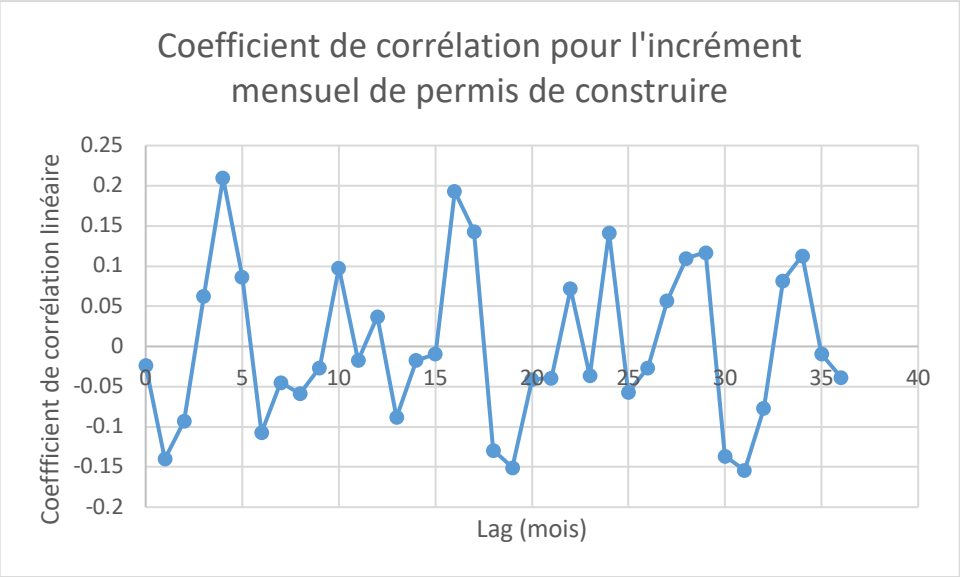


Figure 3-5 Étude du retard de la variable de permis de construire

À aucun moment, la corrélation ne devient significative, mais nous pouvons tout de même constater deux valeurs maximales pour un délai de 4 mois et de 16 mois (coefficient = 0,2). Une légère influence peut être relevée, mais cela ne permet pas de conclure sur un retard pertinent. Ces deux pics sont espacés d'un an ce qui souligne l'effet lié à l'aspect saisonnier de la demande.

Dans la mesure où aucun résultat n'est significatif, il a donc été décidé de scinder la variable en deux sous-groupes : les permis pour des bâtiments avec peu de logements et ceux pour des logements collectifs importants. Arbitrairement la limite de ces deux groupes est prise à 5 logements par bâtiments. Deux variables sont donc créées :

- L'incrément de logements autorisés par un permis de construire pour des bâtiments comprenant de 1 à 4 logements.
- L'incrément de logements autorisés par un permis de construire pour des bâtiments comprenant plus de 5 logements.

De manière similaire, l'autocorrélation des variables est étudiée.

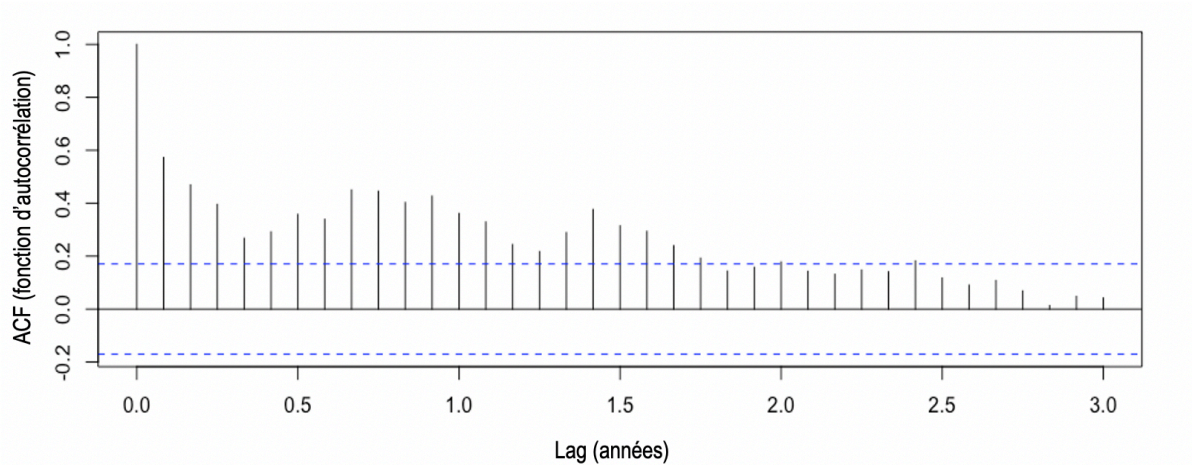


Figure 3-6 Test d'autocorrélation sur les données de permis de construire basé sur une unité annuelle (1 à 4 logements par bâtiment)

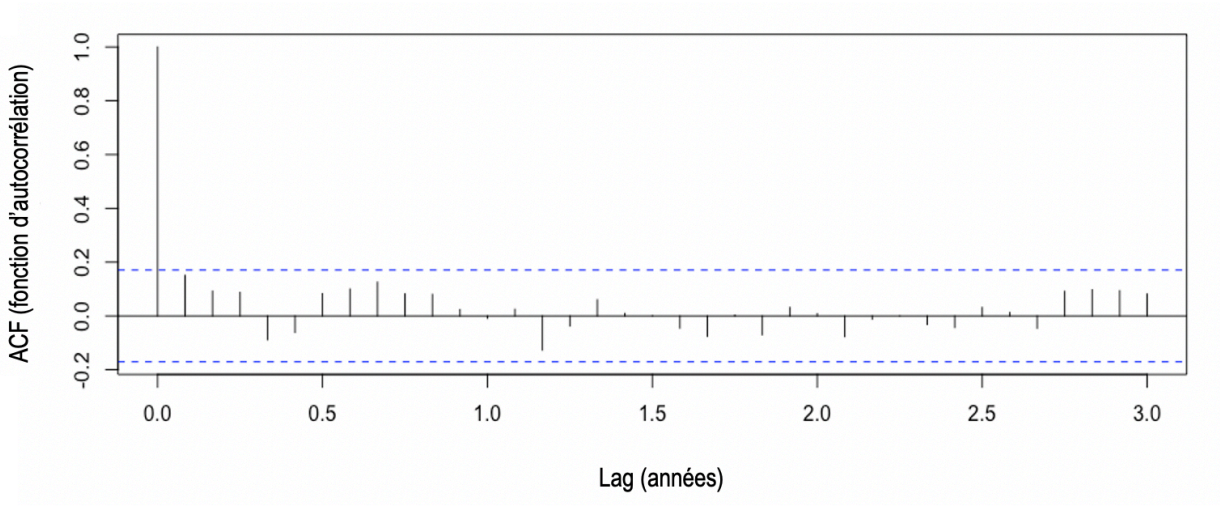


Figure 3-7 st d'autocorrélation sur les données de permis de construire basé sur une unité annuelle (5 et plus logements par bâtiment)

La même étude que pour le total des permis de construire est réalisée pour ces deux sous-groupes. Les résultats sont illustrés par les Figure 3-8Figure 3-9.

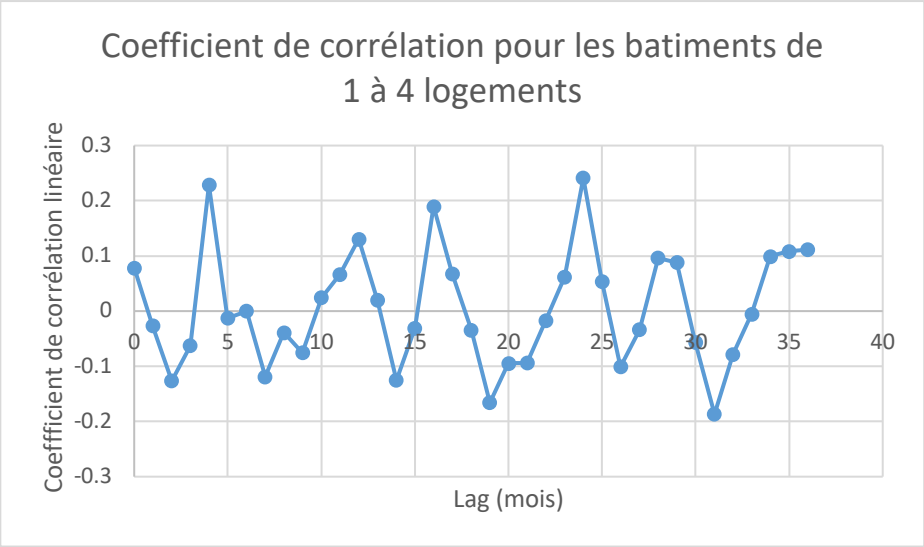


Figure 3-8 Étude de corrélation pour les bâtiments de 1 à 4 logements

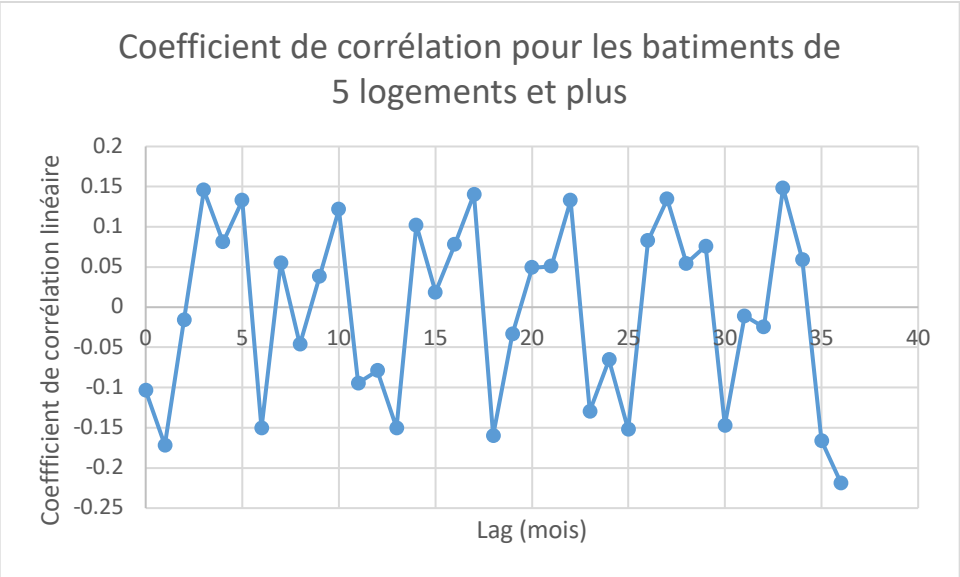


Figure 3-9 Étude de corrélation pour les bâtiments de 5 logements et plus

En prenant en considération la forte autocorrélation annuelle des validations de cartes à puce, le motif saisonnier répété chaque année dont font preuve les tracés des coefficients de corrélation est expliqué. Il est donc difficile d'avoir une idée du délai exact. Un modulo un an est donc à intégrer dans ce retard. Mais il est possible de constater que pour les bâtiments de 1 à 4 logements, le retard de l'effet de la variable est vraisemblablement de 4 mois ou de 1an et 4 mois. Mais aucune corrélation ne s'avère être significative au seuil de 0.4.

Pour les bâtiments de 5 logements et plus, l'analyse est plus difficile, car la corrélation est presque plus souvent négative que positive, ce qui ne semble pas logique dans la mesure ou plus de logements est moteur de la demande.

Dans l'optique de supprimer la composante saisonnière de la demande avant de rechercher une valeur du retard, il est possible de décomposer la demande afin de ne retenir que la tendance. Une décomposition d'une série temporelle de type X11 permet d'extraire les différentes composantes de cette série : tendance, motif cyclique et saisonnier. Ce type de décomposition se base sur une méthode additive ou multiplicative. Elle permet d'avoir des valeurs de la tendance pour toutes les valeurs observées et autorise le motif saisonnier à évoluer lentement. Il paraît donc pertinent d'extraire la tendance de la demande et d'étudier sa corrélation avec le cumul de permis de construire en termes de logements. Cette étude peut permettre de trouver une valeur plus précise du retard de l'effet de la variable.

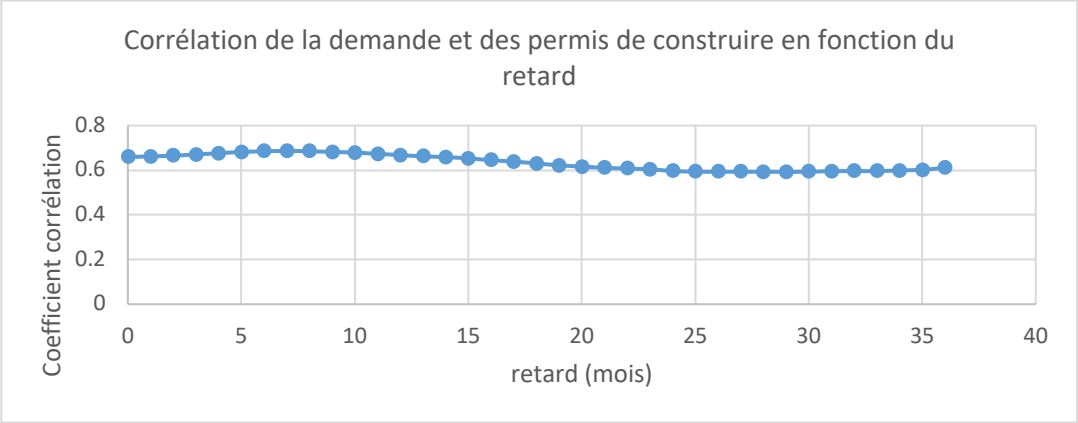


Figure 3-10 Étude de corrélation entre la demande et différents retards du cumul des permis des logements autorisés

Sur la Figure 3-10, il est possible de constater que les résultats ne sont pas concluants. Mise à part une forte corrélation entre les validations moyennes en pointe AM et le nombre de logements autorisés par des permis de construire, aucune valeur du délai de l'effet de la variable ne se distingue significativement. Il n'est donc pas possible de conclure.

3.3.1.2 Revue de la littérature

Les valeurs obtenues précédemment concernant les coefficients de corrélation entre la variable de logements et les validations en pointe AM ne produisent pas des résultats satisfaisants pour pouvoir statuer sur une valeur de ce retard. Il a donc été décidé de se pencher sur la recherche d'un temps approximatif séparant la validation du permis de construire et la prise de possession des lieux par un ménage.

Il est fait comme hypothèse que l'effet sur la demande en TC est immédiat après la prise de possession des lieux par un ménage, que tous les permis donnent lieu à une construction de logements et qu'aucun retard anormal n'est observé durant la phase de travaux ou toute autre phase. Ce délai sera étudié pour chaque type de construction et se verra attribuer un intervalle jugé raisonnable.

Les différents types de construction sont les suivants :

- Habitation unifamiliale contiguë (1 logement)
- Habitation unifamiliale isolée (1 logement)
- Habitation unifamiliale jumelée (1 logement)
- Habitation bifamiliale (2 logements)
- Habitation trifamiliale (3 logements)
- Habitation multifamiliale (4 logements)
- Habitation multifamiliale (5 logements et plus)

Pour les permis pour les habitations unifamiliales, le permis peut perdre sa validité si certains délais ne sont pas respectés. Les travaux doivent commencer dans les 12 mois suivant la délivrance du permis, ils doivent être finis dans les 18 mois suivant la délivrance et ils ne doivent pas être interrompus plus de 6 mois.

Le *United States Census Bureau* a publié une étude sur les temps de construction des nouveaux bâtiments résidentiels (https://www.census.gov/construction/nrc/lengthoftime.html). Une section détaille notamment la durée moyenne écoulée entre l'attribution du permis de construire et la fin des travaux. Ces études sont sectorisées en fonction des 4 grandes régions des US. La région

Northeast, étant la plus proche géographiquement, peut s'avérer produire des résultats similaires pour la région de Gatineau. Il faut tout de même étudier si le type de construction reste similaire entre les deux régions.

Au Canada, la plupart des maisons individuelles ou collectives et les petits bâtiments sont construits sur une dalle de béton sur laquelle est dressée une structure en bois. Certaines maisons sont également construites sur des caves en béton à la place d'une simple dalle. Le revêtement extérieur est souvent fait de bois, de briques ou de pierres. Les bâtiments de plus grande taille comme les immeubles sont souvent une combinaison d'une structure en béton, d'acier et de bois.

Les types de constructions nord-américaines ne sont pas si différents de leurs voisines canadiennes. Il ne faut pas oublier que les permis concernent des constructions récentes dont les méthodes de construction ont tendance à s'universaliser. Dans ce secteur, la majorité des « petites » constructions sont faites sur des structures en bois également. Les structures des immeubles sont d'autant plus similaires que les options pour ce type d'habitations sont moins nombreuses.

Il est donc raisonnable de considérer une similarité dans les temps séparant l'obtention du permis et la fin des travaux. Le Tableau 3-1 montre les résultats obtenus pour le Nord-Est des US. Pour déterminer un retard dans la variable des permis de construire, il faut donc considérer ces valeurs pour chaque type de construction. En considération de la typologie établie par le permis de construire canadien et celle des temps de construction américains, trois segments de logements sont considérés:

- Les logements unifamiliaux
- Les bâtiments comprenant de 2 à 4 unités de logement
- Les bâtiments comprenant plus de 5 unités de logement

Tableau 3-1 Tableau traduit de United States Census Bureau (https://www.census.gov/construction/nrc/lengthoftime.html)

	Nombre moyen de mois entre l'attribution du permis de construire et la finalisation du bâtiment								
	Bâtiments avec un seul logement				Bâtiments avec 2 logements et plus				
	Total	Motif de construction			Total	Nombre de logements dans le bâtiment			
Nord-Est		Construit pour la vente	Construit par une agence immobilière	Construit par le propriétaire		2 à 4	5 à 9	10 à 19	plus de 20
2011	8,9	7,6	9,3	13,0	17,4	15,9	26,1	12,8	21,4
2012	8,5	7,4	8,6	12,6	13,5	10,4	23,8	17,3	20,9
2013	8,4	7,2	8,9	13,8	15,1	13,8	23,1	15,6	16,0
2014	8,6	7,7	8,8	11,9	13,6	12,6	14,8	16,2	14,6
2015	8,9	8,2	9,8	10,8	13,9	13,7	12,1	14,1	16,0
2016	9,6	9,4	9,4	12,3	15,5	13,0	20,5	16,7	18,9
2017	9,5	8,1	11,1	12,9	16,6	14,7	18,3	17,2	19,6
2018	9,3	8,5	9,7	10,9	18,3	13,5	20,2	20,4	23,6
2019	11,1	9,5	11,9	15,5	19,5	15,0	20,4	29,5	22,4

Il faut maintenant déterminer une valeur du délai de la variable pour chacun de ces segments. Pour les logements unifamiliaux, la variabilité temporelle est relativement faible, mais ce n'est pas le cas pour les bâtiments avec deux unités de logement et plus. Le Tableau 3-2 expose les moyennes pour chaque catégorie de logements. Sachant que la variable des permis de construire est mensuelle, il faut avoir un nombre de mois entier. Le choix est donc de prendre l'arrondi supérieur pour chaque valeur afin de tenir compte d'un éventuel délai supplémentaire entre la fin des travaux et la prise de possession des lieux par des habitants.

Tableau 3-2 Moyenne du délai séparant l'obtention du permis et la fin des travaux

SEGMENTS DE CONSTRUCTION	LOGEMENTS UNIFAMILIAUX	BÂTIMENTS DE 2 À 4 LOGEMENTS	BÂTIMENTS DE 5 LOGEMENTS ET PLUS
MOYENNE DE 2012 À 2019 (MOIS)	9,2	13,3	18,8

Les valeurs prises finalement pour chaque catégorie sont les suivantes :

- Les logements unifamiliaux : 10 mois de délai
- Les bâtiments comprenant de 2 à 4 unités de logement : 14 mois de délai
- Les bâtiments comprenant 5 unités de logement et plus : 19 mois de délai

Il peut être également utile d'agréger les données pour avoir une valeur de retard moyen pour toutes les catégories. Pour ce faire, il faut calculer une moyenne pondérée prenant en compte le nombre de permis délivrés pour chaque type de logements. La moyenne sera pondérée en fonction du nombre de logements construits par catégorie entre 2012 et 2019 (pour les données de Gatineau).

Le résultat est de 14,47; par la suite la valeur considérée sera de 15 mois de délai. Une telle valeur permet d'actualiser les données en termes de nombre de logements issues du recensement 2016.

À partir du nombre de logements issu du recensement de 2016, il est possible d'obtenir une estimation de ce nombre dans les années qui suivent 2016 en utilisant les informations sur les permis de construire. Cette estimation est réalisée sur des zones d'agrégation hexagonales qui par la suite seront utilisées pour effectuer la prévision. Pour ce faire, plusieurs hypothèses ont été prises :

- Le retard entre l'attribution du permis de construire d'un logement et la prise de possession est de 15 mois.
- Le nombre de logements détruits est négligeable.
- La distribution des logements dans les îlots de diffusion est considérée comme constante (l'îlot de diffusion est le niveau de résolution le plus précis qui permet d'obtenir les informations sur le nombre de logements issu du recensement).

La méthode utilisée pour changer la zone d'agrégation initiale des logements du recensement (passer des îlots de diffusion aux zones hexagonales) est de faire l'hypothèse d'une distribution constante des logements puis d'évaluer le recouvrement des deux niveaux d'agrégation pour obtenir des valeurs dans les zones hexagonales.

3.3.2 Création de zones d'agrégation

Au fil de l'étude des données, il est apparu comme évident de trouver une unité d'agrégation permettant des résultats précis, mais en s'affranchissant des problématiques liées à la prévision de la demande à chaque arrêt, ceux-ci pouvant changer à travers le temps. Cette partie détaille le travail préliminaire consistant à trouver un système d'agrégation pour les données afin d'obtenir des résultats pertinents sur les futurs modèles de prévision.

3.3.2.1 Nécessité d'agréger les données

Pour prévoir de manière précise la demande, l'idéal serait de pouvoir produire une prédiction à chaque arrêt. Une multitude de modèles réalisent ce type de modèle dans la littérature. Ces modèles

sont appelés DRM (*direct ridership models*), ils se basent sur les caractéristiques d'un arrêt et de son entourage pour prévoir la demande à un arrêt précis. Mais dans le cas présent, ce type de modèle paraît difficile à mettre en plus, car la structure du réseau de la STO est en constante modification.

La Figure 3-11 illustre cette constante transformation du réseau. À savoir que la transformation importante en 2013 est due à l'implantation du Rapibus, une ligne de bus à haut niveau de service.

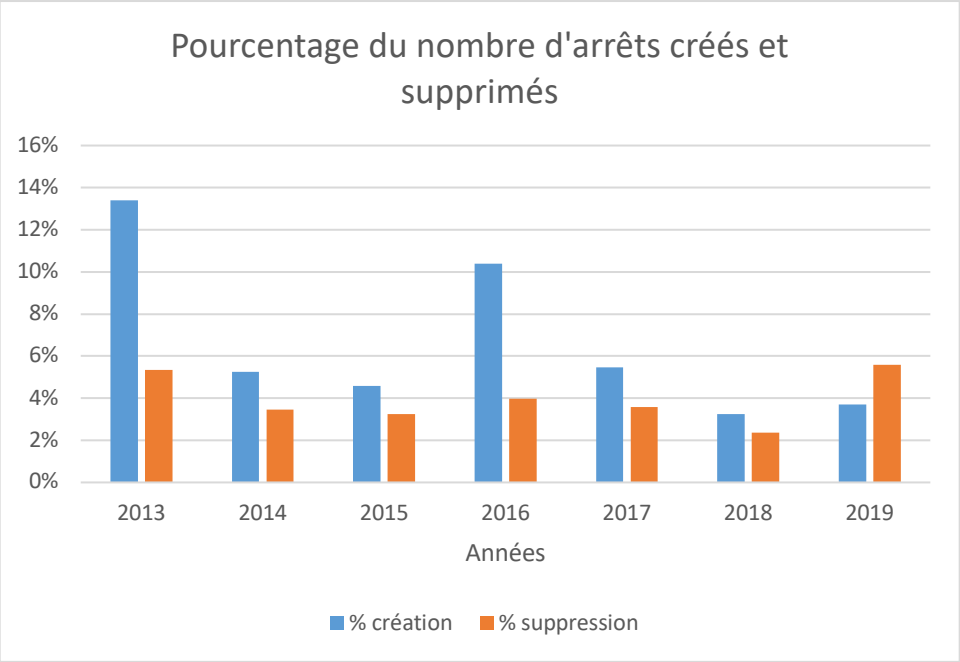


Figure 3-11 Pourcentage de création et suppression d'arrêts

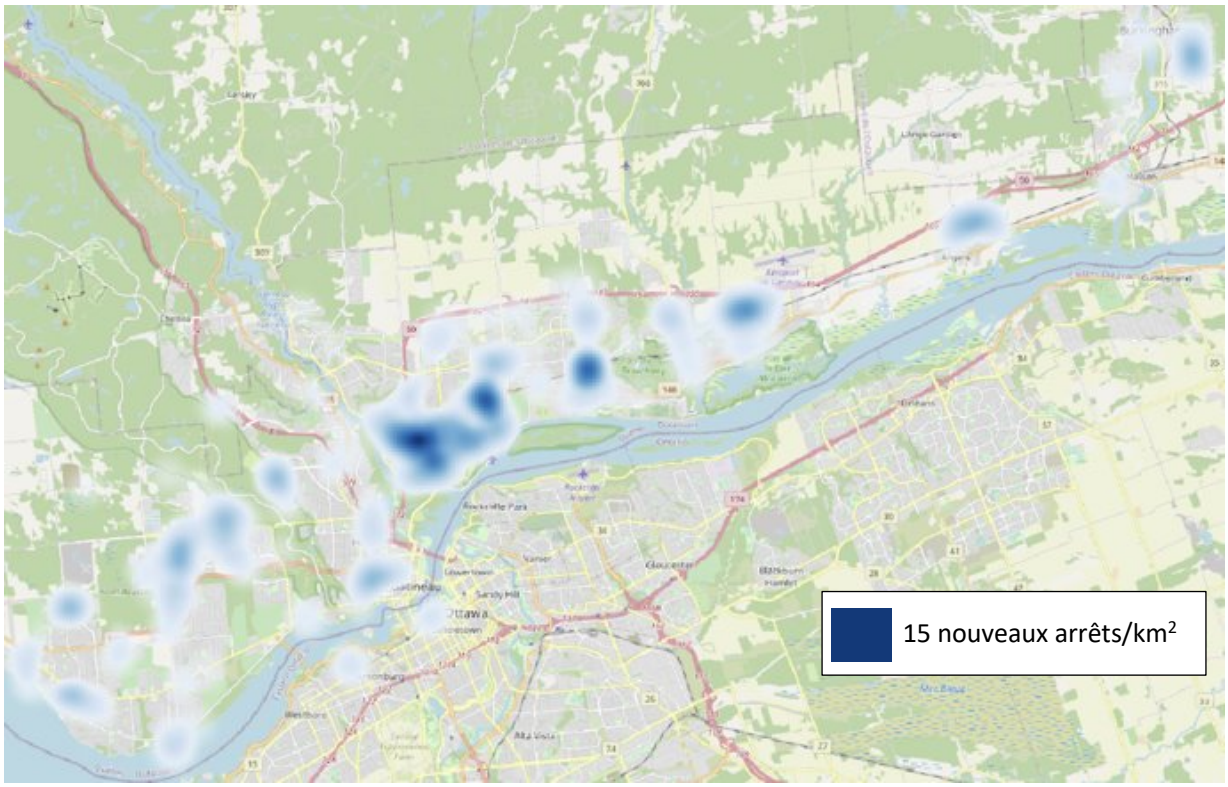


Figure 3-12 Nouveaux arrêts sans les arrêts déplacés

Figure 3-13 Schéma de l'implantation des nouveaux permis de construire

Les Figure 3-12 et Figure 3-13 reflètent la densité d'implantation de nouveaux arrêts ou de logements entre les années 2012 et 2019. Elles montrent que l'implantation des nouveaux arrêts ne peut pas uniquement être justifiée par la présence de nouveaux logements. Une forte restructuration des arrêts est constatée dans le centre de Gatineau. Une corrélation est visible pour les zones de forte densification où il est possible de constater un ajout d'arrêts de bus.

Le problème issu de cette constante modification des arrêts est à la fois la difficulté d'avoir une continuité temporelle de l'élément arrêt. Il y a également un phénomène de concurrence des arrêts entre eux. C'est-à-dire qu'un arrêt nouvellement implanté dans une zone capture une partie de la demande des arrêts voisins. Il est fréquent de constater de fortes baisses de montants à certains arrêts qui ne peuvent être expliquées uniquement par l'implantation d'un nouvel arrêt dans leur voisinage et non par d'autres variables explicatives. Cette concurrence étant difficile à modéliser dans un modèle temporel, le choix est fait de créer un schéma d'agrégation qui a comme but de s'affranchir de ces problèmes.

3.3.2.2 Analyse de sensibilité de la zone d'agrégation

3.3.2.2.1 Indicateurs et paramètres

Les résultats précédents sur les données de carte à puce sont désagrégés. Prévoir la demande en TC à l'échelle de l'arrêt s'avère être très complexe principalement à cause du phénomène de redistribution de l'achalandage sur les arrêts voisins. Cet effet produit une forte corrélation négative entre certains arrêts : l'implantation d'un nouvel arrêt attire des usagers qui par le passé utilisaient des arrêts dans le voisinage. Il est donc nécessaire de procéder à l'agrégation des données. L'enjeu est donc de déterminer le niveau d'agrégation le plus pertinent pour réaliser un modèle de prévision.

Une analyse de sensibilité est donc essentielle pour déterminer quelle agrégation est la plus pertinente pour la création d'un modèle de prévision. Aux vues des données, il est évident que l'agrégation sera spatiale.

Le premier paramètre à étudier est donc la forme que prendront les zones d'agrégation. En étudiant les données de carte à puce, il pourrait paraître pertinent d'agréger les données selon les lignes de bus. Tout au long de ces lignes, la demande est assez similaire. Mais par souci de simplicité et également pour pouvoir déterminer des matrices OD plus pertinentes par la suite, le choix est orienté vers la sélection d'un maillage régulier. Les données agrégées seront déterminées par une

grille d'une certaine forme qui sera superposée aux données désagrégées. Une forme de grille qui limite au maximum l'effet de frontière créé par une telle agrégation est une grille hexagonale. Des cellules d'agrégation hexagonale sont donc testées dans le cadre de cette analyse de sensibilité. À noter également que choisir un maillage régulier comme celui-ci permet une meilleure répétabilité pour le modèle de prévision qui va suivre.

Lorsque l'enjeu est de déterminer un maillage pertinent pour la création d'un modèle de prévision pour un réseau de bus se basant sur les caractéristiques d'un territoire, il faut dans un premier temps se demander quel territoire influe sur la demande. Dans le cas présent, il est désiré de déterminer la répartition de la demande de manière précise sur le réseau de la STO. L'idéal serait de déterminer la demande à chaque arrêt, mais il a été vu que cela présente un grand nombre de complications. Il a été décidé de considérer des zones d'agrégation hexagonale. Mais afin de prévoir la demande dans cette zone, il faut déterminer sa taille. Un des enjeux principaux est de connaître l'étendue des caractéristiques spatiales qui influencent la demande dans un hexagone.

Dans la mesure où les variables qui seront utilisées par la suite prennent compte des lieux de résidence des habitants de Gatineau, il faut se demander quelle distance les usagers de TC acceptent de parcourir pour se rendre à un arrêt de bus. Cette information donnera un ordre d'idée quant à la zone influençant la demande autour d'un arrêt. Cette distance pourra par la suite être adaptée pour les zones d'agrégation.

La STO a publié dans un *Guide d'aménagement et d'accessibilité universelle* dans le cadre de la formulation de son plan stratégique, dans lequel l'implantation des arrêts est justifiée par des distances de marche maximum. Il est convenu que pour 90% de la population de la région de Gatineau, la distance maximale d'accès à un arrêt de bus doit être de 400 mètres (distance réseau) en zone moyennement et fortement dense et de 500 mètres pour les zones faiblement peuplées. L'accès à des arrêts de bus à haut niveau de service doit également être possible pour des distances de moins de 700 mètres en zone urbaine.

Ainsi, les zones potentielles devront prendre en compte ces informations. À savoir que des zones de diamètre inférieur à 800 mètres captureront la relation entre la demande et les variables démographiques de manière plus locale et qu'il faudra considérer par la suite une zone tampon adaptée à cette distance d'accès aux arrêts pour capturer toute la population accédant aux TC dans

la zone en question (en zones moyennement et fortement denses). Des expériences ont donc été réalisées pour un maillage entre 200 et 2000 mètres de diamètre pour les zones hexagonales.

À noter que par la suite, les incréments annuels de montants en pointe du matin font référence à la différence de nombre de montants dans une zone entre les mois de novembre de deux années consécutives.

Plusieurs indicateurs ont donc été testés pour orienter le choix de la taille et de la position de la zone d'agrégation :

- Limite l'effet de redistribution de l'achalandage entre les zones : l'indicateur utilisé est l'indice de Moran I, qui reflète l'autocorrélation spatiale présente entre les zones d'agrégation. Cet indicateur est utilisé sur les incréments annuels de montants en pointe AM. Sous l'hypothèse que la demande a un comportement continu au niveau spatial, il est donc pertinent d'assumer que les zones d'agrégation auront une forte corrélation spatiale entre elles. Un fort indice de Moran I, preuve d'une forte autocorrélation spatiale, signifiera qu'il y a moins de chance que la demande se redistribue entre les zones.

- Le *Local indicator of spatial association* (LISA) est un indicateur d'autocorrélation spatiale. C'est un équivalent de l'indice de Moran, mais à un niveau local. Lorsqu'il est appliqué aux incréments de montants annuels en pointe AM, il permet de détecter l'effet de redistribution de la demande à un niveau local. L'absence de valeur négative de cet indicateur sera un des critères principaux dans la sélection d'un maillage pertinent.

- Le coefficient de corrélation entre les incréments annuels de montants en pointe AM et l'incrément du nombre de logements entre l'année précédente et la présente. En effet, l'objectif de ce maillage est de produire une prévision. La variable explicative principale du futur modèle est les données de logements actualisées sur la région de Gatineau. Afin de maximiser l'effet de cette variable lors de la prévision, il est donc légitime de maximiser la corrélation entre la demande et ces logements.

Il faut noter que l'indice Moran I et le LISA sont, dans l'étude qui suit, utilisés de manière équivalente. Lorsqu'une image globale de l'autocorrélation spatiale est nécessaire l'indice de Moran I est utilisé et le LISA pour l'aspect local. L'aspect local est néanmoins beaucoup plus

important, car pour limiter l'effet de redistribution il ne faut pas de valeurs négatives du LISA peu importe l'indice de Moran I. L'indicateur global permet toutefois de limiter le nombre de paramètres à passer en revue et ainsi limiter le nombre de cartes du LISA à visualiser.

L'étude de ces indicateurs est réalisée de la manière suivante :

- Pour chaque année entre 2013 et 2019, ces indicateurs sont calculés pour plusieurs combinaisons de paramètres (à savoir taille du maillage et position du maillage par rapport aux données)
- Pour la taille du maillage, des cellules hexagonales allant de 200 à 2000 mètres par incrément de 100 mètres sont testées.
- Pour la disposition du maillage par rapport au données, une expérience de Monte-Carlo est réalisée en raison du grand nombre de valeurs possibles. Cette disposition est déterminée par deux paramètres : un taux de décalage selon l'axe vertical et un taux de décalage selon l'axe horizontal. 200 combinaisons de valeurs aléatoires sont donc générées pour chaque taille de maillage. Cette méthode permet donc d'approcher la valeur optimale de la position du maillage en diminuant le temps de calcul.

Après l'analyse des résultats de cette étude de sensibilité, l'enjeu est de déterminer des paramètres d'agrégation pertinents pour chaque indicateur calculé. En prenant en compte uniquement une année et un indicateur, un maillage idéal pourra être déterminé, mais l'enjeu est de déterminer le maillage le plus pertinent pour l'ensemble des années et des indicateurs.

3.3.2.2.2 Poids de chaque indicateur

Lors de la création d'un modèle multiniveau qui sera vu par la suite (Section 6.2.2.1.1), un outil intéressant permettant de quantifier l'effet de plusieurs variables indépendantes sur la variable dépendante emboîtée a été utilisé. C'est un modèle de croissance inconditionnelle, qui consiste à établir une hiérarchie dans les effets des variables indépendantes. Prenons l'exemple des résultats obtenus sur l'année 2019 pour le coefficient d'autocorrélation spatiale. Chaque groupe d'expérience sur une certaine taille de maillage est assimilable à la création de classes, dans le cas présent il y a 20 classes différentes qui représentent les tailles des maillages (entre 200 et 2000 mètres). À l'intérieur de ces classes, il y a les résultats de l'indice de Moran I pour chaque position du maillage (soit 200 valeurs pour chaque taille de maillage). Un modèle de croissance

inconditionnelle permet de calculer différents termes de variance et ainsi de savoir si les variations de la variable dépendante sont liées aux différences entre les classes (c.-à-d.. la taille du maillage) ou aux différences internes aux classes (c.-à-d.. la position de la grille). Il est donc possible de savoir lequel des paramètres déterminant le niveau d'agrégation a le plus d'effet sur l'indicateur.

Tableau 3-3 Résultats étude de variance

Indicateur	Autocorrélation spatiale (Moran I)		Coefficient de corrélation permis de construire/demande (Pearson)	
Années	Corrélation interclasse	Corrélation intraclasse	Corrélation interclasse	Corrélation intraclasse
2019	0,46	0,54	0,27	0,73
2018	0,56	0,44	0,41	0,59
2017	0,64	0,36	0,52	0,48
2016	0,65	0,35	0,40	0,60
2015	0,62	0,38	0,13	0,87

Il apparaît donc que la taille du maillage comme la disposition influencent toutes deux de manière significative les indicateurs. Pour le coefficient d'autocorrélation spatiale, la taille du maillage a un peu plus de poids que la position. Mais concernant le coefficient de corrélation permis/demande, il s'avère que la position influence plus le résultat que la taille du maillage.

Pour l'indice de Moran, il est donc acceptable de trouver une plage de taille de maillage acceptable et d'identifier une position pertinente par la suite. Pour le coefficient de corrélation de Pearson, il influencera probablement plus la position du maillage que sa taille. L'étude est plus compliquée, car il faut étudier les positions pour un nombre plus élevé de tailles de maillage.

3.3.2.2.3 Expérience sur les paramètres

Dans un premier temps, la méthode consiste à étudier le paramètre de la taille du maillage indépendamment de la position de la grille. Pour ce faire, deux paramètres sont étudiés, le

maximum et la moyenne des valeurs observées par taille de maillage. Le calcul de la moyenne et du maximum est réalisé pour chacun des deux indicateurs. L'objectif est d'avoir une idée de la répartition des valeurs de ces indicateurs pour chacune des observations réalisées et regroupées par taille de maillage. Par la suite, une sélection de tailles de maillage sera sélectionnée pour étudier l'influence de la position de la grille est ainsi conclure sur la position et la taille la plus pertinente aux vues des indicateurs.

Comme mentionné précédemment, le LISA est l'indicateur le plus important pour cette étude de sensibilité. Pour une combinaison de paramètres, il est possible de produire une carte du LISA pour chacune des zones. Passer en revue 200 cartes par taille de maillage est une tâche impossible. Il faut donc utiliser des indicateurs de la distribution des valeurs du LISA pour chacune des combinaisons des paramètres. Les cartes suivantes font l'inventaire de plusieurs indicateurs de la distribution du LISA, rassemblés par taille de maillage :

- Sur chacune des combinaisons uniques des paramètres du maillage, une valeur moyenne du LISA est générée et ensuite pour l'ensemble des combinaisons correspondant à une taille de maillage, une valeur moyenne de ces valeurs moyennes est générée. Cela est réalisé pour chaque année sur la Figure 3-14.

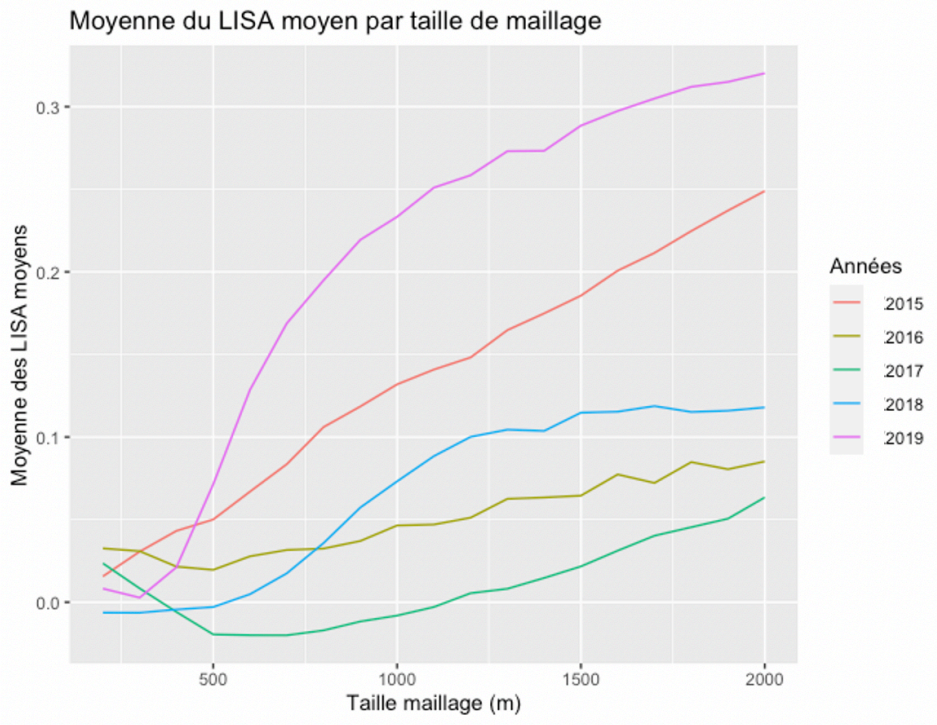


Figure 3-14 Moyenne du LISA moyen par taille de maillage

- Dans la mesure où la recherche de la position optimale du maillage par rapport au territoire d'étude est également une donnée importante, il est pertinent d'aller voir le maximum par taille de maillage du LISA moyen généré par une seule expérience. L'idée est de connaître la valeur du LISA moyen pour la position la plus optimale (au regard du LISA moyen). Le graphe est disponible sur la Figure 3-15.

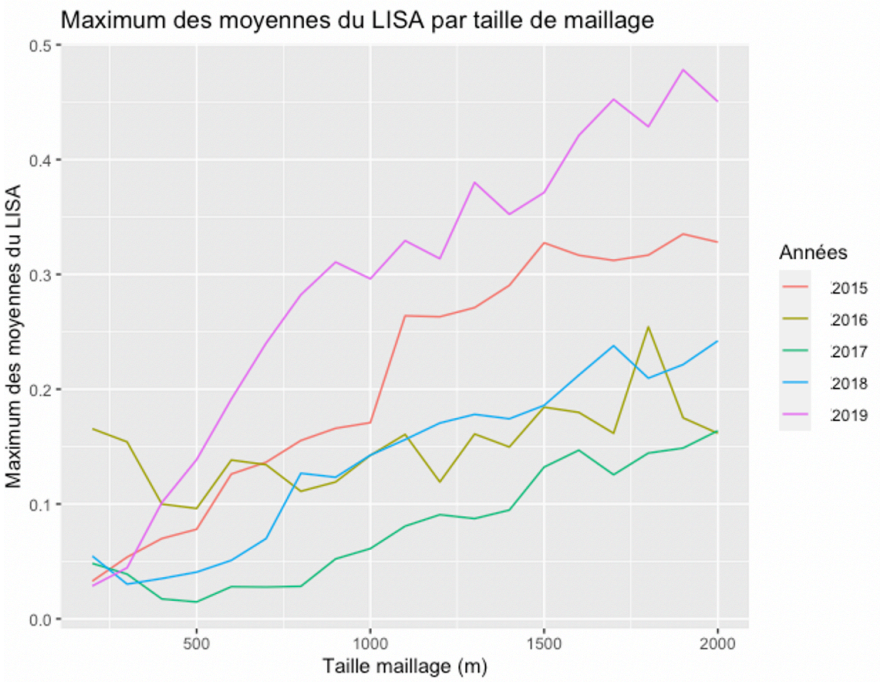


Figure 3-15 Maximum des LISA moyen par taille de maillage

- La même méthode est appliquée au premier quartile plutôt que la moyenne du LISA pour une combinaison de paramètres donnée. Le premier quartile permet de situer plus précisément la distribution des valeurs du LISA. La Figure 3-16 et la Figure 3-17 donnent respectivement la moyenne et le maximum pour chaque taille de maillage.

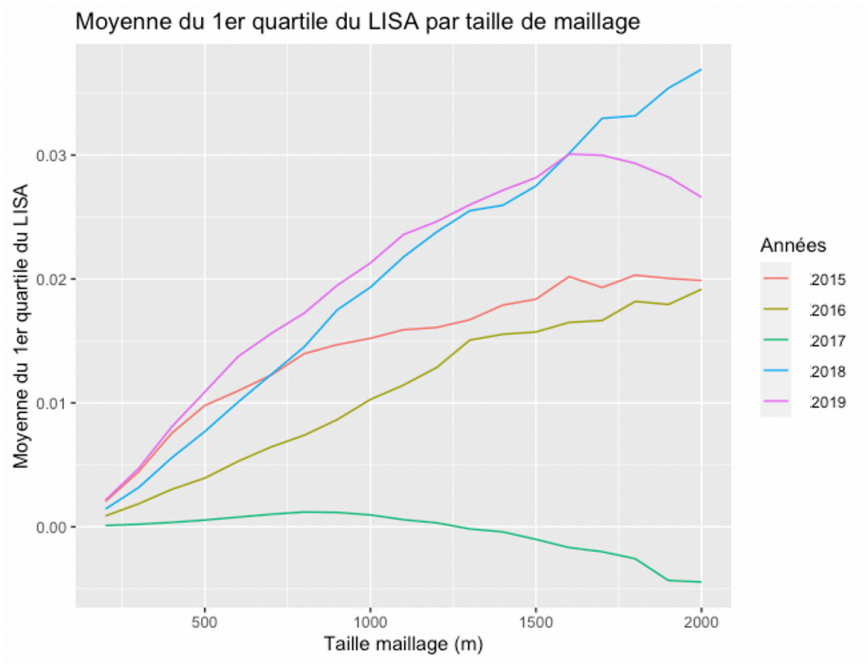


Figure 3-16 Moyenne du 1er quartile du LISA par taille de maillage

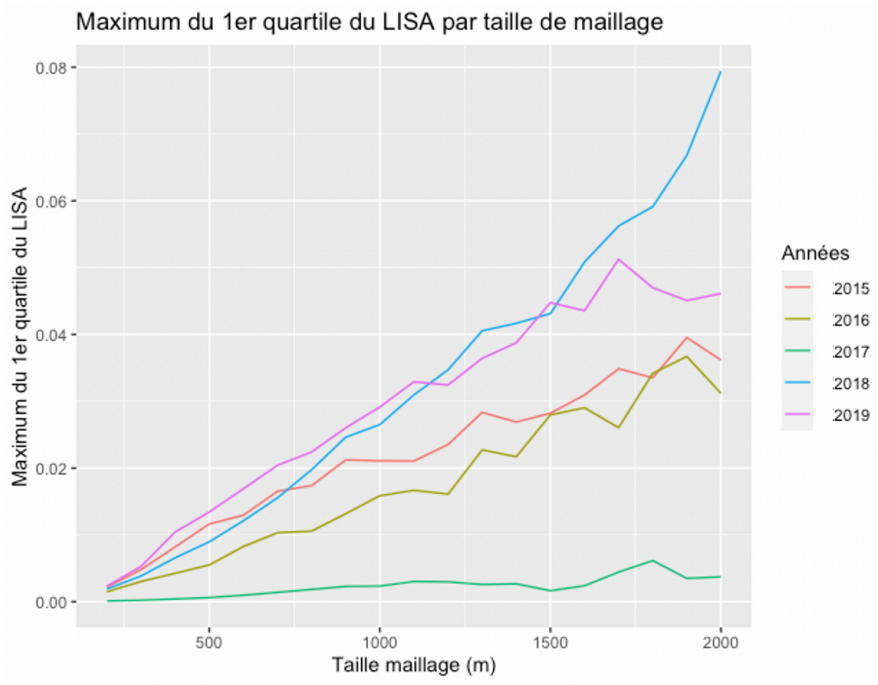


Figure 3-17 Maximum du 1er quartile du LISA par taille de maillage

L'objectif est d'obtenir une taille de zone d'agrégation pertinente pour la prévision à un niveau de résolution spatiale proche de l'arrêt. Il faut donc déterminer un seuil à partir duquel les indicateurs sont acceptables aux vues des limitations qu'ils impliquent.

Sur certaines courbes, il est possible de remarquer un point d'inflexion à environ 900 mètres de diamètre de maillage. Le point d'inflexion peut signifier que pour des diamètres supérieurs à 900 mètres, le gain lié à l'agrandissement d'un incrément de 100 mètres des zones d'agrégation est moindre que pour les valeurs inférieures à 900 mètres. Il est donc intéressant de pousser plus loin l'étude de ce point.

La position du maillage par rapport au territoire est le prochain paramètre à déterminer. Cette position est déterminée par deux paramètres : un décalage selon l'axe x (est-ouest) et un autre selon l'axe y (nord-sud). À noter que la taille du maillage peut changer par la suite si aucune des positions n'est jugée pertinente. Pour définir les dimensions d'un hexagone, le diamètre du cercle inscrit est utilisé. Après avoir statué qu'une étude de la position du maillage est nécessaire pour 900 mètres de diamètre, il est possible de réaliser une interpolation des combinaisons de valeurs déterminant la position du maillage par rapport au territoire obtenu aléatoirement par l'expérience de Monte-Carlo pour une taille de maillage donnée. L'interpolation est nécessaire, car elle permet d'avoir des valeurs du LISA de manière continue en fonction des deux paramètres de la position du maillage. L'interpolation est une pondération inverse à la distance de coefficient de décroissance $1/d^2$ où d est la distance. La distance représente celle entre le point interpolé et un point mesuré voisin. C'est une méthode linéaire permettant de donner une valeur aux points de l'espace à partir d'un échantillon de points connus.

L'interpolation est réalisée sur les valeurs du LISA moyen de chaque combinaison de paramètres unique. Par la suite, plusieurs combinaisons sont sélectionnées et la carte des LISA est étudiée plus en détail. La Figure 3-18 donne, pour chaque année, la valeur du LISA moyen en fonction du décalage du maillage sur x et y.

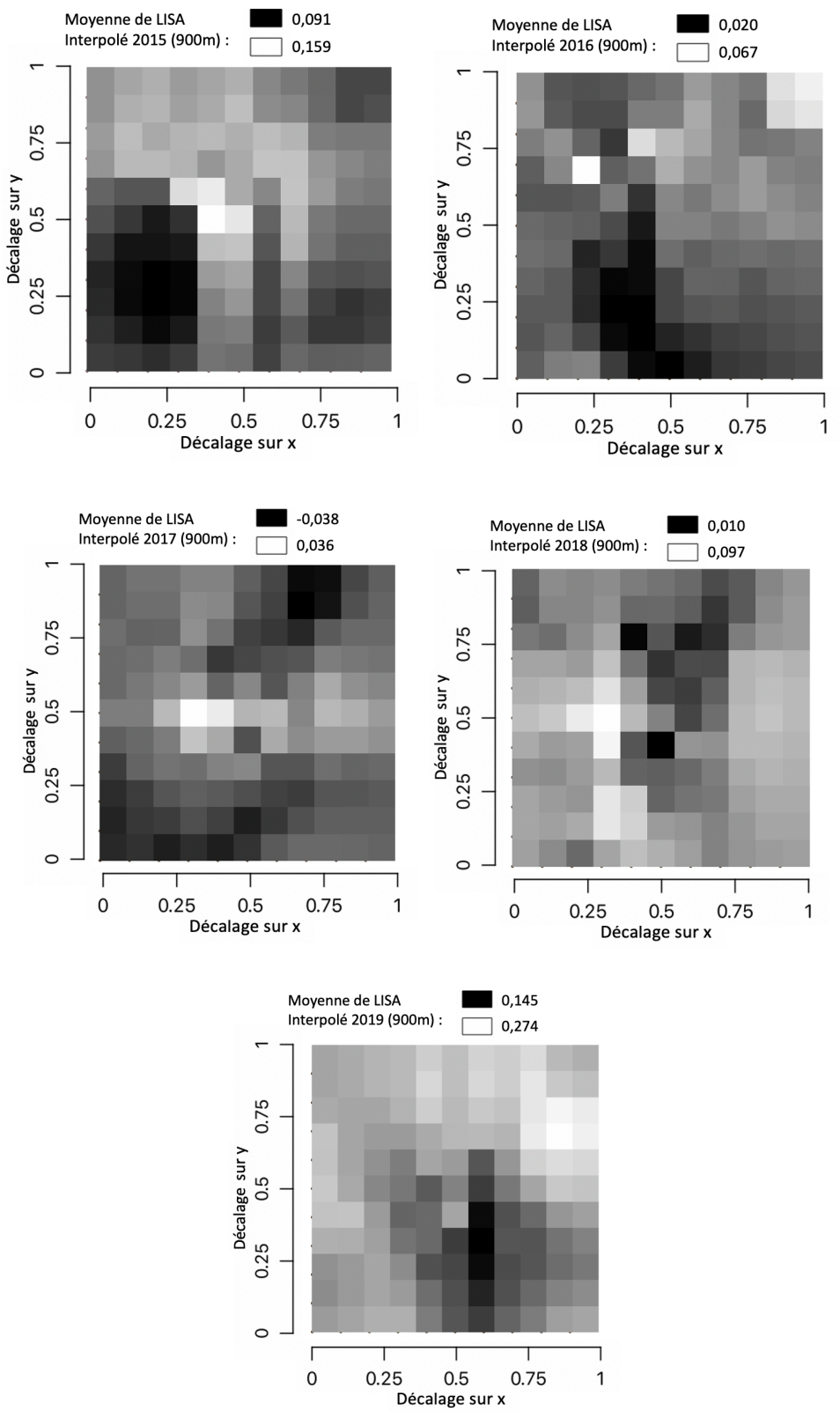


Figure 3-18 Valeurs du LISA moyen interpolées en fonction de la position du maillage

Ces résultats sont satisfaisants au regard des résultats d'autres maillages passés en revue. Un exemple d'un résultat peu pertinent est illustré par un maillage pris au hasard sur la Figure 3-19. Il y figure plusieurs zones où la corrélation spatiale est très forte négativement et beaucoup de zones sont corrélées négativement d'un point de vue spatial.

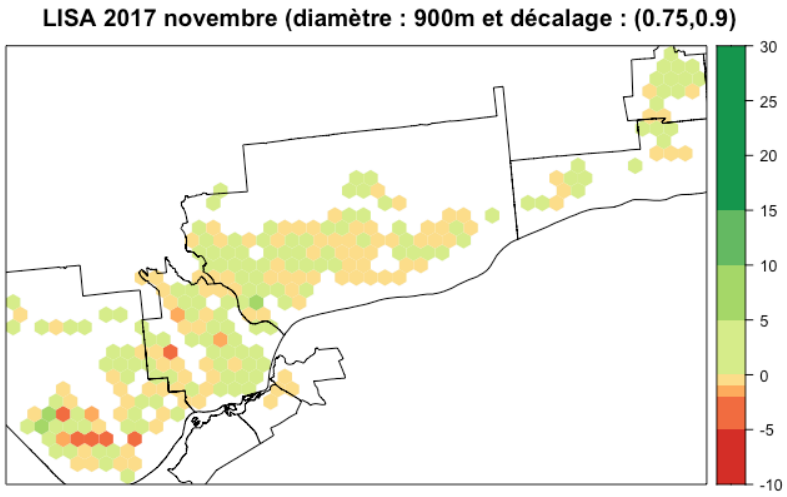


Figure 3-19 Exemple d'un maillage peu pertinent

L'objectif ici n'est pas de trouver un maillage sans aucune corrélation spatiale négative, car cela n'existe pas dans le cas présent. En effet, aucun maillage ne permet de modéliser pertinemment tout le territoire de Gatineau et toutes les années. L'objectif est de minimiser l'apparition de cette corrélation négative. Le maillage de 900 mètres associé à un décalage de 0.3 sur x et 0.5 sur y répond à ces attentes. C'est le maillage de plus pertinent passé en revue, les valeurs du LISA sont illustrées dans les cartes de la Figure 3-20 pour les années 2016-2019.

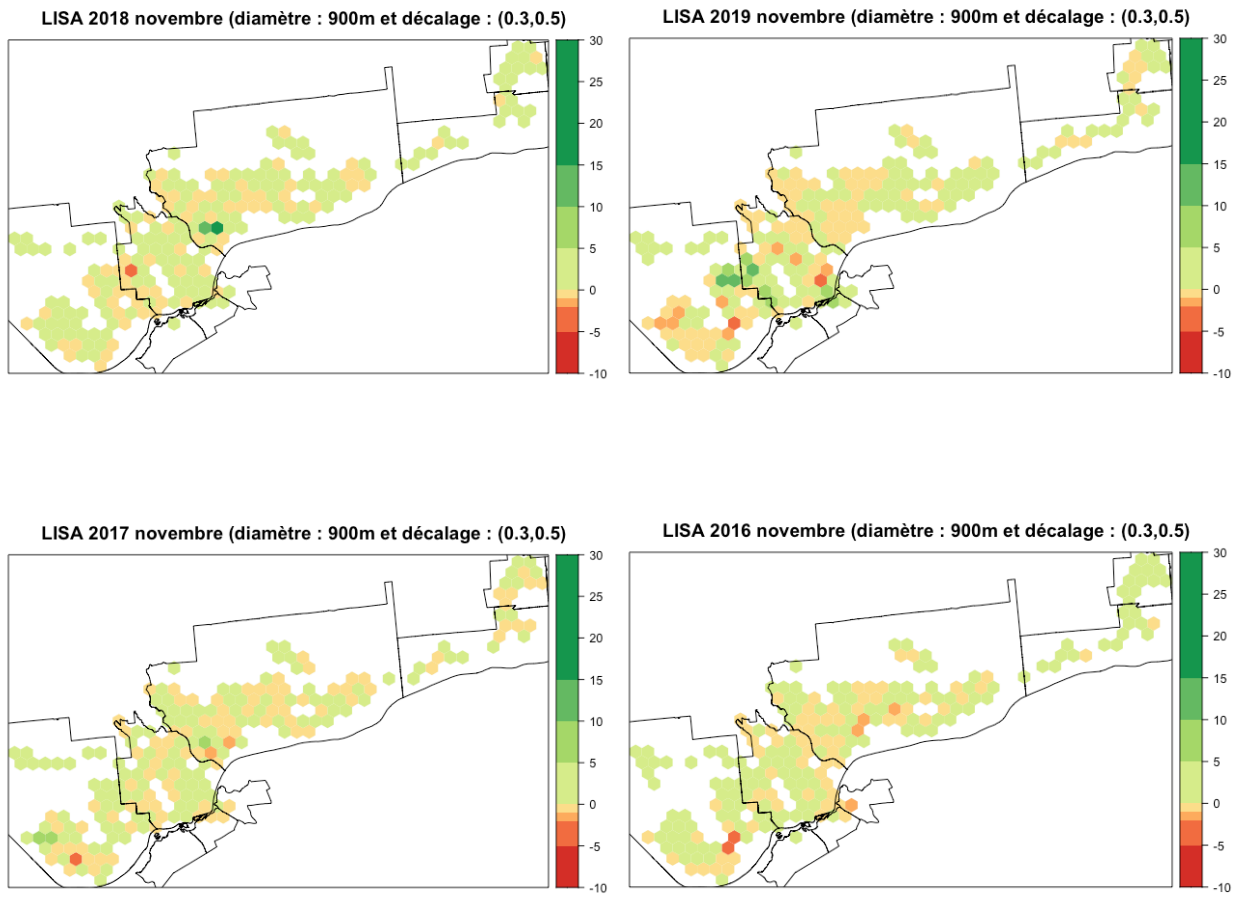


Figure 3-20 LISA du maillage le plus pertinent pour les années 2016-2019

Par souci de simplicité, les zones d'agrégation du recensement ont également été testées avant le maillage hexagonal. Les données les plus désagrégées possibles issues des recensements sont regroupées en îlots ou aires de diffusion. C'est la dénomination spatiale la plus fine qui peut être obtenue; il est donc plus facile d'extraire les variables explicatives des futurs modèles sur ces zones d'agrégation. Une analyse du LISA pour les îlots et les aires de diffusion est donc réalisée et les cartes associées sont disponibles sur les Figure 3-21 et Figure 3-22. Il est possible de constater le grand nombre de valeurs négatives du LISA ce qui indique que ces zones d'agrégation ne sont pas pertinentes. Le maillage déterminé précédemment est donc celui conservé pour la suite du projet.

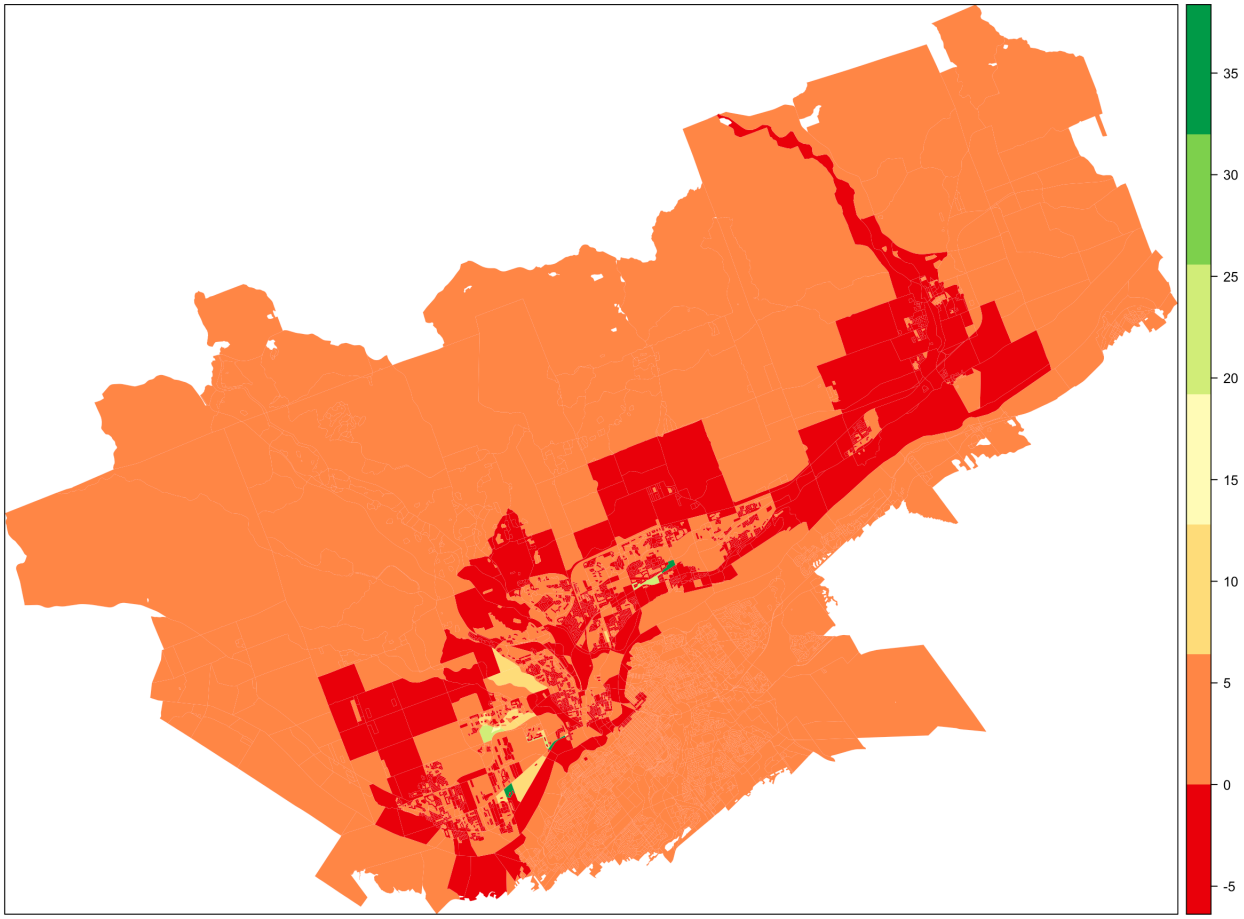


Figure 3-21 LISA des incréments de validations en pointe AM pour le mois de novembre 2019 agrégé selon les îlots de diffusion

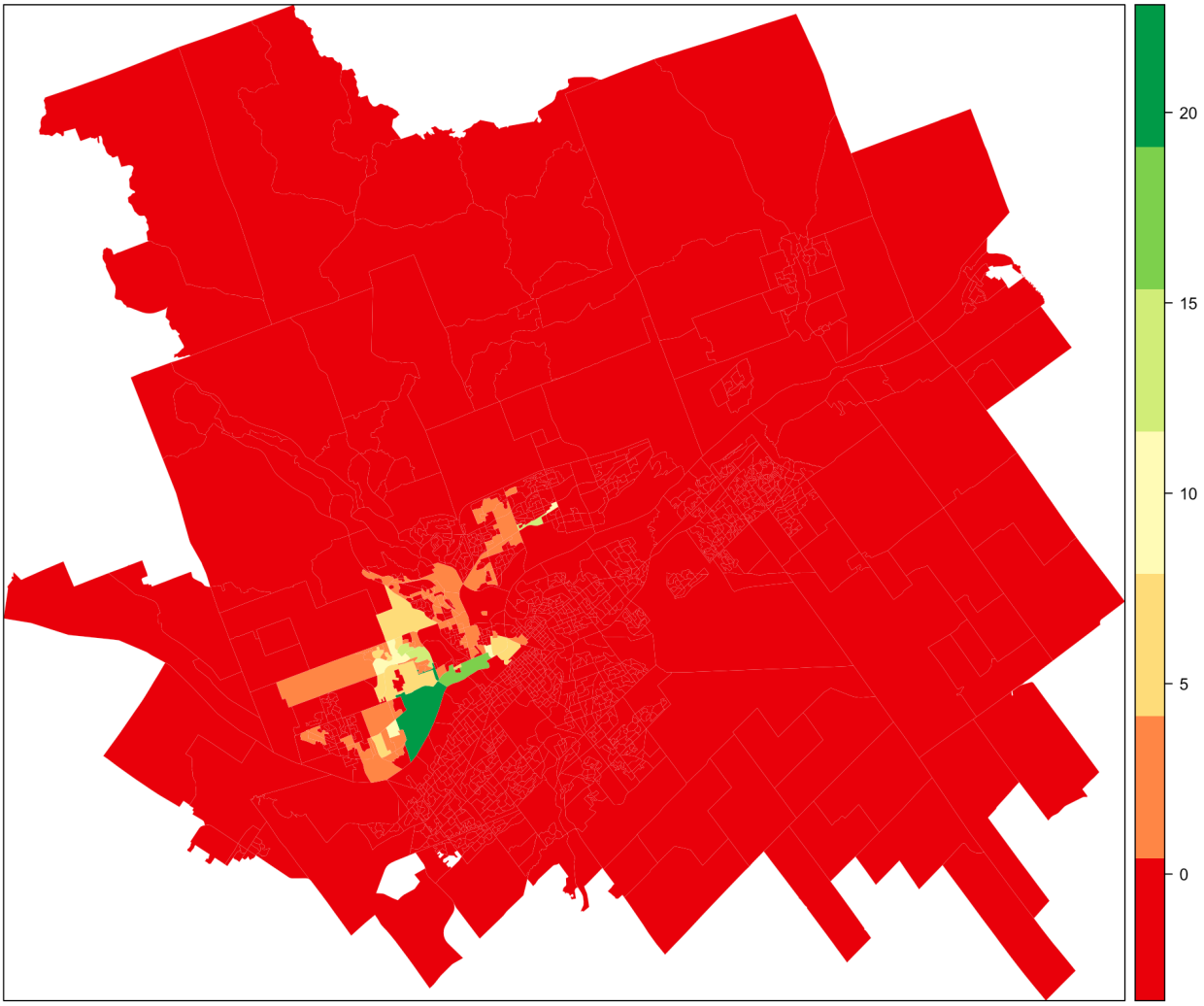


Figure 3-22 LISA des incréments de validations en pointe AM pour le mois de novembre 2019 agrégé selon les aires de diffusion

La présentation des deux travaux majeurs préliminaires aux études qui suivent étant terminée, il est maintenant possible de comprendre la structure du mémoire. Les chapitres suivants se baseront également en partie sur ces travaux.

CHAPITRE 4 ANALYSE RÉTROSPECTIVE

La compréhension des mécanismes régissant l'évolution de la demande en transport en commun est la première étape dans la mise en place de modèles de prévision. La revue de la littérature permet d'obtenir des informations sur les évolutions de la demande en fonction de différentes variables explicatives. Mais il a également été constaté que ces constats dépendent fortement du territoire d'étude. Le chapitre précédent permet d'obtenir des pistes pour identifier des variables explicatives pertinentes à intégrer dans le modèle de prévision. Mais, elle ne remplace en aucun cas une étude rétrospective du réseau de la STO. Étudier de manière approfondie les données décrivant la demande en TC permet de s'approprier le territoire et d'identifier des caractéristiques majeures de la zone d'étude influençant la demande.

Ce chapitre est organisé à l'image de la méthodologie générale de ce projet consistant à étudier plusieurs niveaux d'agrégation en allant du plus agrégé au moins agrégé. Dans un premier temps, des données concernant la globalité du territoire d'étude sont analysées. Les secteurs municipaux sont le suivant niveau d'agrégation étudié. Les zones d'agrégation hexagonale déterminées précédemment sont passées en revue par la suite et enfin, la demande la plus désagrégée possible, la demande au niveau des arrêts de bus permettra de prendre des décisions concernant le niveau d'agrégation le plus pertinent à adopter pour la suite du projet. L'analyse de la demande et des caractéristiques du territoire au niveau des arrêts justifiera le fait qu'aucun modèle n'est calibré à cette échelle.

4.1 Analyse rétrospective du territoire global

L'analyse du territoire global est la première étape, car outre le fait que la majorité des données sont disponibles uniquement à ce niveau d'agrégation, elle permet d'obtenir une vue d'ensemble des mécanismes régissant la demande en TC. Elle permettra d'orienter les analyses sur les niveaux d'agrégation plus fins.

Afin de mener l'analyse rétrospective sur le territoire global, trois études sont menées. En premier, une étude de corrélation de toutes les variables disponibles est réalisée. Cela a pour but de déterminer les dynamiques qui lient les différents indicateurs entre eux et notamment leur relation avec la demande en TC. Par la suite, une synthèse de l'évolution de ces variables est réalisée pour se rendre compte des changements que le territoire a subis. Enfin, une étude est menée sur le taux

d'utilisation du réseau de bus de la STO en fonction de plusieurs segments de la population. Le but est de savoir si la demande de certains segments peut être uniquement expliquée par une variation de la population composant ce segment.

4.1.1 Analyse de corrélation des variables du rapport

En s'appuyant sur les variables les plus utilisées dans la littérature, une analyse de corrélation croisée doit être menée pour essayer de mieux comprendre les facteurs expliquant la hausse de l'achalandage. Un catalogue de l'évolution d'une multitude de variables a été réalisé pour servir de contexte et ainsi comprendre la dynamique de la région. Pour mieux comprendre leur interdépendance et leur effet sur la demande, un lot de 10 variables est proposé dans les Tableau 4-1 et 4-2 pour effectuer une analyse de corrélation croisée. Chaque variable est sélectionnée en gardant en tête que la comparaison doit être possible (type de données, unité spatiale et unité temporelle). C'est pour cela que sont associées à chaque donnée, ses bornes spatio-temporelles.

Tableau 4-1 Proposition de variables pour l'analyse de corrélation croisée

Variables	**Bornes Spatiales**	**Bornes Temporelles**	**Sources**
Véhicule.km des transports en commun	Réseau de la STO	De 2012 à 2019 par incrément d'un an (sauf en 2016)	Données GTFS
Taux de croissance de la validation des titres de TC	Réseau de la STO	De 2001 à 2018 par incrément d'un an (2008 à 2012 réunis) pour un jour moyen de ces années	Données de la STO
Nombre de véhicules par permis	Secteur Gatineau	1986 – 1995 – 2005 - 2011	Enquête OD

Tableau 4-2 Proposition de variables pour l'analyse de corrélation croisée

Variables	**Bornes Spatiales**	**Bornes Temporelles**	**Sources**
Population	RMR Ottawa-Gatineau (partie Québec)	De 2006 à 2019 par incrément d'un an	Statistique Québec
Pourcentage de déplacements pour motif travail	Secteur Gatineau	1986 – 1995 – 2005 - 2011	Enquête OD
Longueur du réseau routier	Secteur Gatineau	2001 – 2006 – 2011 - 2016	Fichiers principaux de routes
Nombre d'unités de logement	RMR Ottawa-Gatineau (partie Québec)	De 1996 à 2016 par incrément de 5 ans	Données recensement
Prix essence	Outaouais	De 2008 à 2019 par incrément d'un an	Régie de l'énergie du Québec
Durée de l'activité travail	Secteur Gatineau	1986 – 1995 – 2005 - 2011	Enquête OD
Part modale TC	Secteur Gatineau	1986 – 1995 – 2005 - 2011	Enquête OD

4.1.1.1 Recouvrement temporel des variables

Pour mener l'étude de corrélation, des données comparables sont essentielles. Les données disponibles font référence à la même zone géographique donc, d'un point de vue spatial, les données sont comparables. Le recouvrement temporel des données est également important. La

chronologie (Figure 4-1) suivante représente la disponibilité de chaque donnée sélectionnée. Pour chaque variable, une valeur est disponible pour chaque année mentionnée sauf en cas d'indication contraire.

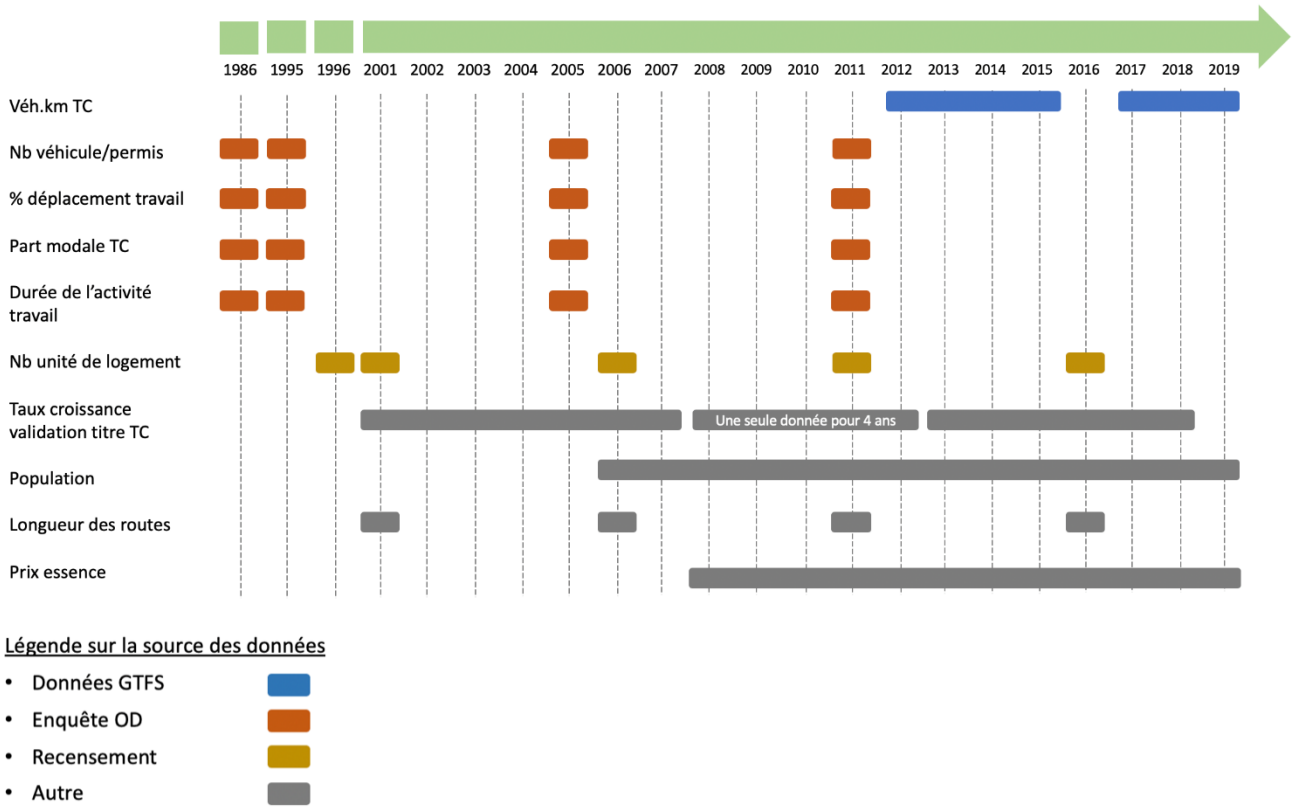


Figure 4-1 Disponibilité des données dans le temps

Par la suite, d'autres données sont ajoutées pour tenter de percevoir plus de relations entre les variables. Mais cette figure met en avant la nécessité de trouver un moyen d'avoir un recouvrement temporel des variables.

4.1.1.2 Préparation des données

Les données obtenues ont comme précisé précédemment un problème de recouvrement temporel. Un travail est donc nécessaire pour pouvoir réaliser une analyse de corrélation. Il faut noter que toutes les données ont été fournies sur une zone géographique comparable : la région métropolitaine de recensement Ottawa-Gatineau (rive Gatineau uniquement).

L'objectif de la préparation des données est d'obtenir des valeurs comparables à la fois spatialement et temporellement pour chaque variable et exécuter une analyse de corrélation linéaire sur ces données. Plusieurs ensembles de données sont générés afin d'obtenir des résultats sur chaque méthode et ainsi limiter le biais lié aux différentes approximations.

Pour résoudre ce problème, plusieurs méthodes sont envisagées :

- Prendre uniquement les années disponibles pour la plupart des variables : seulement trois années peuvent être conservées pour cette analyse, ce qui réduit beaucoup les informations contenues dans les données.
- Estimer chaque variable par une courbe de tendance linéaire et compléter les années manquantes avec l'équation de la courbe linéaire : lisse beaucoup les variations et ne prend en compte que la tendance générale.
- Interpoler linéairement les années comprises entre deux valeurs disponibles : cette méthode tient mieux compte des variations à court terme. Mais pour prévoir les valeurs extérieures à l'intervalle des années disponibles, il faut avoir recourt à une autre méthode : on utilise la courbe de tendance linéaire obtenue grâce à la méthode précédente.

Les résultats de chacune de ces méthodes seront comparés par la suite en considérant les limites induites par leur méthodologie. Les méthodes et les limites vont être détaillées dans les prochaines parties.

4.1.1.2.1 Méthode des valeurs réelles

Cette méthode est la plus simple, mais également celle comprenant le plus de limitations. Devant le manque de recouvrement des années disponibles, seuls trois groupes d'années peuvent être conservés pour comparer les données. Ces données comparées sont celles issues des enquêtes OD et des recensements.

Les trois groupes d'années observées sont : 1995-1996, 2005-2006 et 2011.

Le problème ici est que l'analyse de corrélation est appliquée à un ensemble de 3 valeurs par variables ce qui est extrêmement faible pour ce genre d'étude. Des résultats sont produits, mais leur pertinence n'est pas garantie. Ils sont uniquement là pour être comparés aux autres méthodes et appuyer ou non les constats qui pourront être faits.

4.1.1.2.2 Méthode de la courbe de tendance linéaire

Pour résoudre le problème dû aux années manquantes pour certaines données, il faut projeter les données disponibles pour estimer les lacunes. Dans la mesure où l'analyse qui suit est une analyse de corrélation linéaire, il peut être suffisant d'estimer les données de manière linéaire. Pour cela, il a été décidé de tester une modélisation de régression linéaire des variables en fonction des années (y = a*années + b). Les données disponibles sont estimées grâce à des courbes de tendance linéaire. L'étude est réalisée avec le logiciel Excel (minimisation de l'erreur pour trouver la courbe de tendance). Les figures suivantes font état des régressions linéaires pour chaque variable.

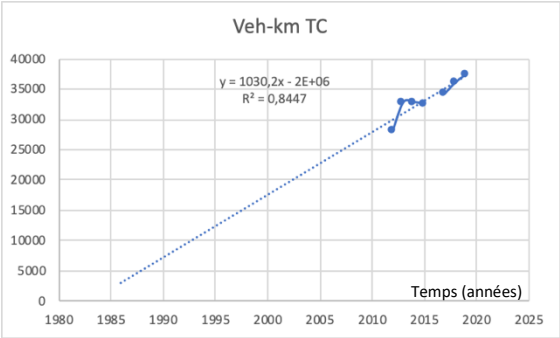


Figure 4-2 Véhicule km pour les TC

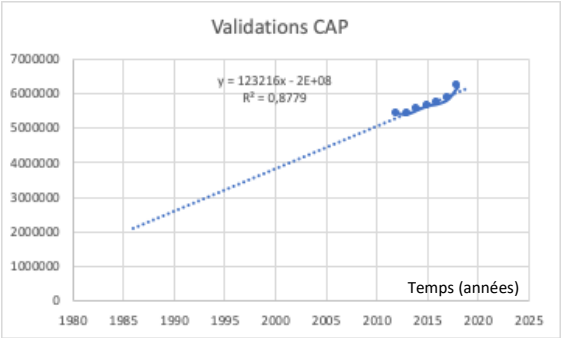


Figure 4-3 Validations carte à puce

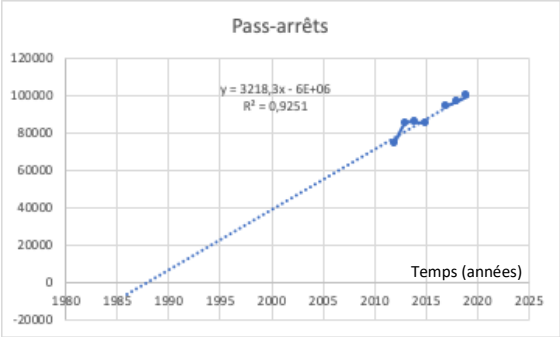


Figure 4-4 Passage arrêt TC

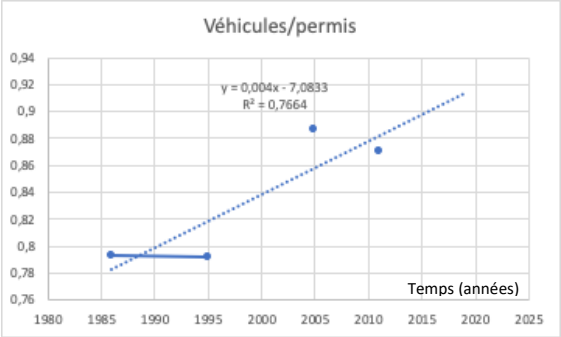


Figure 4-5 Véhicules par permis

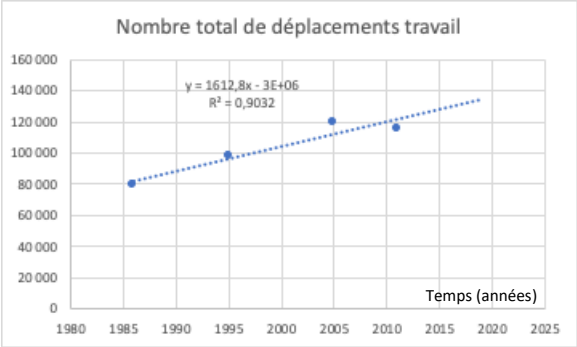


Figure 4-6 Nombre de déplacements total pour motif travail

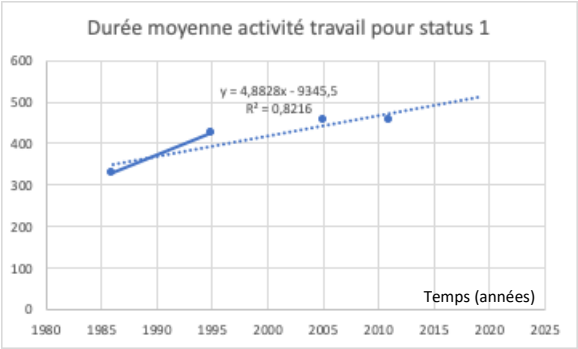


Figure 4-7 Durée moyen de l'activité travail pour les travailleurs

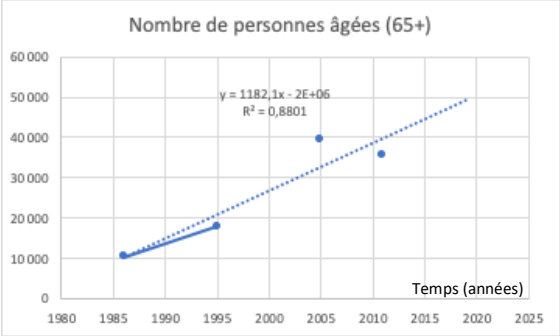


Figure 4-8 Nombre de personnes âgées

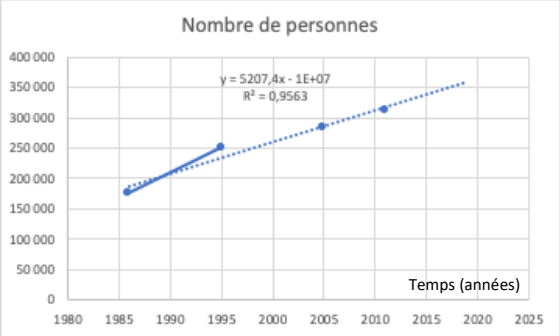


Figure 4-9 Nombre de personnes

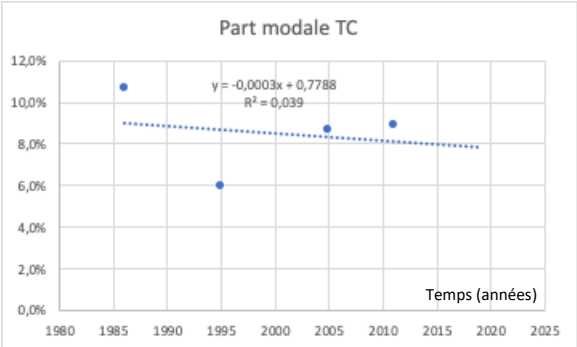


Figure 4-10 Part modale TC

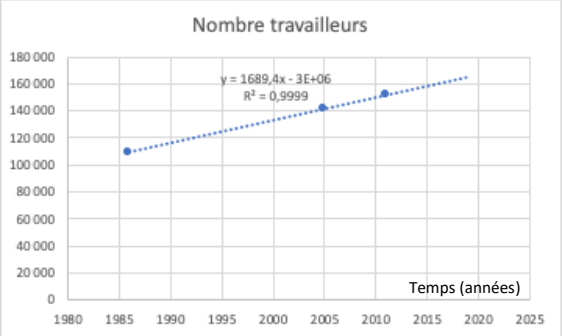


Figure 4-11 Nombre de travailleurs

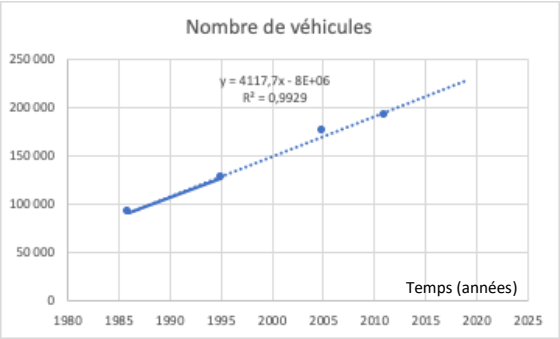


Figure 4-12 Nombre de véhicules

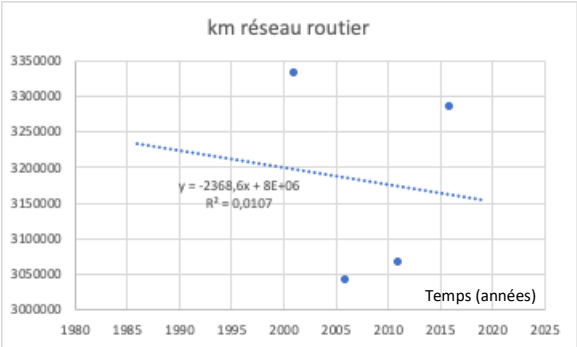


Figure 4-13 Longueur du réseau routier

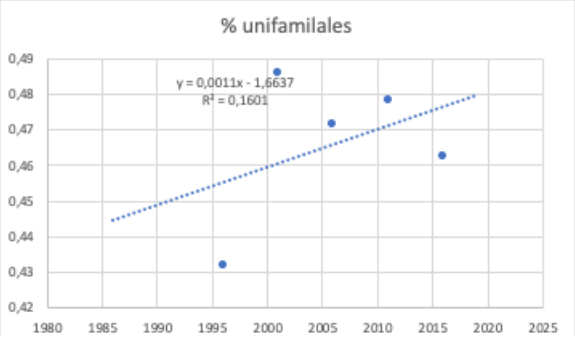


Figure 4-14 Pourcentage de logements unifamiliaux

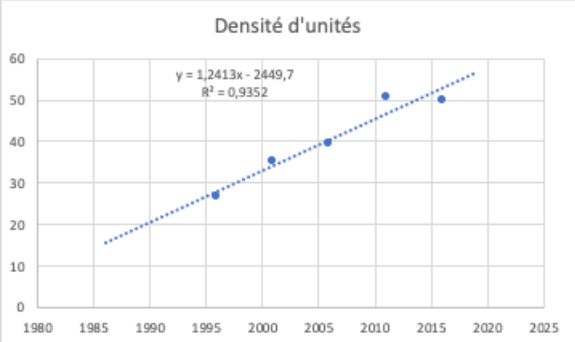


Figure 4-15 Densité d'unités de logements

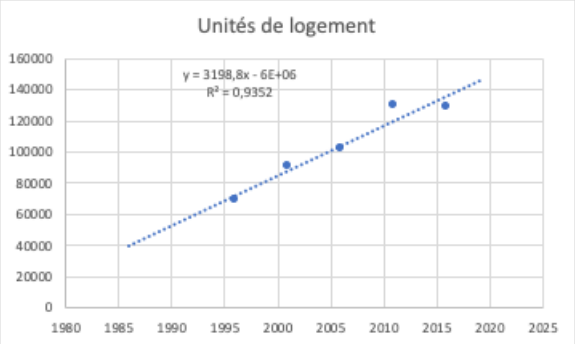


Figure 4-16 Nombre d'unités de logements

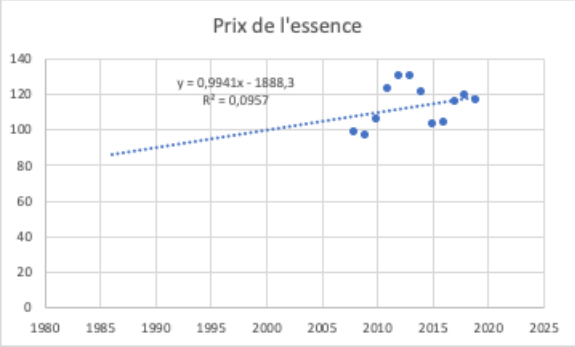


Figure 4-17 Prix du litre d'essence en cent

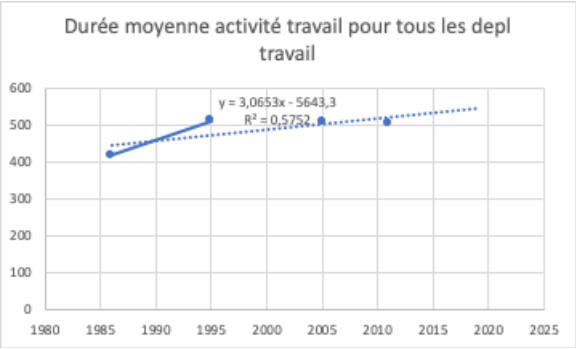


Figure 4-18 Durée moyenne de l'activité travail pour tous les déplacements travail

Avec les équations des courbes de tendance, un tableau allant de 1986 à 2019 avec toutes ces données projetées linéairement a été complété, mais en gardant les valeurs réelles pour les années disponibles. Certaines variables comme la part modale TC, le prix de l'essence, le pourcentage d'unifamiliales et la longueur du réseau routier ne suivent clairement pas une évolution temporelle linéaire. Leur coefficient de corrélation R^2 est inférieur à 0,2. Leurs résultats sont tout de même inclus dans l'étude de corrélation, mais leurs résultats sont à prendre avec du recul. Une telle approximation se justifie néanmoins au regard de l'analyse de corrélation qui est effectuée : elle est linéaire donc elle ne prend pas en compte les comportements non linéaires. Les autres approximations sont satisfaisantes avec un résultat de R^2 supérieur à 0,75.

La grande limitation de cette méthode est le fait qu'approximer la variable par une courbe de tendance lisse totalement les variations de courts et moyens termes. Seule la tendance à long terme est constatée à savoir la croissance ou la décroissance. Les variations non linéaires sont omises, mais il faut savoir que les corrélations non linéaires seront également absentes de l'étude comme cette dernière étudie les liens linéaires. Une simple étude linéaire est justifiée par les modèles qui seront utilisés par la suite pour produire une prévision.

4.1.1.2.3 Méthode d'interpolation linéaire

Pour tenter de combler les limitations de la projection précédente, il faut trouver un moyen de tenir compte dans les projections des variations à plus court terme. Pour cela, il est possible de réaliser une interpolation linéaire entre les données des années disponibles. Ainsi, entre deux années disponibles, si des valeurs sont manquantes, l'hypothèse que la variation entre ces deux années est linéaire a été faite. Par exemple, si des données pour les années 2001 et 2011 sont disponibles, mais

pas entre, la variation de 2001 à 2011 est supposée linéaire et il est donc possible d'estimer les valeurs des années manquantes pour la variable en question.

Le problème de cette méthode est qu'il n'est pas possible d'estimer les valeurs des années extérieures à l'intervalle des données disponibles, l'interpolation linéaire permet uniquement de combler les lacunes. Ainsi, pour extrapoler à l'extérieur de cet intervalle, nous utilisons la méthode de la courbe de tendance.

Les limitations de cette méthode sont donc similaires à celle de l'autre méthode au détail près que les variations de court terme sont estimées linéairement. Une estimation linéaire est faite, mais cette fois sur des segments précis ce qui peut rendre compte de plus de variations.

La Figure 4-19 résume le fonctionnement des trois méthodes utilisées. Il prend un segment de temps arbitraire pour illustrer la construction de la base de données.

Méthode	2001	2002	2003	2004	2005	2006	2007	2008	2009
Valeurs réelles									
Courbe de tendance									
Interpolation linéaire									

Aucune donnée		Calculée à partir de la courbe de tendance	
Donnée réelle		Interpolée linéairement	

Figure 4-19 Schéma des méthodes de projection

4.1.1.3 Résultats

Une analyse de corrélation linéaire est réalisée sur les trois bases de données générées. Ces analyses sont réalisées à l'aide du logiciel Excel. C'est une analyse de corrélation de Pearson qui est une étude de corrélation linéaire.

Les trois tableaux de corrélation sont disponibles en annexe B (Analyse de corrélation).

4.1.1.3.1 Analyse des résultats

Pour pouvoir exploiter les résultats, il faut initialement comparer l'effet des différentes méthodes. Tout d'abord, celles utilisant les valeurs réelles ne comportent pas assez de données, car beaucoup

de coefficients sont égaux à 1 pour deux variables distinctes. Les résultats de cette étude sont incertains, car trois points ne peuvent pas refléter correctement une corrélation entre des variables. Cela soutient la décision de réaliser une projection des variables pour obtenir plus de valeurs. Les résultats des deux méthodes faisant l'usage de projection seront donc utilisés pour identifier des tendances sur la corrélation des variables entre elles.

La méthode de la courbe de tendance produit des coefficients de corrélation plus élevés que ceux issus de l'interpolation linéaire. Ces résultats étaient attendus, car l'interpolation tient compte des variations de court terme qui sont, aux vues des graphes, non linéaires. Mais ne connaissant pas le comportement des variables entre les points disponibles, aucune de ces deux méthodes ne peut être jugée comme plus pertinente. Ainsi les tendances qui vont être exposées par la suite sont identifiées aux vues des résultats des deux méthodes.

Le fait le plus marquant ressortant de cette analyse est que beaucoup de variables sont corrélées entre elles et cela fortement. Ce fait peut, outre une réelle corrélation, être amplifié par la modélisation des données par des modèles linéaires.

Une première famille de variables corrélées peut être définie par celles semblant être associées à une augmentation de la population. Une augmentation de la population s'accompagne d'une augmentation des variables démographiques comme le nombre de véhicules, de travailleurs, de personnes âgées, d'unités et de densités de logements. Cette augmentation est assez intuitive, car plus de personnes sont présentes sur un même territoire au fil des années. La comparaison de ces paramètres fera l'objet d'une étude en termes de taux par la suite. Ensuite, il y a des variables caractérisant la demande en déplacements comme les validations de carte à puce, et le nombre de déplacements pour motif travail qui sont aussi associés à la variation de la population. Pour les motifs travail, il serait donc intéressant d'étudier la proportion des déplacements effectués pour ce motif. Également deux variables d'offre sont fortement corrélées avec la population et les autres variables mentionnées précédemment. Pour mieux comprendre les mécanismes liant cette famille de variables, d'autres études doivent être menées notamment en utilisant des taux. La Figure 4-20 permet de se rendre compte de cette évolution.

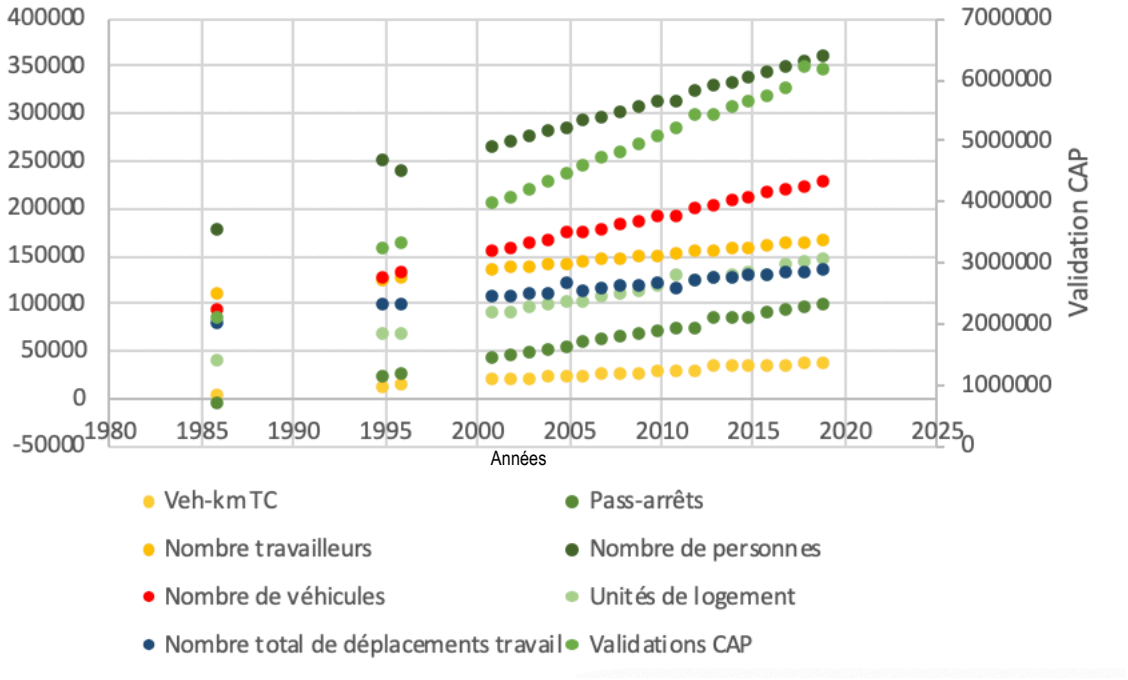


Figure 4-20 Évolution des variables associées à l'accroissement de la population

D'autre part, une corrélation est constatée entre le pourcentage de logements unifamiliaux et le nombre total de logements. Ce fait est évidemment intuitif. Ce pourcentage semble également être légèrement positivement corrélé avec la demande en TC, le nombre de véhicules et de travailleurs. Un facteur qui augmenterait probablement le besoin en déplacements.

Le prix de l'essence ne semble pas être clairement corrélé à d'autres variables. Notamment, il est totalement indépendant de la part des TC et du nombre de véhicules par permis. Ainsi, il ne semble pas avoir d'effet notable sur les habitudes de mobilité et la légère corrélation positive avec d'autres paramètres comme le nombre de véhicules peut être due à des phénomènes comme l'inflation couplée à une croissance de la population ou encore du simple effet de l'offre et la demande sur le pétrole.

Les effets sur la part modale des TC sont à prendre avec du recul en raison de l'approximation qui est faite grâce à la projection linéaire. Mais nous pouvons tout de même constater une forte corrélation négative entre le nombre de voitures par permis et la part modale TC. Des résultats comme la corrélation négative entre la part modale des TC et le nombre de validations sont assez étonnants. Ils voudraient dire que malgré l'augmentation de la fréquentation des TC, celle des autres modes augmente plus vite et ainsi, la part modale du transport public recule. La Figure 4-21

illustre ce phénomène. Les valeurs dites projetées sur le graphique sont les résultats obtenus à partir de la méthode de la courbe de tendance. Nous pouvons également noter que les temps de travail longs s'accompagnent d'une baisse de la part modale TC.

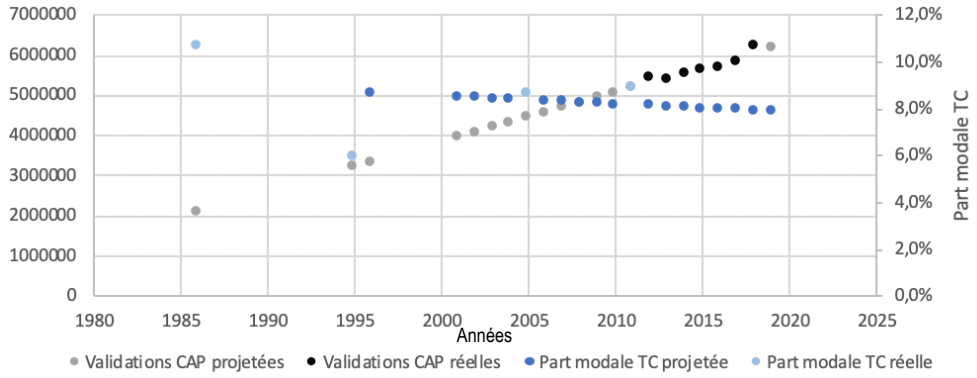


Figure 4-21 Évolution de la part modale TC et des validations de CAP

Comme mentionné précédemment, une grande famille de variables utilisant les valeurs absolues sont fortement corrélées. Cette corrélation peut être liée à une augmentation de la population. Ainsi, il faut enlever l'effet d'échelle induit par cette augmentation. Les résultats suivants utilisent des taux pour mieux comprendre les relations entre ces variables.

Un segment qui pourrait expliquer l'augmentation de la demande en TC est les personnes âgées. Le pourcentage de personnes âgées est corrélé positivement et significativement au nombre de validations de CAP par personne. Un autre segment a été étudié dans cette étude, ce sont les travailleurs. Dans la littérature, les travailleurs sont un segment qui explique une grande partie de la fréquentation des TC. Mais les résultats de cette étude sont étonnants et peuvent être remis en cause. Le pourcentage de travailleurs est corrélé négativement avec le nombre de validations de CAP par personne.

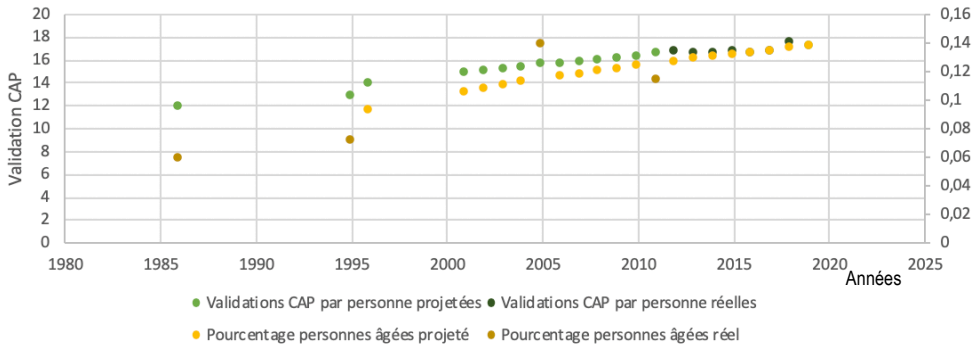


Figure 4-22 Évolution des validations de CAP par personne et des personnes âgées

Il faut également mentionner que les variables associées à l'offre sont fortement corrélées à celles relatives à la demande. L'offre suit étroitement les variations de la demande.

4.1.1.4 Conclusion de l'analyse de corrélation

Cette analyse nous permet de déceler certaines tendances et liens entre les variables. Des variables comme la population, le pourcentage de logements unifamiliaux et de personnes âgées, le nombre d'unités de logement ou encore le nombre de travailleurs semblent être moteurs d'une augmentation de la demande en transport en commun. Ce sont autant de paramètres qui pourront par la suite être considérés dans la création d'un modèle de prévision de cette demande. Il faudra tout de même faire attention à la forte corrélation qui lie tous ces paramètres entre eux.

Cette étude a également mis en lumière certaines caractéristiques des données. La part modale TC n'a pas un comportement linéaire et ne possède pas une tendance stable. La longueur du réseau routier est impossible à exploiter en vue du changement des conventions de calcul qui affectent les résultats.

Dans la mesure où cette analyse fait usage d'une approximation pour projeter les données dans le temps, il faut prendre certains résultats avec du recul. D'autant plus que certaines incohérences habitent les résultats comme la corrélation négative entre la demande et le pourcentage de travailleurs. Les résultats apportent néanmoins un éclairage important sur les relations entre les paramètres qui caractérisent la zone géographique étudiée. Il faut également garder en tête que l'étude réalisée ne tient compte que des corrélations linéaires. D'autres liens non linéaires pourraient exister entre les variables, mais cela nécessiterait un autre type d'analyse.

4.1.2 Synthèse de l'évolution des variables caractérisant le territoire global

Cette section avait pour but de réaliser une analyse rétrospective d'un certain nombre d'indicateurs permettant de caractériser la démographie, l'occupation du territoire, la demande et l'offre en transport en commun. Ce rapport permet d'avoir une idée plus précise des évolutions subies par le territoire de Gatineau. Pour aider la compréhension de ces transformations, un classement de toutes les variables évoquées dans le rapport suivant leur taux de croissance annuelle est réalisé. Trier les variables de cette manière permet de se rendre compte de celles qui subissent une évolution décroissante, celles qui sont stables et celles qui croient. Le taux de croissance permet d'avoir une idée de la magnitude de ces transformations.

Le taux de croissance est calculé pour les valeurs disponibles dans le rapport et se base sur l'évolution subie entre la première année disponible et la dernière. Le taux de croissance est par rapport à la première année disponible et il est annualisé pour obtenir une moyenne qui peut être comparée pour toutes les variables. La démonstration de la méthode de calcul est disponible dans l'annexe B. La Figure 4-23 illustre les résultats obtenus.

Il est possible de constater que la majorité des variables ont un comportement globalement croissant. Notamment les validations de carte à puce subissent la plus forte augmentation que nous pouvons constater. La demande est la variable subissant l'évolution la plus extrême ce qui laisse penser que cette croissance est expliquée par plusieurs facteurs différents qui doivent être clairement identifiés.

De l'autre côté du spectre, les parts modales AP (auto-passager) et TC sont décroissants ce qui peut signifier que l'automobile continue à prendre de l'ampleur et cela accompagné d'une baisse du taux d'occupation des véhicules. Cela est confirmé par la croissance de la part modale AC et également de paramètres images de la possession d'automobiles.

Le taux de mobilité est en baisse : les personnes se déplacent moins, mais la fréquentation des TC augmente en flèche.

Le pourcentage d'utilisateurs de modes actifs est en baisse, mais la part modale de ce mode augmente. Cela pourrait être l'image d'une segmentation de la population : moins de personnes utilisent les modes actifs, mais ceux les utilisant sont de fidèles usagers des modes actifs.

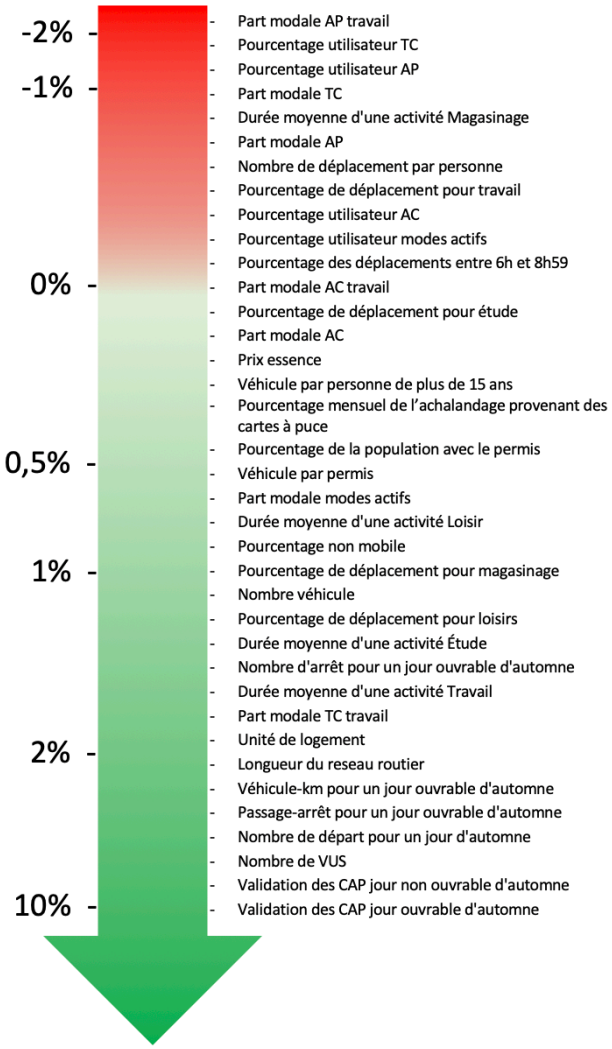


Figure 4-23 Taux de croissance annuelle moyen des variables du rapport

4.1.3 Étude du taux d'utilisation des TC

L'étude du taux de fréquentation du réseau de la STO en termes de déplacement TC par personne est un indicateur important pour obtenir des informations sur l'origine de la variation de la demande en TC. Pour obtenir cet indicateur, le mois de novembre en tant que le mois le plus achalandé est sélectionné. Il s'agit donc de déterminer le taux moyen de fréquentation des TC pour un jour ouvrable moyen de novembre. L'évolution de ce taux est calculée pour les années 2012-2020.

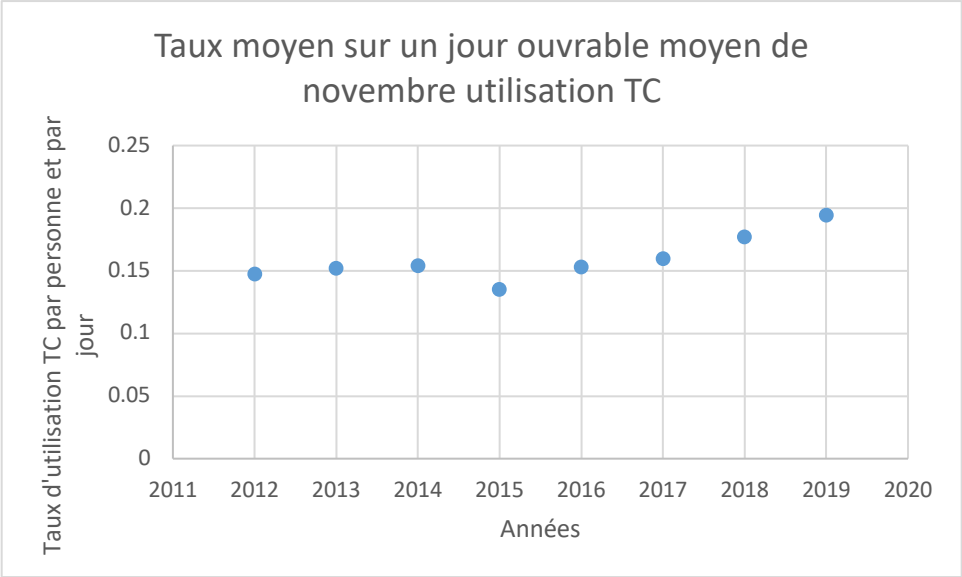


Figure 4-24 Évolution du taux moyen de fréquentation des TC

Une constante aurait indiqué que l'évolution de la demande peut être uniquement liée à une croissance de la population. Mais le fait que ce taux n'est pas constant indique qu'un autre paramètre influe sur la demande. Un tau non constant indique des évolutions différentes. Pour mieux comprendre l'origine de ce paramètre inconnu, il est possible de segmenter la population faisant l'usage des TC et d'étudier les taux segmentés.

Dans les données de validations de carte à puce, il figure le type de tarif utilisé par l'usager. En analysant les différents tarifs proposés par la STO, il est possible d'utiliser cette caractéristique comme proxy pour segmenter la population. Les segments suivants seront donc analysés :

- Les scolaires
- Les étudiants
- Les adultes
- Les aînés
- Les usagers d'un titre de transport OC (réseau de TC d'Ottawa)
- Les titres pour évènement ponctuel

Les titres mis en service pour un évènement ponctuel ne seront pas étudiés dans la mesure où ils ne représentent pas une demande habituelle. L'autre enjeu de cette segmentation est de déterminer qui sont les usagers d'un titre de transport d'Ottawa. L'analyse de ce segment est importante, car leur proportion n'est pas négligeable soit environ 4% de la demande en 2018. Il s'agit de déterminer le lieu de résidence de ces personnes. Pour ce faire, la demande est analysée de manière plus précise. La Figure 4-25 représente la distribution temporelle de ces validations pour un jour moyen de novembre 2018. La majorité de ces validations s'effectuent en début de pointe du soir et une minorité en pointe du matin. Cette indication porte à penser que ces personnes ne résident pas dans la région de Gatineau, mais vraisemblablement à Ottawa. La Figure 4-26 présente la distribution spatiale. Il est possible de constater que la majorité des usagers ont pour lieu de départ le centre-ville d'Ottawa. Il est donc vraisemblable d'établir l'hypothèse que ces usagers ne résident pas dans la région de Gatineau, mais bien à Ottawa et utilisent le réseau de la STO pour se rendre à Gatineau. Ce segment de tarification ne sera donc pas considéré, car la population étudiée est celle de la région de Gatineau.

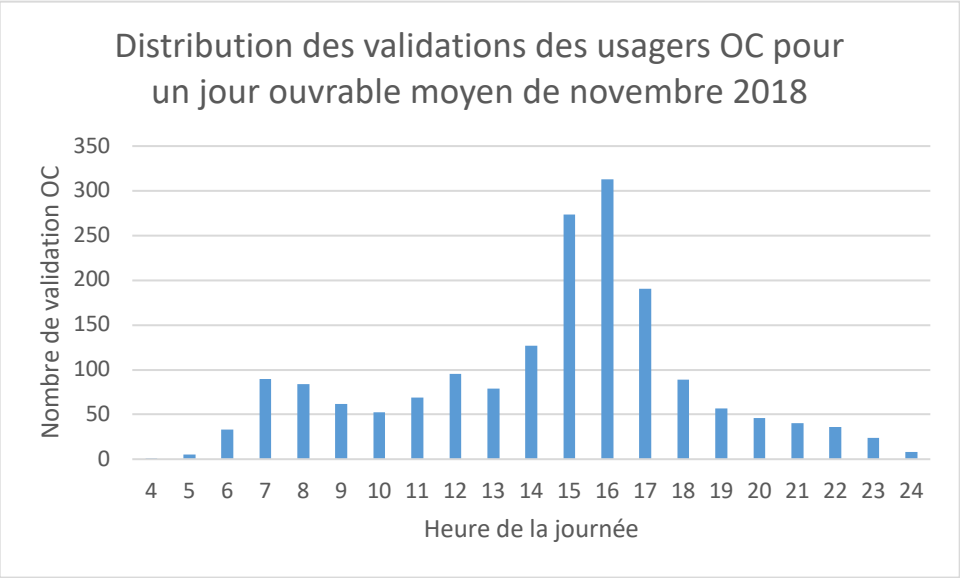


Figure 4-25 Distribution temporelle des validations de CAP du réseau d'Ottawa

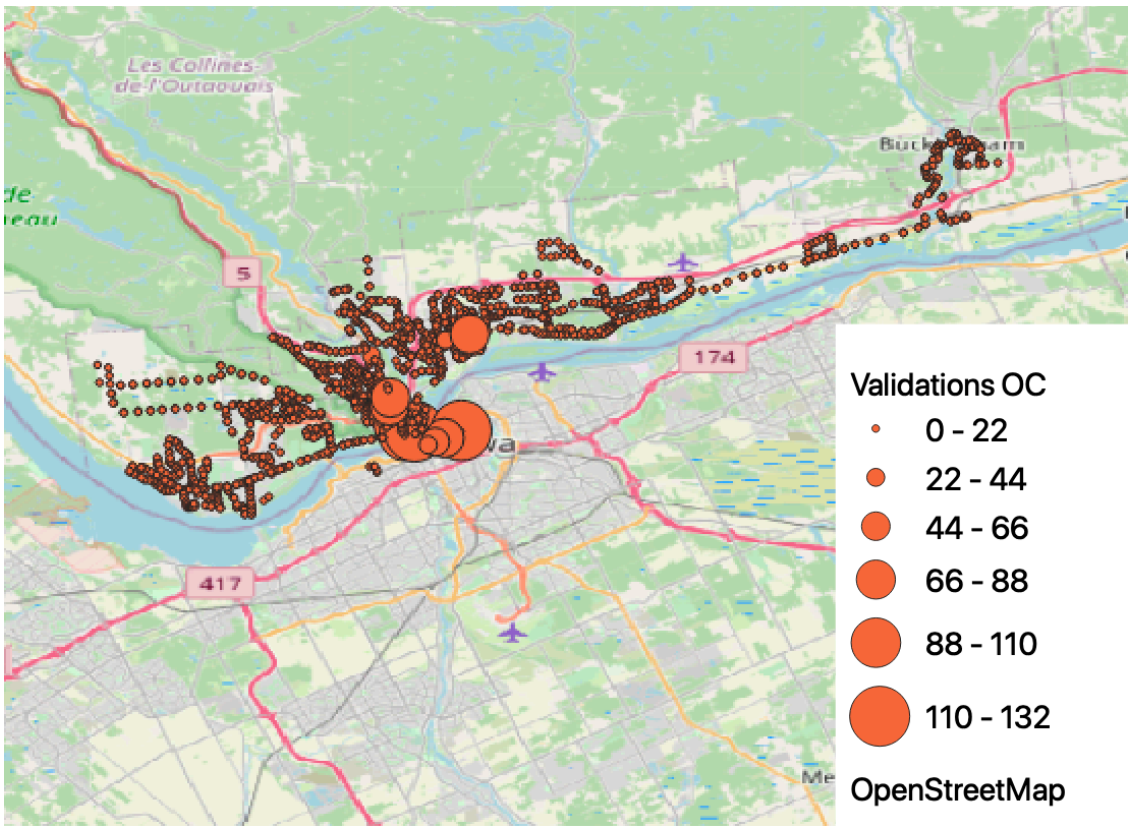


Figure 4-26 Distribution spatiale des validations de CAP du réseau d'Ottawa

Les taux de déplacement TC sont donc générés en utilisant la population de la RMR O-G partie Québec. Pour chaque segment les tranches de population suivantes sont utilisées :

- Scolaires : 5 à 14 ans.
- Étudiants : 15 à 19 ans.
- Adultes : 20 à 65 ans.
- Aînés : 65 ans et plus.

Les résultats sont illustrés par la Figure 4-27. Il est tout d'abord possible de constater que le taux des aînés est plutôt constant. Cela signifie que la demande des aînés peut être uniquement modélisée par la croissance de la population des plus de 65 ans. La tranche des adultes est relativement stable sauf en 2018. Un évènement ponctuel peut avoir influencé cette demande en 2018 ce qui reste à déterminer. Les usagers scolaires sont le segment le moins prévisible grâce aux données démographiques, il faut donc trouver une autre manière d'expliquer la demande des 5-15 ans.

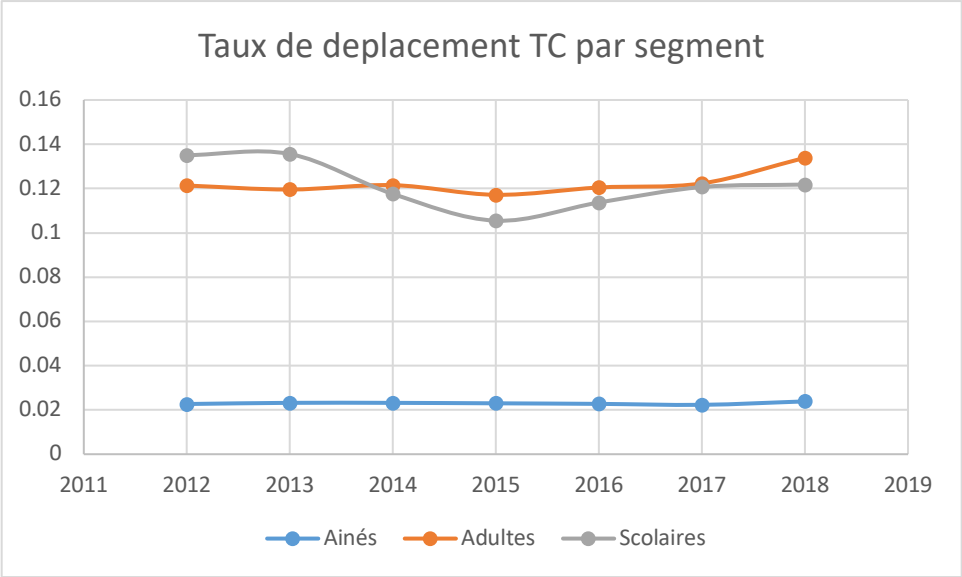


Figure 4-27 Taux de déplacement TC par segment de population

4.2 Analyse rétrospective : secteurs municipaux

Ayant obtenu des informations sur la dynamique de la demande en fonction de plusieurs variables explicatives à l'échelle du réseau global, il est pertinent de réduire la zone d'agrégation sélectionnée pour tenter d'identifier l'effet de caractéristiques plus locales sur les évolutions de la demande en TC.

Pour avoir une vue d'ensemble sur chaque secteur, une fiche synthétique est générée pour chaque secteur. Sur chaque fiche, les variations de la demande et de l'offre en TC sont illustrées. Les variations du nombre de logements sont également présentes pour être mises en parallèle des variations de la demande. De plus, les résultats de certaines caractéristiques démographiques issues des recensements sont affichés.

L'objectif de ces fiches est de décrire les caractéristiques des secteurs et de visualiser les évolutions de la demande en TC associée à ces secteurs. Ainsi, sur chaque fiche, une partie est dédiée à la caractérisation de ces secteurs. Ces fiches sont disponibles en annexe D. Une mise en parallèle de certains indicateurs est rapidement présentée dans la suite de cette partie pour remplir les fiches synthétiques.

4.2.1 Demande TC

Il est également intéressant d'étudier les différents secteurs les uns par rapport aux autres. Comparer certains indicateurs permet de mieux comprendre les interactions entre ces zones géographiques. Tout d'abord, la demande s'avère être très inégale entre les secteurs municipaux. La Figure 4-28 et la Figure 4-29 permettent de se rendre compte de ce phénomène. En pointe du matin, les secteurs de Gatineau, Hull, Aylmer et Plateau sont les lieux d'un nombre conséquent de départs de trajets. Ils sont vraisemblablement des secteurs où résident la majorité des usagers des TC. Les montants en pointe du soir sont les plus importants dans les secteurs d'Ottawa et Hull. Ce sont dans ces secteurs que doivent se situer une part importante des lieux d'emplois des habitants de la région de Gatineau.

La Figure 4-30 classe les secteurs en fonction de la répartition des validations de carte à puce dans la journée. Les secteurs apparaissant en premier sont ceux dont les validations de type origine se situent majoritairement sur la pointe du matin. Cela signifie que le secteur est résidentiel.

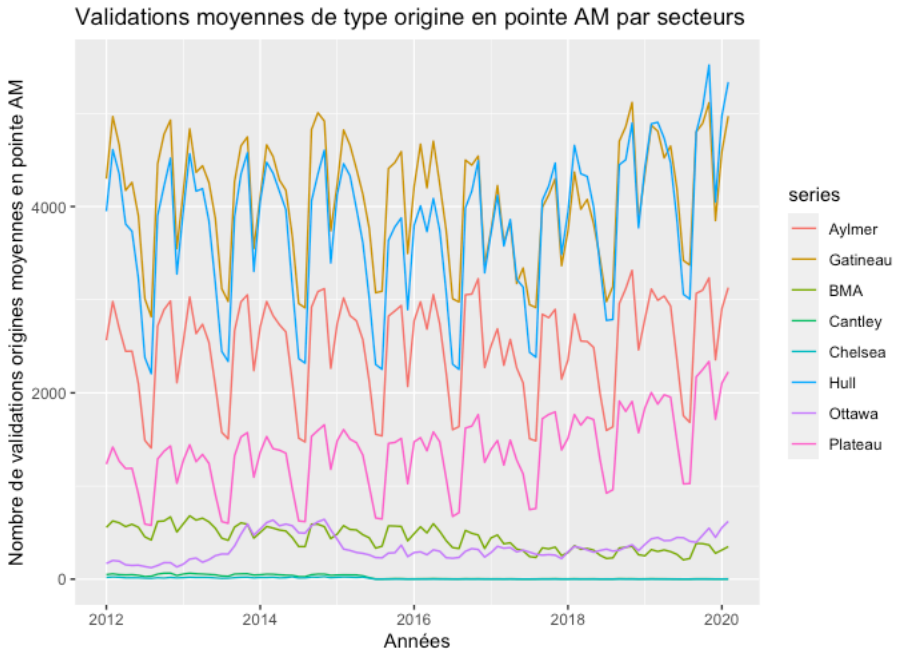


Figure 4-28 Montant par secteurs en pointe AM

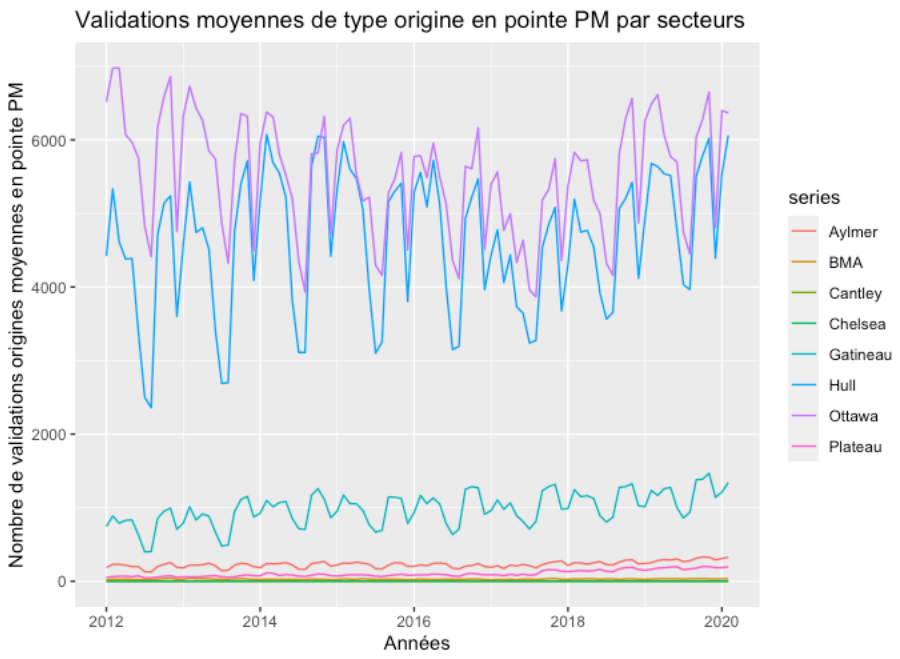


Figure 4-29 Montants par secteur en pointe PM

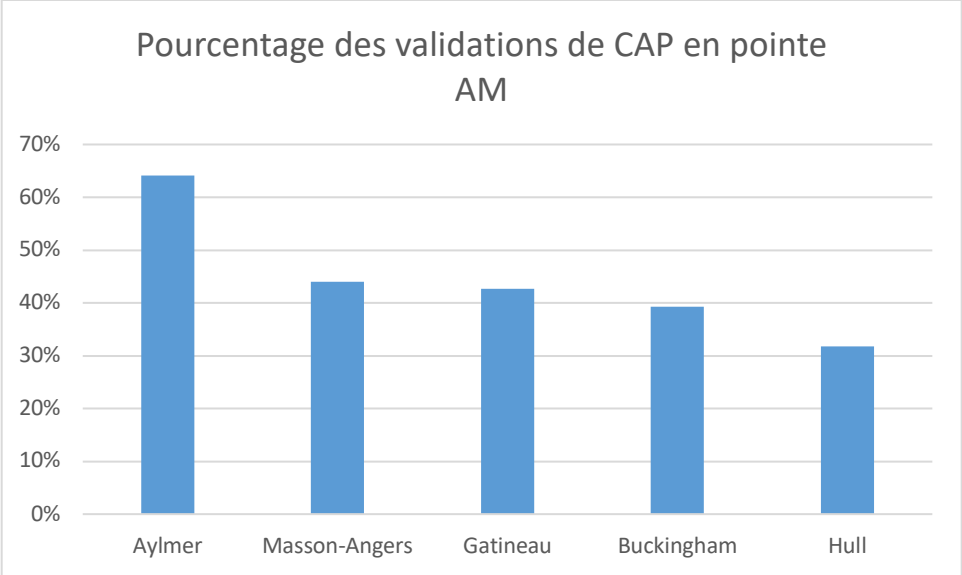


Figure 4-30 Pourcentage de validations de CAP en pointe du matin

4.2.2 Démographie et offre TC

La Figure 4-31 donne les densités de population et les figures Figure 4-32 et Figure 4-33 sont des images de l'offre et de la demande sur le réseau de la STO. Les données démographiques permettent de confirmer les assertions exposées précédemment.

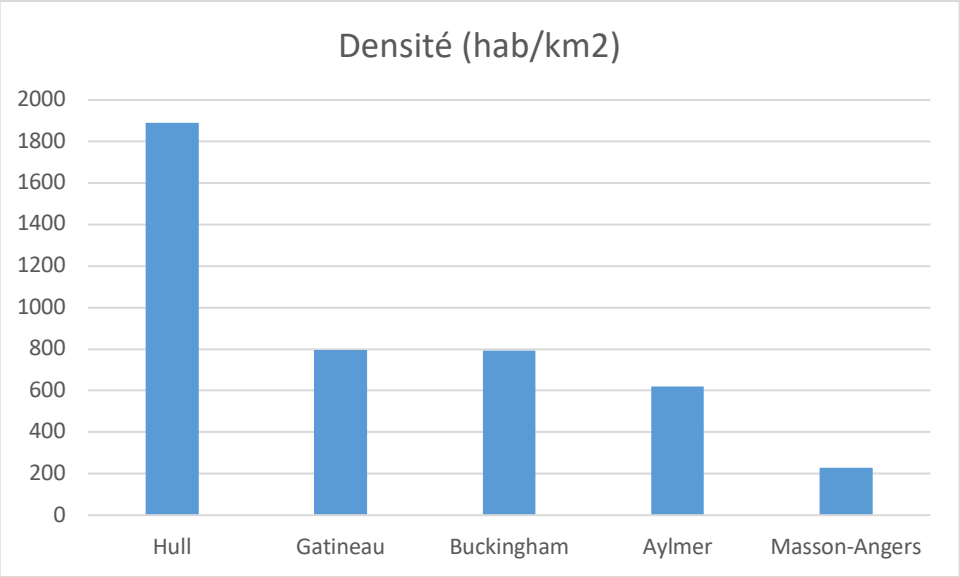


Figure 4-31 Densité de population par secteur

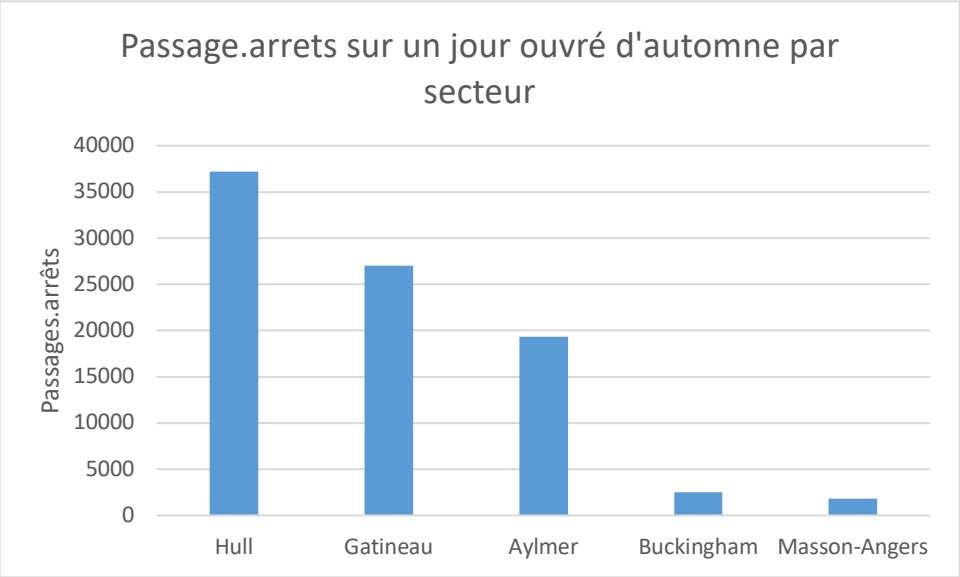


Figure 4-32 Indicateur de l'offre en TC par secteur

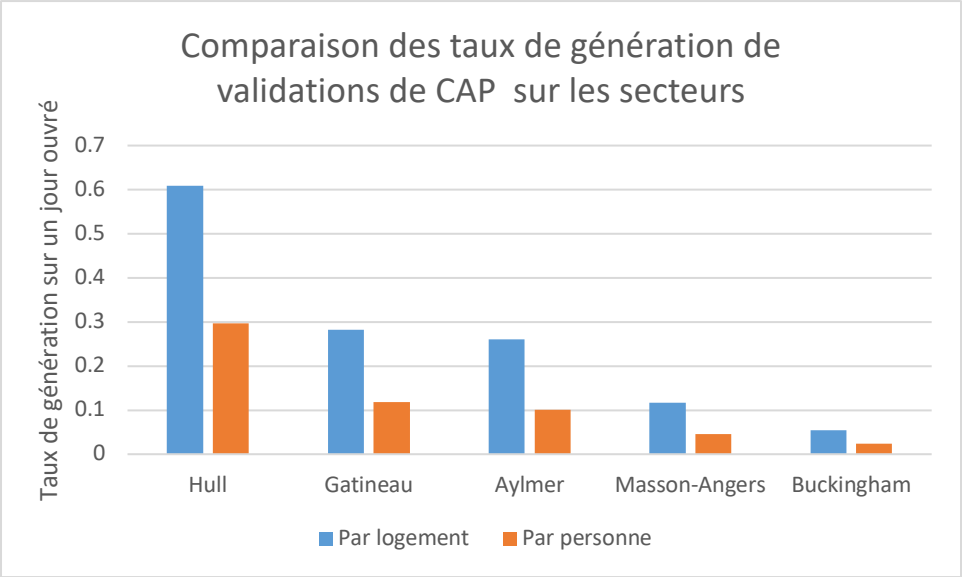


Figure 4-33 Taux de génération de validations de CAP par secteur

Ces informations permettent donc de créer 3 typologies de secteurs :

- Secteurs loin d'Ottawa, faiblement peuplé, utilisation faible et en déclin des TC : Masson-Angers et Buckingham
- Secteur à moyenne distance d'Ottawa utilisant moyennement les TC mais l'utilisation est fortement pendulaire : Gatineau et Aylmer
- Secteur proche d'Ottawa utilisant fortement les TC : Hull

Sur tous les secteurs, seuls les secteurs de Masson-Angers et de Buckingham voient leur fréquentation des TC baisser. Partout ailleurs, l'achalandage est en augmentation depuis les deux dernières années. Ces secteurs en croissance sont également des foyers de population importants.

L'offre de transport est également cohérente avec la densité de population. Un secteur dense obtient également une offre plus dense. Cela confirme bien que la population est motrice de la demande dans la mesure où la STO adapte l'offre à la demande.

4.3 Analyse rétrospective : agrégation au niveau des arrêts

Avoir un niveau de résolution plus fin pour l'analyse rétrospective est un besoin essentiel qui a été déclenché par un manque de pouvoir explicatif de certaines données agrégées. L'agrégation au niveau des secteurs ne fournit pas des résultats satisfaisants. De plus, l'une des volontés du projet

est de pouvoir quantifier l'effet, sur la demande, de modifications de l'aménagement du territoire au niveau d'une zone géographique précise. Analyser les évolutions de la demande à l'échelle des arrêts s'impose donc comme un passage obligé.

La demande en TC est tout d'abord analysée pour tenter de mieux comprendre quelles sont les zones où les transformations sont les plus importantes. Par la suite, ces résultats seront mis en parallèle d'indicateurs pouvant aider à expliquer ces transformations.

4.3.1 Demande TC

Les graphiques qui suivent reflètent l'évolution des taux de croissance annuels de 2013 à 2019 (nombre de montants moyen sur la pointe AM d'un jour ouvré de novembre chaque année). Il est également possible d'y voir les arrêts ayant été créés chaque année.

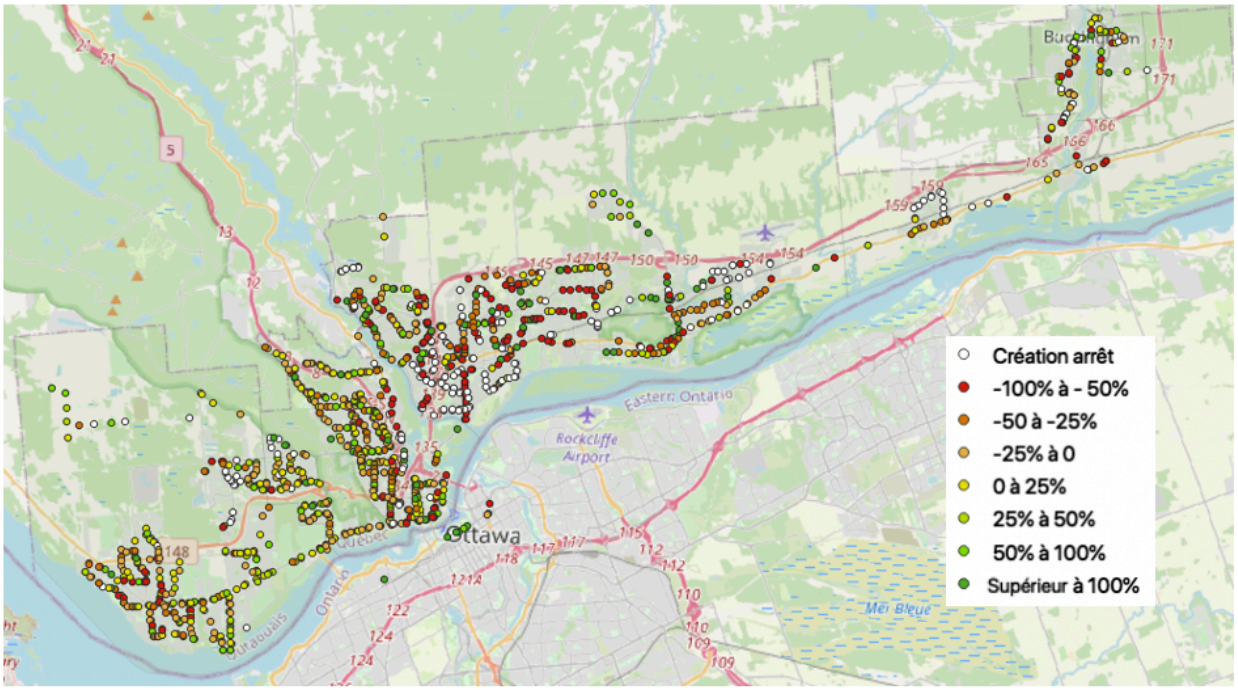


Figure 4-34 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2013

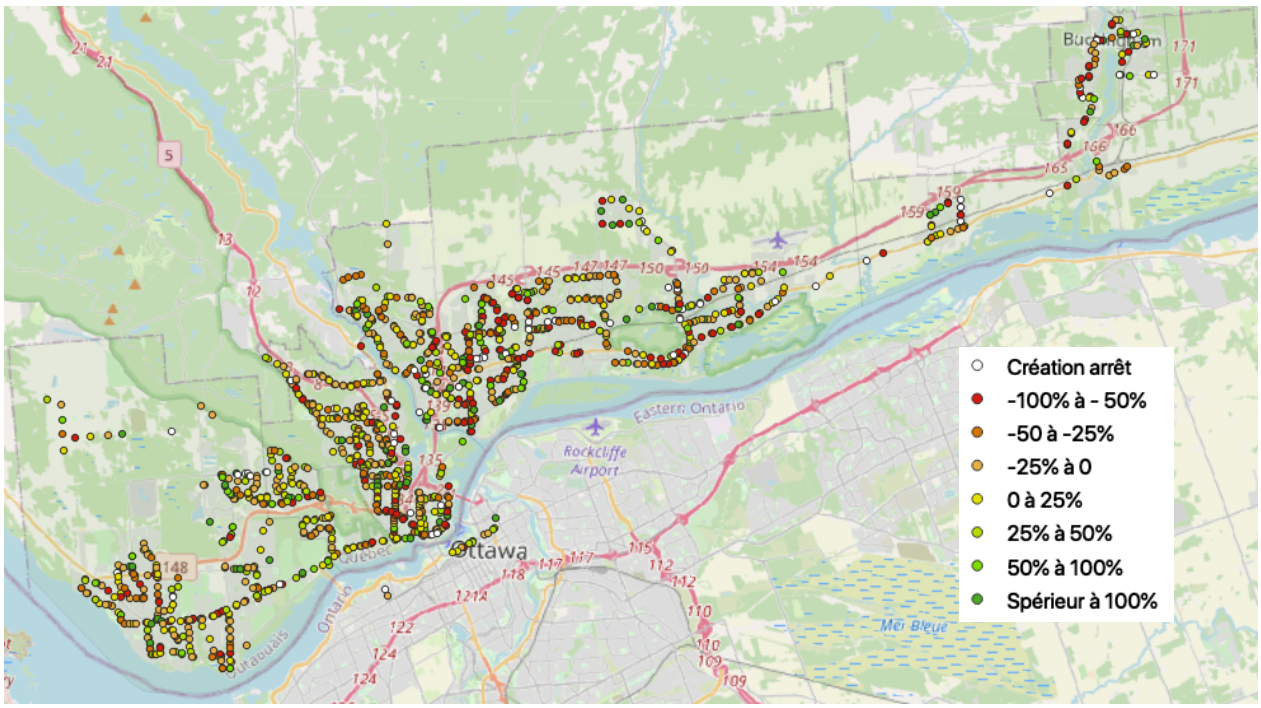


Figure 4-35 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2014

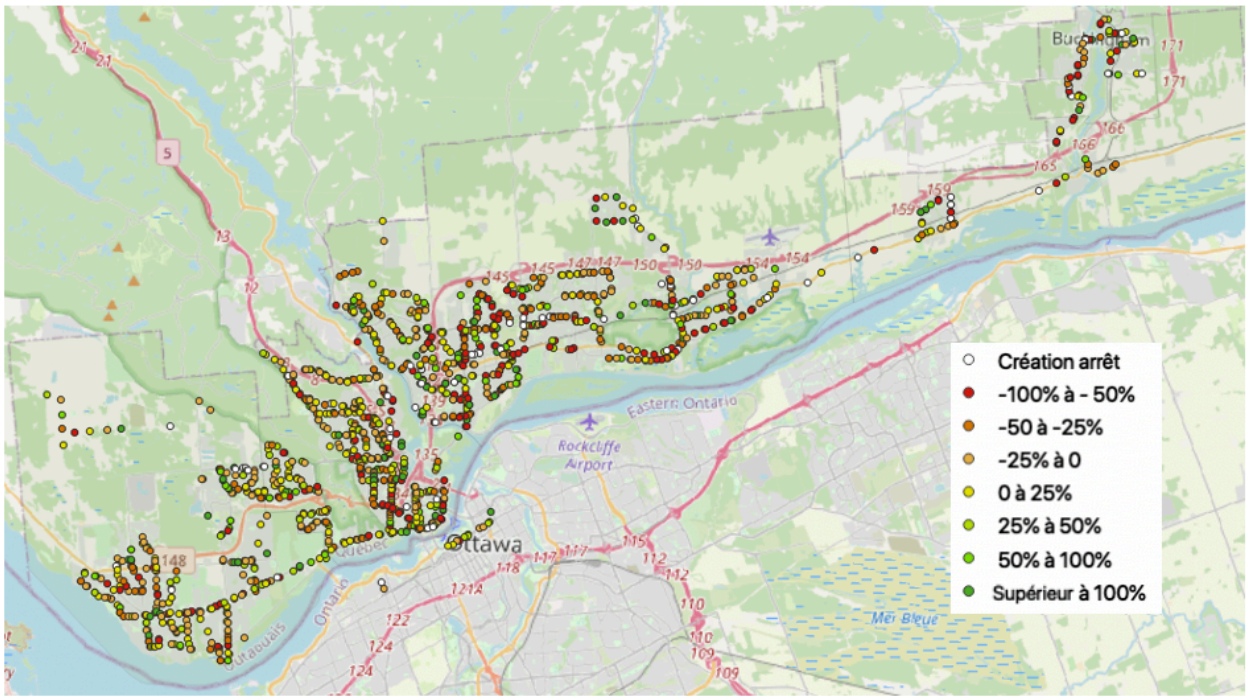


Figure 4-36 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2015

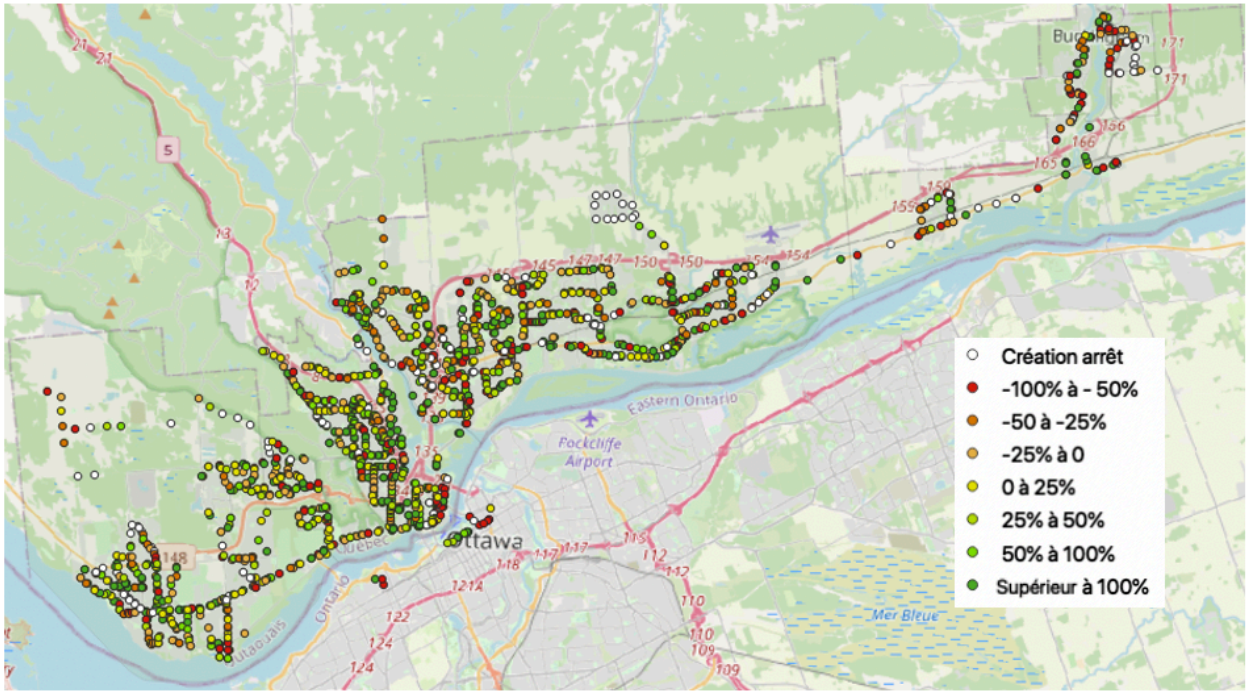


Figure 4-38 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2016

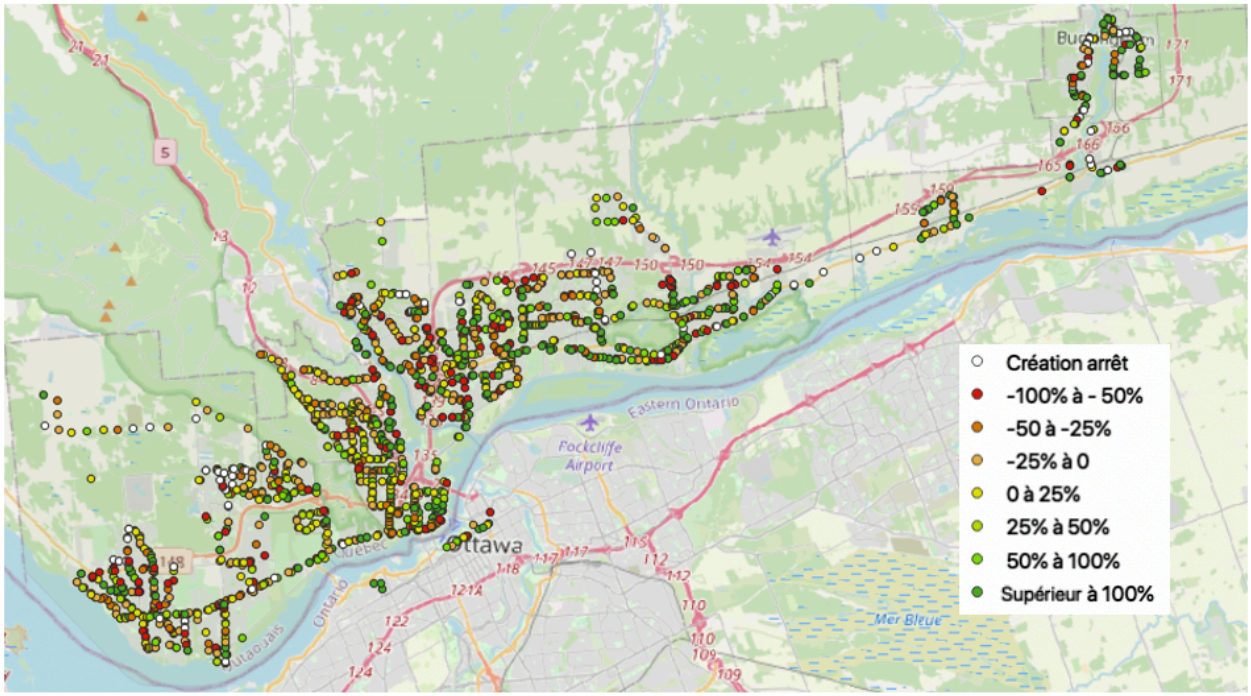


Figure 4-37 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2017

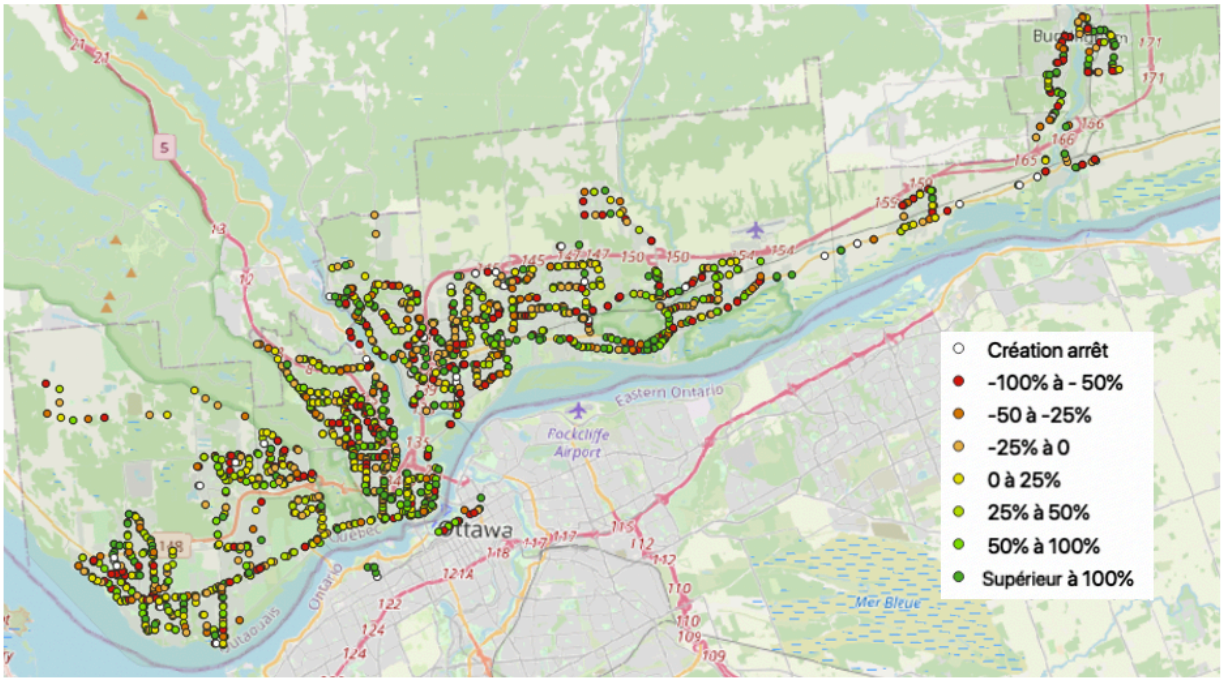


Figure 4-40 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2018

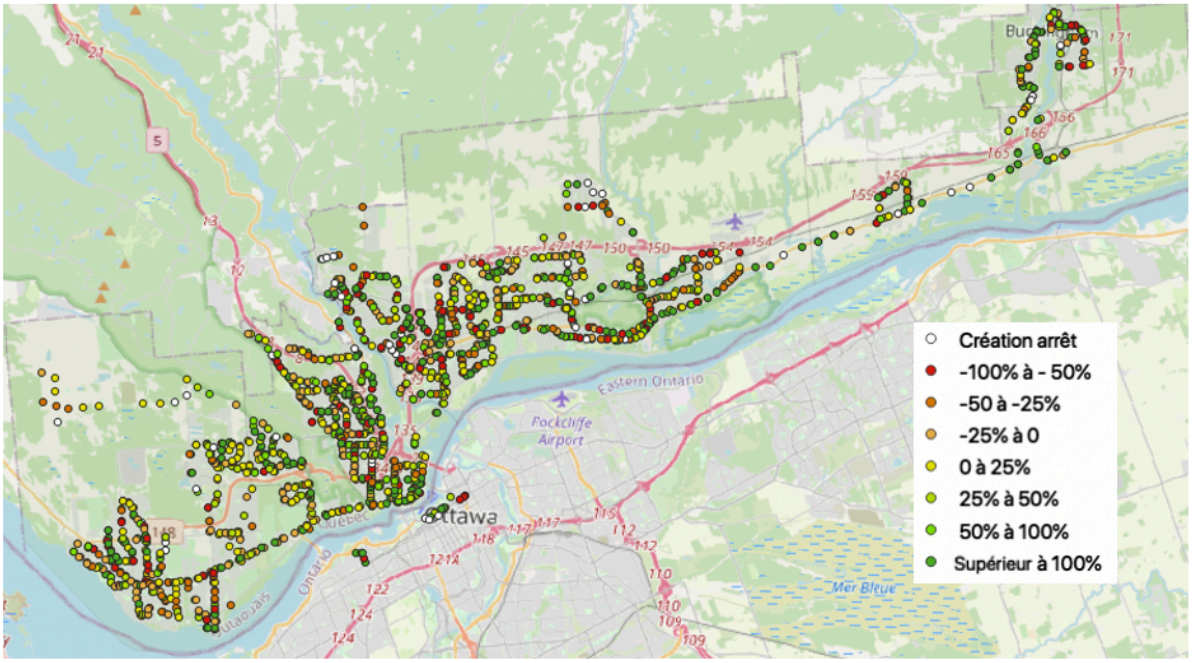


Figure 4-39 Taux de croissance de la fréquentation des arrêts sur la pointe AM un jour ouvré moyen de novembre 2019

Il est déjà facile de constater l'implantation du Rapibus en 2013 : c'est une ligne de bus rapide (BRT) mise en place sur les secteurs d'Aylmer, de Hull et de Gatineau ayant pour but de diminuer la congestion liée à la croissance des parties est et nord de Gatineau. Une multitude de nouveaux arrêts sont donc implantés sur le tracé.

Un autre phénomène est constatable à la même période : beaucoup de nouveaux arrêts sont implantés dans le centre-ville de Gatineau. Cela a donc pour effet de diminuer l'achalandage des arrêts voisins par un effet de concurrence. Il faut tout de même garder en tête que la demande a augmenté dans cette zone.

En 2016 et 2017, un grand nombre d'arrêts ont été mis en activité dans la zone nord du secteur d'Aylmer. Le même phénomène se produit dans l'est du secteur de Buckingham et sur Masson-Angers en 2013. Il semble donc que récemment, la STO a amélioré son offre de transport dans les zones périphériques de Gatineau.

Le phénomène de concurrence entre les arrêts rend la lecture de ces cartes difficiles, notamment lors de l'implantation de nouveaux arrêts ou lors de déplacement d'arrêts. Certaines caractéristiques sont visibles à cette échelle, telle que la modification du réseau. Phénomène qu'il n'est pas possible de constater si une agrégation est réalisée. Mais prévoir la demande à cette échelle n'est pas une option pour ce projet.

4.3.2 Permis de construire résidentiels

Le nombre de permis de construire autorisés par la ville de Gatineau est une variable permettant d'identifier les zones qui se densifient. Ces données sont agrégées pour pouvoir être exploitées, mais le but ici est de les comparer aux données désagrégées de cartes à puce. La cartographie de la Figure 4-41 permet d'identifier de nouveaux foyers de résidence. Le centre du secteur d'Aylmer est la zone où le plus de nouvelles constructions voient le jour. D'autres foyers sont également à constater : sud et ouest d'Aylmer, le voisinage de l'aéroport, le secteur d'Hull et le centre de Buckingham. Il faut également noter que le centre-ville de Gatineau n'est pas le lieu de nouvelles constructions résidentielles, tout comme le centre d'Aylmer.

Cette variable est importante pour la suite des travaux, car c'est une des rares variables explicatives qui est spatialisée de manière précise et dont la date d'acceptation du permis de construire est connue.

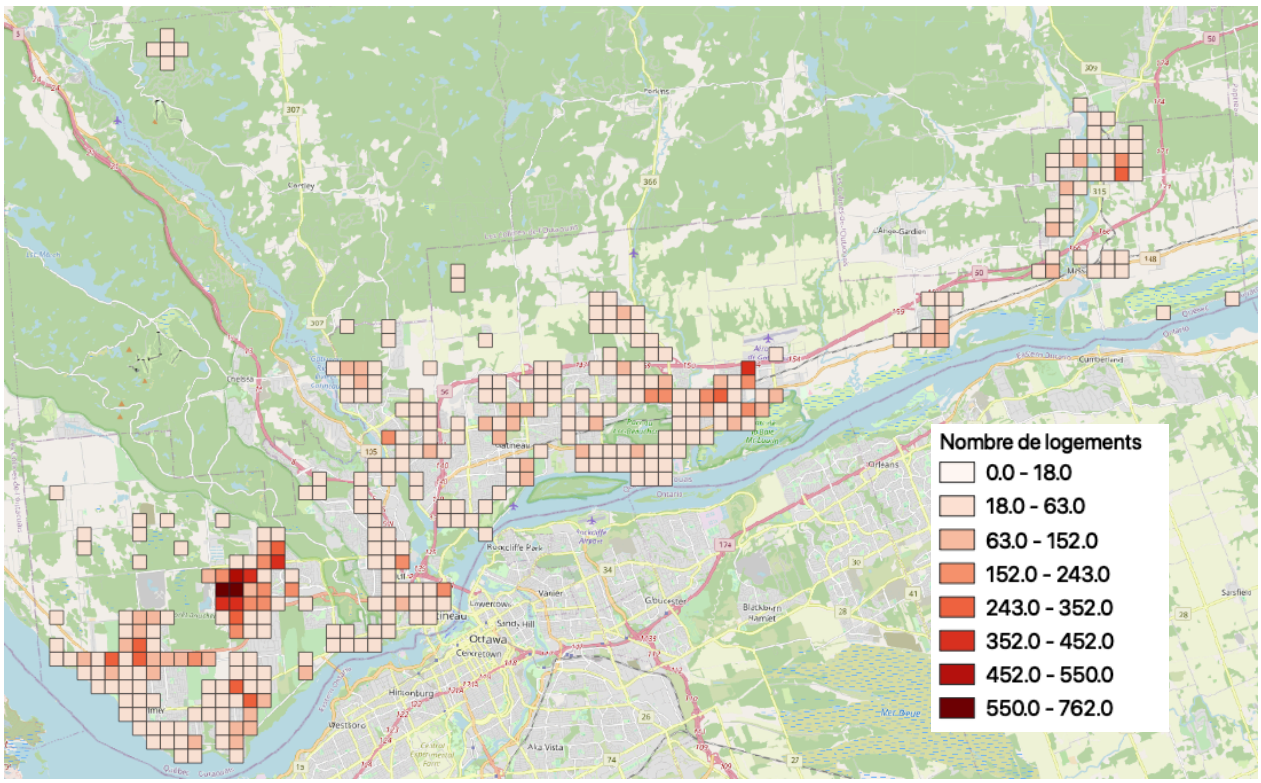


Figure 4-41 Permis de construire autorisés entre 2009 et 2019

En comparant les montants en pointe du matin avec les autorisations de permis de construire, il est évident de constater une forte corrélation spatiale entre ces deux ensembles de données. La zone de forte densification située au nord-est du secteur d'Aylmer fait l'objet d'une forte augmentation du nombre de montants en pointe du matin ainsi que la densification du nombre d'arrêts. Un tel phénomène est directement explicable par la hausse du nombre de logements qui introduit une demande supplémentaire et ainsi une augmentation de l'offre de la part de la STO.

4.3.3 Variables démographiques

Le site de la ville de Gatineau constitue un recueil de données important, mais la plupart d'entre elles ne sont disponibles qu'à un instant précis dans le temps, notamment les données du recensement. Ci-dessous, deux exemples de données qu'il est possible d'obtenir sont illustrés sur les Figure 4-42 et Figure 4-43: le taux d'activité et la densité de population constatés par le recensement de 2016.

Les données de demande exposées précédemment étant celles des montants en pointe du matin, il est pertinent de les comparer à la densité et au taux d'activités. Les mouvements en TC étant

principalement pendulaires pour se rendre au lieu d'activité, ce sont deux indicateurs pertinents pour se rendre compte de la génération de montants en pointe du matin. A noter que suite à la pandémie de Covid19, ces résultats et hypothèses sont susceptibles d'évoluer.

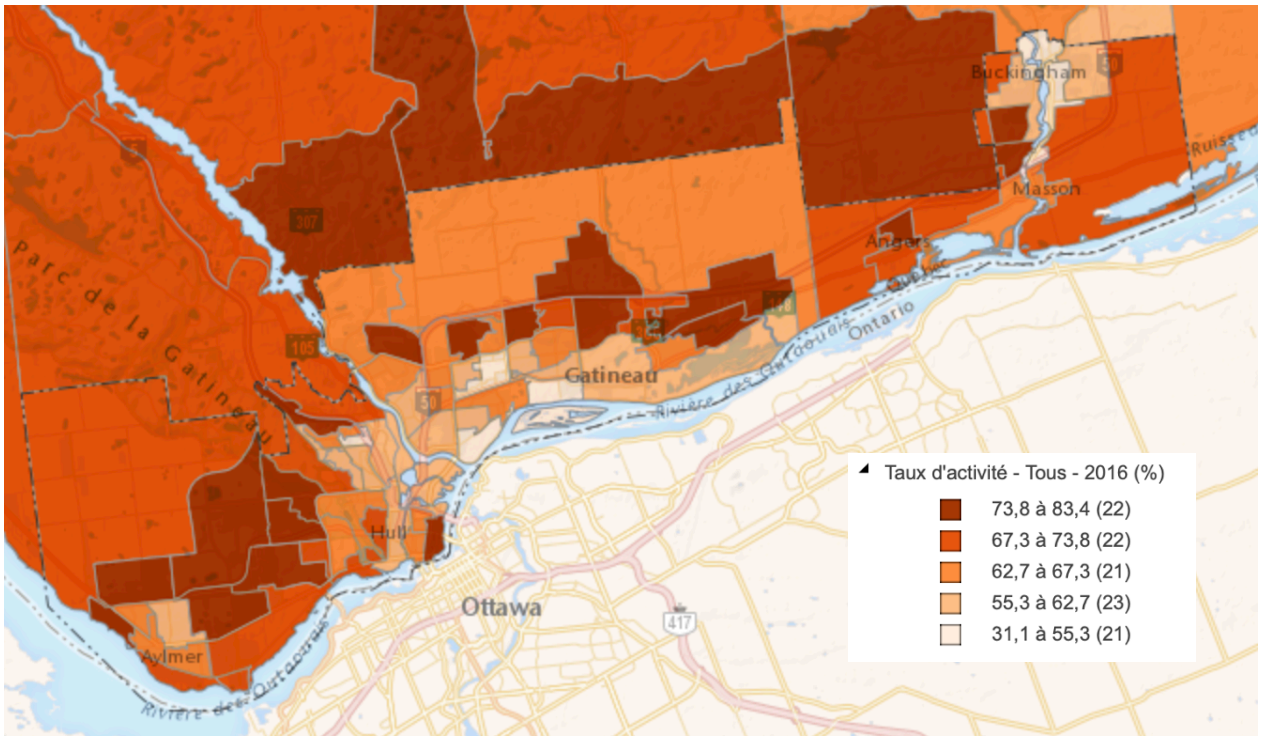


Figure 4-42 Taux d'activité du recensement de 2016

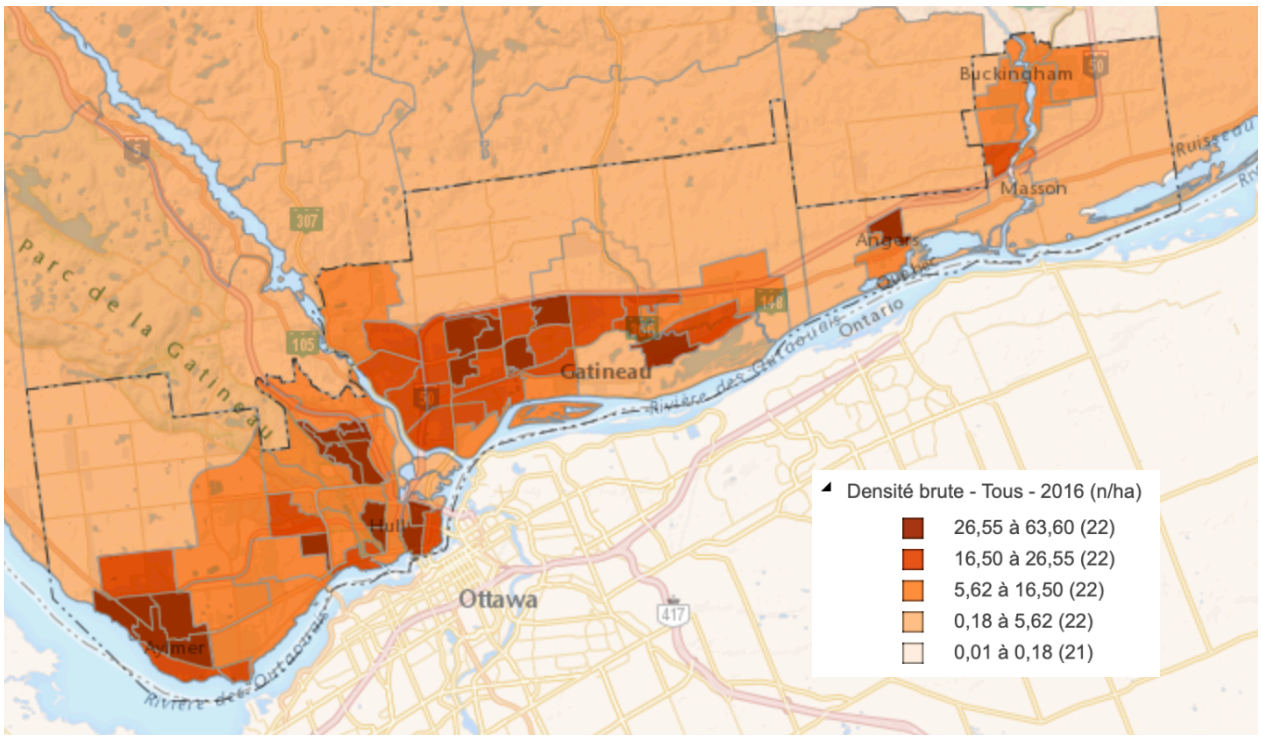


Figure 4-43 Densité de population en 2016 (recensement)

Ces données étant fixes dans le temps, elles sont difficilement directement comparables aux taux de croissance des montants aux arrêts. Il est tout de même possible de constater que la tendance penche généralement vers une augmentation plus forte de la demande dans les zones résidentielles denses et également sur celles d'emplois.

Une corrélation entre l'apparition de nouveaux logements et la densité d'emplois est forte. Les habitants de Gatineau ont donc tendance à vouloir plus de proximité entre les habitations et les lieux de travail.

L'étude rétrospective sur les arrêts s'avère ne pas fournir les résultats voulus. Seules les informations sur la transformation du réseau sont pertinentes. Il est possible de constater la restructuration du réseau dans le centre de Gatineau ou encore l'implantation du Rapibus en 2013. La densification résidentielle est également cause d'une augmentation de l'offre, matérialisée par l'implantation de nouveaux arrêts. Mais il est très difficile de constater les dynamiques qui régissent la demande ou encore de les mettre en parallèle avec d'autres indicateurs. L'effet de concurrence des arrêts est la seule caractéristique importante à retenir sur la demande au niveau des arrêts. Elle justifie l'utilisation des zones hexagonales; l'analyse est donc réalisée à cette échelle dans la partie suivante.

4.4 Analyse rétrospective : agrégation en zones hexagonales

L'étude de la demande au niveau des arrêts fournissant des résultats difficiles à exploiter, il prend sens d'étudier le territoire par le biais des zones hexagonales. Elles permettent d'obtenir une zone d'agrégation géographique comparable, chose qui n'était pas possible au niveau des arrêts : comparaison de la demande désagrégée à des indicateurs agrégés de diverses manières. Ces zones se dessinent donc comme une dénomination spatiale pertinente pour réaliser toutes les études qui suivent.

Dans un premier temps, la demande en TC est passée en revue à l'échelle des zones hexagonales. Elles sont par la suite comparées à deux indicateurs clefs dans la mise en place de modèles de prévision : les données de logements et les données sur les stationnements incitatifs. Ces deux variables sont les seules possédant une dimension spatiale et temporelle précise, caractéristique importante pour obtenir un modèle précis dans l'espace et également dans le temps. Les données de stationnements incitatifs sont nouvelles et introduites dans les parties qui suivent.

L'avantage de ce niveau d'agrégation est également la capacité de créer des paires OD entre des territoires comparables. L'étude de matrices OD réalisées dans cette partie permet de se rendre compte du comportement des usagers sur le réseau et non uniquement leur lieu d'origine comme cela a été étudié jusqu'à maintenant.

4.4.1 Demande TC

Les Figure 4-44 illustre la demande en pointe AM sur les zones d'agrégation hexagonales. Il est possible de constater une augmentation globale du nombre de montants. Certaines zones comme le centre-ville de Gatineau connaissent une croissance plus forte que le reste du territoire. Un nombre limité de zones font l'objet d'une très forte croissance qui peut être expliquée par l'implantation des stationnements incitatifs. Ces stationnements incitatifs sont étudiés plus en détail dans la partie suivante.

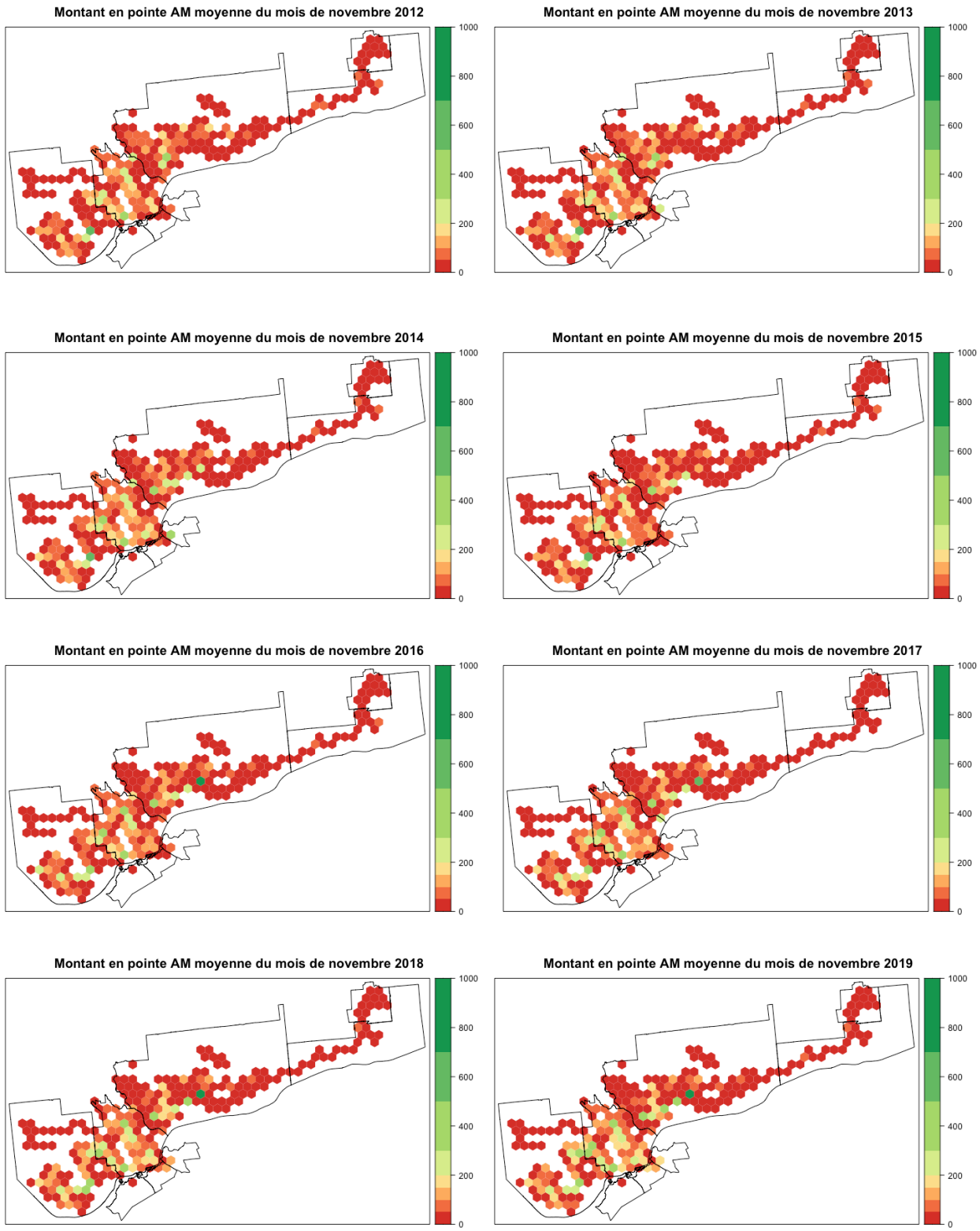


Figure 4-44 Moyenne des montants en pointe AM pour les mois de novembre des années 2012-2019

4.4.2 Logements

La Figure 4-46 représente le nombre de logements sur toutes les zones hexagonales. La donnée est actualisée via les attributions de permis de construire résidentiels. Les logements évoluent peu au cours du temps et l'évolution n'est pas bien visible à l'aide de ce type de carte. C'est pour cette raison qu'uniquement l'état du nombre de logements est uniquement illustré pour l'année 2012 et 2019. Il est tout de même possible de constater une zone dans le secteur de Hull étant passé de 100 logements à plus de 1500. L'évolution est tout de même pertinente pour être intégrée à un modèle.

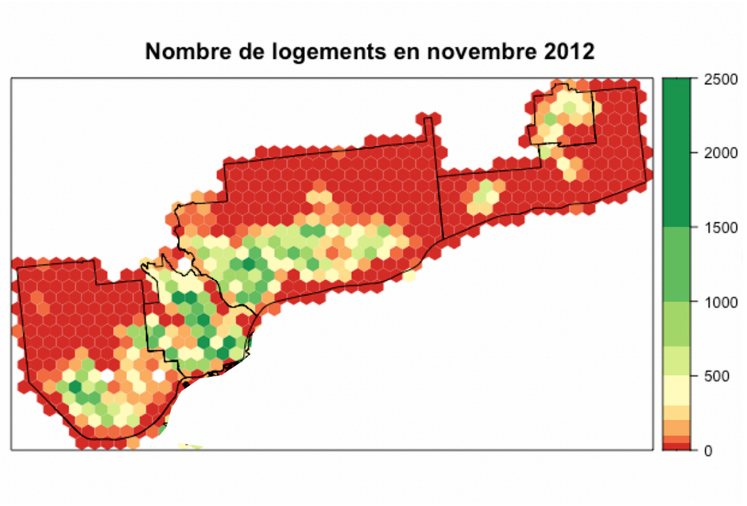


Figure 4-45 Nombre de logements par zone pour novembre 2012

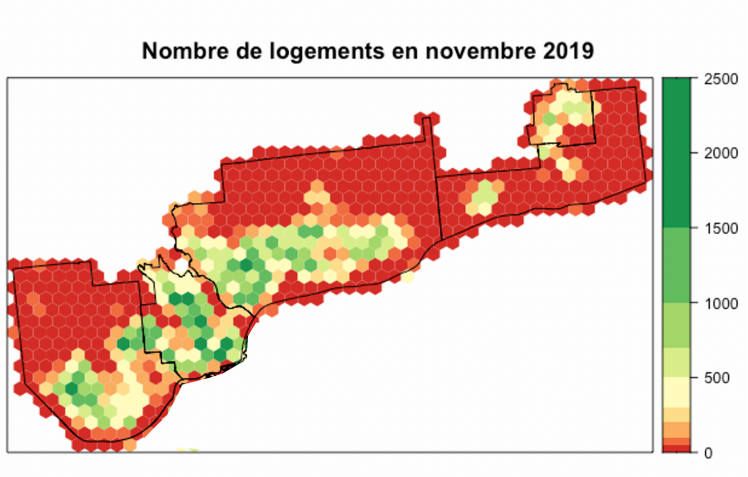


Figure 4-46 Nombre de logements par zone pour novembre 2019

4.4.3 Analyse des parcs-o-bus

Les parcs-o-bus sont des stationnements incitatifs disposés à proximité de stations de bus pour permettre aux usagers du réseau de transport en commun de garer leur véhicule pour prendre le bus. Ces stationnements sont destinés aux personnes titulaires d'un titre de transport, mais ne disposant pas de station de bus à proximité de leur domicile. Ces stationnements sont disposés au niveau de stations stratégiques; il est possible de visualiser ces parcs-o-bus sur la Figure 4-47.

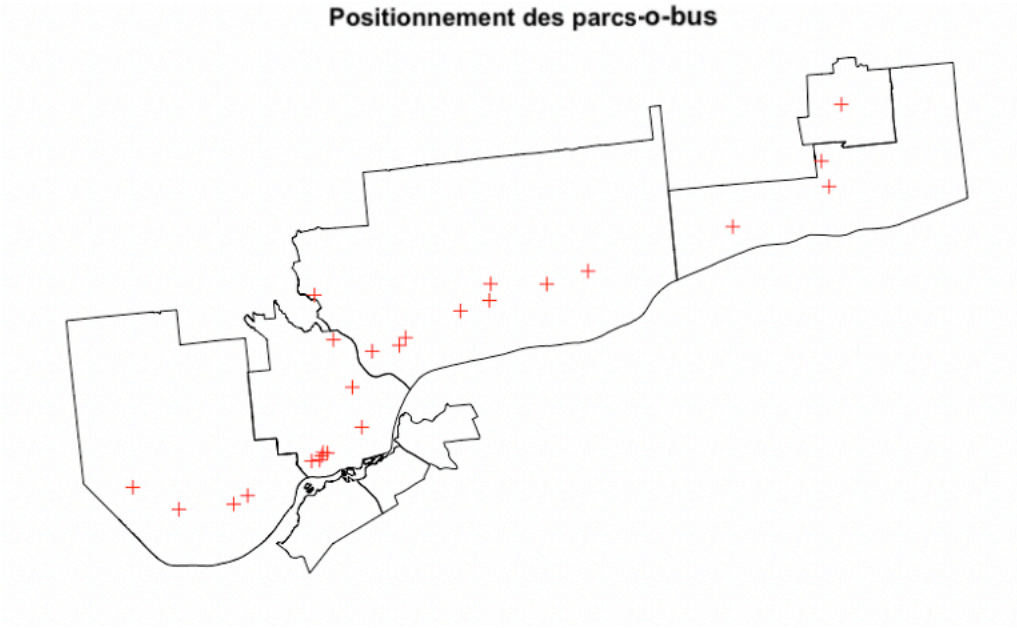


Figure 4-47 Positionnement des parcs-o-bus en 2019

Les informations disponibles pour ces parcs-o-bus sont leur capacité ainsi que des données sur les usagers des parcs-o-bus et leur provenance. Les données de capacité agrégée par zones hexagonales sont disponibles sur la Figure 4-48 pour l'année 2019.

Les données de provenance des utilisateurs sont obtenues grâce à des relevés de plaques effectués une à deux fois par ans dans la majorité des parcs-o-bus. Ces données permettent d'extraire différents indicateurs comme les taux d'utilisation ou encore les rayons d'attraction moyens des différents parcs de stationnement. La Figure 4-49 évalue l'utilisation de ces stationnements. Il est possible de constater que certains dépassent leur capacité. La STO explique que ces comptages peuvent également être effectués sur les stationnements à proximité lorsque que les parcs-o-bus sont pleins. Les véhicules ayant une plaque d'immatriculation appartenant à un secteur différent du stationnement en question sont donc comptabilisés dans les relevés.

La Figure 4-50 représente les rayons d'attraction moyenne des parcs-o-bus. Il est possible de constater que certains stationnements possèdent une grande zone d'influence. Ils sont généralement situés à proximité de grands axes d'accès à la ville. D'autres possèdent des rayons d'attraction très faibles alors qu'ils sont situés dans des zones dont l'offre en transport est forte. Ces stationnements attirent donc une clientèle qui n'utilise pas les lignes de rabattement pour se rendre sur des lignes comme le Rapibus. Ces usagers préfèrent s'y rendre en voiture. Ce phénomène pose plusieurs problèmes : les parcs-o-bus font de la concurrence à certaines lignes du réseau de la STO et le taux de possession de véhicules ne baisse pas, car des usagers ayant accès au réseau de TC continuent à utiliser la voiture. D'un point de vue de la prévision, un problème important émerge également : il faut réussir à quantifier ces usagers qui n'utilisent pas les lignes de rabattement pour se rendre sur le réseau principal.

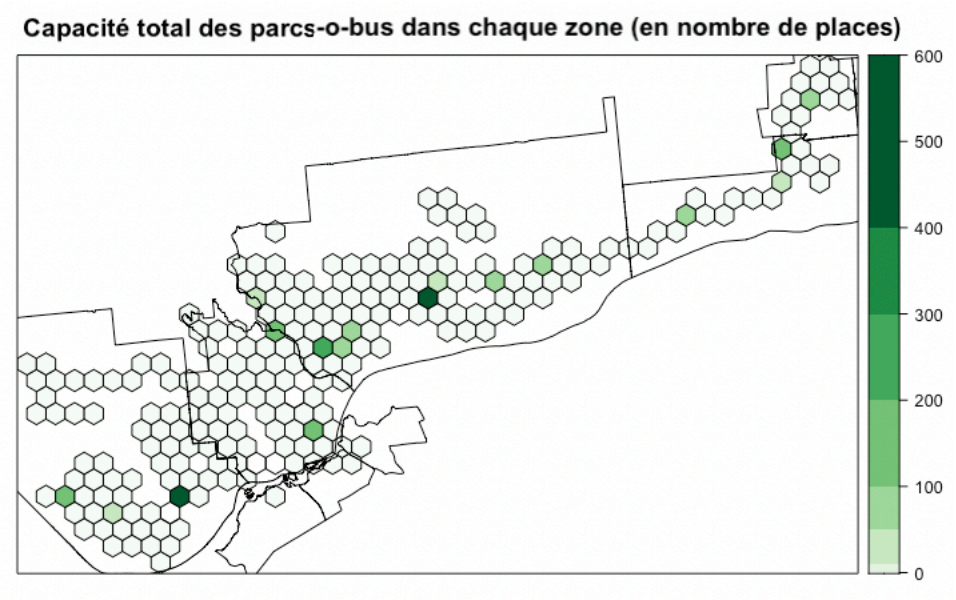


Figure 4-48 Capacité de parcs-o-bus par zone hexagonales en 2019

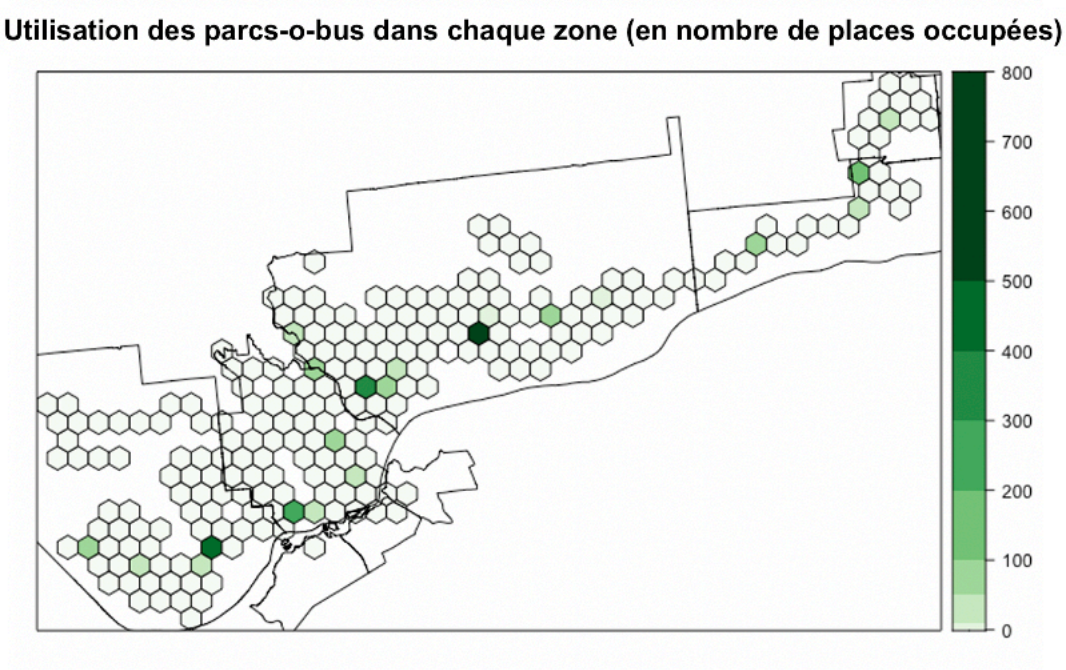


Figure 4-49 Utilisation des parcs-o-bus en 2019 selon les relevés de plaques

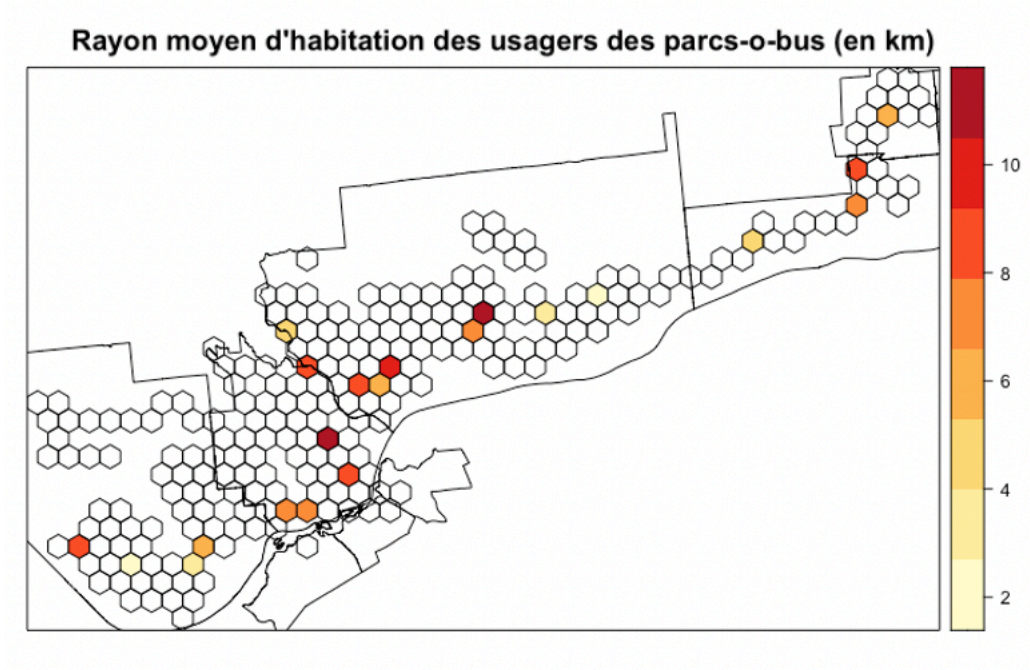


Figure 4-50 Rayon d'attraction moyen des parcs-o-bus en 2019

En revenant à la revue de la demande de la partie précédente, il est possible de constater des zones de forte demande en pointe du matin. Avec les informations de stationnement incitatif, le lien entre cette demande et les parcs-o-bus est constaté. Les capacités de stationnements incitatifs sont

vraisemblablement une variable explicative à inclure dans les modèles de prévision. De plus cette donnée est facile à prévoir, car l'implantation d'un parc-o-bus peut être facilement anticipée dans un plan stratégique. Il faut tout de même rester prudent, car la demande latente attirée par un parc-o-bus n'est pas connue et peut être inexistante dans certain cas ou simplement délocaliser une demande actuelle.

4.4.4 Étude des données OD

Cette partie propose une étude des matrices origine-destination (OD) pour se rendre compte de la distribution. Les matrices OD étudiées dans cette partie sont les OD regroupées selon les zones hexagonales déterminées dans la section précédente. C'est-à-dire qu'une paire OD est le nombre de déplacements s'effectuant sur une dénomination temporelle entre deux zones. Plusieurs dénominations temporelles seront étudiées et leur évolution à travers le temps est constatée.

La manière d'étudier ces matrices passe par une représentation graphique. Pour représenter graphiquement ces matrices, il faut ordonner les zones sur les lignes et les colonnes de la matrice. Cet ordre est déterminé par le classement des zones hexagonales selon leur distance à un point central qui doit être choisi. Une fois les zones ordonnées sur les lignes et les colonnes, la matrice peut être représentée grâce à une coloration des cellules. Plus une part OD est importante, plus la couleur est intense. Une telle représentation permet de mieux percevoir la distribution des déplacements.

Classer les zones selon leur distance à un point qui peut être assimilé à un centre permet de visualiser le caractère attractif d'un point tel le centre-ville : mouvements en provenance de la banlieue vers le centre en pointe du matin et l'inverse lors de la pointe du soir. Prendre le centre gravitaire des déplacements est généralement une option pertinente. La Figure 4-51 représente les 100 paires OD (entre arrêts) les plus importantes, qui sont tracées en fonction de leur intensité. Cela permet de constater des déplacements fortement centrés autour du parlement d'Ottawa. C'est pour cette raison que le centre qui est utilisé pour l'étude des matrices OD est l'arrêt d'autobus situé devant le parlement.

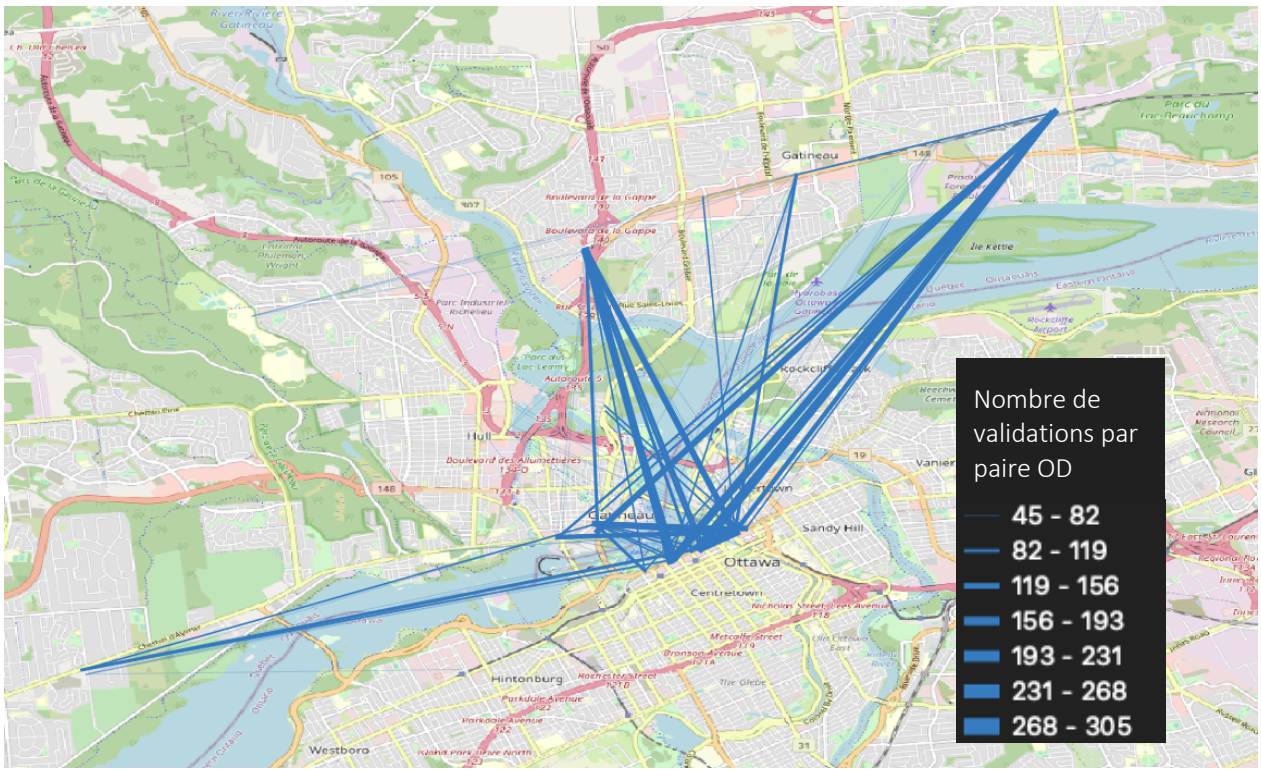


Figure 4-51 Étude des 100 trajets référencés les plus empruntés pour un jour ouvrable moyen de novembre 2019

La première dénomination temporelle étudiée est un jour ouvrable moyen de novembre des années 2019 et 2018. Les Figure 4-52 et Figure 4-53 concernent donc un jour ouvrable moyen du mois de novembre. Il peut y être constaté que la majorité des déplacements sont effectués avec comme origine ou destination le centre. Peu de déplacements s'effectuent entre les banlieues. À noter que l'échelle est logarithmique à cause d'un nombre restreint de zones concentrant une majorité des parts OD.

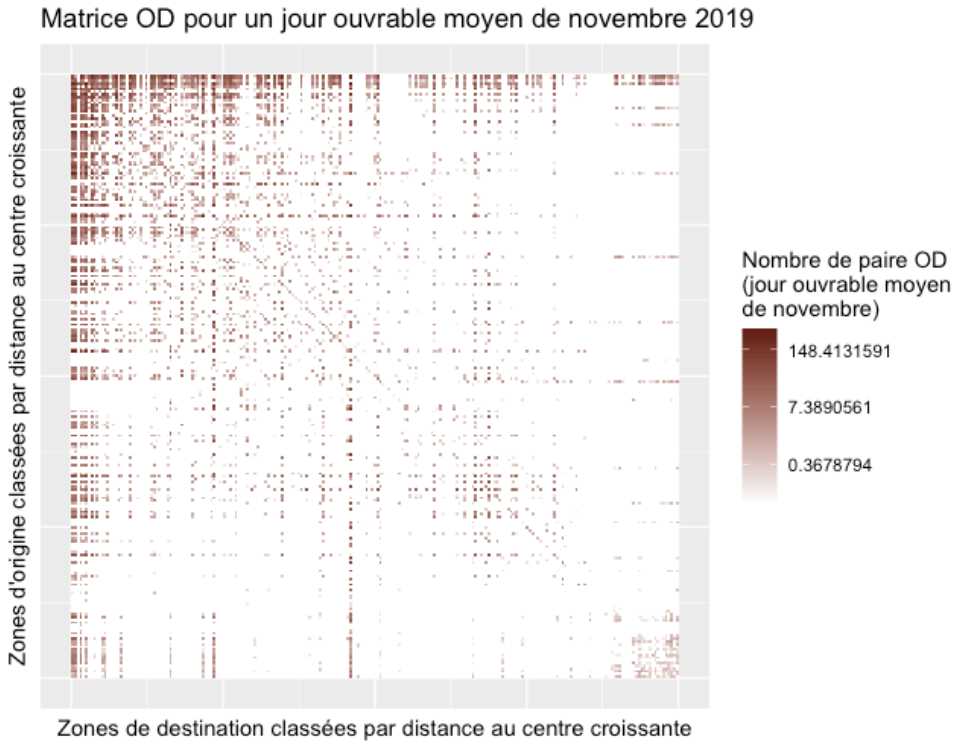


Figure 4-52 Matrice OD d'un jour ouvrable moyen de novembre 2019

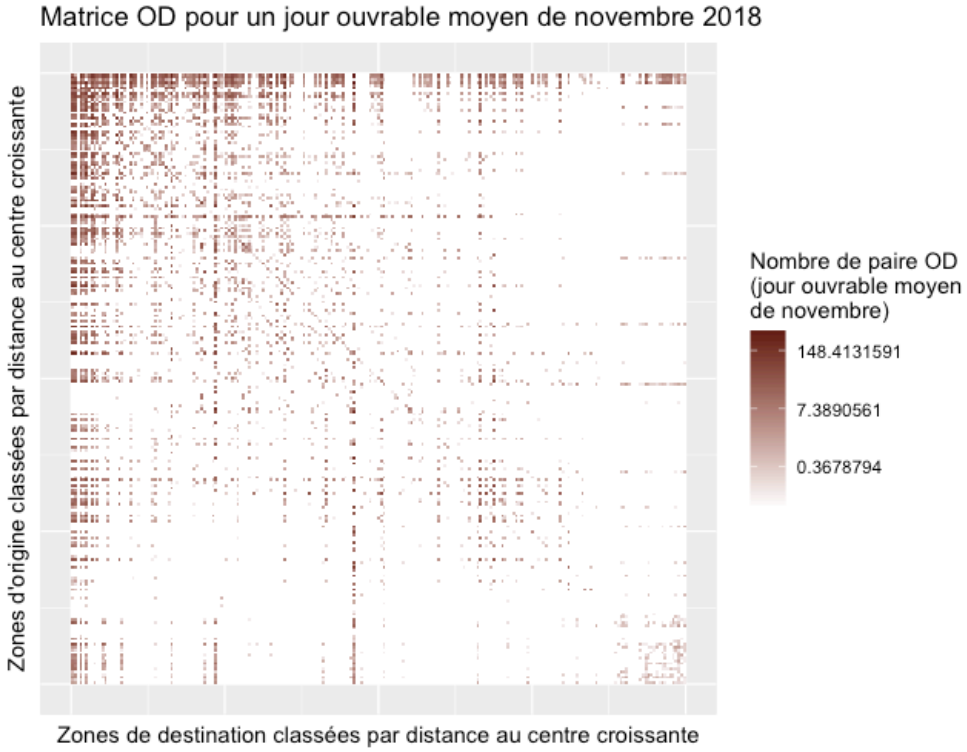


Figure 4-53 Matrice OD d'un jour ouvrable moyen de novembre 2018

Étudier ces matrices sur une journée entière ne permet pas de constater le caractère pendulaire des déplacements. Isoler les pointes AM et PM souligne le caractère pendulaire des déplacements sur le réseau. Les Figure 4-54, Figure 4-55 et Figure 4-56 concernent uniquement ces pointes. En pointe du matin, les parts OD les plus importantes se concentrent vers la gauche de la matrice indiquant des origines distribuées sur tout le réseau, mais des destinations concentrées au centre. En pointe du soir, elles se concentrent vers le haut de la matrice pour indiquer des déplacements à l'origine du centre et à destination de tout le réseau. Il est intéressant de noter que certains motifs se dessinent sur ces matrices : comme une ligne verticale au milieu de la matrice apparaissant constamment sur toutes les pointes. Cette zone de destination a été identifiée comme celle contenant le parc-o-bus de Labrosse. Cette zone est donc une destination conséquente tout au long de la journée. Cela peut s'expliquer tout d'abord par la présence d'un parc-o-bus, mais également par les établissements d'enseignement qui y sont situés.

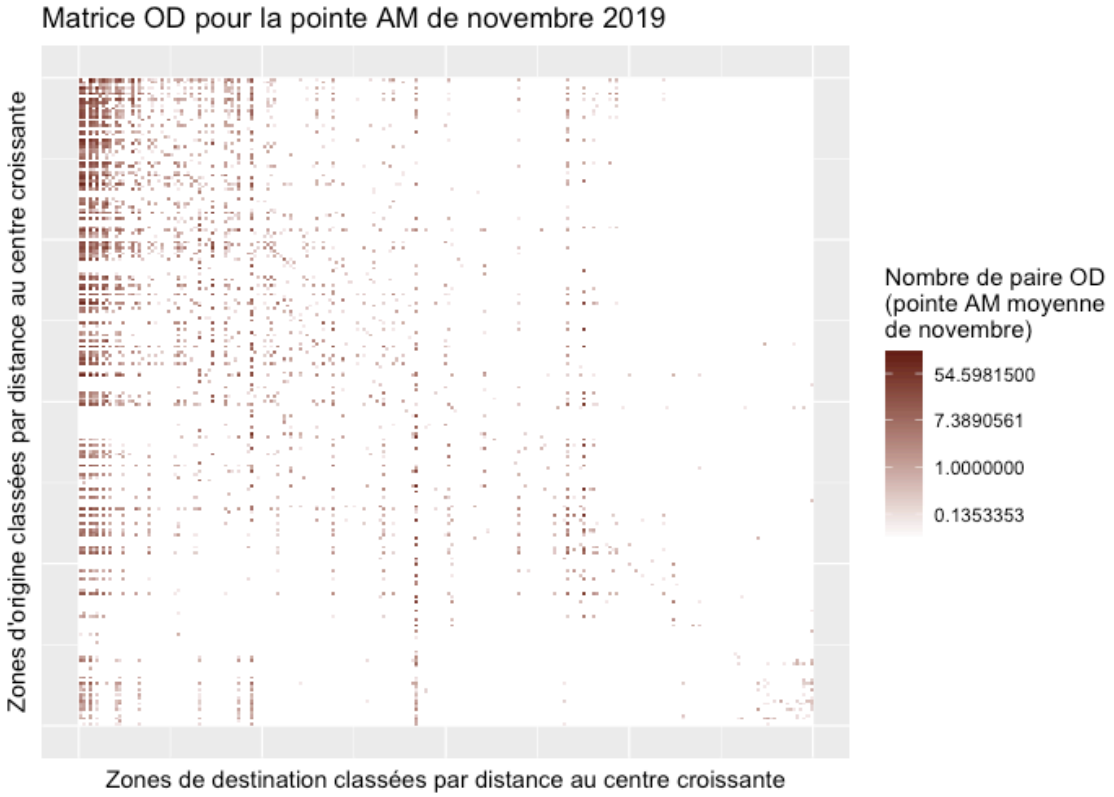


Figure 4-54 Matrice OD de la pointe AM d'un jour ouvrable moyen de novembre 2019

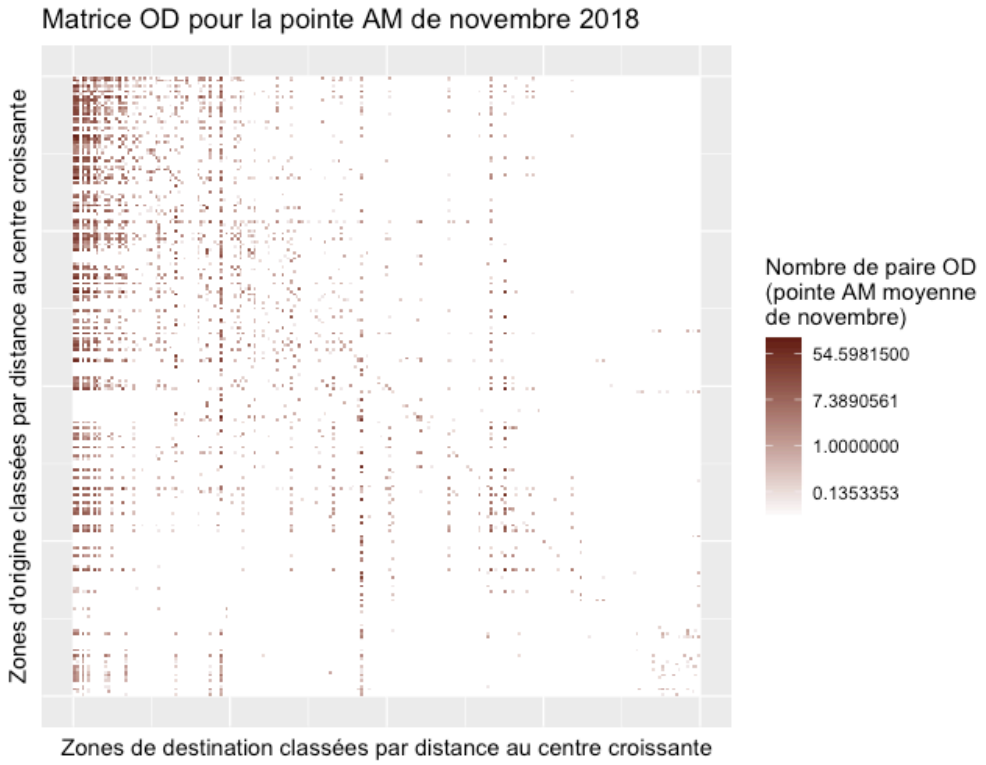


Figure 4-55 Matrice OD de la pointe AM d'un jour ouvrable moyen de novembre 2018

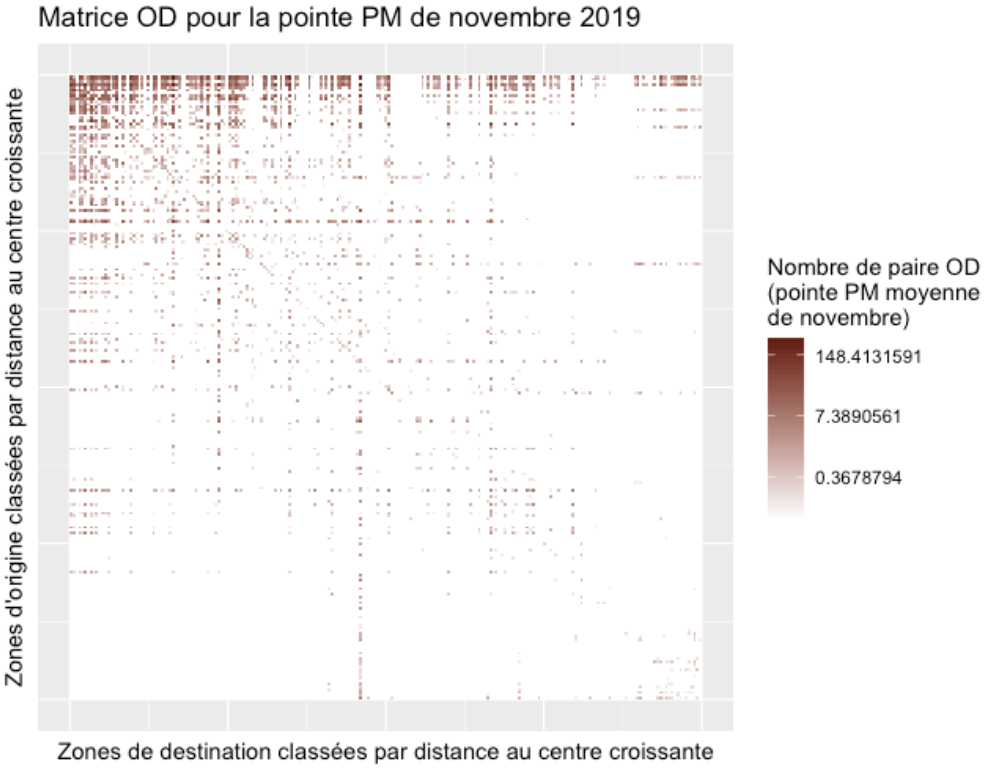


Figure 4-56 Matrice OD de la pointe PM d'un jour ouvrable moyen de novembre 2019

L'analyse rétrospective permet donc de comprendre les dynamiques qui influencent la demande sur les différentes échelles d'agrégation spatiale. Au niveau global, des variables socio-économiques font preuve d'une influence non négligeable. Les chapitres suivants permettent d'évaluer leur performance dans des modèles de prévision. Mais à plus petite échelle, la demande est sensible à d'autres phénomènes. D'autres variables sont donc susceptibles de devoir être intégrées aux modèles pour effectuer des prévisions performantes à cette échelle. L'évolution du nombre de logements ou la présence de stationnements incitatifs permettent d'expliquer en partie les variations de l'achalandage constaté. Les variables socio-économiques ne sont néanmoins pas disponibles au niveau d'agrégation le plus fin donc il est impossible d'observer leur influence à l'échelle locale. Ce chapitre a relevé un enjeu majeur dans la réalisation de modèle de prévision : il est difficile d'obtenir des variables explicatives précises spatialement et temporellement. Le but des modèles qui suivent est donc de maximiser la performance des modèles avec des données qui sont disponibles. Tous les phénomènes ne peuvent pas être pris en compte dans un modèle, mais une sélection de variables explicatives réfléchie peut permettre d'obtenir des modèles fiables.

CHAPITRE 5 PRÉVISION DE LA DEMANDE : RÉSEAU GLOBAL

L'analyse rétrospective a su faire souligner des corrélations de la demande en TC avec des variables explicatives. Le poids de la démographie ou encore des parcs-o-bus est conséquent, mais l'enjeu présent est d'intégrer ces variables dans des modèles de prévision et ainsi de quantifier leur effet sur la demande.

Le projet se base sur la mise en place de modèles de prévision à trois échelles d'agrégation différentes. Ce choix est justifié par l'étude des différentes variables explicatives qui peuvent être utilisées selon le niveau d'agrégation. À l'image des résultats de l'analyse rétrospective, certains phénomènes permettent d'expliquer des variations de la demande à l'échelle globale alors que pour d'autre, la demande ne leur est sensible qu'au niveau local. Ainsi, approcher la création des modèles de prévision de cette manière permet d'associer un poids à chaque variable explicative en fonction des échelles. Toutes les données n'étant pas disponibles à l'échelle des zones hexagonales, leur étude dans des modèles globaux permet de constater leur effet sur la demande sans pour autant les intégrer au modèle final.

Ce chapitre traite de l'ensemble des modèles de prévision de la demande totale sur le territoire de la STO. La structure de ce chapitre se base sur une complexification des modèles au cours des différentes expérimentations. La complexification se fait à la fois en fonction des variables utiliser, mais également en regard de la structure des modèles. Une sélection de modèles issus de la littérature sont calibrés et comparés les uns aux autres pour identifier les faiblesses et les forces de chaque modèle.

Plusieurs modèles sont testés dans ce chapitre, en commençant par les modèles autorégressifs se basant sur l'historique de la demande comme seule variable explicative. Deux structures autorégressives sont étudiées : le modèle ARIMA et un modèle de réseau de neurones. Par la suite, des facteurs explicatifs extérieurs sont intégrés aux modèles. Pour permettre de les prendre en compte, plusieurs déclinaisons de modèles de régression sont étudiées : régression linéaire multiple, régression avec correction de l'erreur par un modèle ARIMA et un modèle de régression sur séries temporelles. Ces types de modèles sont explicités par la suite et leurs différences soulignées.

5.1 Modèle autorégressif

L'objectif de cette première proposition est de créer un modèle simple qui nous permettra d'avoir une référence pour tester par la suite d'autres modèles de prévision. L'idée est de créer un modèle tendanciel de type séries temporelles. La création d'un tel modèle ne nécessite pas la présence de variables explicatives, mais se base uniquement sur le prolongement de la tendance concernant l'achalandage des dernières années.

Les modèles de séries temporelles sont généralement utilisés pour prévoir des évolutions à relativement court terme. Ceci est dû au fait que la précision est relativement fiable sur un nombre limité de valeurs générées lors de la prévision si aucune perturbation sans précédent ne vient perturber la demande. Ainsi, prévoir l'évolution annuelle de la demande avec ce type de modèle peut nous permettre d'avoir un horizon de prévision satisfaisant, car peu d'incréments sont générés (5 ans au moins) si bien sûr la tendance reste la même que par le passé. Ce type de modèle est excessivement simple, mais a pour but dans un premier temps de déterminer si la demande évolue de manière similaire à ses variations antérieures. Par la suite, les résultats (principalement les erreurs) de ce modèle orienteront le type de modèle à développer ainsi que les variables à intégrer.

Dans la suite des expérimentations sur ces modèles, trois dénominations temporelles sont étudiées pour calibrer des modèles de prévision. Ces trois échelles de temps permettent d'avoir une représentation de l'évolution de l'achalandage sur un jour ouvrable. Elles sont constituées des suivantes :

- Mensuel : sommes des validations de carte à puce sur tout le réseau de la STO pour chaque mois disponible.
- Jour ouvrable moyen : achalandage moyen pour un jour ouvrable de chaque mois.
- Pointe moyenne mensuelle : pour chaque mois de l'année, les validations de carte à puce sont sommées pour chaque jour ouvrable du mois sur la pointe AM (7h-9h) puis on prend la moyenne de ces totaux journaliers pour créer l'achalandage d'un jour ouvré moyen. On obtient donc une valeur de cette pointe moyenne pour les mois disponibles.

L'utilisation de données annuelles n'a pas été considérée, car les totaux de montants annuels ne pouvaient probablement pas constituer un ensemble de valeurs suffisantes pour calibrer le modèle. En effet, un modèle de séries temporelles s'appuie sur le fait de répliquer les motifs de variations

constatées dans le passé. Les données agrégées par an ne sont probablement pas assez nombreuses pour remplir ce rôle.

L'étude de ce type de modèle est développée de la manière suivante. Tout d'abord, la méthodologie de calibration du modèle est énoncée et des résultats sont obtenus. Ensuite, une étude est menée pour déterminer l'ensemble de calibration le plus pertinent pour établir le modèle. Cela a pour but d'augmenter la performance du modèle. Enfin, l'horizon de prévision du modèle est déterminé en étudiant l'évolution de l'erreur entre 1 et 5 ans.

5.1.1 Méthodologie de création d'un modèle de séries temporelles

Avant de mettre en place un modèle de série temporelle, il faut être capable de constater les erreurs et les performances des modèles. Pour ce faire, les indicateurs utilisés pour comparer les modèles sont explicités ci-dessous.

Le logiciel R Studio est utilisé pour réaliser la mise en place de ce modèle.

- RMSE (*root mean square error*) : l'erreur quadratique moyenne est définie par la formule suivante :

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$$

  Avec $\hat{y}_t$ est la valeur prévue de la variable dépendante pour l'instant t, y_t la valeur réelle observée à l'instant t et T est le nombre de valeurs prévues.

- MAPE (*mean absolute percentage error*) : le pourcentage d'erreur absolue moyen est défini par la formule suivante :

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right|$$

- Theil U : coefficient qui mesure la qualité de la prévision et qui est défini par la formule suivante :

$$U = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}}{\sqrt{\frac{1}{T}\sum_{t=1}^{T} y_t^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T} \hat{y}_t^2}}$$

5.1.1.1 Étude préliminaire

Avant de calibrer un modèle autorégressif, il faut tout d'abord étudier la variable pour déterminer ses caractéristiques et adapter les caractéristiques du modèle.

Pour étudier ces variations, la Figure 5-1 illustre la décomposition additive de la série temporelle composée des validations de CAP. La donnée étudiée ici est le nombre de montants totaux mensuel. Dans cette figure, nous pouvons observer les valeurs réelles (data), les résidus (remainder), le motif saisonnier (seasonal) et la tendance (trend). Cette décomposition a pour principal but de séparer la tendance de fluctuation de la variable et l'effet saisonnier. Le motif saisonnier est introduit par l'effet de l'année : beaucoup de fréquentation en automne et printemps, mais des baisses au cours de l'hiver et de l'été. L'aspect annuel de la variable est confirmé par l'étude d'autocorrélation disponible sur la Figure 5-2. L'autocorrélation significative avec un délai de 1 (l'unité est une année) montre cette répétition du motif chaque année. La tendance est à l'augmentation de la fréquentation notamment à partir de 2017.

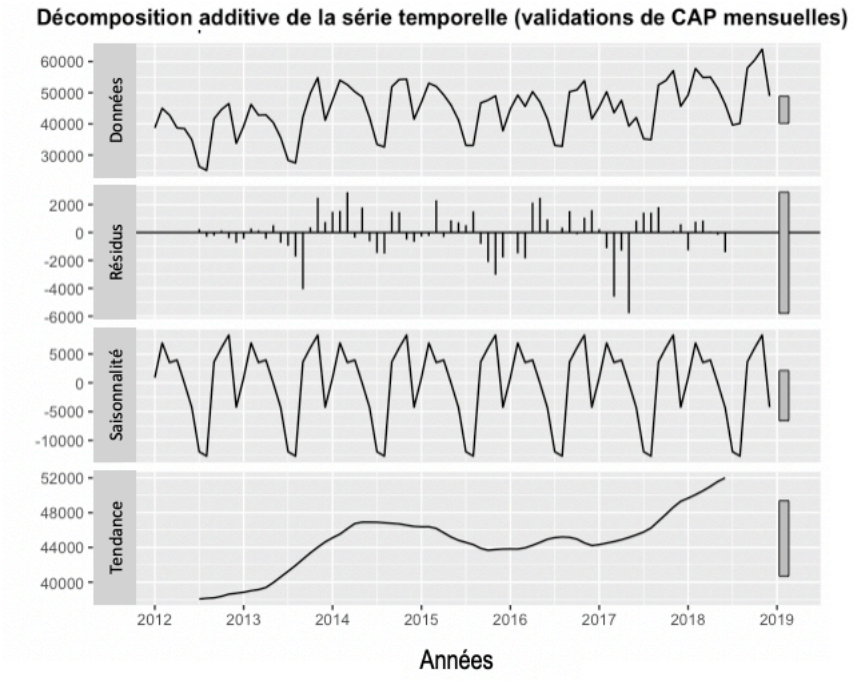


Figure 5-1 Décomposition additive des validations CAP mensuelles

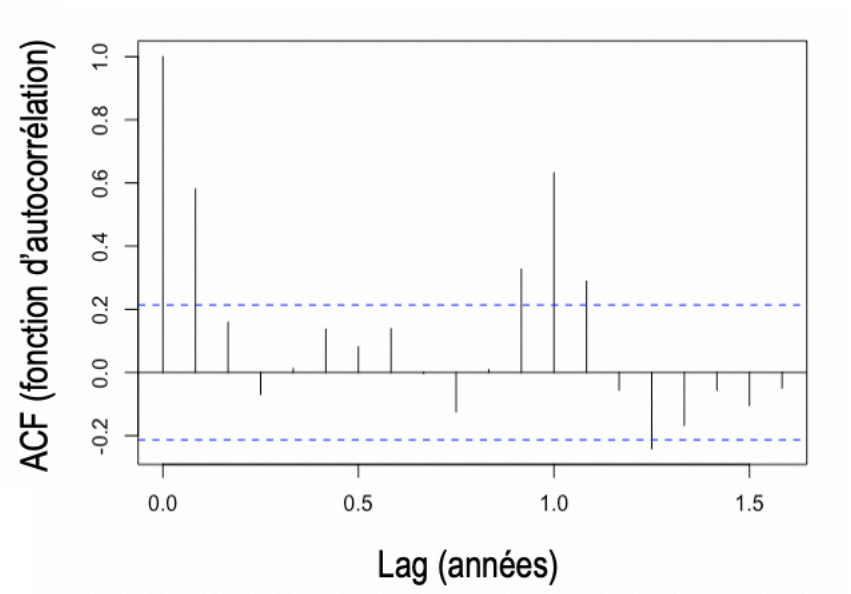


Figure 5-2 Analyse d'autocorrélation des validations de CAP

Cette décomposition indique que pour prévoir de manière pertinente cette série temporelle, il faut une méthode capable de prévoir l'évolution de la tendance et également celle du motif saisonnier. Cette information est importante pour la sélection du modèle.

La méthode de prévision ARIMA semble donc être pertinente pour cette tâche. ARIMA (*autoregressive integrated moving average*) est une approche permettant la prévision de séries temporelles en utilisant l'aspect autorégressif de la série. Elle utilise notamment une combinaison linéaire des valeurs passées. Une méthode autorégressive est appliquée aux valeurs passées de validations de CAP et une méthode de moyenne flottante (moving average) est appliquée aux résidus. Pour ce type de modèle, la série doit être stationnaire et c'est pour cela que préalablement la série temporelle peut être transformée en utilisant, dans le cas d'ARIMA, la différentiation (transformation de la série en la variation entre deux valeurs consécutives). Une telle transformation appliquée à un certain degré (d) a pour effet de rendre la série stationnaire. C'est un des paramètres qui doit être déterminé dans la création d'un tel modèle. Il faut également trouver le paramètre du terme autorégressif (p) et celui de la moyenne flottante (q). Une méthode simple ARIMA est définie par ces trois paramètres : ARIMA (p, d, q). Dans notre cas il faut également rajouter un terme qui définit l'aspect saisonnier et qui est défini de la même manière, mais appliqué à la variation saisonnière.

Mais avant d'appliquer le modèle à la variable, il faut séparer la variable en un ensemble dédié à l'entraînement du modèle est un autre à la validation. De manière générale, tout l'intervalle temporel disponible est divisé en deux suivant les proportions de 80% et 20% respectivement pour les deux usages.

Ici, prendre 2012 à 2017 pour entraîner le modèle et 2018 pour le valider est une option. Par la suite, si le modèle est validé, 2018 sera de nouveau introduit dans le modèle pour pouvoir prévoir à un horizon plus lointain. Une étude de l'effet de la sélection de différents ensembles d'entraînement est réalisée par la suite pour optimiser la calibration du modèle.

Dans R, une fonction permet d'automatiquement déterminer les paramètres énoncés précédemment. La fonction auto.arima() fournit donc les résultats suivants :

```
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:

          ma1     sma1

      -0.2639  -0.5102

s.e.   0.1334   0.1630

sigma^2 estimated as 5362701:  log likelihood=-541.64

AIC=1089.29   AICc=1089.72   BIC=1095.52

Training set error measures:

                    ME     RMSE      MAE        MPE     MAPE      MASE        ACF1

Training set 39.28621 2060.454 1368.618 0.08135521 3.041605 0.4159892 0.003991136
```

Figure 5-3 Résultats prévision ARIMA saisonnier

Le modèle obtenu possède un degré de différenciation de 1, pas de terme autorégressif et une moyenne flottante de degré 1. Le terme saisonnier est bien annuel (signifié par le [12]). Les coefficients d'erreur et de validation indiquent que le modèle est réalisable avec un degré de précision satisfaisant.

La prévision est donc affichée dans la Figure 5-4. L'ensemble de validation y est également présent. Visuellement, le modèle semble fournir des résultats satisfaisants, la prévision suit globalement la tendance et le motif saisonnier. Par la suite, des indicateurs de performance du modèle sont étudiés pour vérifier la validité de cette hypothèse. Les zones en bleu et bleu clair sont respectivement les intervalles de confiance à 80% et 95% pour une distribution supposée normale.

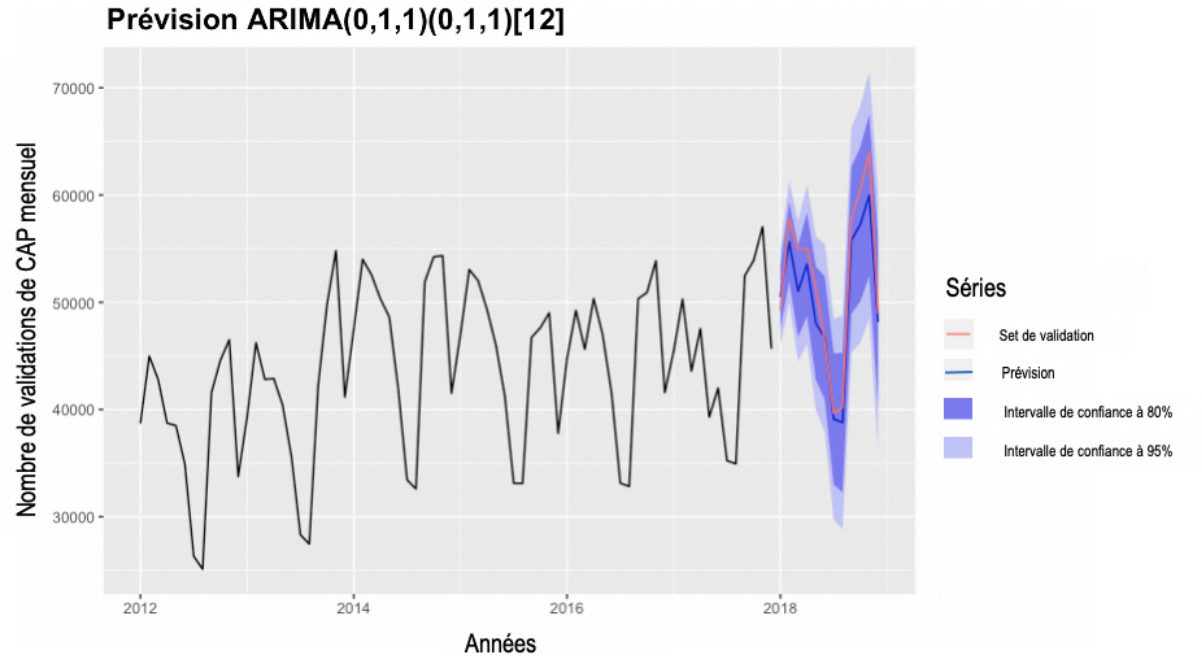


Figure 5-4 Prévision validations CAP sur 1 an du modèle ARIMA

Un résumé des erreurs et d'indicateurs de la pertinence du modèle sont les suivants :

Tableau 5-1 Performances du modèle ARIMA

Indicateurs	RMSE	MAPE	Theil's U
Valeurs	2375.5	3.7%	0.2637

Les résultats sont issus de la comparaison des données prédites et de l'ensemble de validation. Il y a une erreur moyenne (RMSE) de 2375,5, ce qui se retranscrit en un pourcentage d'erreur absolue moyenne (MAPE) d'environ 3,7%. Ces erreurs sont donc relativement faibles et indiquent qu'un tel modèle est pertinent sur un horizon de prévision d'un an. Le coefficient de Theil est un indicateur visant à évaluer la précision d'un modèle de prévision. Il s'avère y avoir une valeur relativement basse ce qui indique que le modèle est précis.

Pour évaluer la pertinence d'un modèle, il est également intéressant d'étudier les résidus. En effet, les résidus donnent les erreurs de la modélisation et leur structure communique beaucoup d'informations. Il est possible de savoir si le modèle utilisé permet d'intégrer toutes les variations

de la variable dépendante. Par exemple, une forte erreur à un instant défini suggère la non-intégration d'un évènement influençant la demande ponctuellement.

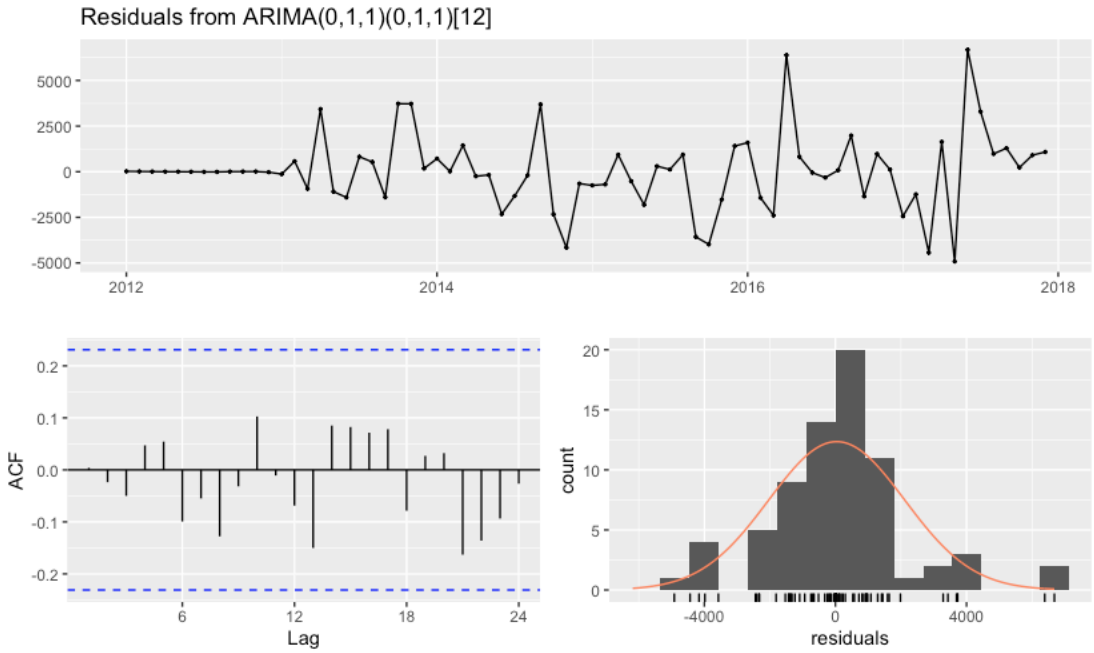


Figure 5-5 Résidus du modèle ARIMA des validations de CAP

Un diagnostic des résidus est donc proposé dans la Figure 5-5. La p-value du Ljung-Box test est de 0,8453 ce qui indique que l'autocorrélation n'est pas significative pour les résidus. Ainsi, le modèle retranscrit bien les différents phénomènes observables dans les fluctuations. De plus, la moyenne des résidus est très proche de 0 ce qui indique que le modèle est bien calibré. Les résidus pourraient tout de même avoir une distribution plus normale autour de 0 pour que le modèle soit parfaitement adapté.

5.1.1.2 Déterminer l'ensemble de données d'entraînement

Dans la littérature, il est possible de constater que la période de l'ensemble d'entraînement utilisé pour créer un modèle pouvait avoir un effet sur la performance. Notamment pour des données qui possèdent un régime transitoire. Un régime transitoire est une phase qui suit l'instauration d'un nouveau service. Ici, le nouveau service est l'utilisation de carte à puce comme titre de transport. Lors de l'ajout de nouveaux services, que ce soit la création d'une nouvelle ligne ou un nouveau type de perception, la fréquentation n'arrive jamais immédiatement à un régime stationnaire. Les usagers des transports en commun doivent l'adopter et cela prend du temps.

La Figure 6-5 (série : set de validation) montre bien que pour la période 2001-2004, la tendance augmente de manière importante qu'après 2004 où le nombre de validations semble commencer à adopter une tendance à peu près constante. 2001- 2004 est donc notre régime transitoire. Dans la suite de cette partie, un test sera réalisé pour savoir si intégrer la période transitoire est une source d'erreur ou non. Cette étude est réalisée sur le nombre de montants en pointe AM et cette valeur est une moyenne mensuelle.

La Figure 5-6 illustre les résultats obtenus pour le modèle prenant comme données de calibration toutes celles disponibles incluant la période transitoire. Le modèle surestime le nombre réel de validations.

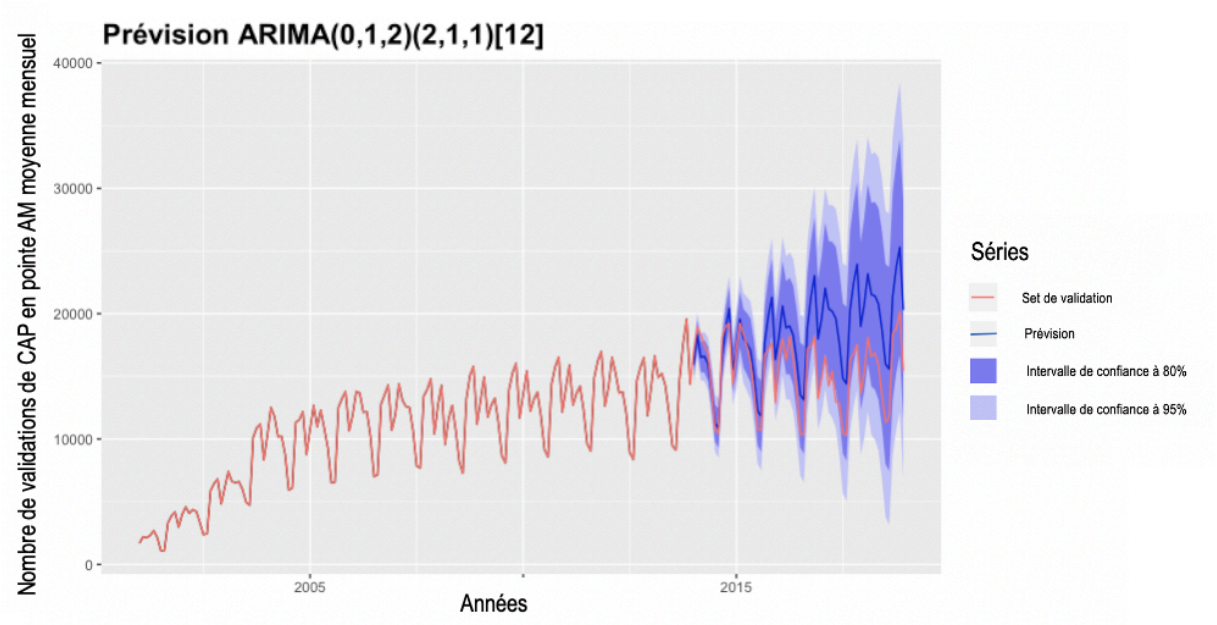


Figure 5-6 Prévision des CAP pour pointe AM d'un jour ouvré moyen mensuel à 5 ans

Pour tenter d'augmenter la précision, il est possible de supprimer les 3 premières années du segment de calibration du modèle. La Figure 5-7 donne le résultat pour une prévision sur 5 ans.

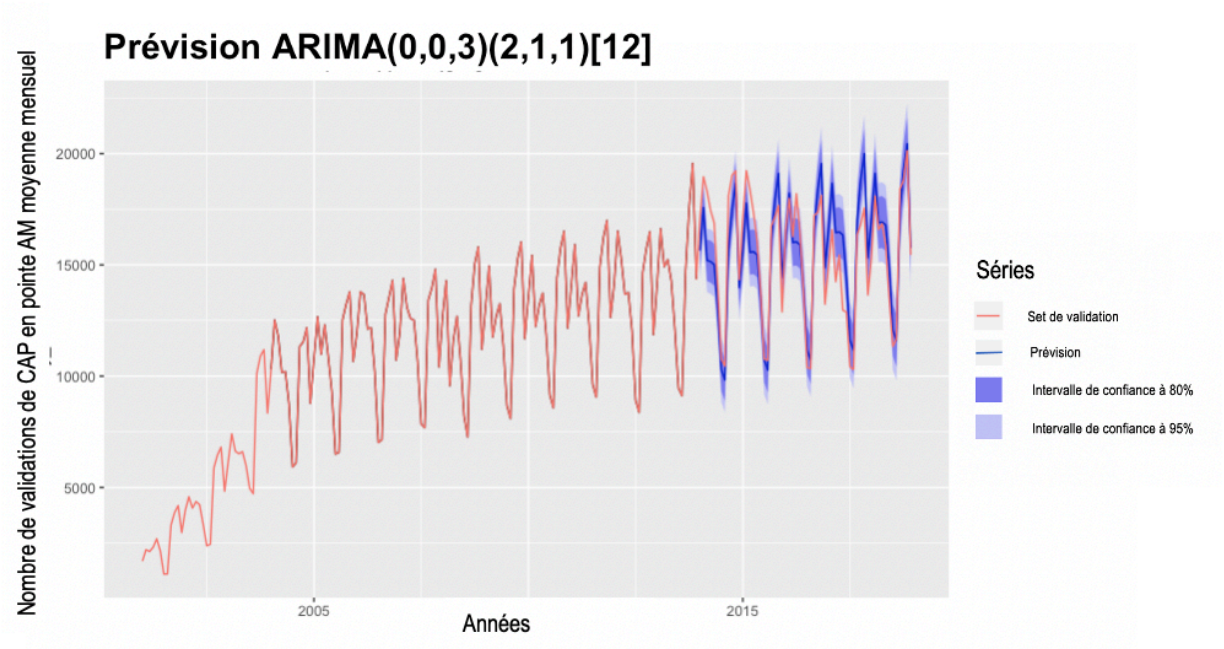


Figure 5-7 Prévision des CAP pour pointe AM d'un jour ouvré moyen mensuel à 5ans calibré sans régime transitoire

Le modèle proposé par la fonction auto.arima() n'est pas le même pour les deux ensembles de données d'entraînement. En effet, sans le régime transitoire, la série temporelle est déjà stationnaire donc n'a pas besoin d'être différenciée. L'ordre de la partie de moyenne flottante est également plus élevé. Visuellement, le modèle semble beaucoup plus précis. Seules les pointes durant les mois d'automne sont sous-estimées. C'est sûrement dû au fait que le degré de la partie saisonnière est plus faible, car elle est considérée plus stable par le modèle.

Le Tableau 5-2 fait la synthèse des erreurs du modèle et d'un indicateur quand ils sont comparés à un set de données de validation (2014 – 2018)

Tableau 5-2 Comparaison de l'effet de l'ensemble de données d'entraînement pour un modèle ARIMA

Validations CAP mensuelles pour pointe AM d'un jour ouvré moyen		**Horizon de prévision**				
Ensemble de calibration	**Indicateur**	**1 an**	**2 ans**	**3 ans**	**4 ans**	**5 ans**
2001 - 2013	**RMSE**	980,8	1455,7	2105,4	3196,2	3565,3
	MAPE	5,1%	7,3%	11,1%	17,5%	20,1%
	Theil U	0,2519	0,3968	0,5729	0,9526	1,0641
2004 - 2013	**RMSE**	1645,8	1498,3	1355,4	1500,7	1394,4
	MAPE	8,4%	7,6%	6,8%	8,1%	7,3%
	Theil U	0,4312	0,4061	0,3576	0,4192	0,3999

Le modèle excluant la période transitoire est beaucoup plus précis à long terme. L'erreur de ce modèle reste relativement constante dans le temps. La tendance est bien mieux estimée que pour le modèle incluant la période transitoire. En effet, durant cette période la tendance est à une forte augmentation; les reliques de cette forte augmentation sont donc responsables de la surestimation des validations après 2 ans de prévision.

L'observation à noter également est l'amplitude des intervalles de confiance (à 80% et 95%) qui varient de manière conséquente entre les deux méthodes. Pour le modèle excluant le régime transitoire, il est très faible ce qui indique une variance faible des résidus observés sur l'ensemble de calibration. Alors que pour l'autre modèle, l'amplitude est plus forte d'autant plus que la série est différenciée ce qui devrait contribuer à réduire cette amplitude. La tendance du modèle sans régime transitoire est donc beaucoup plus constante.

Ainsi, ces résultats orientent fortement la conclusion vers la non-intégration du régime transitoire dans la calibration d'un modèle autorégressif.

5.1.2 Détermination de l'horizon de prévision d'un modèle ARIMA

Le modèle de prévision voulu par la STO doit fournir des données à l'horizon de 5 ans. Mais dans la littérature, les modèles de séries temporelles sont souvent utilisés pour prévoir à court terme, mais assez peu à moyen terme. Le but de cette partie est donc de déterminer dans notre cas quel est l'horizon de prévision d'un tel modèle.

Les modèles de séries temporelles possèdent la caractéristique de pouvoir prévoir de manière fiable un certain nombre d'itérations ou de valeurs pour l'unité temporelle utilisée. Ainsi, prévoir l'achalandage annuel possèdera une précision plus forte sur un certain nombre d'itération donc sur un certain nombre d'années. Un modèle mensuel aura une précision sur un nombre similaire d'itérations mais l'unité est le mois. Un horizon de prévision plus grand est donc susceptible d'être observé sur un modèle ARIMA annuel plutôt que sur un mensuel. Le nombre d'itérations fiables sera peut-être semblable, mais pas l'horizon. D'autres paramètres entrent évidemment en compte quand il s'agit de déterminer l'horizon de prévision comme le nombre de données disponibles pour la calibration du modèle ou encore la conservation de la tendance d'évolution dans le temps. Dans le cas présent, il y a un nombre plus faible de valeurs pour calibrer un modèle annuel que pour un modèle mensuel. Ainsi potentiellement moins d'itérations seront fiables pour un modèle annuel.

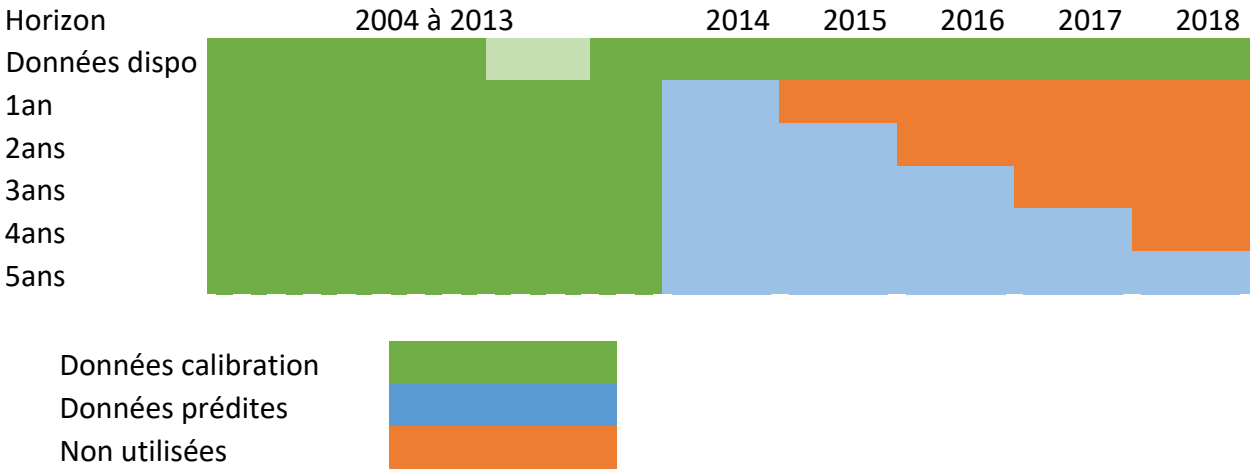
Dans un premier temps, il faut se concentrer sur l'évaluation de l'horizon de prévision pour un modèle mensuel. Pour réussir à déterminer cet horizon de prévision, il s'agit de procéder comme il s'en suit :

- Tester des modèles ARIMA pour trois variables différentes :
    - Total des validations de carte à puce mensuelles
    - Nombre de validations moyen pour un jour ouvrable moyen de chaque mois
    - Nombre de validations moyen en pointe AM pour un jour ouvrable moyen de chaque mois
- Pour chacune de ces variables, déterminer un ensemble de valeurs de calibration du modèle ARIMA. 2004 à 2013 d'après les recommandations de la partie précédente.

- Prévoir à plusieurs horizons les différentes variables (pour tous les horizons d'une variable, les paramètres du modèle ARIMA sont stables) : 1 an, 2 ans, 3 ans, 4 ans et 5 ans.
- Recueillir les erreurs et les indicateurs de validations des différents modèles en comparant la prévision à un ensemble de valeurs dédiées à la validation.

Ainsi pour chaque variable, le modèle utilise les données illustrées par le Tableau 5-3.

Tableau 5-3 Schématisation de la méthode de détermination de l'horizon de prévision



Pour l'instant, les données de 2009 à 2011 ne sont pas disponibles. Pour une étude préliminaire, un modèle ARIMA est utilisé pour combler les 3 années manquantes. La méthode utilisée produit des résultats qui « récupèrent » bien la tendance en 2012. Mais le motif saisonnier est tenu constant ainsi que la tendance. Le fait de combler les années manquantes de cette manière lisse la variation sur ces 3 années, mais n'est là que pour permettre d'avoir assez de données pour réaliser la calibration des modèles.

Cette partie est à reprendre une fois que les données de CAP seront disponibles. Les valeurs qui apparaissent par la suite sont là à titre indicatif.

Tableau 5-4 Indicateurs pour la validation des modèles en fonction de leur horizon

		Horizon				
Variable	**Indicateur**	**1 an**	**2 ans**	**3 ans**	**4 ans**	**5 ans**
CAP en pointe AM (moyenne mensuelle)	**RMSE**	1645,8	1498,3	1355,4	1500,7	1394,4
	MAPE	8,4%	7,6%	6,8%	8,1%	7,3%
	Theil U	0,4312	0,4061	0,3576	0,4192	0,3999
CAP jour ouvré (moyenne mensuelle)	**RMSE**	5659,8	4785,1	4144,2	3847,9	4124,6
	MAPE	10,6%	9,2%	7,6%	6,8%	7,3%
	Theil U	0,6062	0,5652	0,4825	0,4538	0,4902
CAP totale mensuelle	**RMSE**	111727	104925	97511	96027	96210
	MAPE	9,1%	8,4%	7,7%	7,7%	7,6%
	Theil U	0,5946	0,6061	0,5682	0,5830	0,5916

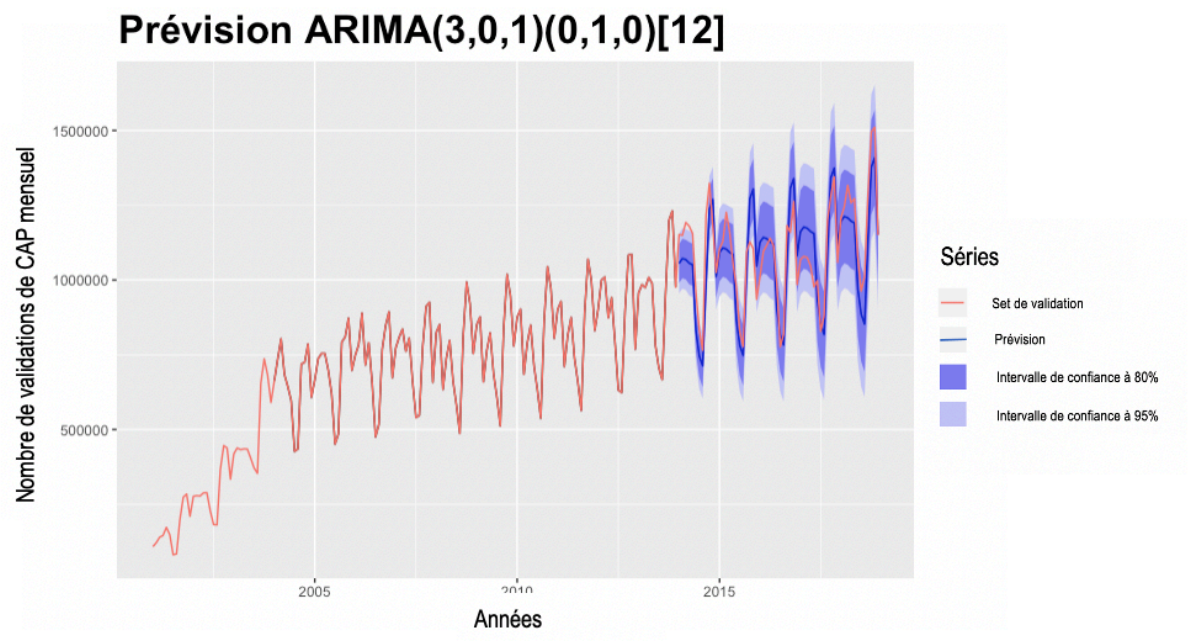


Figure 5-8 Modèle ARIMA validations CAP mensuelles

Grâce à ces résultats, l'horizon de prévision semble pertinent pour 5 ans. La tendance se conserve très bien sur ce segment temporel. Seule une légère augmentation de l'achalandage en 2015 a du mal à être constatée pour tous les modèles. Un modèle tendanciel n'arrive pas à prévoir cette augmentation : cela peut signifier que la tendance est modifiée par des paramètres extérieurs qui ne sont présentement pas inclus dans le modèle. Par la suite, certains paramètres vont devoir être intégrés dans le modèle pour limiter cette erreur. Le motif saisonnier évolue également. Le modèle ARIMA utilisé ne prend pas en compte l'évolution importante de ce motif. D'autres études doivent être menées pour prendre en compte cette évolution.

Il faut savoir que ces variables désignent des quantités différentes, mais elles sont fortement corrélées. Un modèle ARIMA s'adapte assez bien aux fluctuations présentement observées, car la tendance est globalement conservée pour toutes les variables. Mais ce modèle n'est pas adapté si une perturbation majeure des habitudes de mobilité survient.

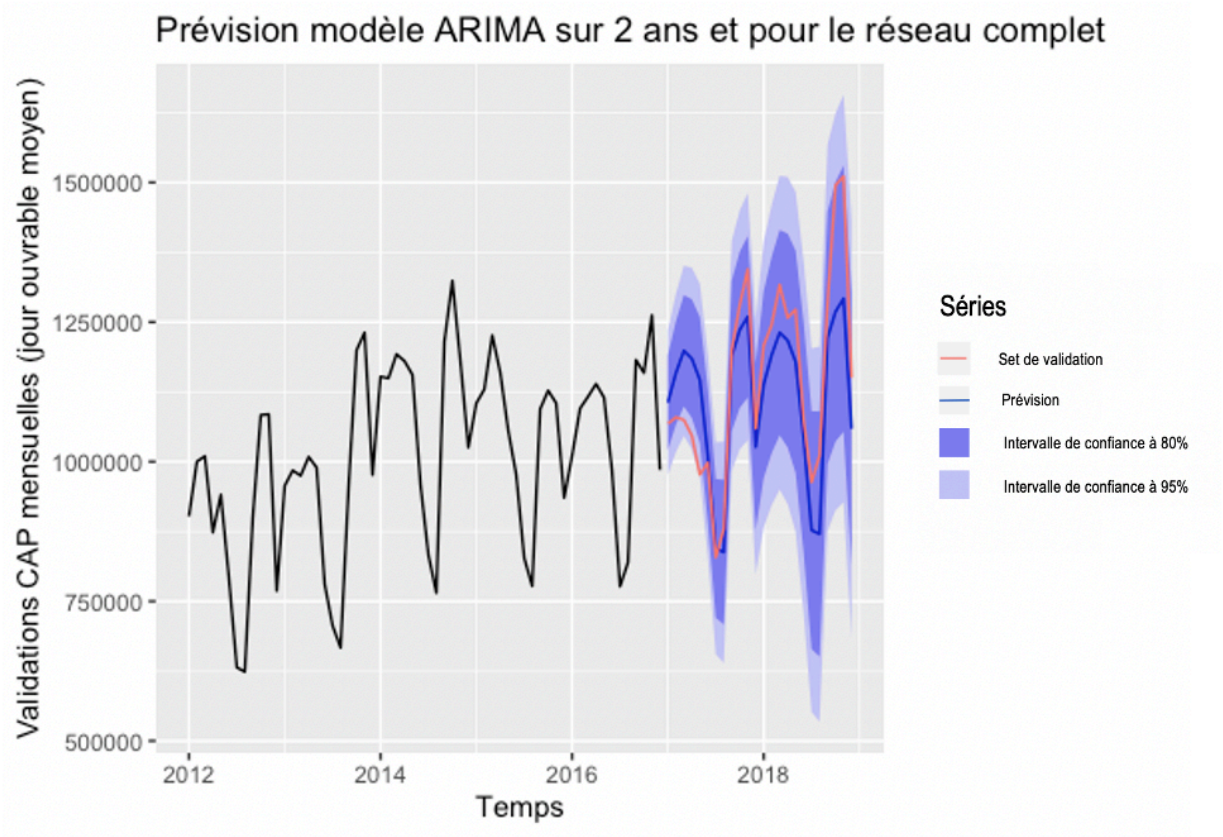


Figure 5-9 Modèle ARIMA à un horizon de 2 ans

5.2 Modèle de réseau de neurones autorégressif

Les modèles de réseau de neurones permettent de prendre en compte des relations complexes et non linéaires dans les données. Cette partie a pour but de déterminer si ce type de modèle peut permettre d'avoir des résultats plus performants qu'un modèle ARIMA.

Un modèle de réseau de neurones est emprunté à l'intelligence artificielle dans la mesure où il utilise une structure simplifiée et similaire à celle du cerveau humain. Un réseau de neurones est composé d'entrées, de couches de neurones composées de nœuds et de sorties. Des données sont attribuées aux nœuds de la couche d'entrée qui sont par la suite modifiés par des nœuds des couches dites « cachées » et les résultats sont donc fournis à la couche de sortie. Chaque lien entre les nœuds se voit associer un poids dans la calibration du modèle et chaque nœud une fonction non linéaire comme une sigmoïde. L'entraînement du modèle est réalisé par un algorithme qui minimise la fonction de coût définie de manière similaire à l'erreur quadratique.

La méthode utilisée par la suite est un réseau de neurones autorégressif (NNAR). C'est un modèle avec une unique couche cachée, k nœuds sur cette couche et prend en compte les p dernières valeurs de la série temporelle. Ces paramètres doivent être déterminés et peuvent l'être automatiquement par la fonction nnetar() dans R. Un terme saisonnier peut aussi est implémenté dans le modèle. La période est de m unités et le modèle remonte à P périodes pour collecter les données. Ces paramètres peuvent également être déterminés automatiquement.

L'algorithme d'entraînement a une dimension aléatoire; il est donc souvent d'usage de faire plusieurs calibrations de ce modèle, de les projeter et de prendre un résultat moyen. Cette méthode permet de produire des résultats plus performants.

Pour déterminer si un tel modèle prenant en compte des composantes non linéaires est plus performant pour prévoir l'achalandage, une étude comparative des résultats produits par un modèle ARIMA et NNAR est réalisée.

Pour permettre une comparaison, plusieurs modèles sont entraînés sur la période 2012-2016 et validés sur 2017-2018 :

- Modèle ARIMA généré automatiquement par la fonction auto.arima().
- Modèle NNAR généré automatiquement par la fonction nnetar().
- Modèle NNAR dont les paramètres sont déterminés manuellement. NNAR(12,2,7)[12]

Les résultats sont rapportés dans le Tableau 5-5 et la Figure 5-10 rapporte la validation des graphes du modèle NNAR manuel et de l'ARIMA.

Tableau 5-5 Comparaison des modèles NNAR et ARIMA

Modèles	**RMSE**	**MAPE**	**THEIL U**
AUTO ARIMA	1319,2	6,92%	0,4124
AUTO NNAR	1699,6	9,15%	0,5288
MANUEL NNAR	1422	7,34%	0,4686

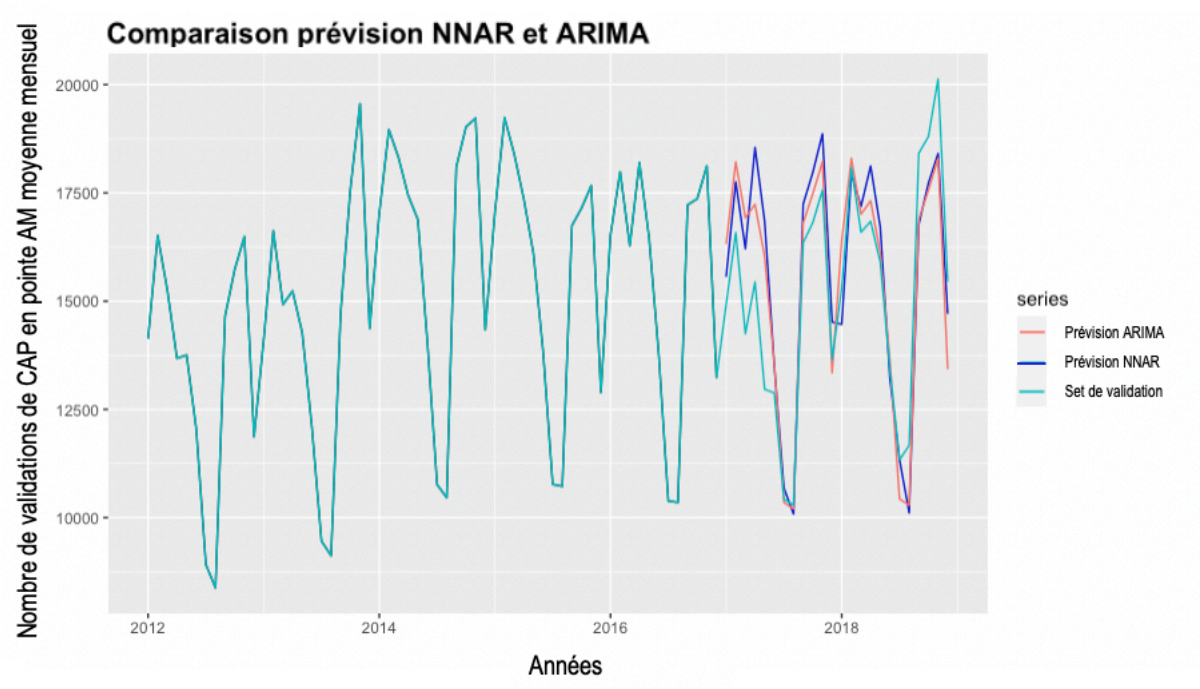


Figure 5-10 Validation des modèles NNAR manuel et ARIMA

Les deux modèles de prévision ARIMA et NNAR produisent des résultats relativement proches, mais ARIMA produit une erreur moins grande. Pour cette variable, un modèle de série temporelle linéaire produit des résultats plus performants qu'un modèle de réseau de neurones. Il ne semble donc pas pertinent d'introduire un modèle d'intelligence artificielle pour prévoir cette variable. Il a de plus été montré dans la littérature que les modèles ARIMA étaient performants pour décrire la demande en transport.

5.3 Modélisation d'une année par 3 mois

Pour répondre plus pertinemment aux attentes du projet, il n'est pas nécessaire de modéliser l'année au complet. Pour simplifier les données, il est possible d'effectuer la prévision sur uniquement 3 mois de l'année. Ces trois mois sont sélectionnés comme ceux les plus achalandés en moyenne sur un jour ouvré et depuis 2012. Le Tableau 5-6 expose le nombre moyen de validations de CAP pour un jour ouvrable pour chaque mois depuis 2012.

Tableau 5-6 Sélection des mois les plus achalandés pour la modélisation

Mois	CAP Jour ouvré moyen
janv	45526
févr	50790
mars	47746
avr	47735
mai	44458
juin	40514
juil	32744
août	32317
sept	49026
oct	51654
nov	54213
déc	41485

Il est possible de considérer la sélection des mois de février, octobre et novembre pour la modélisation de chaque année. Ainsi, la prévision de la demande se fera par le biais de la prévision des validations de CAP sur ces 3 mois chaque année. Néanmoins, cette analyse montre que la période de l'automne est la plus achalandée. Il est commun dans les prévisions de planification de ne considérer que cette période. Malgré le fait que février est le troisième mois le plus achalandé de l'année, la sélection est faite sur les trois mois d'automne : septembre, octobre et novembre. Il y a plus de continuité temporelle entre ces mois et la pratique courante préconise l'utilisation de cette année type. À noter que la différence entre l'achalandage de février et septembre n'est pas significative.

La modélisation de l'année en 3 mois d'automne permet d'avoir un modèle simplifié contenant toutes les informations permettant le dimensionnement et la prévision des ressources nécessaires à l'opération du réseau. Dans cette partie, l'objectif est de comparer la prévision sur cette année type d'un modèle ARIMA calibré sur 12 mois précédemment.

Le problème avec cette modélisation est que l'ensemble de calibration des données est réduit. Les modèles doivent dans certains cas être calibrés manuellement, à savoir choisir manuellement les paramètres du modèle ARIMA.

5.3.1 Prévision

Pour attester du gain en performance obtenu par la modélisation de l'année grâce aux trois mois les plus achalandés, le modèle sur 3 mois est comparé aux modèles précédents. Il s'agit de prévoir les montants sur la pointe AM d'un jour ouvré moyen pour les trois mois de l'année type. L'objectif fixé ici pour ce type de modèle est de prévoir la demande à un horizon de 2 ans. De manière identique aux modèles précédents, les modèles possèdent un horizon de prévision de 2 ans (2017-2018) et sont calibrés sur les années 2012-2016.

Le modèle ARIMA généré sur les validations en pointe AM de cette année type de septembre, octobre et novembre est disponible sur la Figure 5-11.

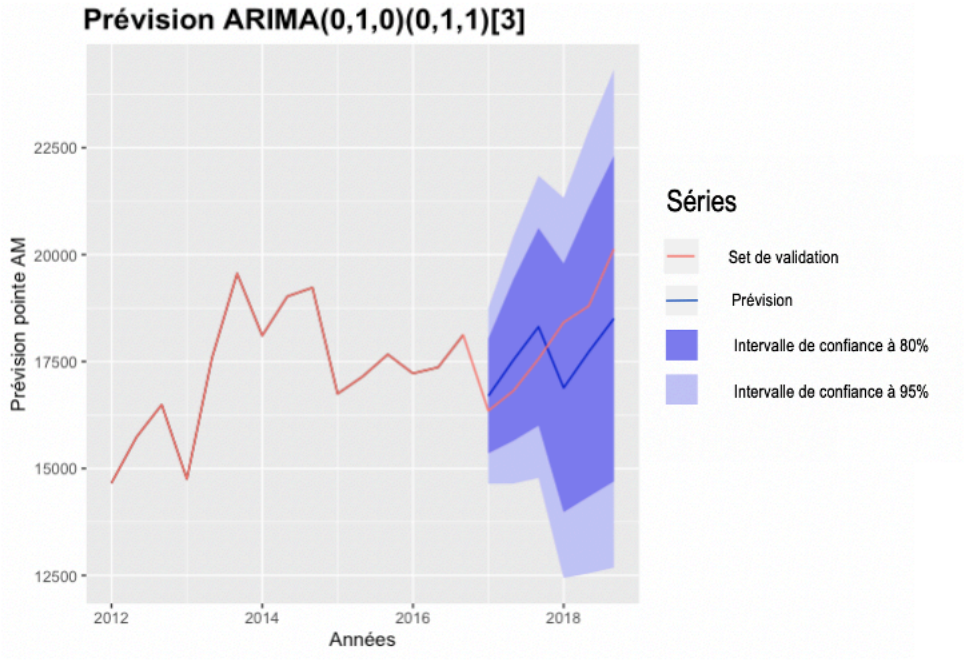


Figure 5-11 Prévision ARIMA pointe AM moyenne d'un jour ouvré

Le modèle est calibré pour prévoir la pointe AM moyenne mensuelle. Les résultats des indicateurs de performance sont disponibles sur le Tableau 5-7. L'année type des mois d'automne permet d'obtenir un gain en précision sur le modèle ARIMA. L'omission des autres mois dans le modèle permet de considérer uniquement les dynamiques régissant la demande à cette période. Le gain en précision représente 16% de l'erreur quadratique moyenne. Utiliser cette année type permet de gagner en performance sans s'encombrer de prévision sur des mois non pertinents pour soutenir la planification stratégique.

Tableau 5-7 Résultats de la comparaison de l'année type d'automne avec l'année complète

Découpage de l'année	**Année type de 3 mois (septembre, novembre et octobre)**	**Année de 12 mois**
Pointe AM moyenne mensuelle	ARIMA (0,1,0)(0,1,1)[3]	ARIMA (0,1,1)(0,1,1)[12]
RMSE	1105,5	1319
MAPE	5,46 %	6,91 %
Theil U	1,4549	0,4124

Par la suite, dans la mesure où il y a assez de données pour calibrer les modèles, la demande est prédite sur ces trois mois d'automne. Le gain en performance est important et ces mois sont suffisants pour remplir les objectifs du projet.

5.4 Introduction de la population comme variable explicative

Il a été vu dans les parties précédentes que la demande ne peut pas uniquement s'expliquer par ses variations passées. Il faut donc trouver des facteurs explicatifs. Dans l'analyse de corrélation, le facteur semblant avoir l'effet le plus important est l'augmentation de la population. La revue de la littérature confirme aussi l'importance de la population. Pour un premier essai d'intégration d'une variable explicative dans le modèle, il a été choisi d'intégrer la population comme variable explicative pour effectuer un modèle prédictif sur un an. De plus, la donnée la plus impactée par les évolutions de la population est les montants en pointe du matin. Ceci est d'autant plus vrai que la zone d'agrégation est fine. Mais pour pouvoir comparer les résultats avec les autres modèles, cette variable est sélectionnée pour effectuer de la prévision même si cela concerne le territoire global.

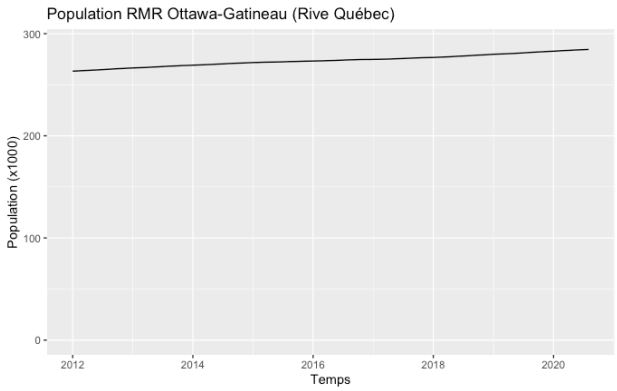


Figure 5-12 Évolution de la population de la RMR O-G (rive Québec)

Le modèle utilisé dans cette partie est un modèle de régression linéaire avec correction de l'erreur par un modèle ARIMA. Il permet donc l'introduction d'une variable explicative ainsi que de conserver le caractère autorégressif du modèle ARIMA. Le modèle est calibré de la manière suivante : un modèle de régression linéaire est calculé tout en autorisant les résidus à être autocorrélés (paramètre non voulu dans une régression linéaire simple), puis ces résidus sont corrigés par un modèle ARIMA. La combinaison de ces deux étapes permet donc de produire une prévision.

L'enjeu est de déterminer si cette structure de modèle apporte plus de performance. Il faut noter que l'avantage de l'intégration d'une variable explicative est de pouvoir par la suite créer des scénarios d'évolution de la population est ainsi de constater l'impact sur la demande. De plus en cas de modification brusque et sans précédent de la tendance de la variable explicative, la prévision sur la demande prend ce paramètre en compte.

Le modèle développé par la suite met en place la structure énoncée précédemment. La population est la variable explicative, mais deux berceaux de population sont testés. En effet, la zone d'influence du réseau de la STO sur la région n'est pas connue avec précision. La population de Gatineau et la population de la RMR Ottawa-Gatineau (rive Québec) sont testées pour déterminer le pouvoir explicatif de chacune de ces variables. La RMR Ottawa-Gatineau (rive Québec) regroupe Gatineau et les zones alentour sans compter la partie Ottawa. Les résultats des deux modèles calibrés sont présents dans le Tableau 5-8 ainsi que le modèle ARIMA comme référence.

Tableau 5-8 Comparaison de la finesse des données et de la population de référence

Variable	**CAP de la pointe AM d'un jour ouvré (moyenne mensuelle)**		
Modèle	**Régression avec correction d'erreur ARIMA (Population annuelle Gatineau)**	**Régression avec correction d'erreur ARIMA (Population recensement RMR O-G)**	**ARIMA**
RMSE	1226,5	1399,2	1207,6
MAPE	6,42%	7,38%	6,03%
Theil U	1,5251	1,7623	1,4443

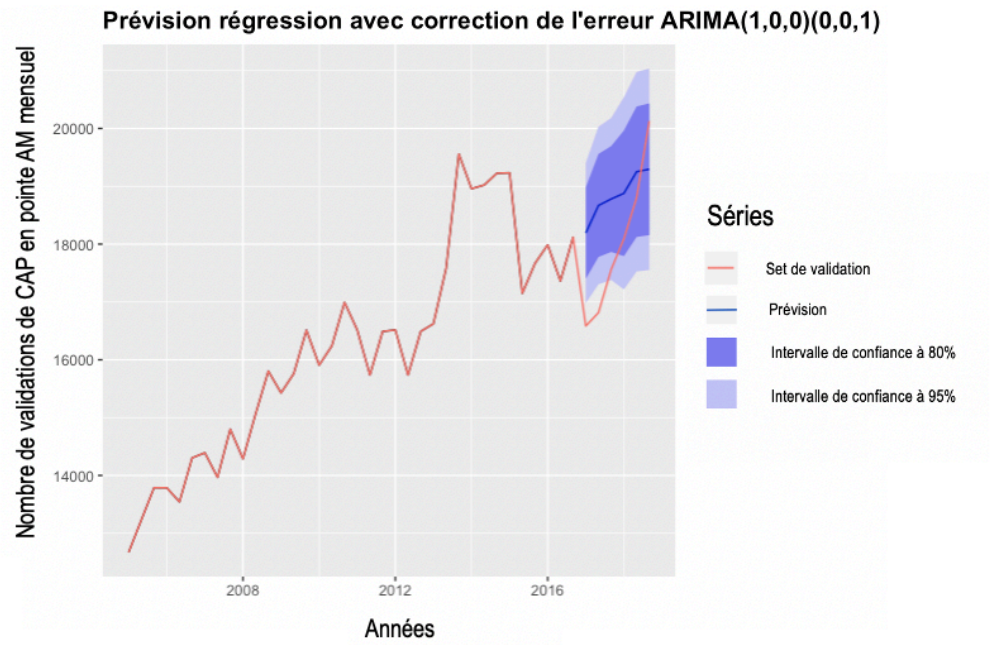


Figure 5-13 Régressions avec correction ARIMA pour les validations de CAP en pointe d'un jour ouvrable moyen

L’utilisation de la population de Gatineau produit des résultats presque aussi précis que le modèle ARIMA. L’utilisation d’une variable explicative est bénéfique, car elle permet de tenir compte de modifications sans précédent du territoire de Gatineau (ici de la population) ce qui n’est pas possible avec un modèle ARIMA. La zone d’emprise fournissant des résultats plus pertinents est la ville de Gatineau.

Malgré une erreur faible sur une prévision d’un an, les variations ne sont pas aussi bien décrites que pour le modèle ARIMA. La tendance stable de l’évolution de la population ne permet pas d’apporter plus d’information que le modèle ARIMA. Un modèle purement autorégressif est capable de très bien prédire si la tendance est stable. L’avantage d’une variable explicative dans ce type de modèle est de souligner des perturbations sans précédent. Dans le cas présent, la population ne remplit pas ce rôle. Ainsi, le gain en performance n’est pas constaté.

Les perturbations non expliquées par le modèle sont visibles lors de l’étude des résidus. C’est un outil très important pour améliorer les modèles, car il est possible d’identifier des périodes où le modèle possède une erreur importante et par la suite trouver la cause de cette perturbation.

L’étude des résidus est illustrée par la Figure 5-14. Le modèle n’est pas mauvais (pas d’autocorrélation significative et résidus centrés autour de zéro), mais certaines valeurs extrêmes dans les résidus montrent que certains phénomènes ne sont toujours pas pris en compte (mi 2013 et 2015). Il faut donc trouver une variable capable d’expliquer ces variations.

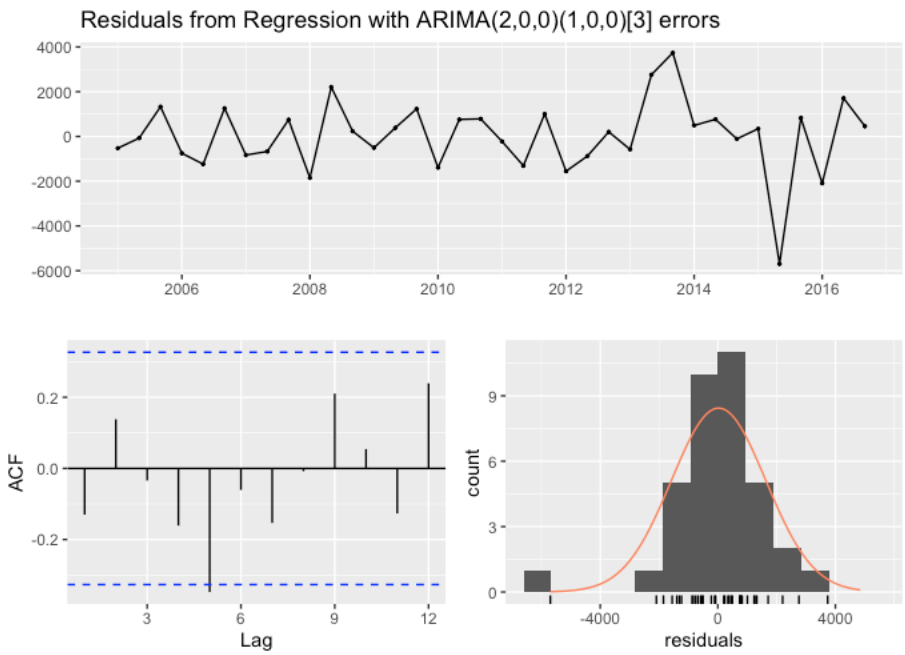


Figure 5-14 Étude des résidus du modèle de régression avec correction de l'erreur ARIMA pour les validations moyennes d'un jour ouvré

5.5 Modèle économétrique

Dans la littérature, certains modèles font usage de variables uniquement socio-économiques pour prévoir les variations de la demande. Ce type de modèle peut être pertinent pour la prévision de la demande sur le réseau global dans la mesure où les variables démographiques et économiques concernent la globalité du territoire de Gatineau. L'objectif de cette section est donc de tester si l'usage unique des variables économétriques permet d'obtenir des résultats pertinents pour prévoir la demande globale sur le réseau.

5.5.1 Variables

Les permis de construire se présentaient comme une variable idéale aux regards de ces besoins. Le permis de construire possède une localisation spatiale et temporelle très précise ce qui permet de pouvoir agréger ces données comme bon nous semble. Mais il s'est avéré qu'il manquait de pouvoir explicatif dans cette variable notamment concernant les facteurs socio-économiques. L'objectif est donc de trouver d'autres phénomènes qui influencent la demande en TC. Les variables de

population, d'emplois ou encore de chômage disponibles mensuellement pour l'ensemble du territoire sont étudiées afin de déterminer si il est possible de les intégrer dans un modèle de prévision.

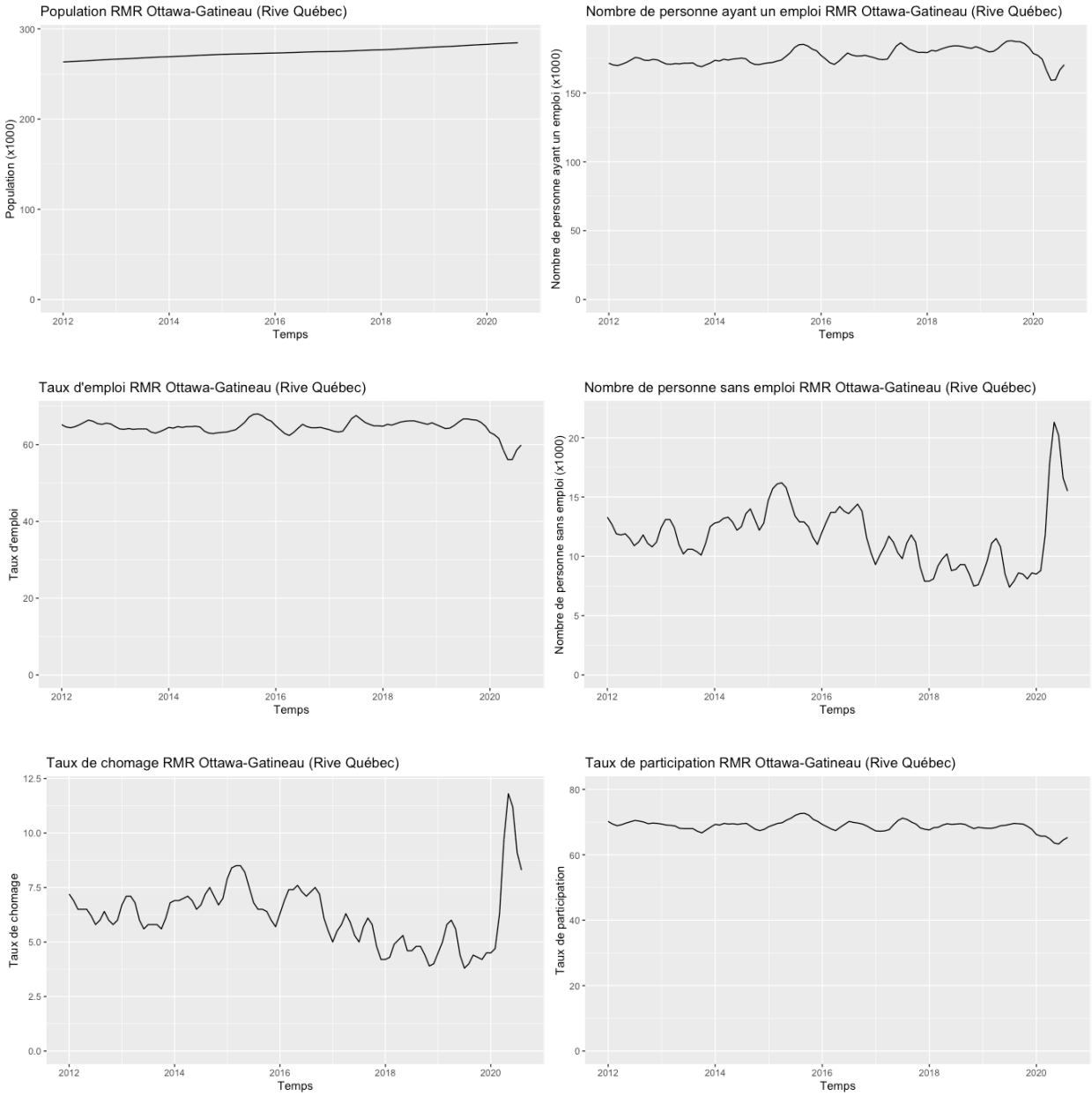


Figure 5-15 Évolution des variables socio-économiques

Ces variables sont candidates dans la création d'un modèle socio-économique. Mais afin de déterminer quelles variables sont pertinentes et peuvent être intégrées au modèle, il faut tout d'abord effectuer une analyse de corrélation. Les résultats de l'étude de corrélation linéaire (coefficient de Pearson) sont disponibles sur la Figure 5-16.

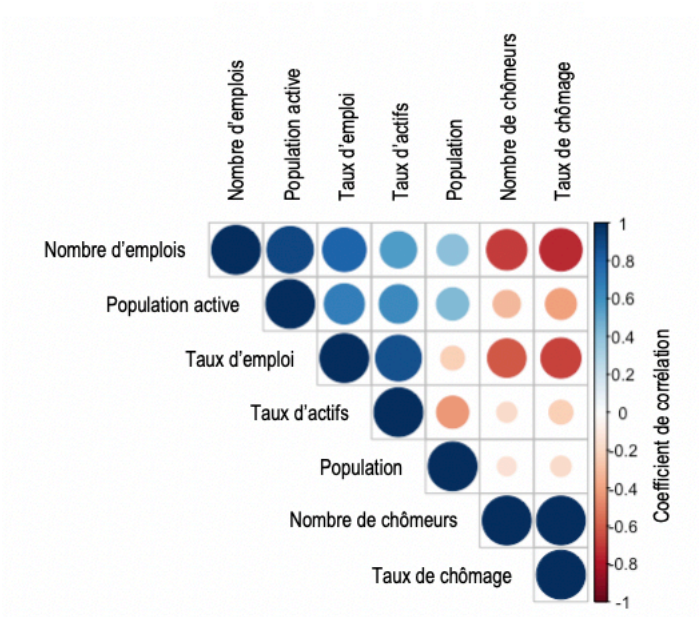


Figure 5-16 Corrélogramme des variables trouvées pour la RMR O-G (rive QC)

Aux vues des résultats de l'analyse de corrélation linéaire (Pearson) des potentielles variables de contexte, deux variables qui pourraient être intéressantes à intégrer dans les modèles sont :

- La population pour tenir compte de l'effet d'échelle.
- Le taux d'emploi, car les travailleurs sont connus pour générer une grande part de l'achalandage TC en pointe.

Ces deux variables ne sont pas significativement linéairement corrélées et peuvent donc être mises en parallèle de la variable locale qui est représentée par les permis de construire pour déceler une potentielle corrélation.

5.5.2 Modèle

Les variables candidates sont celles déterminées dans la section précédente. L'analyse de corrélation a déterminé que le taux d'emploi et la population étaient indépendants et ainsi pouvaient

être intégrés simultanément à un modèle de régression. Chacune de ces variables concerne la rive québécoise de la RMR d'Ottawa-Gatineau; ces données sont actualisées pour obtenir des données mensuelles.

Deux modèles sont testés ici afin de déterminer la performance de l'usage de ces variables. Ces deux modèles ont pour but de prévoir le nombre de validations de carte à puce de type origine sur un jour ouvrable moyen, cela calculé mensuellement et sur tout le réseau de la STO. Les modèles sont les suivants :

- Un modèle de régression linéaire simple : les variables explicatives sont constituées de la population et du taux d'emploi. La formulation à l'instant t est la suivante:

  $$\mathrm{Y}_t = \beta_0 + \beta_1 * P_t + \beta_2 * E_t$$

  Où les β sont des constantes à déterminer,

  Y est la variable dépendante, P la population et E le taux d'emploi

- Un modèle de régression linéaire sur série temporelle : les variables explicatives sont les mêmes au détail près qu'un terme autorégressif est ajouté. La formulation à l'instant t est la suivante :

  $$Y_t = \beta_0 + \beta_1 * P_t + \beta_2 * E_t + \beta_3 * Y_{t-i}$$

  i est un retard à déterminer.

Après avoir vérifié la stationnarité et l'indépendance des variables, la variable de population est transformée à l'aide d'une différenciation d'ordre 2 pour vérifier les hypothèses des différents modèles. Les modèles sont calibrés et présentés ci-dessous.

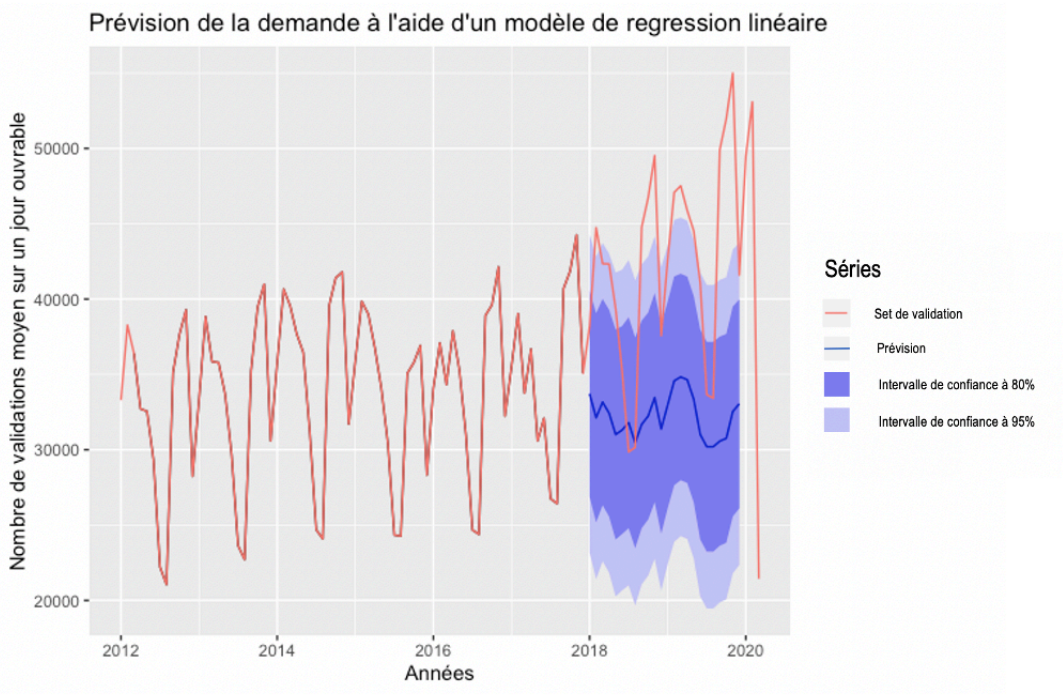


Figure 5-17 Prévision régression linéaire

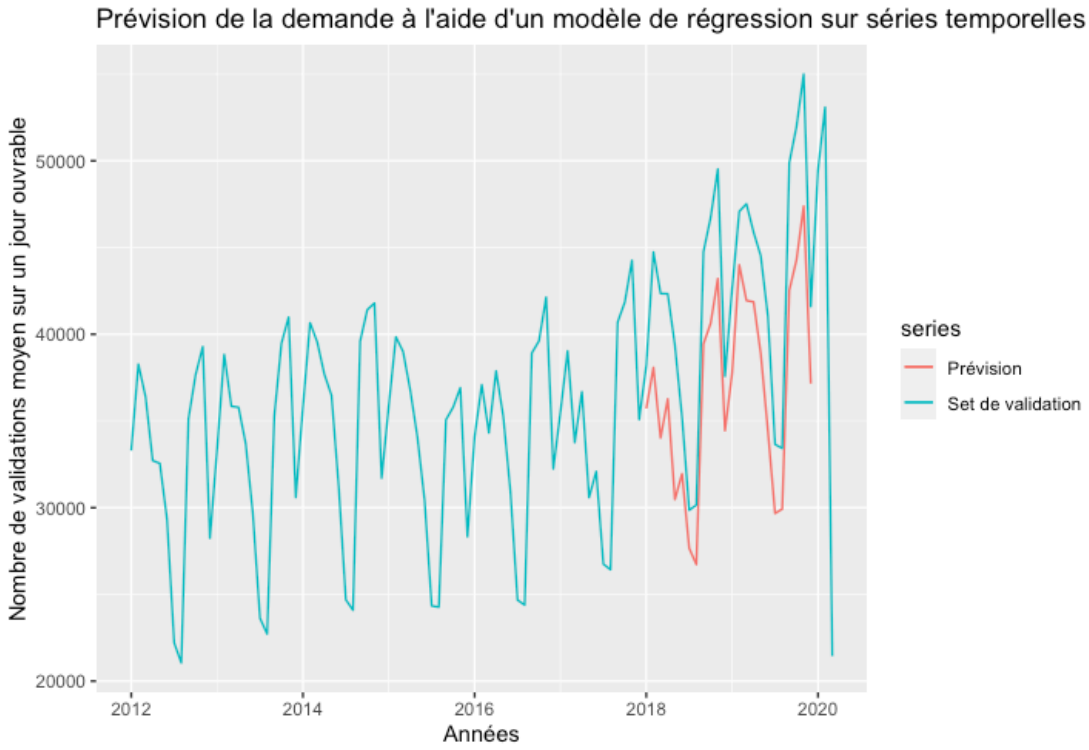


Figure 5-18 Prévision régression linéaire sur séries temporelles

Tableau 5-9 Performance des modèles de régression

Modèles	Régression linéaire	Régression linéaire sur séries temporelles (terme autorégressif)	ARIMA
RMSE	11751.9	5601.7	4144,2
MAPE	22.7%	12.3%	7,6%
Theil U	1.5596	0.7552	0,4825

Les résultats de ces deux expérimentations montrent que la prise en compte du caractère autorégressif dans les modèles de prévision de la demande est nécessaire. Une information importante émanant de ces résultats est que la considération unique de la population et du taux d'emploi apporte assez peu de performance lorsque le modèle est linéaire. La Figure 5-17 montre que les variations de la demande ne sont pas explicables par le biais de ces deux variables. Les variations sont en effet lissées et peu cohérentes lorsqu'elles sont comparées aux valeurs réelles. Un modèle ARIMA constitue un modèle plus performant que ces deux modèles considérant des variables sociodémographiques.

L'objectif de ce projet ne réside pas dans la prévision de la demande au global. Dans la mesure où les modèles faisant l'usage de variables sociodémographiques ne produisent pas des résultats pertinents et que ces variables ne sont pas disponibles pour des subdivisions géographiques plus précises, il ne paraît pas pertinent de continuer à utiliser uniquement ces variables dans les expérimentations sur les modèles.

Ce chapitre a pour but d'identifier un modèle permettant la prévision des montants sur le réseau de transport en commun. Plusieurs modèles ont été passés en revue. Si la sélection se basait uniquement sur les indicateurs d'erreur, le modèle ARIMA serait retenu comme le plus pertinent. Mais d'autres paramètres rentrent en compte, comme la capacité du modèle à s'adapter à des changements brusques de la demande. Cela n'est pas possible pour le modèle ARIMA qui ne prend en compte que les variations passées de la demande. La dimension historique se montre comme importante pour établir un modèle de prévision. En se basant uniquement sur cette dernière, l'erreur de la prévision est très faible, mais insuffisante d'un point de vue explicatif. Les variables socio-

économiques associées au caractère autorégressif permettent de décrire la demande sans constater des évolutions aberrantes. L'erreur est plus importante, mais le pouvoir explicatif doit être plus fort. Les modèles présentement calibrés ne permettent pas de faire de la prévision sur un horizon de 5 ans, mais ils apportent des pistes pertinentes concernant la marche à suivre pour la mise en place des modèles aux autres niveaux d'agrégation.

CHAPITRE 6 PRÉVOIR LA DEMANDE AU NIVEAU DES SECTEURS

Les secteurs municipaux sont au nombre de cinq pour la ville de Gatineau. La STO offre du service dans toutes ces zones. La majorité des usagers se concentre dans les secteurs de Gatineau, Hull et Aylmer. À noter que le réseau de la STO est également présent pour effectuer la liaison avec le centre-ville d'Ottawa. La Figure 6-1 décrit le découpage qui servira dans ce chapitre pour agréger les données.

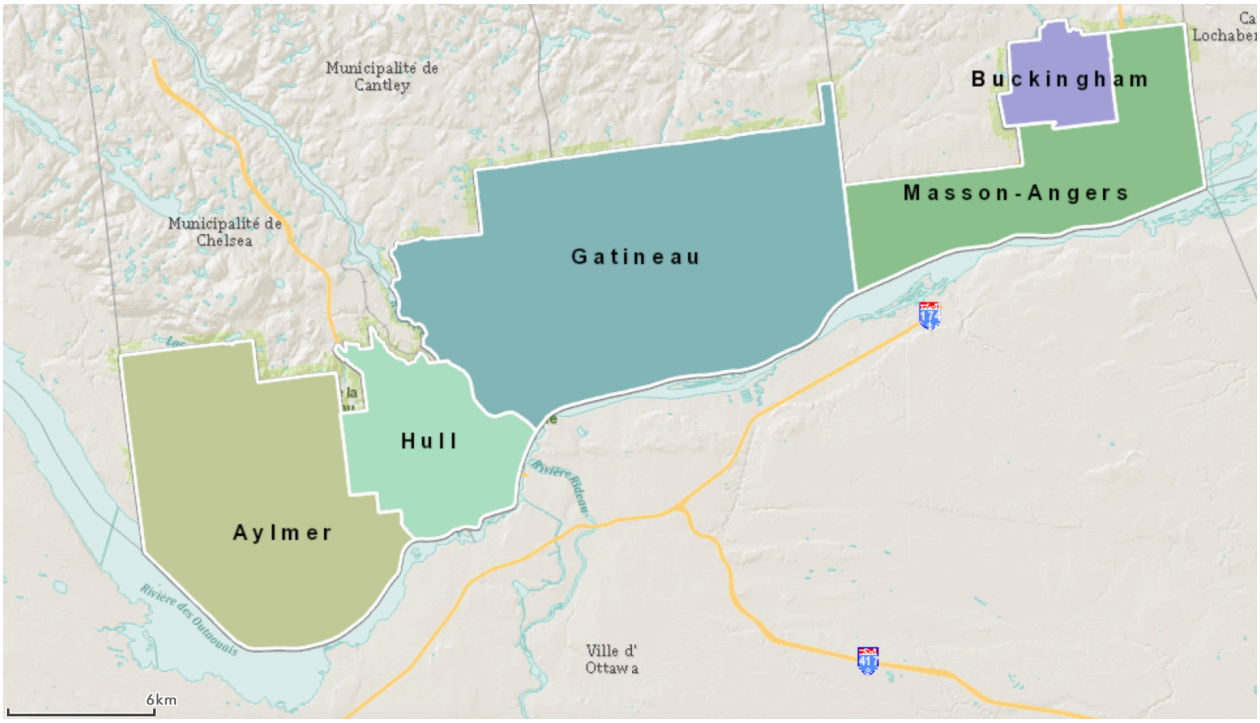


Figure 6-1 Secteurs municipaux de Gatineau

Effectuer des prévisions à l'échelle des secteurs municipaux permet de gagner en précision spatiale. Le découpage reste assez grossier ce qui permet dans un premier temps d'obtenir des modèles relativement simples tout en gagnant en précision. Cela permet de se familiariser avec les enjeux du changement d'agrégation avant de passer aux zones hexagonales.

Précédemment, les résultats ont montré que prévoir globalement la demande ne permettait pas d'avoir des résultats satisfaisants. Une des causes possibles de ces erreurs est le fait que les phénomènes influençant la demande ne sont pas étudiés de manière assez précise. Ainsi, aller étudier ces phénomènes au niveau des secteurs peut être un moyen d'éliminer ce biais. Pour déterminer s'il est possible de capturer ces phénomènes plus locaux, plusieurs modèles sont testés.

Étant donné que peu de nouvelles variables sont disponibles pour les secteurs, seuls deux modèles sont testés dans cette partie. Dans un premier temps, l'effet du changement de niveau d'agrégation est capturé sur l'étude du modèle ARIMA. Par la suite, un modèle multiniveau est calibré. Ce type de modèle permet de capturer l'effet de variables sur l'ensemble du réseau et également sur les secteurs indépendamment. L'objectif est de mesurer la performance d'une telle modélisation sur un niveau d'agrégation simple pour déterminer s'il est pertinent par la suite de calibrer un modèle multiniveau sur les zones hexagonales.

6.1 Modèle ARIMA par secteur

Certaines fluctuations ne sont pas prises en compte par le modèle ARIMA. Commence ici une investigation pour réussir à déterminer comment résoudre ces erreurs. Une des techniques possibles est dans un premier temps de tenter de localiser géographiquement d'où viennent ces erreurs. Cela va permettre de savoir si elles sont dues à des paramètres de certaines zones géographiques ou si elles sont généralisées à tout le territoire d'étude. Cette information est importante, car elle peut permettre de constater si une zone possède une erreur importante et ainsi par une analyse du territoire d'identifier des caractéristiques qui diffèrent des autres zones.

Dans la littérature, il est souvent montré que plus les données sont désagrégées, plus les modèles tendanciels ont tendance à fournir des résultats erronés. Ainsi cette démarche est là pour nous permettre d'analyser les erreurs et non pour tester un potentiel modèle de prévision.

Sept zones sont différenciées dans cette étude : Aylmer, Buckingham, Gatineau, Hull, Masson-Angers, Ottawa centre-ville et les validations dont la zone est inconnue. Sept modèles sont donc réalisés en calibrant les données sur la période 2012-2016 et en les validant sur 2017 et 2018. La donnée prédite est le nombre de validations de CAP pour un jour ouvrable moyen par mois. Les résultats sont disponibles dans le Tableau 6-1 ainsi que des graphes soulignant les faits saillants.

Tableau 6-1 Résultats des erreurs pour les modèles ARIMA par secteur

Secteur	**RMSE**	**MAPE**	**Theil U**
Aylmer	801,2	11,55%	0,5246
Buckingham	104,5	29,50%	2,0301
Gatineau	856,5	5,60%	0,4457
Hull	3303,0	15,22%	0,9060
Masson-Angers	59,87	11,46%	0,8515
Ottawa centre-ville	2372,23	19,47%	1,5946
Inconnue	1135,8	154,22%	8,6721

Le modèle le plus erroné est celui concernant les validations dont le secteur est inconnu. Cela était attendu, car ces problèmes ne répondent pas à une tendance constante. Le résultat pour ces validations n'est pas exploitable à la vue de sa nature, mais il est à garder en mémoire, car ces validations sont en réalité associées à un des six autres secteurs.

En considérant uniquement les secteurs réels, Buckingham est celui dont le modèle est le plus erroné. Mais son erreur moyenne est seulement de 104,5 ce qui signifie que peu de validations sont effectuées dans ce secteur. Ce n'est donc pas un des secteurs les plus influents et donc l'erreur constatée ici n'a que peu de répercussions sur la prévision du réseau global de la STO. De plus, comme c'est un secteur peu fréquenté, les fluctuations relatives du nombre de validations sont beaucoup plus sensibles à un faible changement d'achalandage. Ce n'est donc pas un des secteurs à étudier en priorité.

Ottawa centre-ville et Hull contribuent en revanche à une part conséquente de l'achalandage du réseau de la STO et les modèles réalisés sont beaucoup moins précis que pour les autres secteurs. Dans les deux cas, l'achalandage est significativement sous-estimé, ils sont au-dessus de la borne supérieure des intervalles de confiance à 95%. Le motif saisonnier reste similaire, mais la tendance

réelle augmente fortement. Dans ces deux secteurs, un ou plusieurs facteurs sont sûrement responsables de cette hausse. Il faut donc étudier plus en détail l'évolution des indicateurs qui caractérisent le centre-ville d'Ottawa et Hull. Les Figure 6-2 et Figure 6-3 illustrent ces variations.

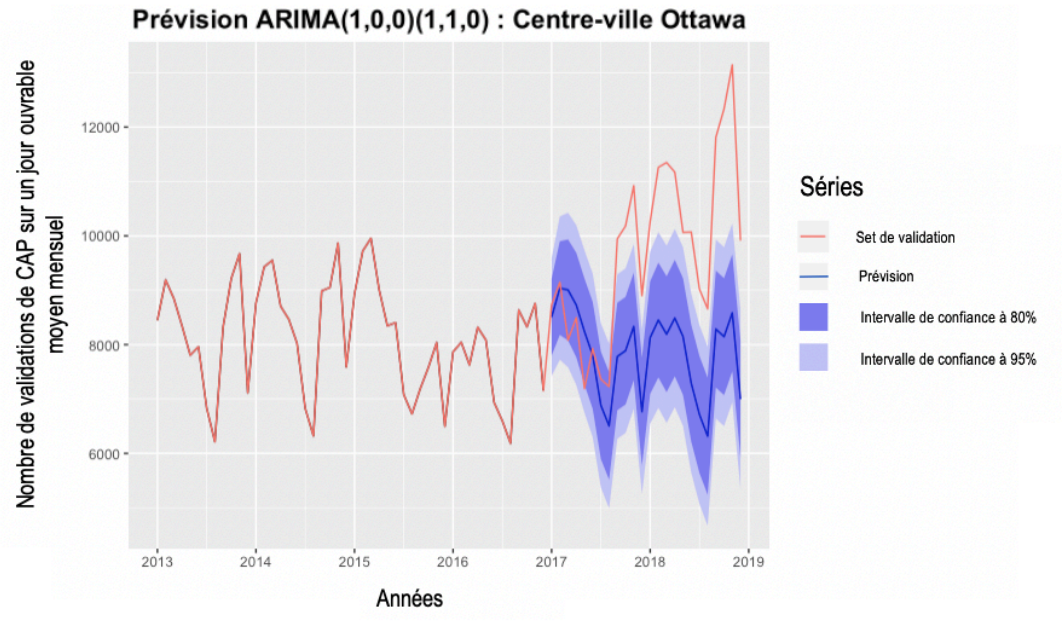


Figure 6-2 Modèle ARIMA pour CAP d'un jour ouvré moyen par mois pour le centre-ville d'Ottawa

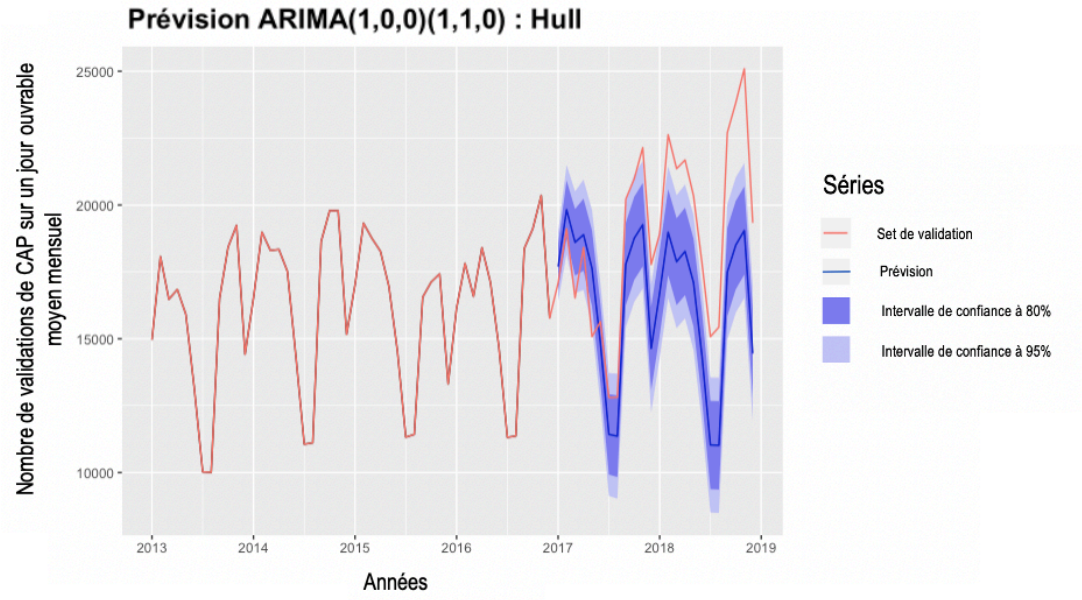


Figure 6-3 Modèle ARIMA pour CAP d'un jour ouvré moyen par mois pour Hull

Une comparaison entre l'application d'un modèle de prévision sur le total des validations et la combinaison des modèles pour chaque secteur a également été réalisée. Combiner des modèles dont le niveau d'agrégation est spatialement plus faible ne produit pas des résultats plus précis, au contraire. Les résultats sont d'autant plus sous-estimés. Cela est dû à la difficulté d'un modèle autorégressif à prévoir l'achalandage à plus petite échelle, qui est plus sujet à des perturbations extérieures.

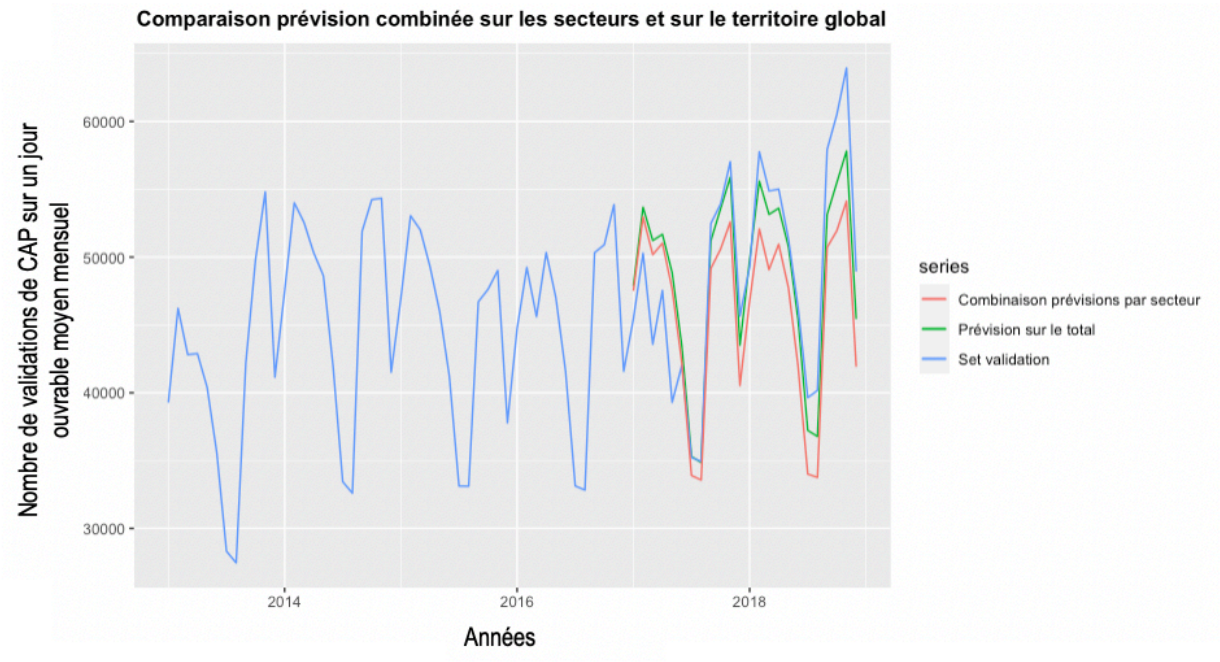


Figure 6-4 Comparaison du modèle combiné par secteur

La prévision mensuelle n'est pas la plus pertinente pour les besoins de prévision stratégique. Un modèle de régression linéaire avec correction de l'erreur ARIMA a également été calibré, mais aux vues des résultats, il n'est pas nécessaire de l'inclure dans ce chapitre. Un tel modèle a précédemment été calibré sur le réseau global et n'a pas fourni de résultats performants. Les erreurs sont encore plus importantes au niveau des secteurs municipaux.

6.2 Modèles multiniveaux

6.2.1 Le modèle « idéal »

Les données actuellement disponibles ne possèdent pas toutes la même résolution géographique. Certaines données concernent uniquement les secteurs municipaux de Gatineau et d'autres la totalité de la RMR O-G (rive Québec). Sachant que l'objectif de ce chapitre est de prévoir la demande à la résolution du secteur, il faut trouver un moyen de prendre en compte les effets de « contexte » (ceux dont les données sont disponibles pour toute la RMR) et les effets de secteurs (ceux dont les données diffèrent entre les secteurs).

Dans une prise de recul, une structure de modèle « idéal » est explicitée. L'objectif est de déterminer des caractéristiques importantes que doit posséder le modèle de prévision. Par la suite, les modèles de la littérature correspondant au mieux à cette structure sont testés.

Il est donc envisageable de créer des modèles indépendants pour chaque secteur prenant en compte à la fois ces effets de « contexte » et ceux propres à leur secteur. Mais il faut garder en tête que les variables de contexte n'auront pas la même influence sur chaque secteur. De plus, il est démontré dans la littérature que des modèles indépendants sur chaque secteur ne produisent pas des prévisions fiables, car ils sont victimes de surcalibration. Il n'est pas envisageable de créer un modèle par secteur. C'est pour cette raison que déterminer le degré d'influence de ces variables sur chaque secteur peut être intéressant. Le modèle illustré par la Figure 6-5 permettrait de prendre en compte les niveaux d'agrégation différents des données.

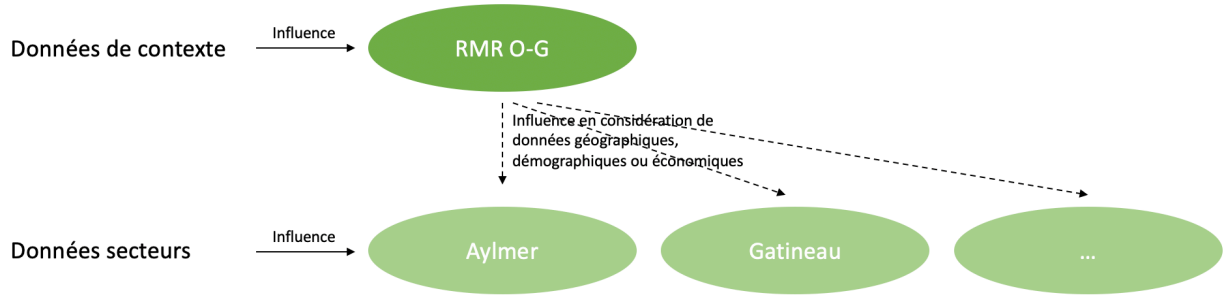


Figure 6-5 Schéma d'un modèle idéal

Dans la revue de la littérature, certains modèles s'approchent de cette structure. Notamment, les modèles multiniveaux. Ils permettent de prendre en compte un effet de contexte qui se retranscrit

de manière différente sur chaque secteur. D'autres modèles comme ceux nommés par *Post processing models* consistent en la réalisation des modèles indépendants sur chaque secteur en prenant en compte uniquement les données de secteur et la génération de résultats concernant la demande. Par la suite, un modèle global est réalisé en prenant en compte les données de contexte et des résultats sont également générés. Ces résultats dits globaux permettent par la suite d'effectuer une nouvelle calibration des résultats des secteurs pour prendre en compte les effets de contexte.

La structure multiniveau est la plus proche de ce modèle, il est donc testé dans la partie suivante.

6.2.2 Modèles multiniveaux

La structure des modèles multiniveaux est très intéressante, car elle permet de modéliser la demande à plusieurs échelles en prenant compte des influences à ces multiples niveaux. Dans cette partie, un modèle multiniveau dit « classique » est mis en place dans un premier temps. La méthodologie de création du modèle est explicitée au fils des résultats. Dans un deuxième temps, un modèle de classe latente, un type de modèle multiniveau, est calibré afin de tenter de régler certains problèmes constatés sur le modèle « classique ».

6.2.2.1 Modèle classique

Les modèles multiniveaux ont une structure proche de la réalité; ils sont donc susceptibles de remplir une partie des besoins exprimés. Ces modèles ont la capacité de pouvoir discerner des effets de contexte et des effets individuels (équivalent d'effet RMR et de secteurs dans notre cas). L'intérêt d'utiliser un tel modèle est de pouvoir quantifier l'effet d'une variable de contexte sur les différents secteurs, chose qui ne pourrait pas être réalisée par des modèles indépendants sur chaque secteur. De plus, il permet d'avoir des modèles similaires sur les secteurs, mais les effets de variables de secteurs ne sont pas quantifiés de la même manière. Dans le cas présent, cela voudrait dire que la variable de permis de construire peut avoir plus ou moins d'effet sur un secteur plutôt qu'un autre.

Après une familiarisation avec les modèles multiniveaux, il est apparu que toutes les variables ne pourraient pas facilement avoir une dépendance temporelle. Il est possible de se rapprocher d'une dépendance temporelle à l'aide de méthodes plus complexes et d'approximation. Mais dans un premier temps, il faut créer un modèle multiniveau « classique » qui sera par la suite adapté pour répondre de manière plus pertinente aux exigences.

Le modèle simple sera composé de deux niveaux et des variables suivantes :

- Niveau 1 : le contexte dont les variables explicatives varient mensuellement.
    - E_t : le taux d'emploi mensuel de la RMR O-G (rive QC)
    - P_t : la population mensuelle de la RMR O-G (rive QC)
- Niveau 2 : le secteur dont les variables explicatives varient entre les secteurs.
    - L_i : un indicateur reflétant l'évolution des permis de construire entre les secteurs

La mise en équation du modèle est donc de la forme suivante :

- Niveau 1 :

$$y_{ti} = \beta_{0i} + P_t * \beta_{1i} + E_t * \beta_{2i} + \epsilon_{ti}$$

- Niveau 2 :

$$\beta_{0i} = \beta_{00} + L_i * \beta_{01} + \alpha_{0i}$$
$$\beta_{1i} = \beta_{10} + L_i * \beta_{11} + \alpha_{1i}$$
$$\beta_{2i} = \beta_{20} + L_i * \beta_{21} + \alpha_{2i}$$

Avec,

y_{ti} le nombre de montants moyen en pointe AM par secteur et moyenné mensuellement.

β double indicé représente une constante qui sera calculée par le modèle.

ϵ_{ti} est un terme d'erreur qui suit une loi normale (0, \sigma_{ti}^2).

α_{ni} est un terme d'erreur qui suit une loi normale (0, \sigma_{ni}^2) pour n allant de 0 à 2.

Les résultats concernant les modèles multiniveaux suivants sont obtenu à l'aide du logiciel R Studio et plus particulièrement avec l'utilisation de package « nlme ».

Le modèle explicité précédemment est le modèle final souhaité. Mais avant toute chose, plusieurs étapes sont nécessaires pour effectuer les bons choix dans la calibration du modèle final. Un modèle de moyenne inconditionnelle est la première étape. C'est un modèle sans variable explicative, mais qui permet d'identifier la variabilité de la demande en fonction des effets du secteur ou globaux. Par la suite, les variables d'influence globale sont introduites dans le modèle. Cette structure

s'appelle modèle de croissance inconditionnelle. Enfin, le modèle final peut être mis en place en tenant compte des résultats des modèles précédents.

6.2.2.1.1 Modèle de moyenne inconditionnelle

Avant toute chose, lors de la mise en place d'un modèle multiniveau classique, il faut passer par certaines étapes pour comprendre la dynamique des données qui sont traitées. La calibration d'un modèle de moyenne inconditionnelle est la première étape. Cela s'assimile à un modèle multiniveau sans variables explicatives.

L'équation est assez simple :

$$y_{ti} = \beta_{0i} + \epsilon_{ti}$$

$$\beta_{0i} = \beta_{00} + \alpha_{0i}$$

Le modèle suppose que l'évolution de la demande est constante pour un secteur (niveau 1) et que ces valeurs de demande se répartissent autour d'une moyenne globale de tous les secteurs (niveau 2).

Le résultat est le suivant :

Performances			
AIC	5318.205		
BIC	5330.111		
Effets aléatoires	~1 | id_secteur		
	Ordonnée à l'origine	Résidus	
Variance	1664.64	207.0465	
Effets fixes	CAPO_JOAM_X11~1		
	Valeur	Erreur quadratique	p-value
Ordonnée à l'origine	2710.148	832.3856	0.0012

Figure 6-6 Résultat du modèle de moyenne inconditionnelle dans R

On retient donc que la variance du niveau 1 soit ε_{ti} est de 1664,64 et que la valeur de la moyenne globale est de 2710,15 validations pendant la pointe AM par secteur. La variance intersecteur est de 42868 validations.

Le coefficient de corrélation intragroupe est de 0,998 ce qui signifie que presque la totalité de la variance est attribuable à des différences entre les secteurs. Cela nous donne beaucoup d'informations sur le type de modèle à générer. Il faut donc avoir des variables qui nous permettent de différencier les secteurs entre eux. Ainsi les variables initialement considérées ne sont peut-être pas extrêmement pertinentes à cet égard. Ajouter une variable comme la distance au centre-ville (ou à Ottawa) ou encore la densité peut donner de bons indicateurs (sous réserve que les deux ne soient pas corrélées).

6.2.2.1.2 Modèle de croissance inconditionnelle

L'étape suivante dans la construction d'un modèle de croissance est l'introduction des variables de niveau 1. Dans le cas présent, cela est constitué des variables de population et de taux d'emploi. Le passage par ce modèle change la signification des variances qui peuvent être recensées. La variance observée au niveau 1 maintenant la variance de l'erreur sur la modélisation obtenue par une combinaison de la population et du taux d'emploi. Au niveau 2, trois variances sont maintenant intéressantes à étudier, σ_0^2 qui reflète la variabilité interpersonne à l'instant initial, σ_1^2 qui reflète la variabilité due au changement de population et σ_2^2 du taux d'emploi.

La mise en équation du modèle est donc de la forme suivante :

- Niveau 1 :

$$\mathrm{y}_{ti} = \beta_{0i} + P_t * \beta_{1i} + E_t * \beta_{2i} + \epsilon_{ti} \quad \text{avec} \quad \epsilon_{ti} \sim \mathcal{N}(\mu,\, \sigma_\epsilon^2)$$

- Niveau 2 :

$$\beta_{0i} = \beta_{00} + \alpha_{0i} \quad \text{avec} \quad \alpha_{0i} \sim \mathcal{N}(\mu,\, \sigma_0^2)$$

$$\beta_{1i} = \beta_{10} + \alpha_{1i} \quad \text{avec} \quad \alpha_{1i} \sim \mathcal{N}(\mu,\, \sigma_1^2)$$

$$\beta_{2i} = \beta_{20} + \alpha_{2i} \quad \text{avec} \quad \alpha_{2i} \sim \mathcal{N}(\mu,\, \sigma_2^2)$$

Les résultats sont donc les suivants :

Performances du modèle			
AIC	5278.471		
BIC	5318.183		
Effets aléatoires	~population + taux d'emploi|id_secteur		
	Variance		
Ordonnée à l'origine	1.66E-05		
Population	5.281		
Taux d'emploi	3.63E-04		
Résidus	189.46		
Effets fixes	Montants en pointe AM ~ population + taux d'emploi		
	Valeur	Erreur quadratique	p-value
Ordonnée à l'origine	993.18	654.95	0.1302
Population	10.42	3.22	0.0014
Taux d'emploi	-17.41	-2.19	0.0294

Figure 6-7 Résultats du modèle inconditionnel de croissance

σ_0^2 et σ_2^2 sont toutes les deux presque nulles. Cela signifie que la variabilité intersecteur est nulle à l'instant initial et que l'évolution du taux d'emploi influence de manière similaire les secteurs entre eux.

$\sigma_1^2 = 27,9$

Cette valeur est assez faible, mais elle signifie tout de même que la demande de chaque secteur évolue légèrement différemment en réaction aux variations de la population. Cet aspect était attendu dans la mesure où les variations de la population ne sont pas identiques dans l'ensemble des secteurs.

$\sigma_\epsilon^2 = 35896$

La valeur de la variance intrasecteur reste relativement constante par rapport au modèle précédent. Cela signifie que les variables introduites n'aident que très peu à expliquer les évolutions de la demande au sein des secteurs.

Il faut tout de même noter que l'ordonnée à l'origine du modèle final n'est pas significative (p-value > 0,05). L'hypothèse que cette valeur peut être nulle ne peut donc être rejetée. Ce problème

se comprend assez bien quand on regarde les différences d'achalandage entre les secteurs. Pour l'instant, notre modèle suppose une ordonnée à l'origine identique pour tous les secteurs. Une telle constance de ce paramètre n'est pas vraisemblable et peut donc générer cette erreur dans le modèle de croissance inconditionnel. La prochaine étape est donc d'introduire les variables de niveau 2 qui permettent de différencier les secteurs entre eux.

6.2.2.1.3 Le modèle multiniveau classique

Ce modèle est également appelé modèle de croissance, car il traite des données longitudinales. C'est une adaptation d'un modèle multiniveau à ce type de donnée. Après les étapes préliminaires réalisées précédemment, il a été possible de recueillir des informations quant à l'organisation des données. Il est donc temps de mettre en place le modèle initialement voulu (les équations sont celles exposées au début de cette partie).

Les résultats de ce modèle sont donc décevants :

Performances du modèle			
AIC	5282.629		
BIC	5334.256		
Effets aléatoires	~population + taux d'emploi|id_secteur		
	Variance		
Ordonnée à l'origine	2.21E-05		
Population	5.112		
Taux d'emploi	5.58E-04		
Résidus	189.08		
Effets fixes	Montants en pointe AM ~ population*moyenne logements + taux d'emploi*moyenne logements		
	Valeur	Erreur quadratique	p-value
Ordonnée à l'origine	1732.29	1220.13	0.1565
Moyenne logement	-21.72	30.23	0.5471
Population	4.7	5.91	0.4275
Taux d'emploi	-13.46	14.84	0.3652
Moyenne logement*Population	0.17	0.15	0.2519
Moyenne logement*Taux d'emploi	-0.12	0.37	0.19

Figure 6-8 Résultats du modèle classique

Aucun des coefficients n'est significatif ce qui montre un réel problème dans la calibration du modèle. La structure n'est donc pas adaptée aux données.

Ces résultats ne sont vraiment pas étonnants, car le modèle a été créé sans tenir compte des conclusions obtenues lors des étapes préliminaires :

- Les principales différences sont constatées entre les secteurs. Cela implique de trouver des indicateurs permettant de caractériser les secteurs indépendamment les uns par rapport aux autres. Nous avons uniquement la variable des permis de construire pour l'instant. Il pourrait être judicieux d'ajouter une variable comme la distance à Ottawa (en effet il a été montré que plus le secteur était loin d'Ottawa plus la demande se faisait rare).
- La nécessité de différencier les secteurs a été aussi soulignée par le modèle inconditionnel de croissance où l'ordonnée à l'origine n'était pas significative. La population et le taux d'emploi ont leur place dans le modèle.

Après des tests de plusieurs modèles, les conclusions concernant les variables à utiliser sont les suivantes :

- La distance à Ottawa est un ajout pertinent, mais ne produit pas de coefficient significatif lorsqu'il est combiné au taux d'emploi.
- La moyenne des incréments mensuels de permis de construire n'est pas bénéfique pour différencier les secteurs donc doit être retirée.

Le modèle qui apparaît comme le plus pertinent avec les données disponibles est celui avec la population et le taux d'emploi comme variables de niveau 1 et la distance à Ottawa en niveau 2, mais avec la suppression du terme d'interaction entre le taux d'emploi et la distance à Ottawa. La décision de prendre ce modèle comme le plus pertinent est soutenu par les valeurs des critères d'information d'Akaike (AIC) des différents modèles.

La mise en équation est la suivante :

- Niveau 1 :

$$\mathrm{y}_{ti} = \beta_{0i} + P_t * \beta_{1i} + E_t * \beta_{2i} + \epsilon_{ti}$$

- Niveau 2 :

$$\beta_{0i} = \beta_{00} + D_i * \beta_{01} + \alpha_{0i}$$
$$\beta_{1i} = \beta_{10} + D_i * \beta_{11} + \alpha_{1i}$$
$$\beta_{2i} = \beta_{20} + \alpha_{2i}$$

Où D_i est la distance en km du secteur avec Ottawa.

La décision de supprimer le terme d'interaction entre le taux d'emploi et toute autre variable est soutenue par le fait que, dans les modèles préliminaires, la variabilité de la demande TC en réaction à une évolution du taux d'emploi était presque nulle. Donc cela signifie que tous les secteurs réagissent de manière similaire à cette donnée.

Les résultats de ce modèle sont les suivants :

Performances du modèle			
AIC	5201.022		
BIC	5248.677		
Effets aléatoires	~population + taux d'emploi|id_secteur		
	Variance		
Ordonnée à l'origine	4.23E+02		
Population	1.91E-04		
Taux d'emploi	7.35E-04		
Résidus	172.75		
Effets fixes	Montants en pointe AM ~ Taux d'emploi + Population*Distance Ottawa		
	Valeur	Erreur quadratique	p-value
Ordonnée à l'origine	-4175.92	935.78	0
Taux d'emploi	-17.42	7.28	0.0173
Distance Ottawa	430.51	57.21	0.0172
Population	36.21	2.84	0
Distance Ottawa*Population	-2.15	0.19	0

Figure 6-9 Résultats du modèle multiniveau

Les résultats de ce modèle sont intéressants dans la mesure où ils montrent qu'un modèle multiniveau n'est pas vraiment adapté aux données disponibles. Tout d'abord, les variances presque nulles associées aux variables de niveau 1 (population et taux d'emploi) montrent que chaque secteur réagit de manière similaire à leur évolution. De plus, la forte variance intersecteur montre que les différences ne sont pas assez captées par les variables de niveau 1; il y a une grosse différence à l'état initial de la valeur de la demande entre les secteurs. Enfin, la variance intrasecteur qui décrit l'évolution temporelle est également importante ce qui est également signe d'un manque de pouvoir explicatif des évolutions temporelles.

Pour montrer tout de même les résultats générés par ce modèle, voici quelques résultats montrant le niveau de pertinence de cette modélisation. Des indicateurs de performance sont affichés dans

le Tableau 6-2 et la demande modélisée est comparée à leurs valeurs réelles dans les figures qui suivent.

Tableau 6-2 Indicateurs de performance du modèle multiniveau

Secteur	**RMSE**	**MAPE**	**Theil U**
Aylmer	393,6	13,2%	35,822
BMA	44,6	9,1%	4,166
Gatineau	728,1	20,5%	63,176
Hull	386,1	8,2%	30,908

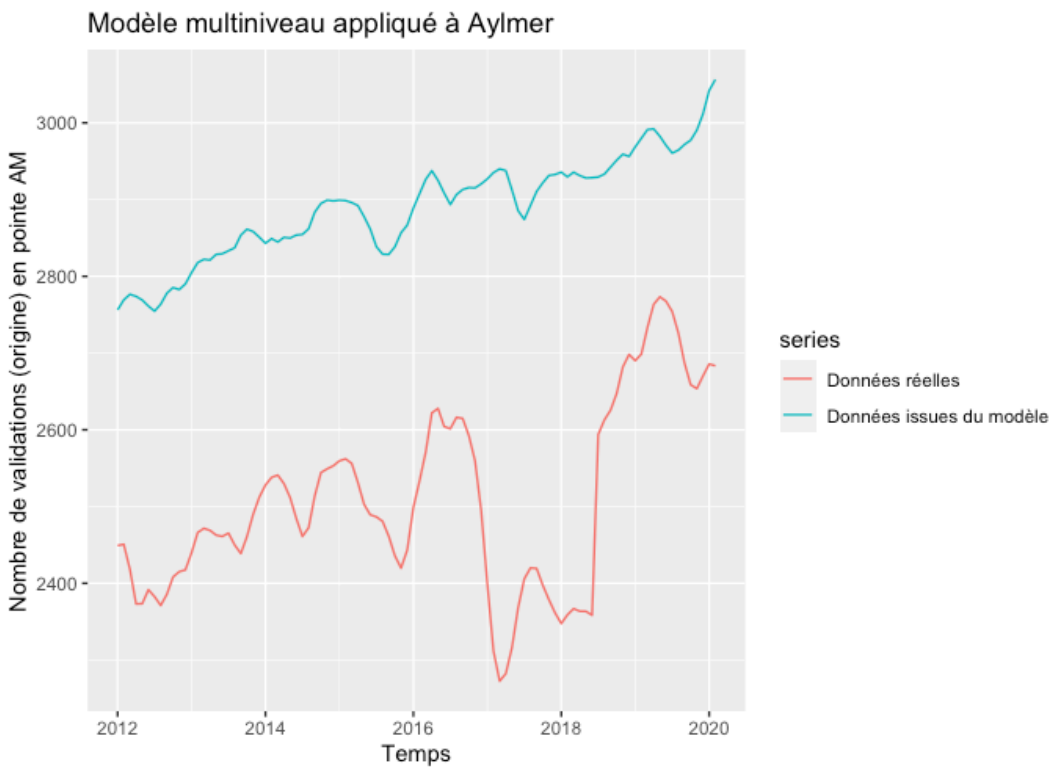


Figure 6-10 Modélisation multiniveaux sur le secteur Aylmer

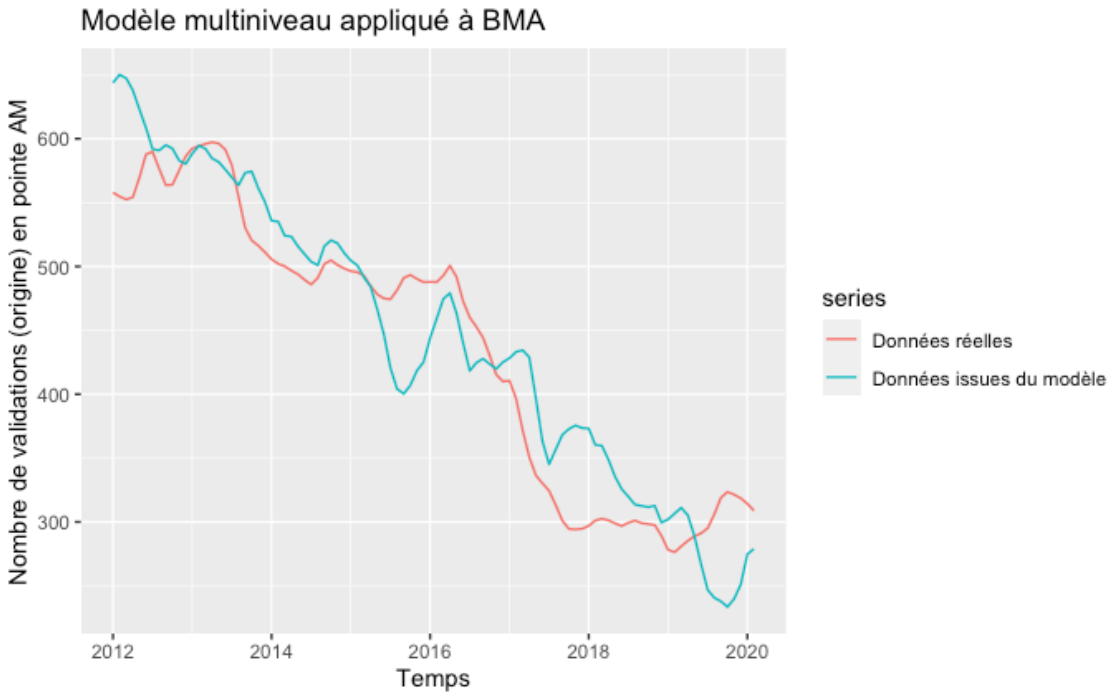


Figure 6-11 Modélisation multiniveaux sur le secteur BMA

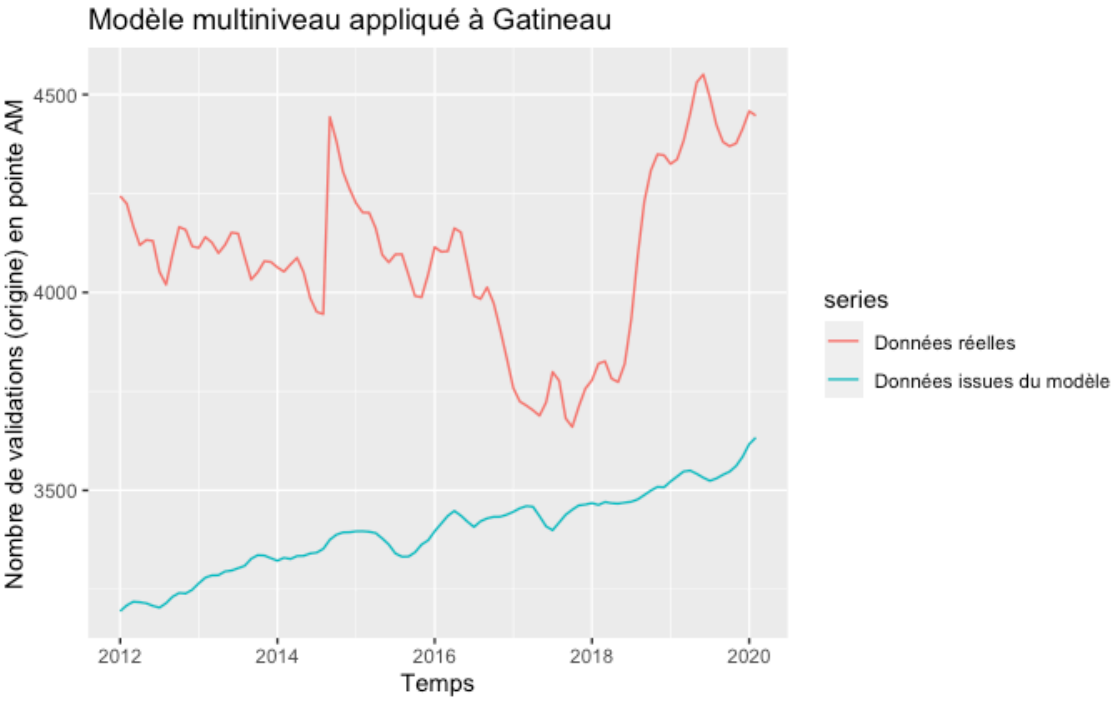


Figure 6-12 Modélisation multiniveaux sur le secteur Gatineau

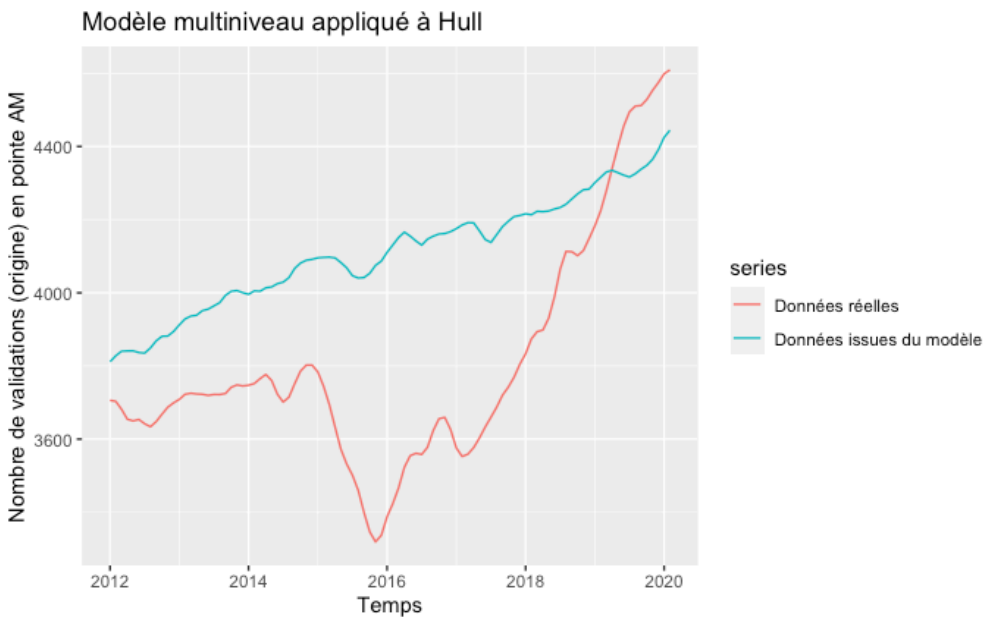


Figure 6-13 Modélisation multiniveaux sur le secteur Hull

Les problèmes de ces modélisations sont le manque de variation dans l'évolution de la demande et les ordonnées à l'origine imprécises. Les variations réelles sont souvent plus complexes que celles modélisées. Cela peut être dû à beaucoup de paramètres comme notamment l'impossibilité de modéliser le phénomène linéairement, le manque de sensibilité du modèle ou encore le manque de valeur explicative dans les variables. Tous ces écarts ne peuvent pas être pertinemment corrigés en continuant avec ce genre de modèle surtout qu'il s'agit ici d'une modélisation et non d'un modèle prédictif. De plus, il est possible de constater que pour certains secteurs, les variations sont cohérentes, mais les ordonnées à l'origine sont décalées. Dans ce modèle, l'ordonnée à l'origine est fonction de la distance au centre-ville d'Ottawa. C'est la variable qui a fourni les meilleurs résultats, mais ce n'est pas encore assez précis. Pour remédier à ce problème, il serait possible de créer un modèle de classe latente pour modéliser différemment ce paramètre en fonction des secteurs.

Il ne faut pas perdre de vue que l'objectif est de prévoir les évolutions temporelles et non les différences entre secteurs.

6.2.2.2 Modèle de classe latente

Comme évoqué précédemment, le modèle multiniveau classique ne permet pas de gérer une dépendance temporelle pour les variables influençant directement le secteur municipal. Cette

impossibilité d'avoir une composante temporelle au niveau des variables sectorielles fait donc perdre en pouvoir explicatif de notre modèle. Les modèles de classes latentes dérivent des modèles multiniveaux en permettant de créer des typologies de secteurs regroupées en sous-groupes. C'est équivalent à créer un niveau intermédiaire entre le contexte (RMR O-G) et les secteurs. Cela peut paraître anodin, mais cette structure permet de prendre en compte les comportements fortement différents de chaque secteur en relation avec les variables explicatives. Il est maintenant possible d'avoir différentes variables explicatives pour différents sous-groupes de secteurs. Il n'est toujours pas possible d'inclure une dépendance temporelle dans les variables de secteurs, mais cette formulation permet de plus fidèlement modéliser la réalité.

Le modèle multiniveau classique suppose une répartition normale des comportements vis-à-vis des variables autour d'une valeur moyenne. Supposer une telle répartition nécessite une hypothèse forte et non vérifiable concernant les variables dépendantes. Les variables de contexte même latentes ne doivent pas être corrélées avec les variables de secteurs. Mais le modèle de classe latente permet de se séparer de cette hypothèse en modifiant la structure de l'estimation des erreurs. Le modèle de classe latente est aussi appelé modèle mixte non paramétrique, car il suppose que la distribution des comportements (équivalent aux erreurs dans la mise en équation) est non paramétrique. Cela revient à estimer des sous-groupes de secteurs aux comportements similaires, la variabilité des comportements n'est pas plus étudiée que cela. Ce modèle a donc l'avantage d'examiner les caractéristiques des secteurs appartenant à un même sous-groupe pour mieux comprendre comment ces paramètres ont un effet sur la demande en TC différents des autres sous-groupes.

Mais garder la même structure donnerait la mise en équation suivante du modèle :

- Niveau 1 :

$$\mathrm{y}_{ti} = \beta_{0i}^{j} + \beta_{1i}^{j} * P_t + \beta_{2i}^{j} * E_t + \epsilon_{ti}$$

- Niveau 2 :

$$\beta_{0i}^{j} = \beta_{00}^{j} + \beta_{01}^{j} * L_i + \alpha_{0i}^{j}$$
$$\beta_{1i}^{j} = \beta_{10}^{j} + \beta_{11}^{j} * L_i + \alpha_{1i}^{j}$$
$$\beta_{2i}^{j} = \beta_{20}^{j} + \beta_{21}^{j} * L_i + \alpha_{2i}^{j}$$

L'indice j représente les sous-groupes qui seront générés par un algorithme. Les erreurs ne supposent donc plus de répartition normale, mais non paramétrique (la génération de ces sous-groupes est gérée par un algorithme).

Mais cette structure de notre modèle de classe latente ne nous apporte pas vraiment de similarité avec la réalité. Le problème résidait dans le manque de dépendance temporelle de la variable des permis de construire. En inversant la structure du modèle : faire passer le contexte (avec la dépendance temporelle) en tant que niveau 2 et le secteur en niveau 1, il est possible de récupérer la dépendance temporelle sur toutes les variables explicatives. Mais initialement, dans le modèle classique, cela signifiait attribuer à chaque secteur les mêmes coefficients de régression à un instant t. Le modèle de classe latente vient ici toujours attribuer les mêmes coefficients, mais cette fois aux sous-groupes de secteurs. Cela vient faire l'hypothèse que les sous-groupes de secteurs ont un comportement homogène en leur sein au regard de l'évolution des permis de construire.

Mettre en équation le modèle donne les équations suivantes :

- Niveau 1 :

$$\mathrm{y}_{ti} = \beta_{0t}^{j} + \beta_{1t}^{j} * L_{ti} + \epsilon_{ti}$$

- Niveau 2 :

$$\beta_{0t}^{j} = \beta_{00}^{j} + \beta_{01}^{j} * P_t + \beta_{02}^{j} * E_t + \alpha_{0t}^{j}$$

$$\beta_{1t}^{j} = \beta_{10}^{j} + \beta_{11}^{j} * P_t + \beta_{12}^{j} * E_t + \alpha_{1t}^{j}$$

Les résultats fournis par ce modèle ne sont guère encourageants, car les classes latentes créée sont fortement déséquilibrées. En effet une des classes possède presque l'intégralité des observations qui sont au nombre total de 97. Ce chiffre est expliqué par le nombre d'observations mensuelles réalisées entre janvier 2012 et février 2020.

Tableau 6-3 Classes du modèle de classes latentes

Classes	**N°1**	**N°2**	**N°3**	**N°4**
Nombre d'observations par classe	5	86	1	5
Pourcentage des observations	5,15	88,66	1,03	5,15

Après avoir réalisé ce modèle de classe latente, il s'est avéré que ce n'est pas pertinent, car la distinction entre secteur n'est plus assurée, les indices j distinguent des sous-classes de périodes, mais pas de secteurs. Pour prendre en compte cela, il peut être possible d'insérer des indices booléens pour assurer une différence entre secteurs, mais cela signifierait que le modèle de classe latente perd tout son sens.

Peu de modèles sont calibrés sur ce niveau d'agrégation, car il a principalement pour but de calibrer des modèles plus complexes comme le modèle multiniveau afin de déterminer s'il est pertinent de le calibrer sur les zones hexagonales. Les résultats suggèrent le contraire : le modèle ne permet pas de modéliser de manière fiable la fréquentation du réseau de TC à l'échelle des secteurs municipaux. Un tel modèle tente majoritairement de différencier les secteurs les uns par rapport aux autres et non de modéliser l'évolution temporelle. Malgré une structure prometteuse, ce modèle n'est pas adapté aux objectifs du mémoire. Il ne sera donc pas calibré sur les zones hexagonales.

CHAPITRE 7 PRÉVISION SUR LES ZONES D'AGRÉGATION PRÉCISES

Le maillage hexagonal a été déterminé dans le but d'obtenir des résultats spatialement précis concernant les évolutions de la demande. C'est le niveau d'agrégation qui a pour but de fournir des prévisions répondant aux objectifs fixés par le projet. La majorité du travail de recherche a été effectuée à cette échelle. Jusqu'à maintenant, les modèles ont été calibrés sur la globalité du réseau ou sur les secteurs municipaux. Cela a permis de se familiariser avec le changement de niveau d'agrégation, mais également d'identifier certaines composantes importantes à la prévision de l'usage du TC.

Un des enjeux majeurs de vouloir prévoir la demande à l'échelle locale, des hexagones de 900 mètres de diamètre, est la disponibilité de données. L'obtention de variables explicatives permettant d'être agrégées par ce maillage et qui sont également longitudinales, est une tâche difficile. Peu de données répondent aux exigences du projet. Présentement, la variable actualisée des logements et les données de stationnements incitatifs sont les seules variables disponibles et ayant un effet prouvé sur la demande. Ces paramètres sont donc utilisés de manière récurrente dans ce chapitre. Des données de compositions de la population sont également utilisées par la suite. Il s'avère que des projections de ces données sont disponibles. Mais malheureusement, la demande de ces données n'aboutira pas dans le temps imparti pour la réalisation de ce mémoire. Une modélisation est donc réalisée dans l'optique de réaliser des prévisions quand les données seront obtenues.

Ce chapitre se déroule de la même manière que les deux précédents. Les modèles calibrés et jugés pertinents dans les chapitres antérieurs sont calibrés dans un premier temps. Il est possible de constater l'effet du passage à une agrégation fine pour ces modèles et de les comparer avec les autres niveaux. Par la suite, un modèle de régression linéaire est adapté à ce niveau d'agrégation. Finalement, des typologies de zones sont réalisées afin de mettre en place un modèle permettant l'obtention de taux de génération de montants.

7.1 Modèles précédemment mis en place

7.1.1 Modèles autorégressifs

Le niveau d'agrégation ayant été déterminé en considération des indicateurs observés, il faut donc à présent réaliser des modèles de prévision. Tout d'abord, l'objectif est de tester le comportement des modèles précédemment mis en place sur des niveaux d'agrégation plus grossiers. Ces modèles ne sont peut-être pas adaptés à une résolution fine, mais ils permettront de récolter des informations importantes concernant le modèle idéal. De la même manière que précédemment, les zones produisant de fortes erreurs au niveau du modèle ARIMA ne comportent pas une tendance régulière. Il faut donc identifier les phénomènes qui mènent à ces perturbations de l'achalandage.

Deux modèles ARIMA sont appliqués à chaque zone du maillage. Les variables dépendantes sont respectivement le nombre de montants sur une pointe AM moyenne mensuelle et le nombre de montants sur un jour ouvrable moyen mensuellement. L'ensemble d'entraînement utilisé est les données entre 2012 et 2016 et la prévision est effectuée sur 3 ans. Les figures suivantes situent les erreurs de ces deux modèles. Le pourcentage d'erreur moyenne et l'erreur quadratiques sont les deux indicateurs utilisés.

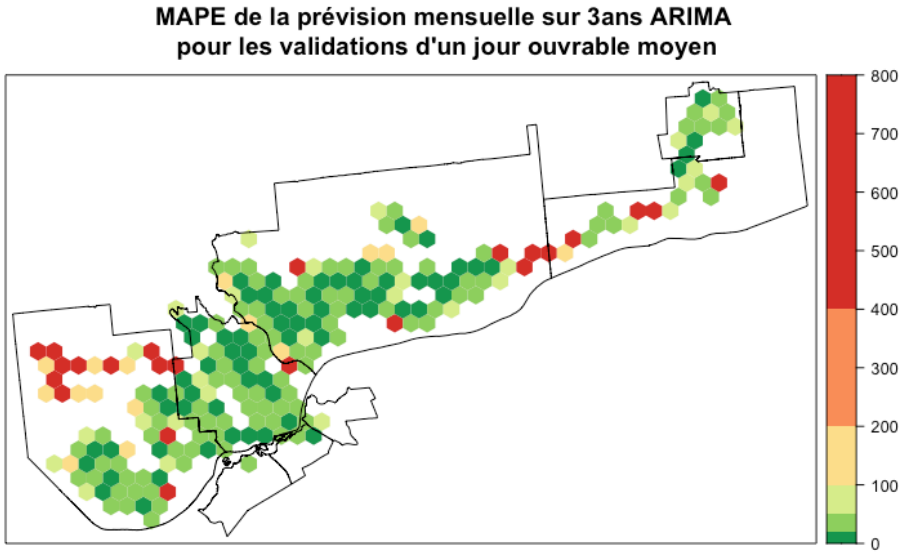


Figure 7-1 Pourcentage d'erreur quadratique moyen pour le modèle sur les jours ouvrés

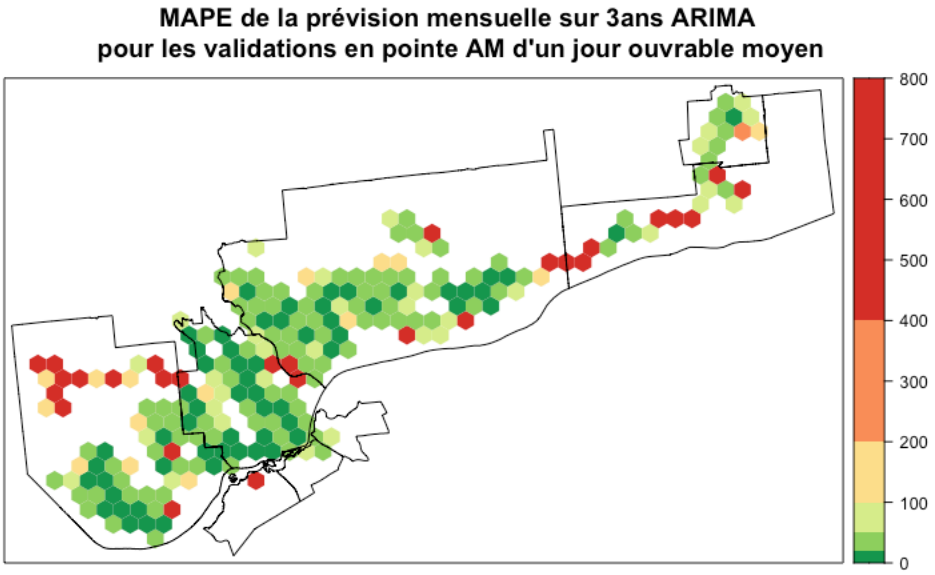


Figure 7-2 Pourcentage d'erreur quadratique moyen pour le modèle sur les pointes AM

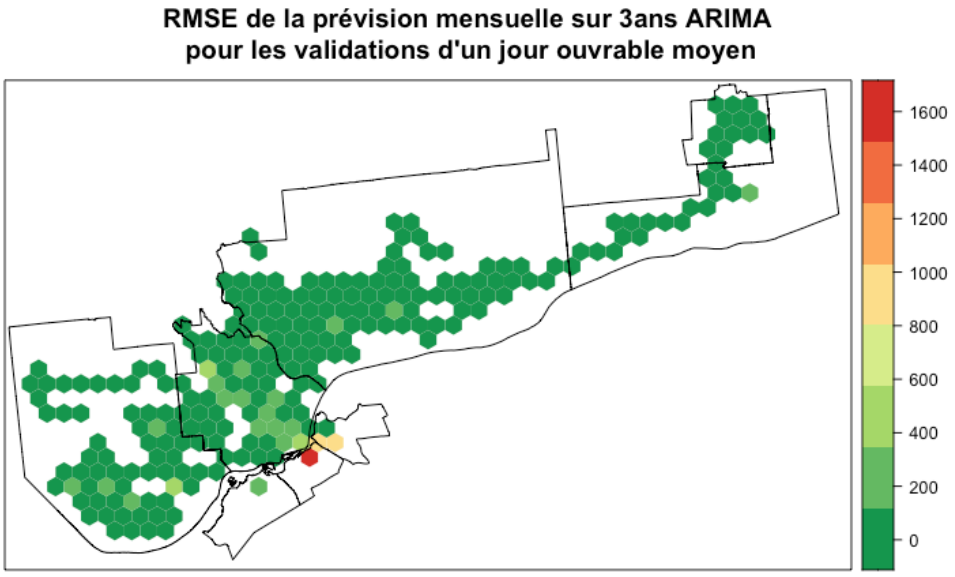


Figure 7-3 Erreur quadratique moyenne pour le modèle sur les jours ouvrés

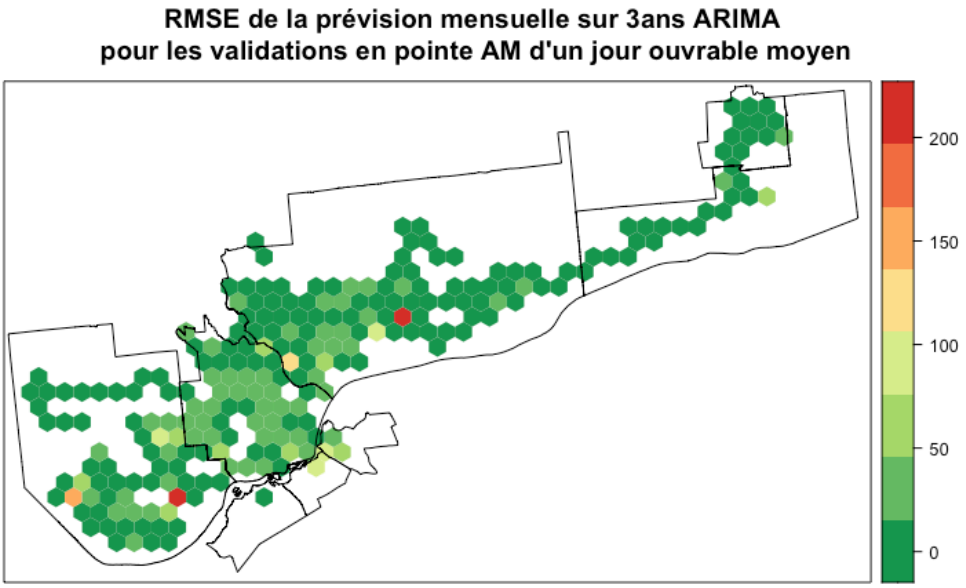


Figure 7-4 Erreur quadratique moyenne pour le modèle sur les pointes AM

Ce premier jeu de résultats se base sur un modèle autorégressif. L'étude des erreurs souligne donc les lieux où la demande ne suit pas une tendance constante. La combinaison du pourcentage d'erreur moyen et de l'erreur quadratique moyenne permet de nuancer les résultats. En effet, une multitude de zones ont un pourcentage d'erreur moyenne fort, mais ne font pas preuve d'une demande conséquente. Ces zones ont souvent subi une transformation importante, comme création d'un unique arrêt ou forte transformation résidentielle.

En parallèle, des zones comme le centre-ville d'Ottawa font preuve d'une faible erreur relativement au nombre de passagers embarquant dans cette zone, mais l'erreur est conséquente quand elle est comparée aux autres secteurs. Ottawa est une zone intéressante à analyser, car une grande partie des usagers embarquant sur le réseau de la STO débarquent à Ottawa. Si la pointe du matin est analysée indépendamment, l'erreur de la prévision est relativement faible comparée à celle si la prévision est faite sur la totalité d'un jour ouvrable. Le problème lié à Ottawa réside donc principalement dans la pointe PM. L'étude des logements qui a été réalisée ne prend pas en compte la zone du centre-ville d'Ottawa donc il n'est pas possible d'ajouter cette variable explicative dans cette zone. Néanmoins, il paraît peu probable que cette variable explique les validations à Ottawa,

car les usagers se rendant à Ottawa résident à Gatineau. Des données d'emplois seraient donc plus cohérentes avec les activités de ces usagers une fois à Ottawa.

La modélisation de la pointe du matin permet d'identifier deux foyers de forte erreur quadratique. Elle est supérieure en moyenne à 200 montants. Ces deux zones ont déjà été identifiées comme le lieu d'implantation des parcs-o-bus. L'ajout de cette variable est donc indispensable pour réduire l'erreur sur la prévision.

Enfin les logements prouvent encore une fois leur pertinence, car la zone située au nord-est comprend un fort pourcentage d'erreur. Cette partie de la ville subit une forte évolution du nombre de logements.

Les figures suivantes reflètent l'erreur brute à trois instants dans le temps : les mois de novembre de 2017, 2018 et 2019. Il est possible de déterminer les zones où l'achalandage est surestimé ou sous-estimé ce qui est également fort en informations sur l'effet des perturbations que la zone a subies. Les figures suivantes ne concernent que le modèle effectué sur les jours ouvrés. Mais ces mêmes figures sont présentes dans la partie sur les modèles sur 3 mois (7.1.3). Elles ne sont pas présentes deux fois afin d'éviter les redondances.

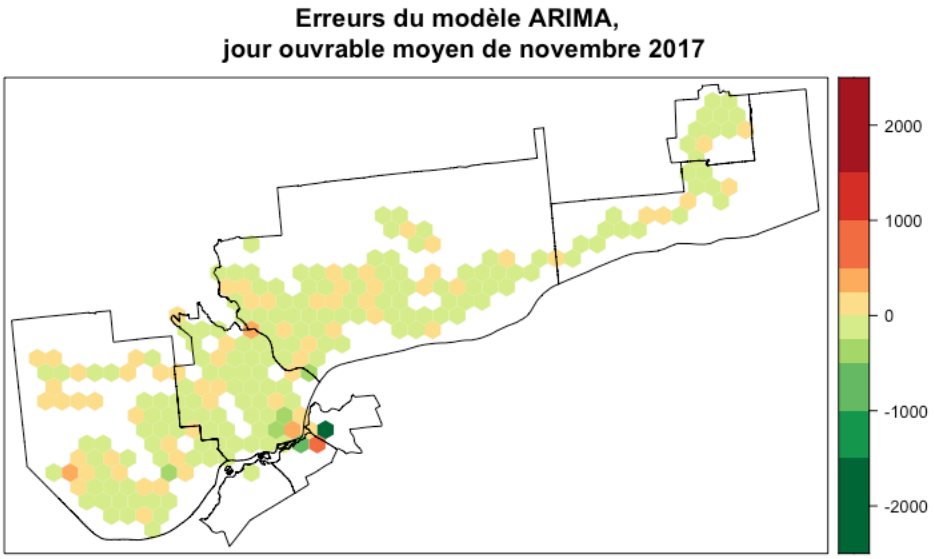


Figure 7-5 Erreurs du modèle ARIMA, jour ouvrable moyen de novembre 2017

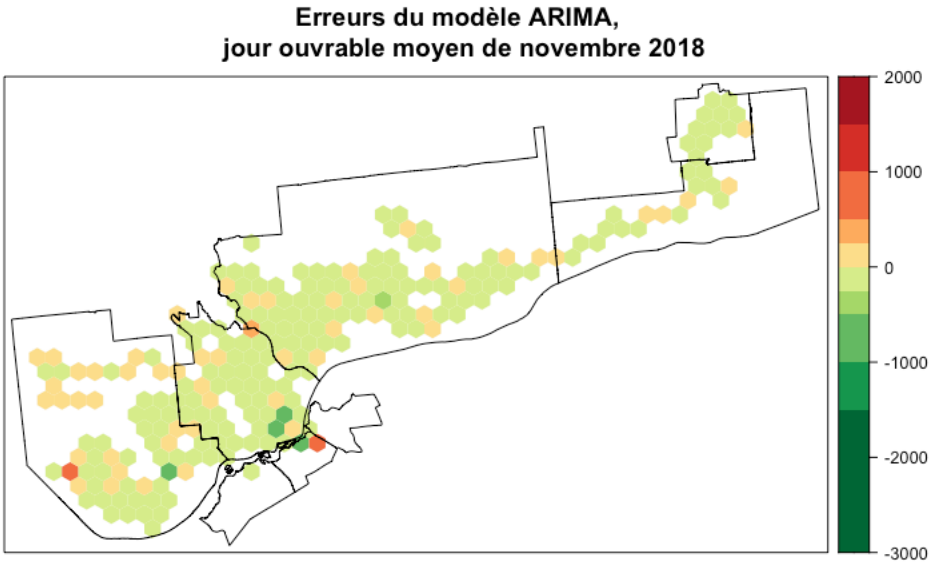


Figure 7-6 Erreurs du modèle ARIMA, jour ouvrable moyen de novembre 2018

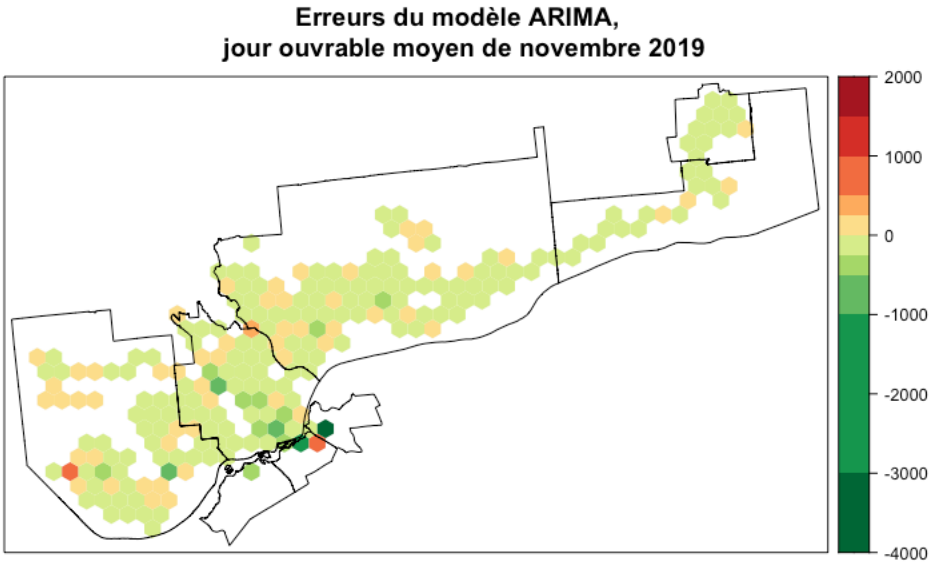


Figure 7-7 Erreurs du modèle ARIMA, jour ouvrable moyen de novembre 2019

7.1.2 Intégration de la variable logement

Ayant identifié que les zones où l'erreur s'avère être plus importante peuvent être assimilées à des zones de transformation résidentielle, il s'agit donc d'ajouter au modèle la variable des logements. Cette variable est issue de l'actualisation des données du recensement 2016 à l'aide des données de permis de construire.

Le premier modèle utilisé pour inclure cette variable est un modèle de régression linéaire avec correction de l'erreur ARIMA. Cela permet donc de garder la dimension de prévision autorégressive et d'y ajouter une variable explicative.

L'utilisation des données de logements implique que la prédiction est effectuée pour la pointe AM. Il a été montré précédemment que les logements avaient majoritairement une influence sur les origines en pointe du matin.

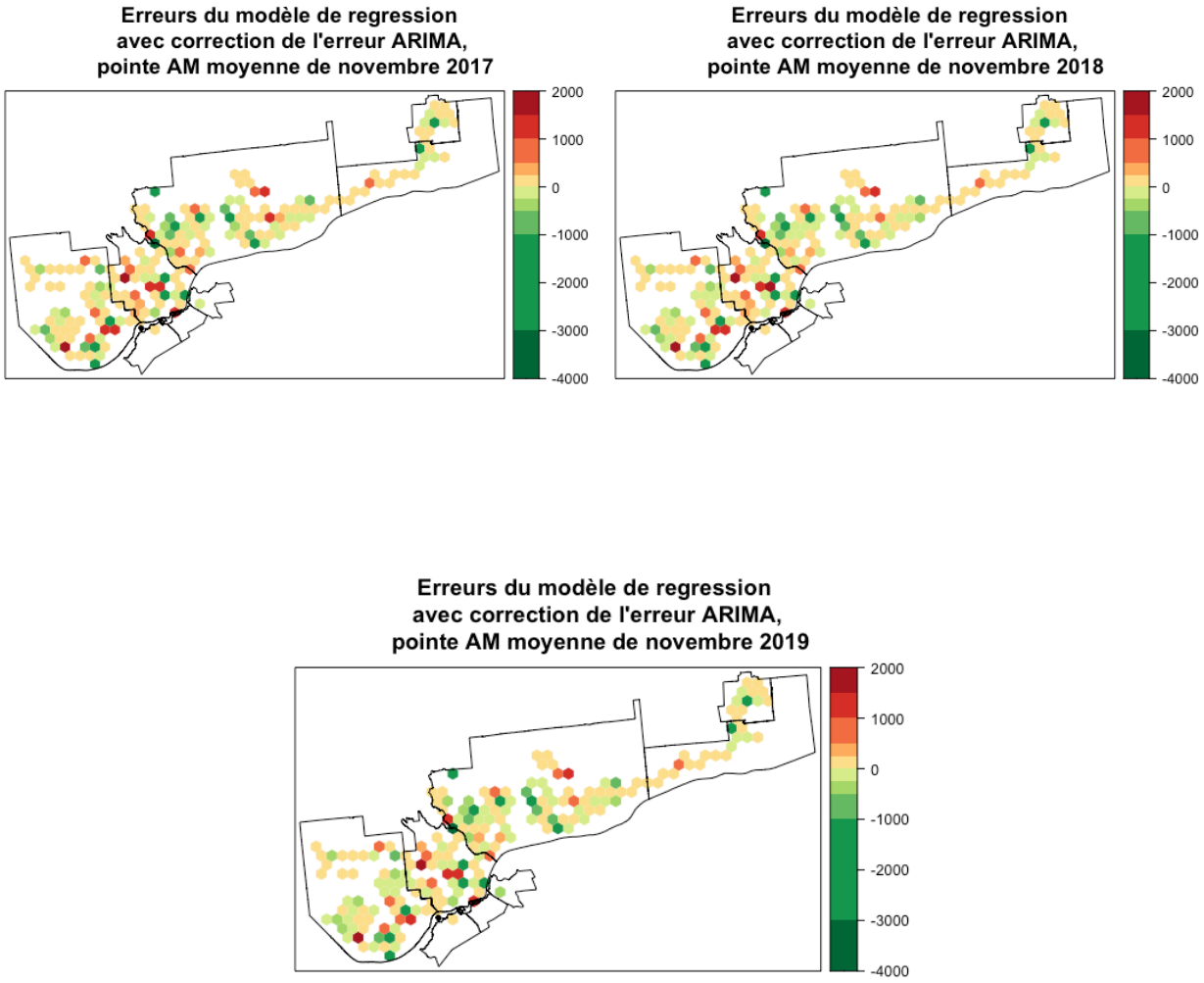


Figure 7-8 Résidus des mois de novembre pour le modèle de régression linéaire avec correction de l'erreur ARIMA

Les erreurs introduites par ce modèle sont trop importantes pour être étudiées. Ce modèle n'est pas approprié donc aucune conclusion ne peut être tirée des résidus.

7.1.3 Modèle sur 3 mois

Lors de la création de modèles sur le réseau global, les modèles calibrés sur les 3 mois d'automne se sont montrés plus performants. Il s'agit donc de tester si ce résultat est constant pour les zones hexagonales. Le nombre de montants sur une pointe du matin moyenne mensuelle est la variable

dépendante. Le modèle est calibré sur les années 2012 à 2016 et à un horizon de 3 ans. La comparaison avec le modèle mensuel classique est effectuée ci-dessous. Les paramètres de calibration sont identiques.

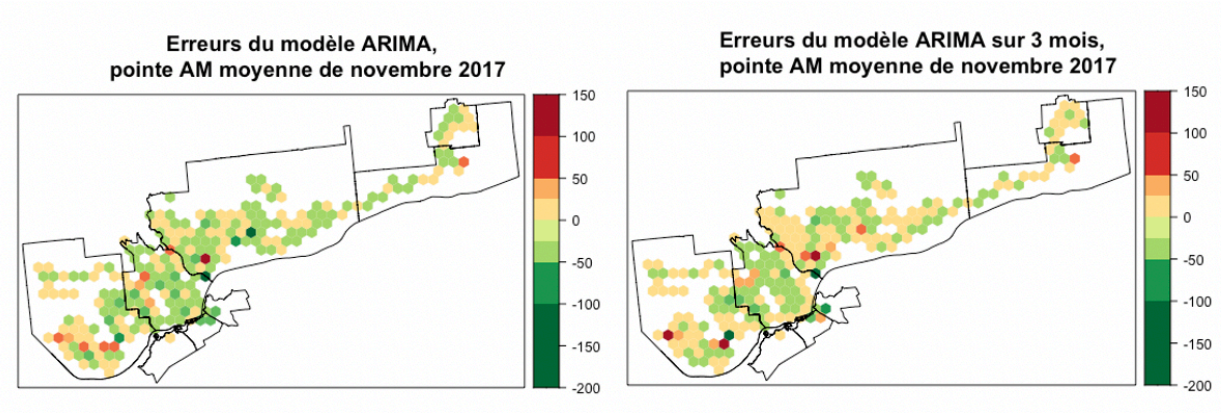


Figure 7-9 Comparaison des deux modèles pour novembre 2017

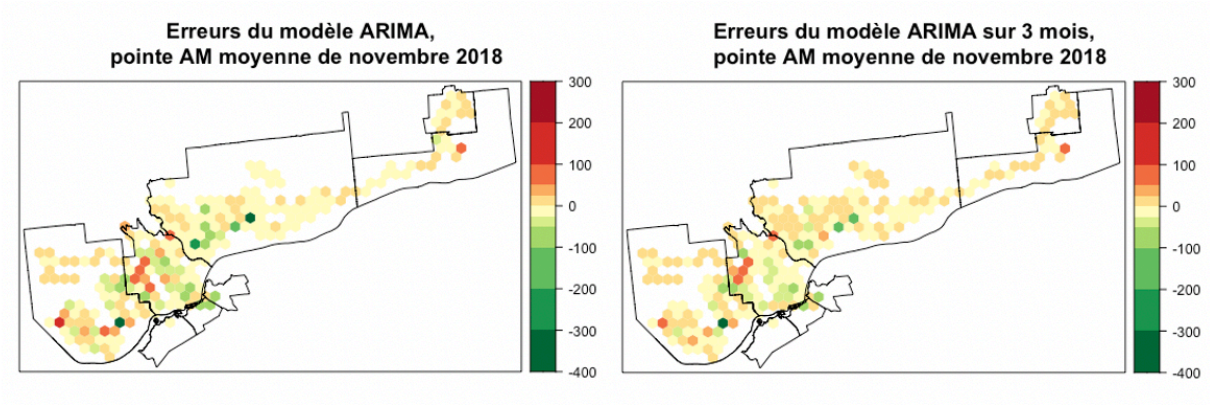


Figure 7-10 Comparaison des deux modèles pour novembre 2018

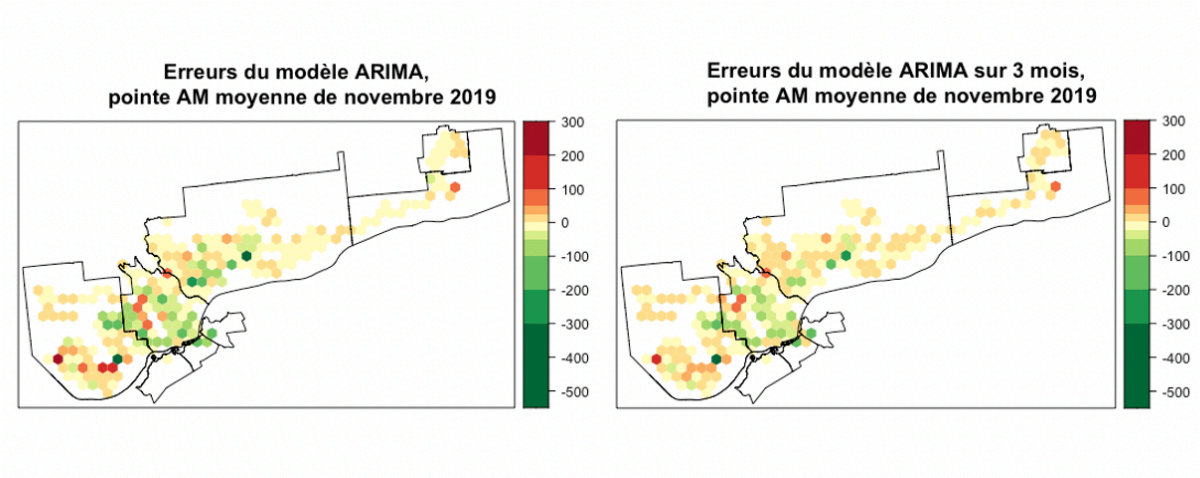


Figure 7-11 Comparaison des deux modèles pour novembre 2019

Il est observé que les erreurs sont globalement plus faibles pour une modélisation uniquement des 3 mois d'automne. Néanmoins, certaines valeurs peuvent apparaître comme légèrement aberrantes sur ce modèle. Ce phénomène peut être dû à un nombre réduit de données pour calibrer le modèle. Il est donc pertinent de privilégier une modélisation uniquement de la période d'automne. La distribution des erreurs est également mise en valeur pour confirmer la validité du modèle.

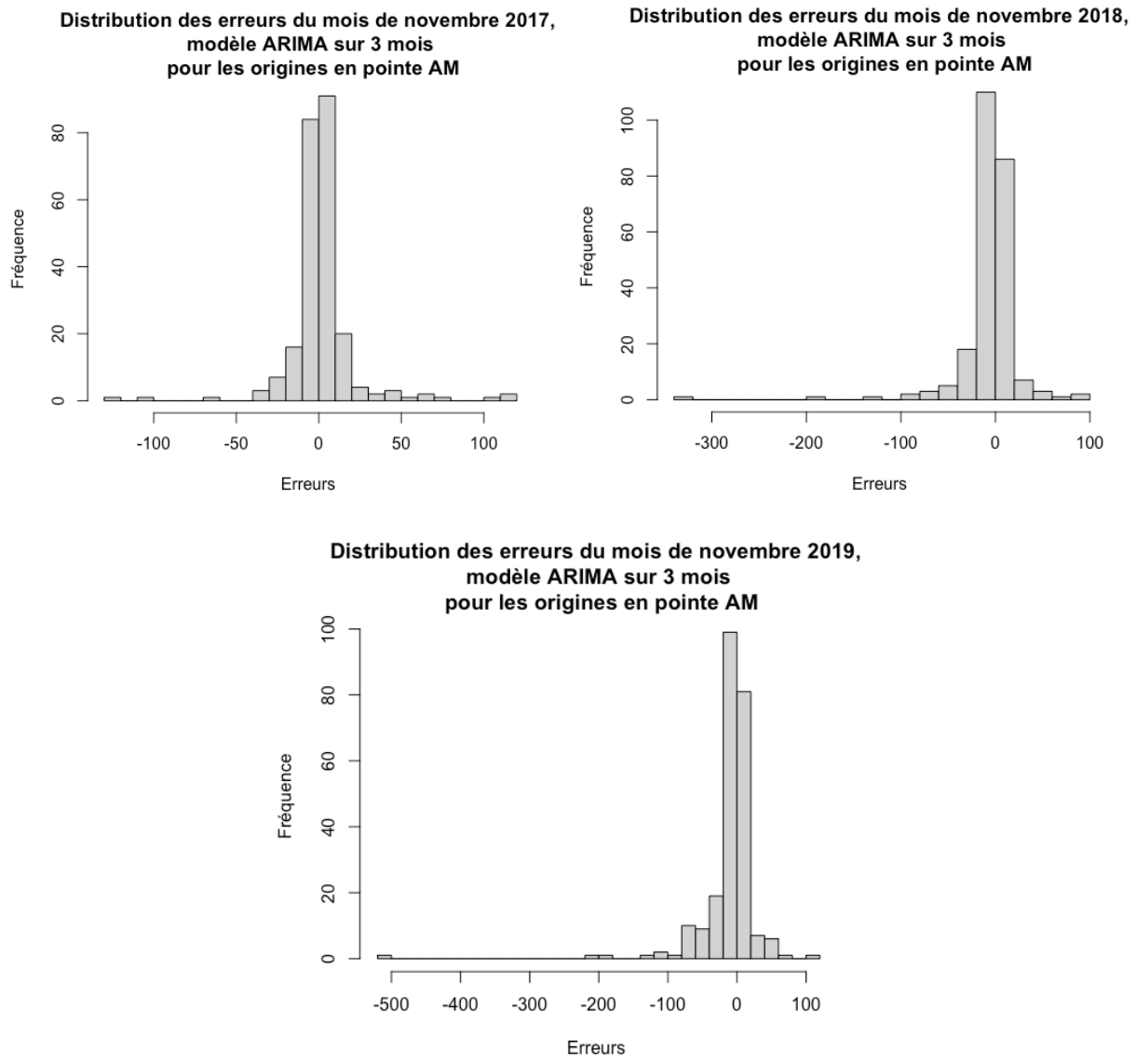


Figure 7-12 Distributions des erreurs pour le modèle sur 3 mois en fonction des années

7.2 Modèle de régression sur série temporelle

À la suite des résultats obtenus sur les résultats précédents, la demande doit être modélisée en prenant en compte son aspect autorégressif, mais également des facteurs d'influence. Un modèle de régression linéaire associant des valeurs du retard de la demande et des valeurs de variables indépendantes semble donc être une solution. Le modèle étudié dans cette section est un modèle de régression multiple pour séries temporelles.

Le modèle s'explicite de la manière suivante :

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \\
&+ \delta_{11} X_{1,t-1} + \delta_{12} X_{1,t-2} + \cdots + \delta_{1q} X_{1,t-q} \\
&+ \ldots \\
&+ \delta_{k1} X_{k,t-1} + \delta_{k2} X_{k,t-2} + \cdots + \delta_{kq} X_{k,t-q} \\
&+ u_t
\end{aligned}
$$

Où Y est la variable dépendante et p représente le nombre de retards pris en compte,

$(X_1, X_2, \ldots, X_k)$ sont k variables explicatives dont q valeurs du retard sont prises en compte,

δ représente des coefficients réels calculés pour calibrer le modèle,

u_t représente l'erreur sur la prévision de Y_t.

Pour que ce modèle soit pertinent, les hypothèses suivantes doivent être vérifiées :

- Le terme d'erreur u doit posséder une moyenne nulle.
- $(Y, X_1, X_2, \ldots, X_k)$ doivent être stationnaires
- $(Y_t, X_{1,t}, X_{2,t}, \ldots, X_{k,t})$ et $(Y_{t-j}, X_{1,t-j}, X_{2,t-j}, \ldots, X_{k,t-j})$ deviennent indépendants lorsque j est grand.
- La probabilité d'obtenir des valeurs absurdes est faible
- Il n'existe aucune multicolinéarité parfaite.

Le modèle sera appliqué pour les données de montants sur une pointe AM moyenne de chacun des mois. La variable explicative utilisée dans un premier temps est constituée des résultats du recensement 2016 concernant les logements privés et actualisés mensuellement avec les données de permis de construire.

7.2.1 Stationnarité et indépendance

La première hypothèse à vérifier pour mettre en place ce modèle est la stationnarité des séries temporelles. Toutes les variables doivent donc être stationnaires. En visualisant les données disponibles, il s'avère que la variable la plus compliquée à rendre stationnaire est la variable dépendante, à savoir les données de carte à puce. Il s'agit donc de trouver une transformation qui

permettra d'obtenir une variable dépendante stationnaire. Cette transformation sera par la suite testée sur la variable de logements pour déterminer si la série est stationnaire après cette transformation commune.

Pour tester la stationnarité des séries temporelles, le test augmenté de Dickey-Fuller (ADF) est utilisé. L'hypothèse nulle de ce test est que la série est non-stationnaire. L'objectif est donc de rejeter cette hypothèse en obtenant une valeur p inférieure à 0.05. La version augmentée du test de Dickey-Fuller est utilisée, car elle permet de tester des processus autorégressifs d'ordre supérieur à 1.

Le test effectué sur la variable dépendante pour chaque secteur confirme bien la non-stationnarité de cette dernière avec une p-value sur la majorité des secteurs de 0.99. Il s'agit donc d'appliquer une transformation pour stationnariser la série. Les transformations dites classiques sont testées. La variance des séries étant relativement constante, une transformation de type log ou racine carrée n'apporte pas de meilleur résultat. Visuellement, il est possible de remarquer que c'est majoritairement le motif saisonnier et la tendance qui rendent la série non stationnaire. Les différenciations simples ou saisonnières sont donc les outils adaptés.

Une différenciation d'ordre 2 est la transformation qui sera utilisée par la suite. C'est la plus simple permettant d'obtenir des séries stationnaires. Elle est donc également testée sur la variable de logements.

Une autre hypothèse importante à vérifier est l'indépendance des données (l'indépendance stricte est plus contraignante que l'hypothèse énoncée précédemment). Un test de corrélation linéaire de Pearson est donc appliqué entre les deux variables. Un échantillon des résultats est donné sur le Tableau 7-1 à titre d'exemple.

Tableau 7-1 Échantillon des tests réalisés sur les séries différenciées d'ordre 2

Zone	**X67**	**X86**	**X66**	**X48**	**X33**	**X65**
p-value du test ADF appliqué aux données de CAP (différenciation d'ordre 2)	0.01	0.01	0.01	0.01	0.01	0.01
p-value du test ADF appliqué aux données de logements (différenciation d'ordre 2)	0.01	0.01	0.01	0.027	0.01	0.01
Coefficient de corrélation linéaire entre les deux variables.	0.016	0.001	-0.034	0.142	0.096	-0.212

Toutes les p-values sont inférieures à 0.05 et la valeur absolue du coefficient de Pearson est inférieure à 0.4. Les hypothèses sont ainsi vérifiées.

7.2.2 Détermination du retard à utiliser

Pour mettre en place un modèle de régression sur séries temporelles, il faut déterminer l'ordre du retard à utiliser pour chaque variable. La manière de procéder est tester plusieurs combinaisons d'ordre et de comparer des indicateurs sur les modèles. L'indicateur utilisé pour cette étude est le critère d'information de Bayes (BIC) :

$$BIC(p) = \log\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{\log(T)}{T}$$

Avec p la longueur du retard, SSR la somme quadratique des résidus du modèle et T le nombre de valeurs utilisées pour calibrer le modèle.

L'objectif est de minimiser cet indicateur.

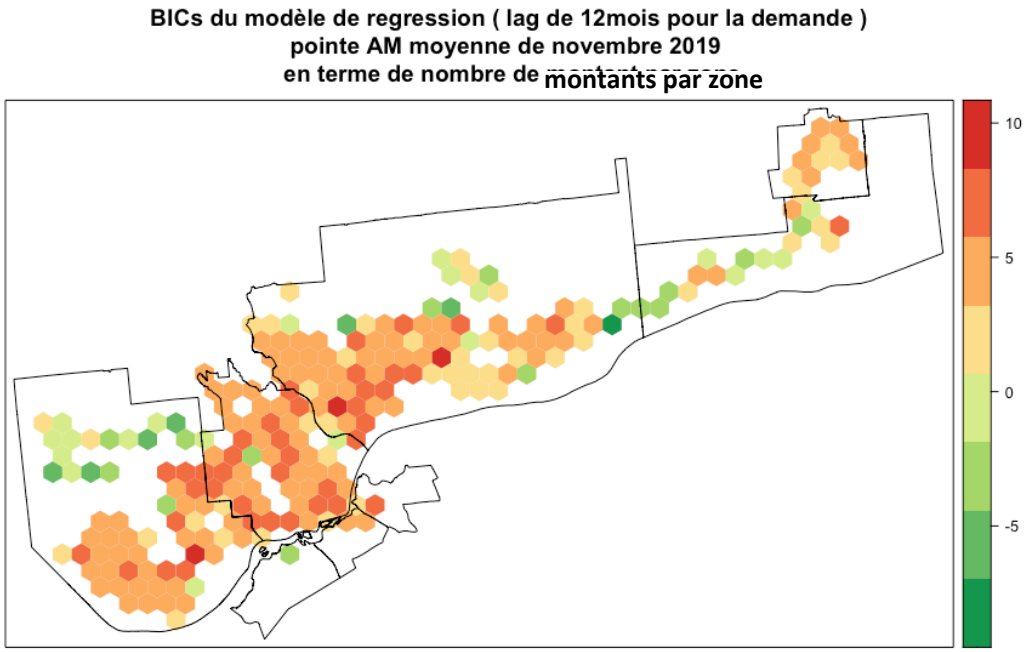


Figure 7-13 BICs du modèle pour un retard de 12mois

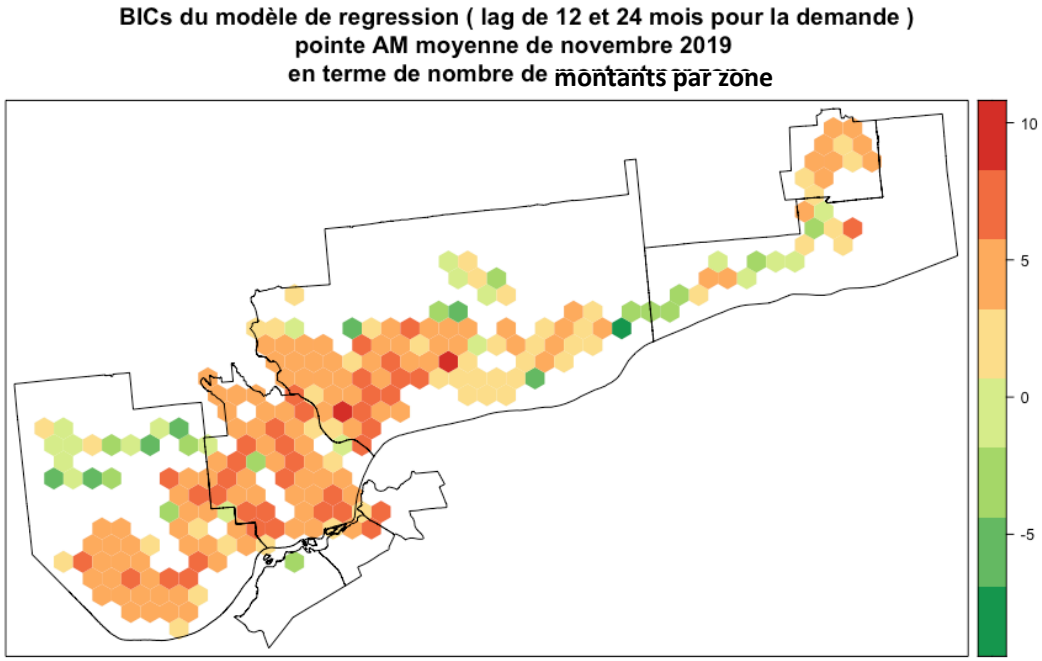


Figure 7-14 BICs du modèle pour un retard de 12 et 24 mois

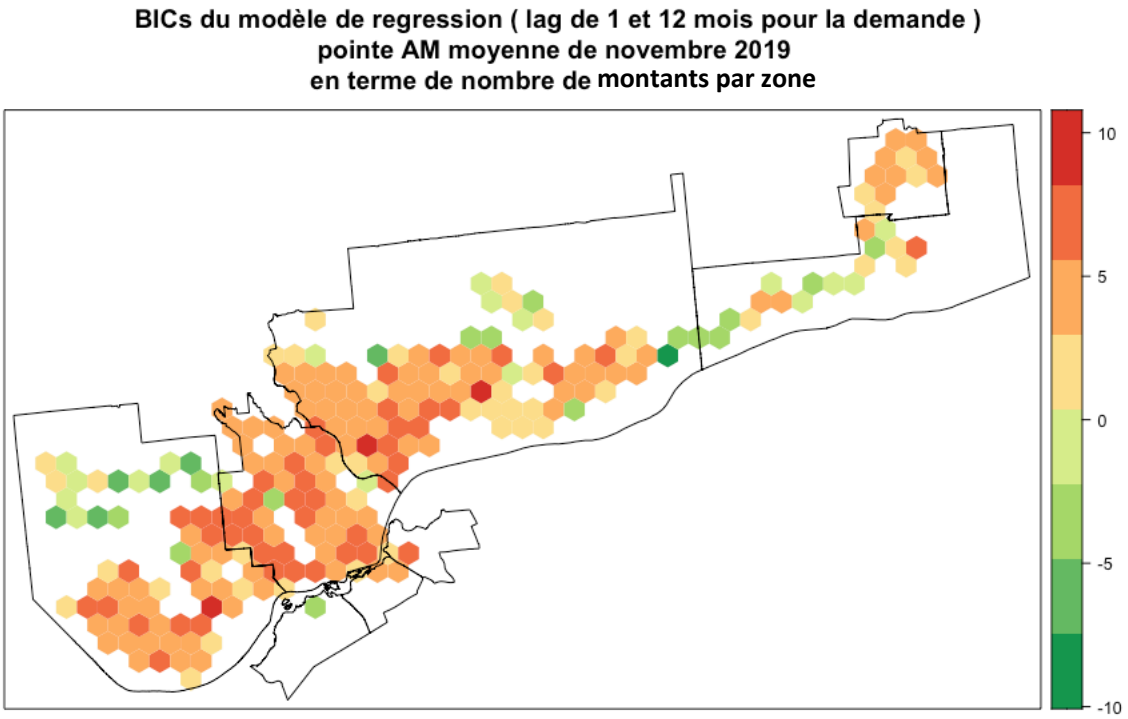


Figure 7-15 BICs du modèle pour un retard de 1 et 12 mois

Le retard de 12 mois pour la variable de fréquentation des TC est le paramètre qui minimise au mieux le critère BIC. Le modèle est donc calibré en utilisant ces valeurs. La Figure 7-16 représente les erreurs obtenues au mois de novembre 2019.

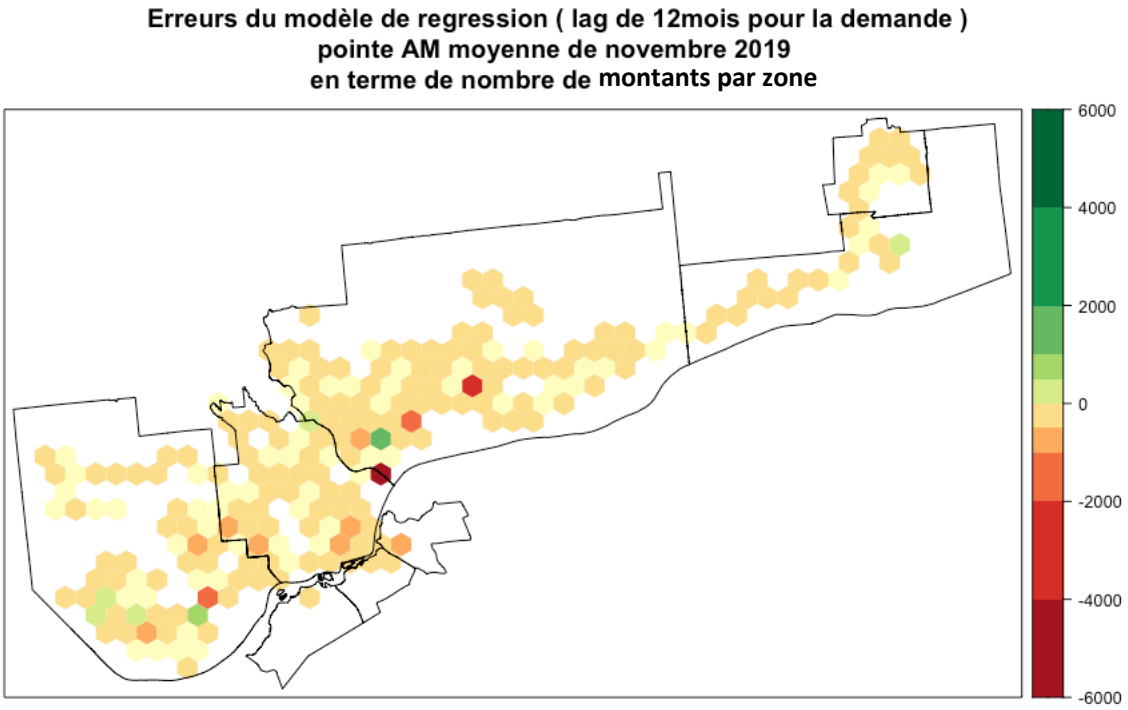


Figure 7-16 Résultat du modèle de régression sur séries temporelles

Il est possible de constater une forte erreur sur la prévision. Les résidus avant transformation inverse disponibles sur la Figure 7-17, montrent des valeurs qui semblent ne pas comporter d'erreurs trop importantes. Ces valeurs sont prises sur une zone à titre d'illustration, mais le comportement est similaire sur toutes les zones faisant preuve d'une forte erreur. Ces valeurs sont centrées autour de zéro et il n'y a pas trop de valeurs extrêmes. Mais une fois la transformation inverse effectuée et la prévision sur le nombre de montants obtenue (

Figure 7-18), il y a une dérive dans les résultats. C'est un problème qui peut émerger des transformations par une différenciation. Il est souvent recommandé de limiter au maximum l'ordre de la différenciation. Ici seulement une différenciation d'ordre deux produit une série stationnaire. Après des recherches pour éliminer ce phénomène, aucune solution n'est trouvée. Le modèle ne peut donc pas être jugé comme adéquat.

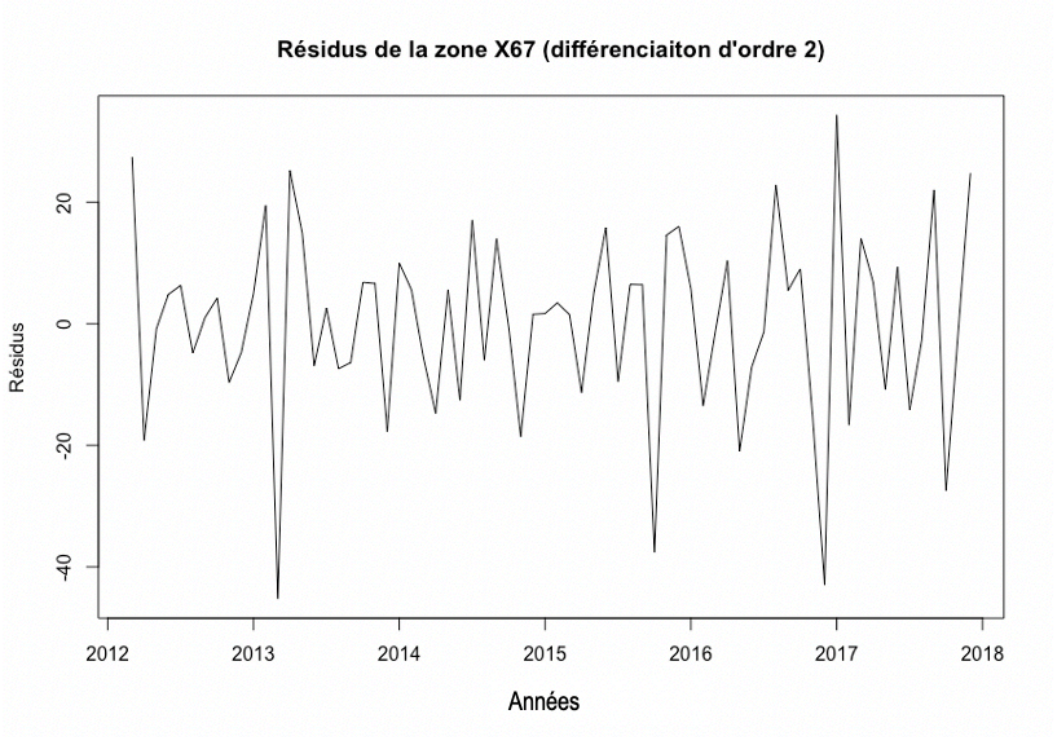


Figure 7-17 Exemple des résidus d'une zone

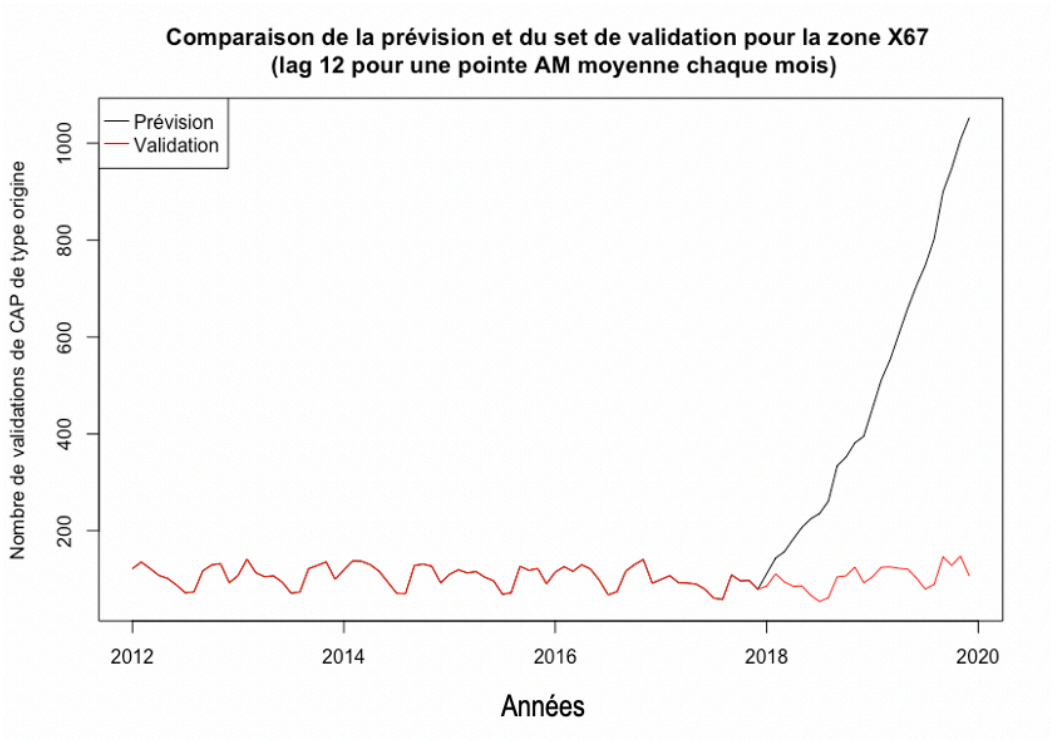


Figure 7-18 Exemple de prévision sur une zone

7.3 Création d'une typologie de zones

La création d'une typologie a pour but d'améliorer significativement la performance des modèles qui seront établis par la suite. En effet, créer des modèles uniques pour chaque secteur peut paraître pertinent, car les indicateurs de performance indiquent souvent une erreur plus faible pour chacune des zones. Mais cette méthodologie n'est pas fiable quand il s'agit de produire des prévisions. Il a été prouvé que de regrouper les zones d'agrégation en typologie et d'appliquer un modèle pour chacune des typologies permet de produire des résultats plus fiables. La typologie sera basée dans notre cas sur les caractéristiques de chaque zone. Ces caractéristiques, lorsqu'elles sont communes, permettront la création d'un type. Pour toute la typologie, un modèle similaire sera appliqué, mais les variables et le paramétrage du modèle pourront varier selon les types.

7.3.1 Typologie basée sur la demande

Dans un premier temps, il paraît intéressant d'étudier une typologie basée sur les variations de la demande. Cette typologie ne sera pas utilisée dans la création de modèles, mais elle pourra permettre d'identifier certaines caractéristiques du territoire qui affectent de manière similaire les variations de la demande. C'est en quelque sorte le travail inverse par rapport à celui qui sera réalisé par la suite : créer une topologie de variation de la demande pour ensuite identifier dans ces typologies certaines caractéristiques du territoire qui peuvent être communes aux zones appartenant à un type.

Dix types sont donc déterminés en fonction des motifs de variation de la demande; les types sont positionnés géographiquement sur la Figure 7-19. Les parcs-o-bus sont également positionnés sur cette carte, car il s'agit d'un élément déterminant dans les variations de la demande.

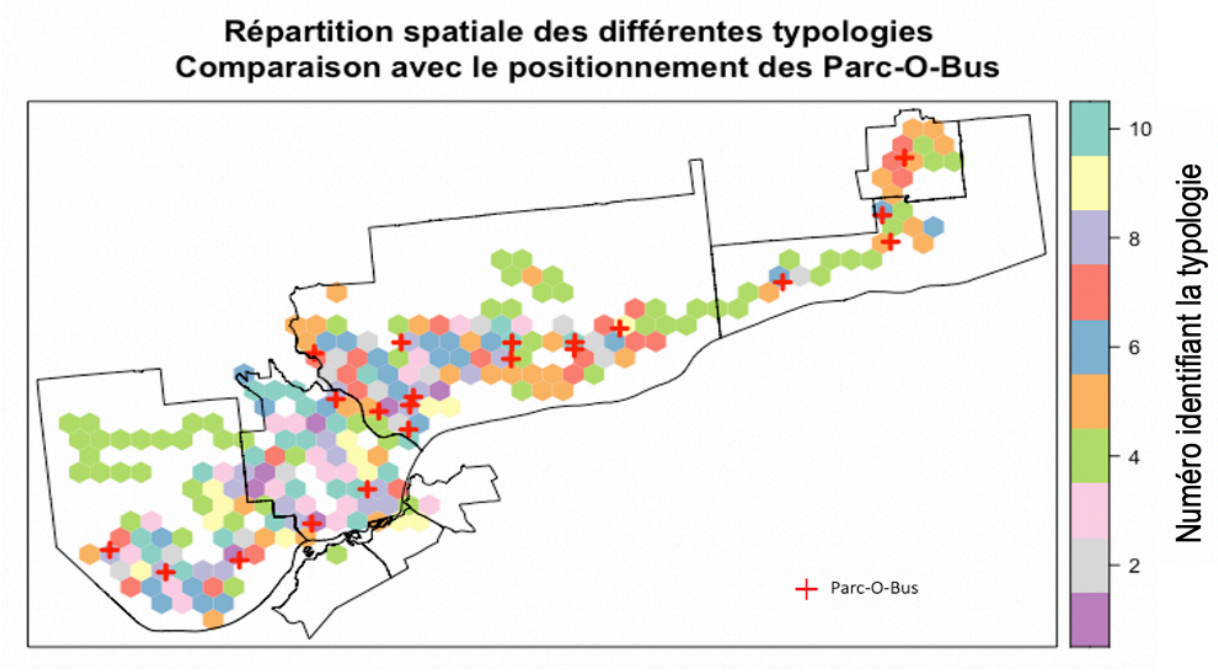


Figure 7-19 Réparation spatiale des types de la typologie basées sur la demande

Les variations types sont présentées dans les figures suivantes. Pour chaque type, une variation moyenne est établie pour une année type des trois mois d'automne.

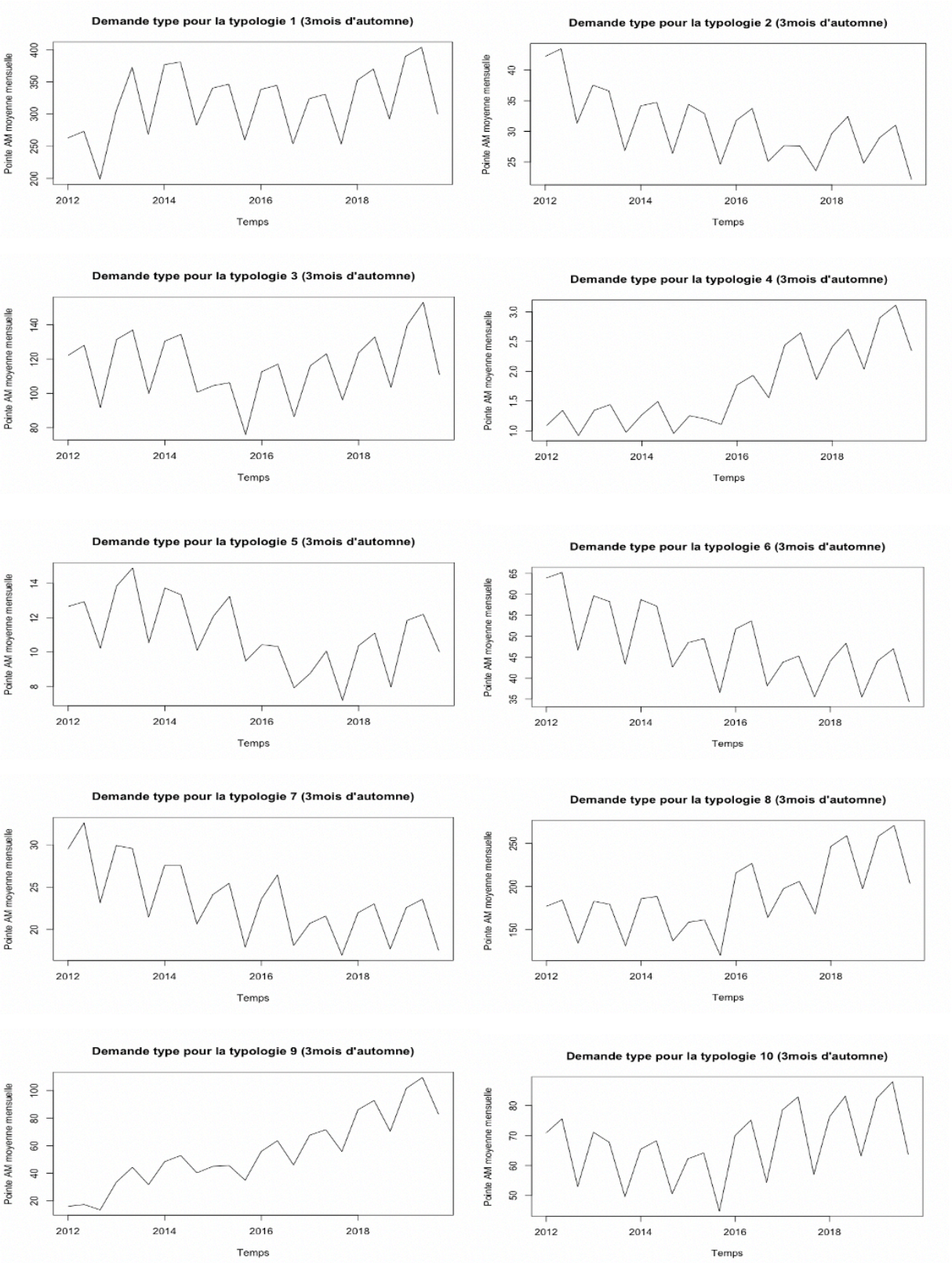


Figure 7-20 Évolution de la fréquentation TC en pointe AM pour chaque segment de la typologie

Le nombre de logements par zone appartenant à une même typologie est étudié dans le Tableau 7-2. Le but est d'identifier si les logements déterminent significativement la demande.

Tableau 7-2 Logements selon la typologie basée sur la demande

Typologie	1	2	3	4	5	6	7	8	9	10
Moyenne du nombre de logements par zone	1041	628	1227	128	313	618	463	985	924	560
Écart-type du nombre de logement par zone	500	318	920	144	235	249	232	575	787	363

La position des établissements scolaires est également répertoriée.

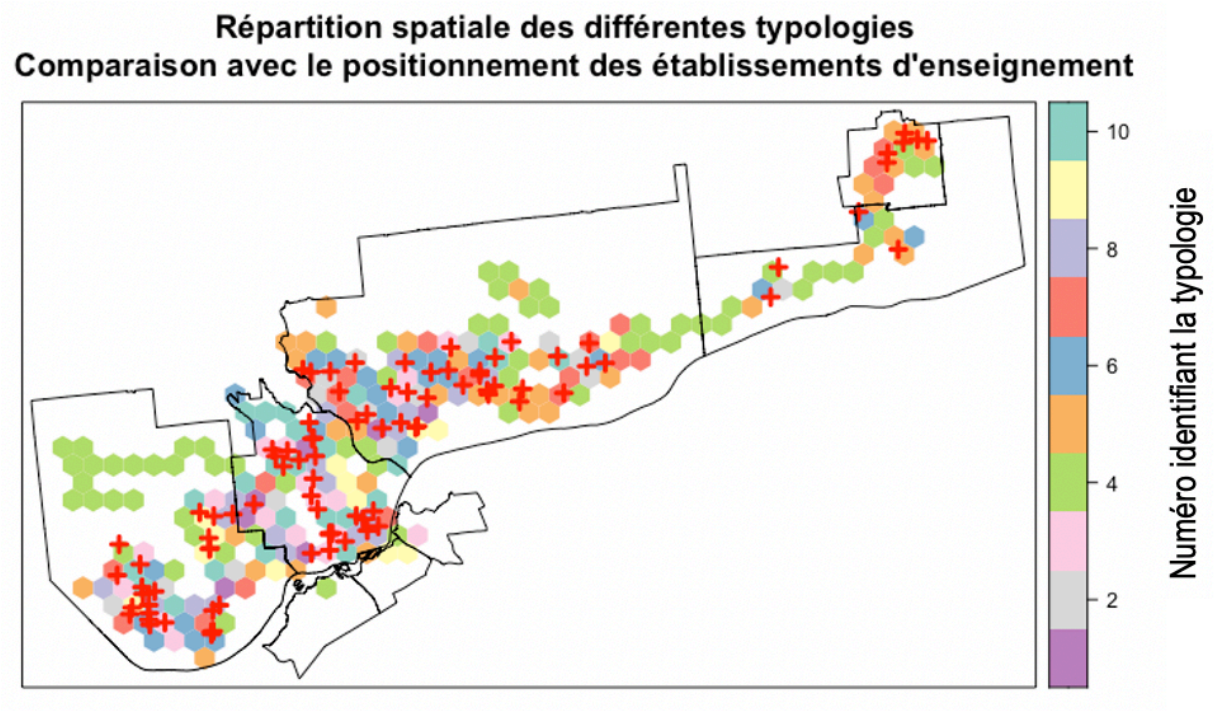


Figure 7-21 Répartition spatiale des segments de la typologie

Grâce à cette typologie basée sur la demande, il est possible de lier des évolutions de la fréquentation du réseau de TC à des paramètres. Les parcs-o-bus ont une influence non négligeable,

la présence est souvent liée à l'appartenance de la zone au type 8. Le nombre de logements est aussi un caractère important dans la différenciation des types. En prenant en compte ces caractéristiques, il est maintenant possible de réaliser le travail inverse. L'objectif est de créer une typologie basée sur les caractéristiques comme les stationnements incitatifs et les logements et de déterminer si ces types peuvent, via un taux de génération, refléter le nombre de montants en pointe AM.

7.3.2 Typologie basée sur le nombre de logements

L'objectif de cette partie est de commencer à bâtir une typologie basée sur les caractéristiques des zones. Étant donné que la variable des logements actualisée est une des seules variables spatialisées et disponibles mensuellement, il s'agit de déterminer si seuls les logements peuvent produire une typologie pertinente pour réaliser un modèle de prévision. Cette méthode de prévision se base sur la création d'une typologie pour ensuite attribuer un taux de génération de montant moyen pour chaque zone d'un même type. Par la suite, les résultats sont déterminés en multipliant le taux de génération du type de la zone en question par le nombre de logements présents dans la zone.

Les types sont créés afin de minimiser la variation du taux de génération de montants par logement dans le temps. Il est également souhaitable d'avoir des taux moyens assez différenciés entre chaque type de zone. Une zone peut changer de type au cours du temps, car les centroïdes sont stables dans le temps. La composition de la typologie est explicitée dans le Tableau 7-3.

Tableau 7-3 Composition de la typologie de logements

Type	**C1**	**C2**	**C3**	**C4**	**C5**
Centroide (nombre de logements)	5	50	100	300	800

Cinq types sont demandés pour cette typologie et les bornes du nombre de logements de chacun sont déterminées avec les données de logements allant de 2012 à 2016. Afin de vérifier cette typologie, les taux de montants générés par logement sont étudiés au cours du temps pour chaque type.

Pour toutes les zones, le taux de montants générés lors de la pointe AM pour chaque mois. La moyenne et l'écart-type de ces valeurs sont visualisés pour attester ou non de leur stabilité au cours du temps.

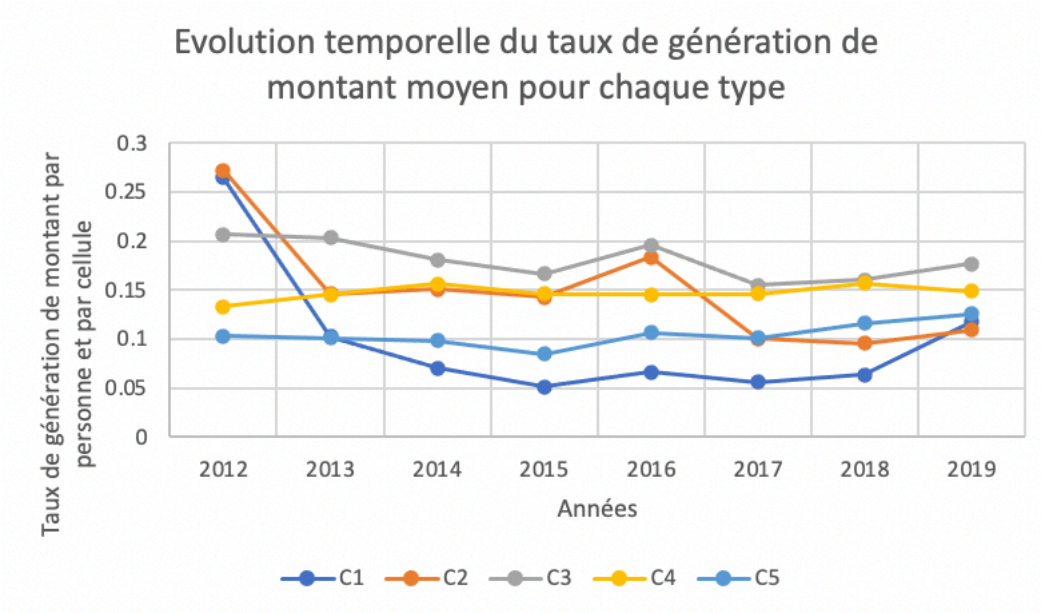


Figure 7-22 Évolution temporelle du taux de génération de montants moyen pour chaque type

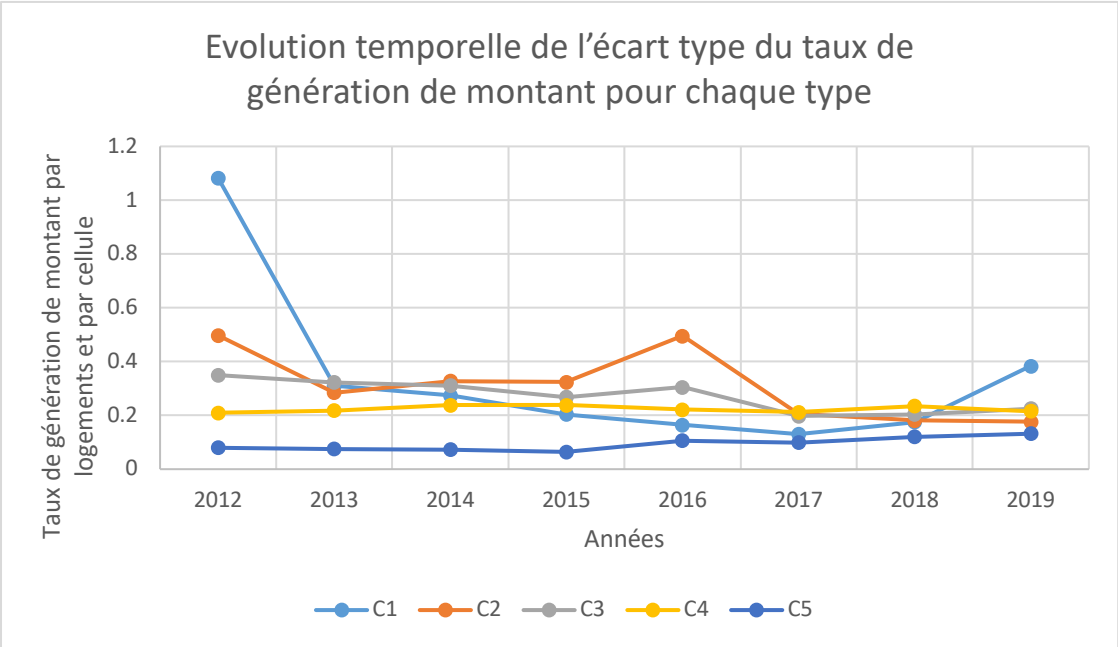


Figure 7-23 Évolution temporelle de l'écart-type du taux de génération de montants moyen pour chaque type

Il est possible d'observer que pour les typologies, certains taux se différencient d'un type à l'autre. Les taux de génération ne sont pas extrêmement stables au cours du temps, mais il est tout de même possible de tenter la calibration d'un modèle de prévision. Sur les figures, les taux des années 2017-2019 sont présentés, mais pour déterminer le taux utilisé dans la modélisation, seules les années de 2012 à 2016 sont utilisées. Un taux moyen est attribué à chaque type et les valeurs de nombre de montants par zone sont calculées en multipliant le taux attribué à la zone par son nombre de logements. Les figures suivantes illustrent les erreurs obtenues par le biais de l'utilisation de cette méthode.

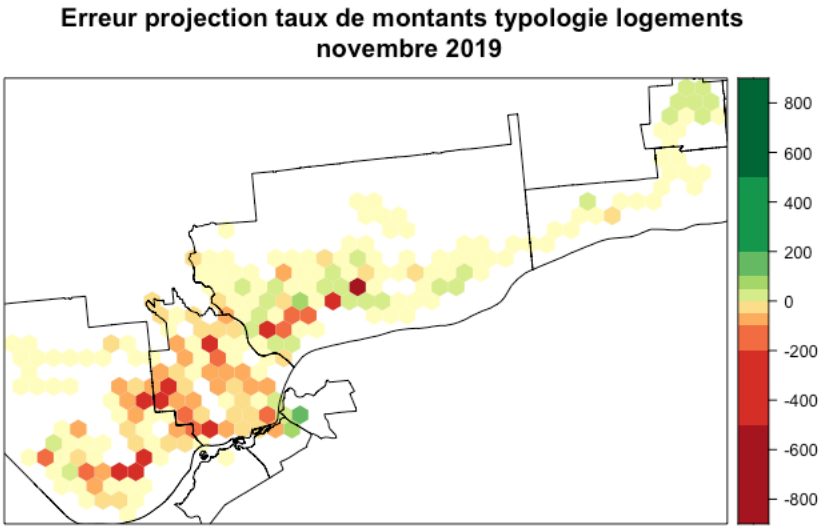


Figure 7-24 Erreurs projection typologie de logements 2019

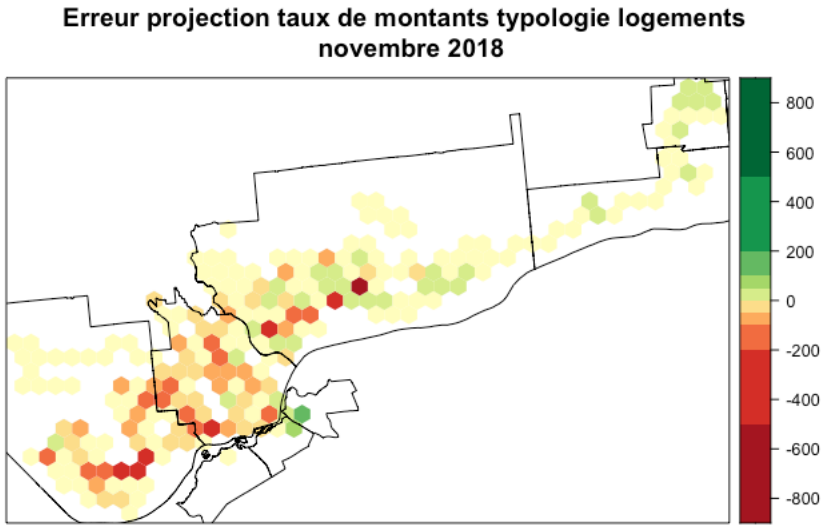


Figure 7-25 Erreurs projection typologie de logements 2018

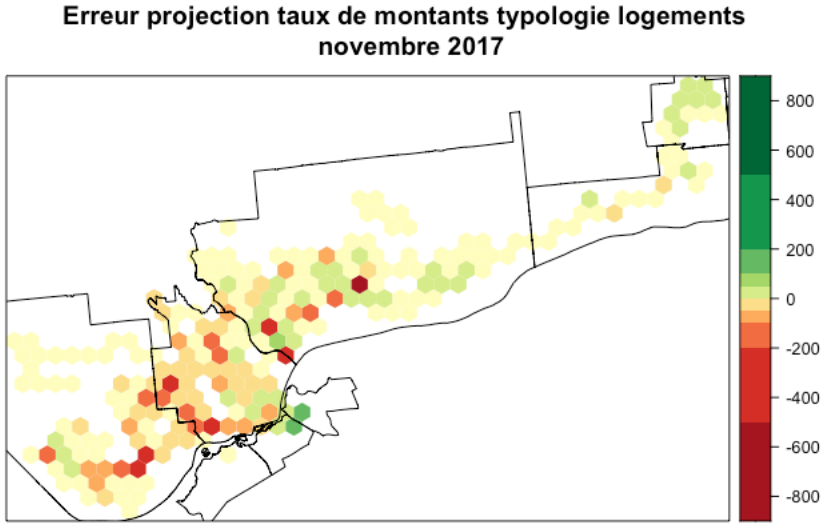


Figure 7-26 Erreurs projection typologie de logements 2017

Tout d'abord, comparé au modèle de régression linéaire sur les séries temporelles, le modèle est beaucoup plus performant. Le modèle ARIMA fournit des erreurs plus faibles, mais ne comporte aucune dimension explicative, car il n'est relié à aucune variable explicative. L'utilisation d'une

typologie pour créer des taux de génération de montants semble donc être une méthode pertinente. La typologie de logements peut être affinée en étudiant la localisation des erreurs du modèle.

Il est possible de constater que les zones comportant des parcs-o-bus ont des erreurs importantes. Ce paramètre doit donc être intégré dans la prochaine typologie. De plus, un problème constaté est la sous-estimation de la demande. Même avec une faible demande, le nombre de montants est sous-estimé sur un grand nombre de zones. Cela est explicable par l'augmentation du taux de génération de montants sur les années 2017-2019, visible sur la Figure 7-22.à

7.3.3 Typologie basée sur les logements et les parcs-o-bus

Dans la typologie basée sur la demande, il a été montré que les parcs-o-bus ont un effet non négligeable sur la demande. Il parait donc intéressant d'essayer de les intégrer dans la typologie de zone. À noter que seul un nombre restreint de zones comprennent un parc-o-bus.

Les données des parcs-o-bus sont intégrées en conservant la même typologie que celle des logements, mais il est supposé que le taux de génération d'une place de stationnement incitatif soit de 1. Cette hypothèse peut être effectuée, car l'analyse des parcs-o-bus a montré que la majorité des zones de stationnement fonctionnaient à leur capacité maximale. De nouveaux taux de génération par logements sont calculés de nouveau en considérant que les montants liés aux stationnements incitatifs n'apparaissent pas dans le taux de génération par logements. Les projections sont effectuées de manière identique au modèle précédent et les montants liés au parc-o-bus sont ajoutés par la suite en supposant un taux de génération de montants par place de stationnement égal à 1. Les résultats sont donc générés en multipliant le nombre de logements par zone par le taux de génération de montants (uniquement lié aux logements) puis en ajoutant les montants liés aux stationnements incitatifs en supposant un taux de génération de 1 par place de parking. Les résultats sont disponibles sur les figures suivantes.

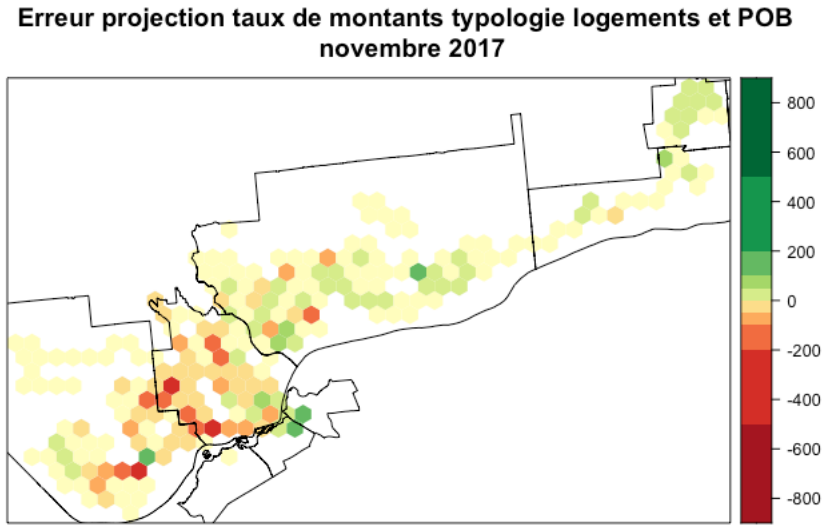


Figure 7-27 Erreurs du modèle pour une projection en novembre 2017

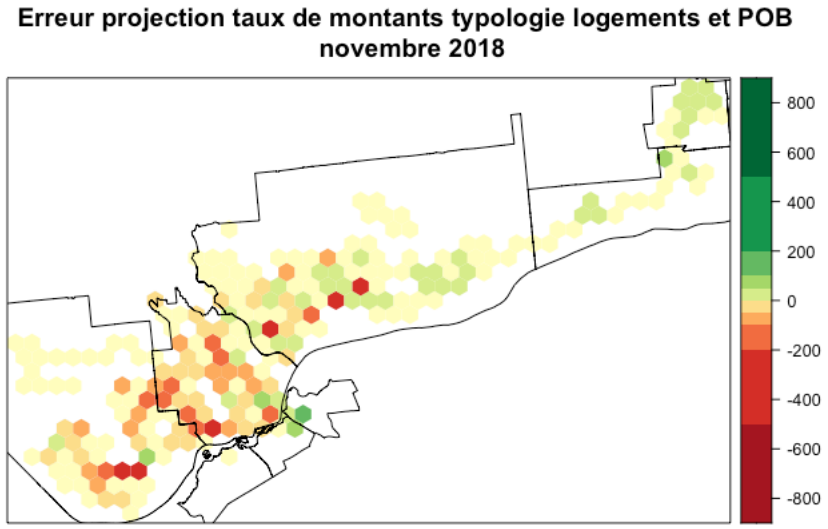


Figure 7-28 Erreurs du modèle pour une projection en novembre 2018

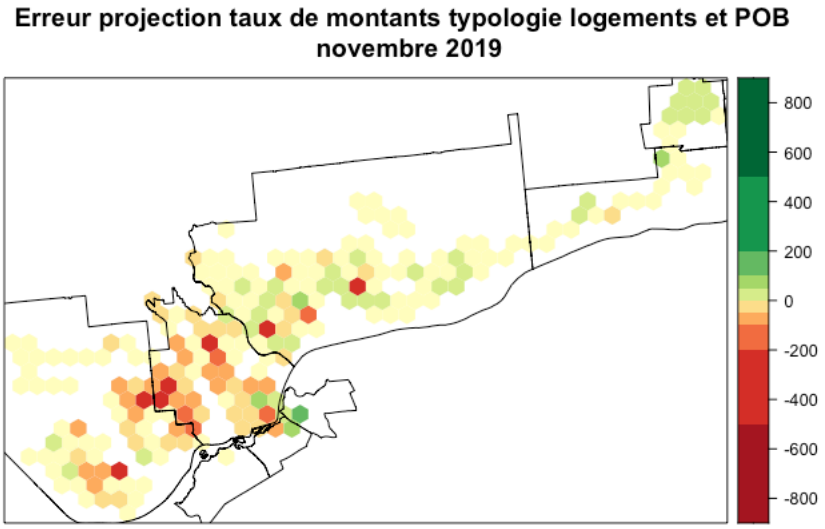


Figure 7-29 Erreurs du modèle pour une projection en novembre 2019

L'échelle est la même que pour le modèle précédent pour mettre en valeur le gain en performance de cette nouvelle structure de modèle où le taux de génération de montants est différencié de celui par place de parc-o-bus. Une erreur plus faible est observable sur les trois années projetées, particulièrement sur novembre 2017. Cette typologie est donc pertinente pour la création d'un modèle de taux de génération de montants. Il apparaît que cette typologie maximise l'influence des variables explicatives longitudinales disponibles.

7.4 Modèle spatial

La création d'un modèle spatial est une tentative de déterminer les caractéristiques du territoire qui influencent significativement la demande dans les zones hexagonales. Le but est de constater quelles différences entre les zones expliquent différents volumes de demande. Ces résultats aideront par la suite à la création d'une nouvelle typologie qui pourra peut-être améliorer les résultats obtenus par la typologie de logements et de parc-o-bus. L'objectif est d'intégrer certaines données du recensement 2016 dans un modèle sans composante temporelle et d'observer quelles caractéristiques différencient significativement les zones entre elles et au regard de la prévision de la demande. Plusieurs modèles spatiaux sont calibrés et la significativité des variables est observée

pour déterminer quelles sont les plus pertinentes pour différencier les zones dans l'objectif de prévoir la fréquentation du réseau de TC.

Étant donné que l'objectif est de modéliser les différences entre les zones, il n'y a pas de composante temporelle. Il est donc possible de choisir un unique point dans le temps. Dans la mesure où le manque de donnée évoluant dans le temps est un problème majeur dans ce projet, il est possible, en supprimant cette dimension temporelle, d'avoir un nombre important de données. Le point le plus pertinent est de prendre le mois de mai 2016, car le dernier recensement a été réalisé à ce moment.

Les données sont obtenues sur les zones hexagonales en supposant une répartition constante des différentes variables sur les îlots de diffusion du recensement. C'est pour rappel le niveau d'agrégation le plus précis sur lequel les résultats des recensements sont disponibles.

Les données potentielles sont composées de :

- Validation de carte à puce : d'un jour ouvrable moyen de mai 2016
- Le recensement 2016 : composition de la population, emploi, chômage, taille moyenne des ménages, et population
- Les parcs-o-bus : capacité et taux d'utilisation.
- Les points d'intérêt de la ville de Gatineau

Les variables explicatives utilisées dans le modèle doivent vérifier l'hypothèse d'indépendance linéaire. La Figure 7-30 donne les coefficients de corrélation linéaire entre les variables.

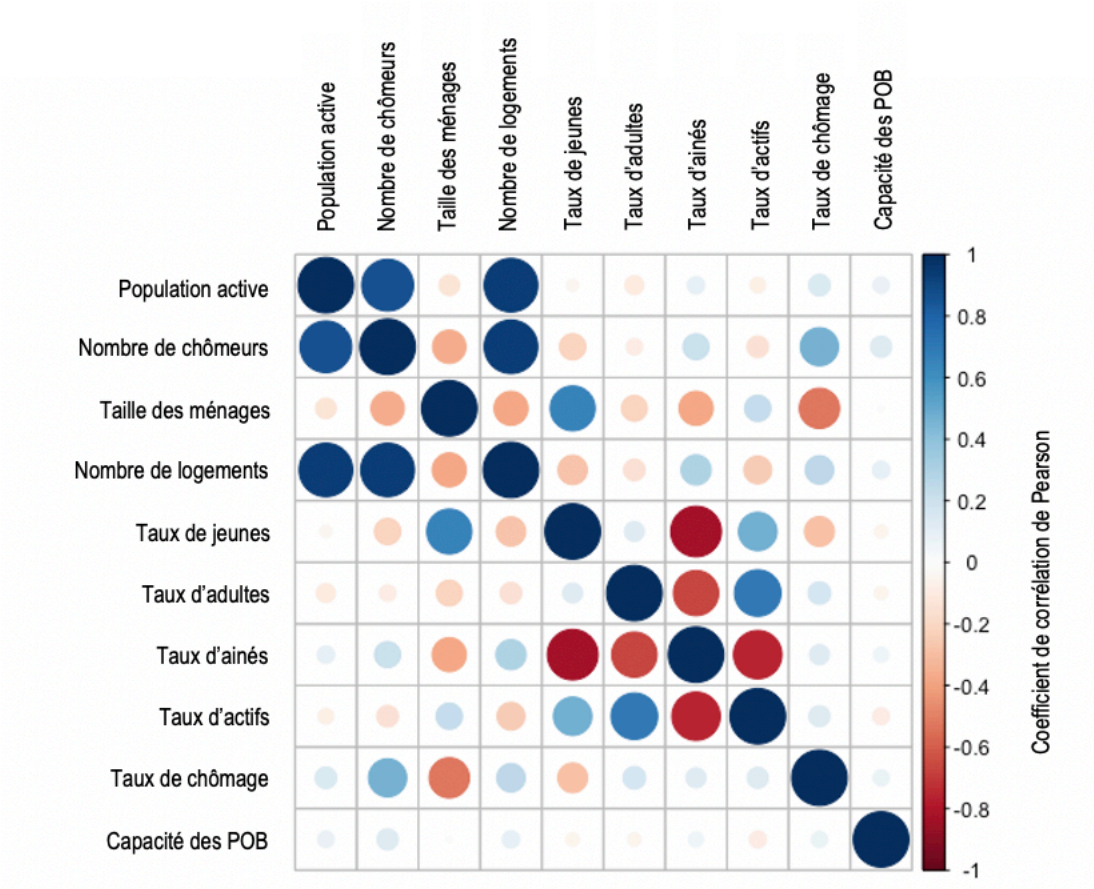


Figure 7-30 Coefficient de corrélation linéaire des variables explicatives considérées (Pearson)

De plus, la variable dépendante doit être normale. Une transformation logarithmique est donc effectuée pour vérifier cette hypothèse. La Figure 7-31 donne la distribution obtenue.

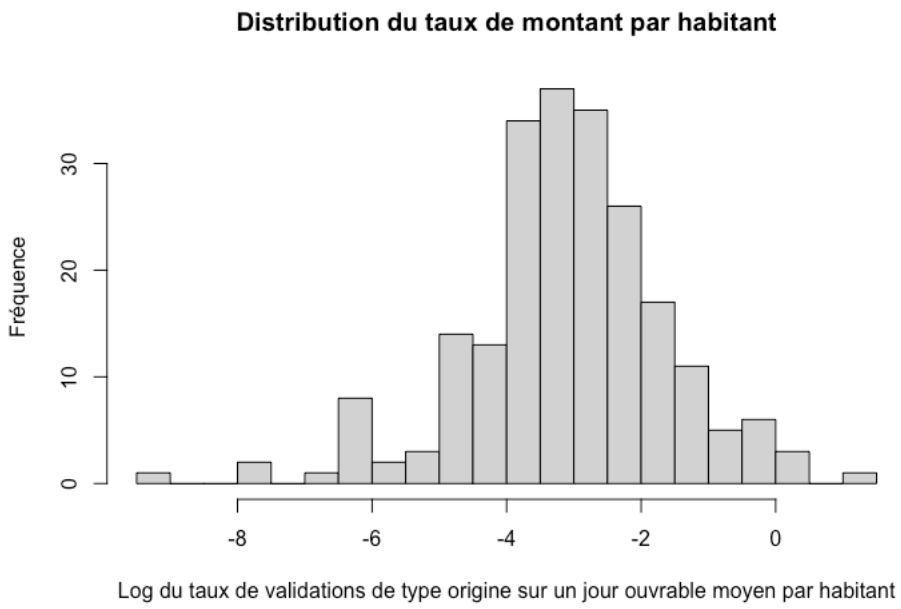


Figure 7-31 Distribution du taux de montants par habitant

Les résultats sur les corrélations permettent de tester plusieurs modèles différents :

- M1 : Log (y) ~ taille_menage + tx_adultes + capacite_POB
- M2 : Log (y) ~ taille_menage + tx_actifs + capacite_POB
- M3 : Log(y) ~ tx_jeunes + tx_chomage + capacite_POB
- M4 : Log(y) ~ tx_aines + tx_adultes + tx_chomage + capacite_POB
- M5 : Log (y) ~ taille_menage + tx_adultes + capacite_POB + nb_travail + nb_ecoles
- M6 : Log (y) ~ taille_menage + tx_actifs + capacite_POB + nb_loisirs + nb_ecoles
- M7 : Log(y) ~ tx_jeunes + tx_chomage + capacite_POB + nb_loisirs + nb_ecoles
- M8 : Log(y) ~ tx_aines + tx_adultes + tx_chomage + capacite_POB +nb_travail + nb_loisirs

Avec y la variable dépendante représentée par le taux de montants par zone sur un jour ouvrable moyen de mai 2016. Les variables indépendantes sont la taille des ménages, le taux de jeunes, le

taux d'adultes, le taux d'aînées, le taux d'actifs, le taux de chômage, et la capacité des parcs-o-bus. Ces variables sont réparties dans les différents modèles comme statué ci-dessus.

Le tableau suivant reflète la performance des modèles testés ainsi que la significativité des variables explicatives du modèle.

Tableau 7-4 Résumé de la significativité des variables pour chaque modèle

Modèles	R^2	Taille ménages	Taux d'adultes	Taux de jeunes	Taux d'ainés	Taux de chômage	Taux d'actifs	Capacité des parc-o-bus	Nombre d'écoles	Nombre de lieux de travail	Nombre de lieux de loisir	Ordonnée à l'origine
M1	0.15	***	NS					***				NS
M2	0.15	***					NS	***				NS
M3	0.10			.		*		***				***
M4	0.09		NS		.	*		***				**
M5	0.18	***						***	NS	*		*
M6	0.13			NS		*		**	NS		**	***
M7	0.13				NS	*		**	NS		**	***
M8	0.19	***						***		.	.	*

Les seuils de significativité dépendent des p-value des variables sur hypothèse nul déterminant si la variable à un impact significatif dans le modèle.

Tableau 7-5 Légende tableau 7-4

Symboles	P-value inférieur à
***	0.0001
**	0.001
*	0.01
.	0.1
NS	Supérieur à 0.1

Comme précisé précédemment, l'objectif est d'étudier la significativité des variables dans les différents modèles. L'impact du taux d'adultes ou d'actifs n'est jamais significatif. Cette variable ne sera donc pas utilisée en priorité dans les modèles. En revanche, la taille moyenne des ménages

s'avère être un critère de différenciation important. Le taux de chômage apparaît également comme significatif.

Néanmoins, la découverte d'un nouvel ensemble de données comprenant des projections de la composition de la population dans les années à venir ne permet pas d'établir une typologie par l'intermédiaire de la taille des ménages ou du taux de chômage. Ces résultats sont donc donnés comme des recommandations pour de futurs travaux. La disponibilité actuelle des données ne permet pas de prendre en compte ces paramètres.

7.5 Typologie de génération de déplacements

Statistique Canada réalise des modèles de prévision de population. Ces prévisions réalisées à l'échelle locale comprennent des données sur la composition de la population. Cet ensemble de données peut permettre de réaliser une typologie basée sur la composition de la population. De la même manière que les modèles basés sur la génération de taux de montants, une typologie est développée et afin d'obtenir des taux de montants. Cette fois ci, le taux est basé sur la population est non les logements comme précédemment. Un premier modèle vient donc prévoir les montants dans le réseau de TC.

Les données disponibles pour ce projet permettent pour la plupart d'être intégrées dans ce modèle : données sur les ménages et sur la composition de la population. Dans un premier temps, le modèle est construit.

Grâce aux résultats issus des modèles spatiaux, il est possible de tester plusieurs typologies de génération de taux de montants. Deux typologies sont étudiées et elles prennent en compte les données suivantes :

- Typologie 1 : densité, taux de jeunes, taux d'ainés et les capacités des parc-o-bus
- Typologie 2 : densité, taux de jeunes, taille des ménages et les capacités des parc-o-bus

Une fois les typologies identifiées, il est possible d'étudier la distribution des taux de génération de montant en pointe du matin pour chacun des types.

Les compositions des différentes typologies sont les suivantes. Elles sont décrites par la valeur des centroïdes de chaque segment :

- Pour la typologie 1 :

Tableau 7-6 Centroïdes typologie 1

Segments	Taux jeunes	Taux ainés	Densité (hab/m^2)
1	0.3791	0.0525	0.00177
2	0.2448	0.2174	0.00198
3	0.3479	0.0901	0.00113
4	0.2902	0.1366	0.00140
5	0.1927	0.3791	0.00225
6	Présence de parc-o-bus		

- Typologie 2 :

Tableau 7-7 Centroïdes typologie 2

Segments	Taille ménage	Taux jeunes	Densité (hab/m^2)
1	1.8779	0.2715	0.00189
2	2.8148	0.3528	0.00132
3	2.4524	0.3189	0.00132
4	1.5374	0.2494	0.00177
5	2.1474	0.2844	0.00167
6	Présence de parc-o-bus		

Les distributions pour chacun des segments figurent sur la page suivante.

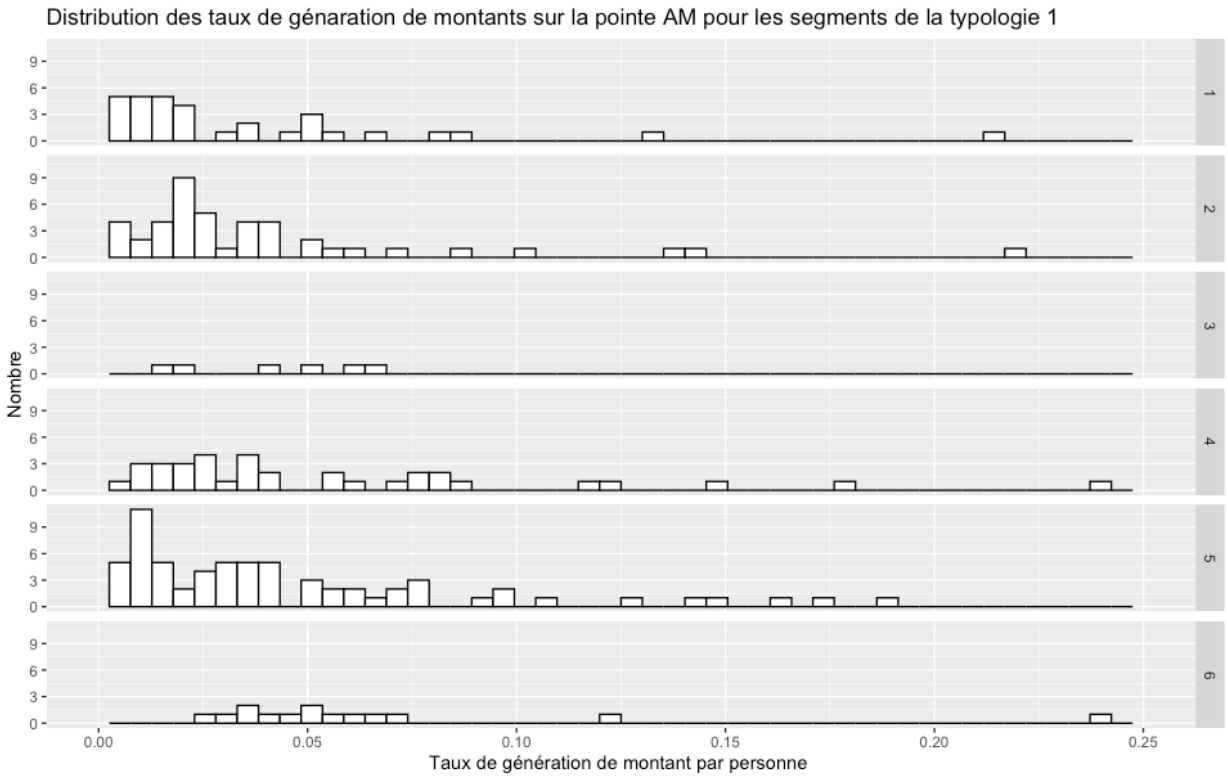


Figure 7-32 Distribution des taux de montants pour la typologie 1

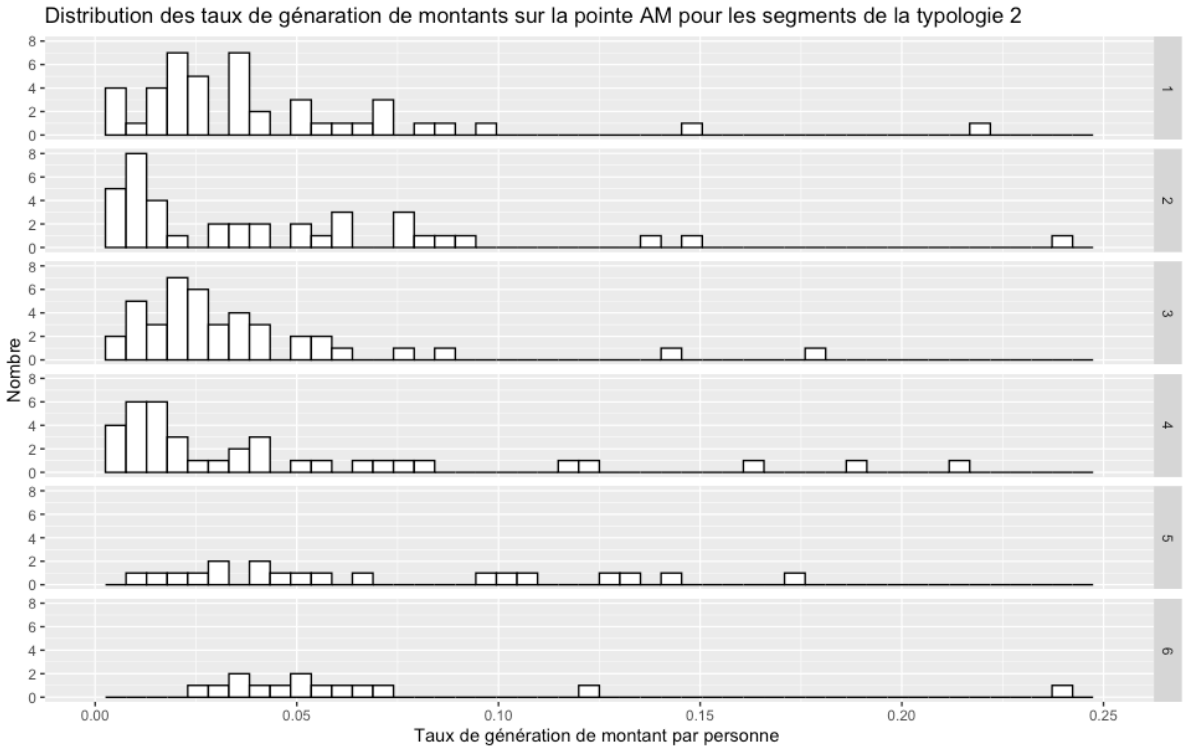


Figure 7-33 Distribution des taux de montants pour la typologie 2

Les taux moyens de génération de montants sont les suivants pour chacune des deux typologies considérées :

Tableau 7-8 Taux de génération de montants moyens pour les deux typologies

Typologies/segments	S1	S2	S3	S4	S5	S6
T1	0.0355	0.0383	0.0415	0.0424	0.0614	0.1782
T2	0.0378	0.0394	0.0413	0.0427	0.0637	0.1782

Les projections de population n'ont pas été disponibles durant le temps imparti à la réalisation de ce mémoire. Des projections ne peuvent donc pas être réalisées mais il est possible d'évaluer la performance de la modélisation sur l'année 2016 qui a permis la calibration des taux de génération de montants. Les erreurs issues de cette modélisation sont disponibles sur la figure suivante pour la typologie 1.

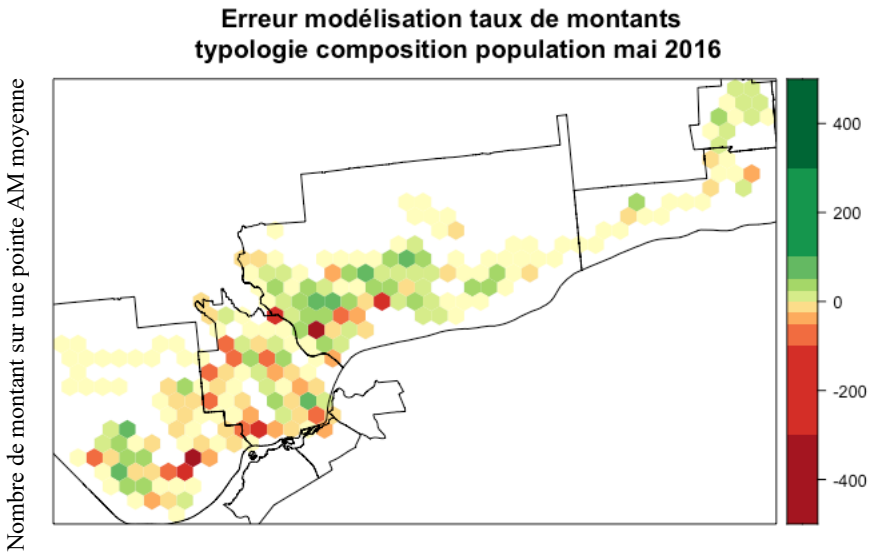


Figure 7-34 Erreurs de la modélisation pour la typologie 1

Les résultats donnés lors de la calibration du modèle sont encourageants. Comparativement à la typologie de logements et de parc-o-bus, l'erreur sur le set de calibration est plus faible sur la typologie de composition de la population. Il est donc possible d'espérer une meilleure précision de ce modèle lors de la projection. Différencier les zones en fonction de la composition semble

permettre de créer des segments ayant des comportements de mobilité comparables. Sans les données de population actualisées, il n'est pas possible d'effectuer des prévisions. Ces résultats sont donc des recommandations pour de futurs travaux. Dès que ces données seront disponibles, il sera possible de valider le modèle et par la suite d'anticiper les variations de la fréquentation de l'usage du réseau de transport en commun.

Les modèles de prévision réalisés dans ce chapitre poussent à considérer les modèles de génération de taux de montants par l'intermédiaire d'une typologie comme l'approche la plus pertinente. La disponibilité des données ne permet pas de mener à bout l'étude de la composition de la population mais les résultats qu'il est possible de générer sont performants. La typologie de logements permet néanmoins d'obtenir des prévisions fiables. L'utilisation de données de composition de la population restent encore à être pleinement testées mais elles pourraient permettre de prendre en compte d'autres facteurs explicatifs de la demande. Utiliser les deux typologies en parallèle peut être une manière de tenir compte des différentes transformations que peut subir le territoire dans le futur.

Les modèles de régression linéaires permettent de prendre en compte une multitude de facteurs différents mais les résultats ne sont pas satisfaisants. L'approche des typologies est plus pertinente et doit être favorisée pour la génération de prévision.

CHAPITRE 8 CONCLUSION (ET RECOMMANDATIONS)

8.1 Synthèse de la recherche

Face à un besoin évoqué par la Société des transports de l'Outaouais, l'objectif de ce mémoire était d'identifier un ou des modèles de prévision permettant d'estimer les évolutions de la fréquentation du réseau de bus à moyen terme. Sans un modèle de prévision de l'achalandage, il est difficile d'identifier et principalement de quantifier l'effet des facteurs influençant la demande. Aucune décision stratégique ne peut donc être prise avec certitude. La problématique majeure concerne l'anticipation d'acquisition de ressources matérielles et humaines. Plusieurs modèles de prévision longitudinale de la fréquentation du réseau de transport sont donc mis en place dans ce mémoire.

La revue de la littérature est la première étape de tout travail de recherche, car elle permet d'acquérir des connaissances supplémentaires sur le sujet d'étude, mais principalement de situer les travaux dans le contexte pour identifier les contributions possibles. Cette partie a abouti sur l'identification de modèles potentiellement pertinents pour remplir les objectifs de ce mémoire. Elle a également identifié un manque de travaux dans la littérature comparant plusieurs modèles sur un même territoire d'étude. La méthodologie est justifiée par cette lacune et a pour but de comparer différentes structures de modèles afin d'identifier les forces et faiblesses de chacun. Cette comparaison aboutit à l'identification du modèle le plus pertinent afin de remplir les objectifs fixés.

L'enjeu étant de déterminer un modèle adapté au territoire de la STO, une analyse rétrospective de la demande et de divers indicateurs permettant de décrire la région d'étude est effectuée. Les principales sources de données sont constituées des données de carte à puce du réseau de la STO, du recensement et de l'étude OD sur la région de Gatineau. Malgré l'identification d'un manque de données longitudinales, ces indicateurs ont permis d'identifier des facteurs d'influence de la demande. Les données de permis de construire ainsi que celles concernant les stationnements incitatifs ont été identifiées comme les données les plus utiles pour le développement de modèles de prévision. L'analyse rétrospective met en valeurs l'influence de multiples facteurs explicatifs comme le taux d'emploi, mais le manque de données dans le temps empêche leur utilisation dans des modèles longitudinaux.

Les résultats de l'analyse rétrospective ont démontré le besoin d'effectuer deux travaux préliminaires majeurs. Tout d'abord l'actualisation des données de logements grâce au registre des permis de construire de la ville de Gatineau. C'est un paramètre important dans la mise en place des modèles de prévision, car c'est une donnée précise spatialement et dans le temps. Elle permet donc de capturer des modifications à l'échelle locale. Pour que la fréquentation à cette échelle locale puisse être capturée, le besoin de créer des zones d'agrégation de petite taille a été constaté. Un maillage pertinent pour l'étude a été développé afin d'agréger toutes les données utilisées. Les prévisions à ce niveau d'agrégation permettent d'obtenir une image de l'utilisation du réseau de bus à l'échelle locale.

Les parties d'expérimentation sont donc l'étape qui suit. Des modèles longitudinaux de prévision de l'achalandage sont calibrés à trois échelles d'agrégation : le réseau global, les secteurs municipaux et les zones d'agrégation précises déterminées antérieurement. Cette structure permet d'identifier quels paramètres permettent d'expliquer la demande aux différentes échelles. Au niveau du réseau global, les variables socio-économiques sont pertinentes pour obtenir une estimation de la tendance décrite par la fréquentation du réseau. À l'échelle des zones hexagonales d'agrégation, la demande est plus sensible à des facteurs locaux. Le nombre de stationnements incitatifs ou l'évolution du nombre de logements dans chaque zone permettent de capturer les évolutions à l'échelle locale. Mais, les variables socio-économiques ne le permettent pas à l'échelle d'un quartier.

Tous les modèles calibrés sont résumés dans le Tableau 8-1. La performance est synthétisée par un indicateur subjectif : Bon (***), moyen (**) et non adapté (*). Un indicateur d'erreur n'est pas donné volontairement, car les modèles n'ont pas tous les mêmes horizons de prévision donc l'erreur sur la variable dépendante n'est pas calculée pour la même durée. De plus, certains modèles ont une erreur très faible, mais le pouvoir explicatif ne permet pas de tenir compte des évolutions du territoires. Ainsi, il ne peut pas être considéré comme performant pour effectuer des prévisions en fonction de différents scénarios. Les différents niveaux d'agrégation sont spécifiés pour chaque modèle. Il est ainsi possible de constater l'évolution de la performance des différents modèles en fonction des niveaux d'agrégation des données. Les variables explicatives utilisées dans les

modèles sont également spécifiées pour souligner les dynamiques influençant la demande mise en œuvre pour fournir des prévisions.

Tableau 8-1 Synthèse de la performance des modèles

Modèles	Niveau d'agrégation	Variables explicatives				Horizon de plannification		Performance
		Logements actualisés	Caractéristiques socio-économiques	Stationnements inscitatifs	Historique de la fréquentation TC	Court terme	Moyen terme	
ARIMA	Global				•	•		***
	Secteurs municipaux				•	•		**
	Zones hexagonales				•	•		*
NNAR	Global				•	•		**
Régression linéaire avec correction de l'erreur ARIMA	Global	•			•	•		*
	Secteurs municipaux	•			•	•		*
	Zones hexagonales	•			•	•		*
Régression linéaire simple	Global		•				•	*
Régression linéaire sur séries temporelles	Global		•		•		•	**
	Zones hexagonales	•		•	•		•	**
Modèles multinivaux	Secteurs municipaux		•			•		*
Typologie de génération de taux de montants	Zones hexagonales	•		•			•	***

Il est intéressant de noter que des modèles comme ARIMA perdent en performance plus le niveau d'agrégation des données est fin. La demande est sensible à des perturbations locales lorsqu'elle est étudiée de manière précise spatialement. Son caractère autorégressif n'est plus suffisant pour expliquer les variations.

Pour expliquer la demande au niveau des zones hexagonales, une structure se dessine comme la plus pertinente : la typologie de génération de taux de montants. Plusieurs typologies sont réalisées pour prendre en compte différents facteurs d'influence. Le modèle fournit des résultats satisfaisants aux vues des enjeux établis par ce projet.

8.2 Contributions

La revue de la littérature a montré une lacune concernant la comparaison de différents modèles sur un même territoire et sur les mêmes bases expérimentales. Ce mémoire a pris comme ligne conductrice le fait de comparer une multitude de modèles déjà existants dans la littérature en les appliquant au même territoire d'étude. Les variables ne sont pas toujours exactement les mêmes

pour chaque modèle, mais la dynamique régissant le territoire reste la même. La pointe du matin est dans la majeure partie des cas la période que les modèles tentent de prévoir dans ce mémoire. Les horizons de prévision sont différents, mais les résultats sur les modèles restent comparables. Il est donc possible de comparer les forces et les faiblesses de chaque modèle passé en revue et de les comparer aux autres. Une telle démarche permet d'identifier le modèle qui est le plus adapté au territoire d'étude. Les résultats sont propres au territoire, car il a été montré par la revue de la littérature qu'aucun modèle ne satisfait les caractéristiques de tous les territoires. Mais la démarche menant à l'identification d'un ou plusieurs modèles adaptés au territoire est une contribution. La comparaison des différents modèles permet de maximiser la performance d'un modèle final au regard des contraintes présentes sur le territoire : dynamiques de la demande et données disponibles.

La détermination du maillage menant à l'agrégation précise des données est également une contribution de ce mémoire. Le fait que la méthode nécessite presque uniquement les données de carte à puce et que le maillage ne soit pas influencé par les caractéristiques géographiques du territoire (maillage régulier) fait que la détermination d'un niveau d'agrégation similaire peut être effectuée sur une multitude d'autres territoires. La méthode est réplicable pour d'autres territoires et permet de s'affranchir de l'effet de redistribution de la demande entre des arrêts voisins lorsque le réseau de TC est en constante évolution. L'étude des distances d'accès au réseau de transport permet d'identifier une fenêtre de taille du maillage à étudier et l'analyse du coefficient d'autocorrélation local (LISA) en fonction de différentes combinaisons de taille de maillage et de position permet d'identifier des valeurs optimales. Cette méthodologie a prouvé fournir des résultats pertinents en s'affranchissant au maximum de l'effet de redistribution des montants entre les arrêts, mais tout en conservant une précision spatiale permettant d'étudier la fréquentation des TC à l'échelle locale.

D'autres contributions telles que l'étude ayant mené à l'actualisation des données de logements sont également importantes, mais les deux principales sont celles mentionnées précédemment.

8.3 Limites

Deux types de limites sont à distinguer dans l'étude qui a été menée : les limites concernant les données et les limites méthodologiques.

Dans un premier temps, les limites concernant les données sont relatives aux données utilisées lors de la création des modèles de prévision. Avoir un ensemble de données ne comportant pas d'imperfection est chose rare. La majorité des ensembles de données utilisés dans ce mémoire comportent des problèmes.

Tout d'abord, les données issues des comptages de cartes à puce posent plusieurs problèmes. Sur l'ensemble des comptages effectués, un certain nombre de validations de carte à puce ne sont pas localisées dans l'espace. Cela ne pose pas de problème lorsqu'il s'agit d'effectuer une prévision sur le réseau global, mais lorsque les prévisions sont spatialisées, un certain nombre de montants n'est pas pris en compte. Cela représente néanmoins une part négligeable aux vues des erreurs constatées sur les modèles.

L'ensemble des données socio-économiques sont issues d'estimations lorsque leur valeur est mensuelle. La population ou encore le taux d'emploi n'est pas relevé mensuellement, mais une valeur est estimée chaque mois par l'Institut Statistique du Québec. La majorité de ces données sont issues du recensement ou de l'enquête OD. Pour chaque variable, une valeur est disponible tous les cinq ans le reste sont des données actualisées. La même chose s'applique aux données de logements utilisées, leur actualisation étant réalisée via le registre des permis de construire de Gatineau. Les hypothèses sont citées dans la partie dédiée à la création de cette donnée, mais des limites en découlent.

Les limitations présentes dans ce mémoire ne concernent pas uniquement les données disponibles pour la réalisation des modèles. Elles émanent également de la méthodologie utilisée.

Premièrement, une limite concernant les données, mais qui est déclenchée par un choix méthodologique : la création de zones d'agrégation hexagonales. Pour réussir à agréger certaines données sur ce maillage, telles que les données du recensement, il faut effectuer un transfert de zone d'agrégation entre les îlots de diffusion (la zone d'agrégation la plus précise issue du recensement) et les zones hexagonales. Pour effectuer ce changement de zones d'agrégation, il a fallu faire l'hypothèse que chaque variable a une répartition constante sur les îlots de diffusion pour transférer la valeur sur les zones hexagonales. Les îlots de diffusion sont petits dans les zones denses par rapport aux zones d'agrégation déterminées, mais une erreur est tout de même introduite à ce niveau.

La comparaison de modèle en elle-même est également une limite du mémoire. Le choix des modèles à tester a été effectué par rapport à certains critères tels que la faisabilité ou encore la capacité de calibrer un tel modèle. Ces choix ont une dimension subjective et ne permettent pas de savoir si le modèle le plus adapté pour le territoire de Gatineau se trouve parmi ceux testés.

Enfin, la comparaison des modèles ne peut pas toujours se baser sur les critères d'erreur qui sont souvent utilisés pour juger de la pertinence d'un modèle tout au long du mémoire. Pour certains modèles, la variable dépendante n'est pas la même, car il n'est pas toujours pertinent de prédire le nombre de montants en pointe AM aux vues de variables utilisées. Le modèle multiniveau ou encore le modèle global se basant sur des variables socio-économiques sont des exemples. De plus, les horizons de prévision de tous les modèles ne sont pas les mêmes. Pour les modèles purement autorégressifs, il n'est pas pertinent d'aller chercher l'erreur moyenne à un horizon de cinq ans. Ces modèles ne sont pas adaptés pour cet horizon de temps. C'est pour cette raison que dans la synthèse des modèles, le critère de performance est subjectif. Cela permet également d'intégrer la dimension explicative du modèle. Comparer les multiples expérimentations effectuées dans ce projet reste tout de même pertinent surtout quand le but est d'identifier les forces et faiblesses des modèles pour construire un modèle final plus performant.

8.4 Perspectives

Les perspectives de ce mémoire se distinguent en deux types. Le premier est constitué de pistes pouvant être exploitées pour résoudre certaines limites constatées lors de cette étude. Tous les problèmes constatés dans ce mémoire ne peuvent pas être résolus, mais certains peuvent voir leur effet atténué. Le deuxième type concentre l'ensemble des travaux futurs pouvant être réalisés pour enrichir ce projet.

Les limites concernant les données ne peuvent guère être résolues. Toute donnée vient avec ses erreurs. Mais il est intéressant de discuter des pistes d'amélioration pour la méthodologie.

L'objectif principal du mémoire est de comparer des modèles longitudinaux de prévision de l'usage du réseau de la STO. Plusieurs modèles sont calibrés pour tenter de créer ces prévisions, mais la comparaison n'est pas toujours simple. Les modèles ont des horizons de prévision différents. Certains impliquent de prévoir la demande sur un jour ouvrable et d'autres sur la pointe du matin aux vues des variables utilisées. Comparer ces modèles peut être une tâche difficile, car l'utilisation

uniquement des indicateurs d'erreurs ou des résidus eux-mêmes ne permet pas toujours de rendre compte de la différence entre deux modèles. Une méthode de comparaison systématique aurait pu être développée afin de pouvoir mettre en perspective tous les modèles entre eux et cela de manière réplicable. Les expérimentations effectuées dans ce mémoire permettent la comparaison de la plupart des modèles entre eux, mais il y a un manque de réplicabilité. L'utilisation des indicateurs d'erreurs entre deux modèles dont les horizons sont les mêmes est une méthode adaptée. Mais lorsque certains paramètres commencent à diverger entre deux modèles, il est alors plus difficile de les comparer sans une méthode déterminée au préalable. Néanmoins, l'étude des résidus d'un ou plusieurs modèles permet tout de même de juger de leur fiabilité et d'obtenir des pistes d'amélioration pour la suite du projet.

Face à la multitude de modèles différents dans la littérature, il est impossible de les tester tous. L'approche de modélisation adoptée sur ce projet reste relativement simple. Des modèles plus complexes sont capables de prendre en compte plus de facteurs d'influence. Néanmoins, elle permet de se rendre compte des facteurs influençant la fréquentation du réseau et de quantifier à un terme d'erreur près leur effet sur l'achalandage. Les données utilisables pour un modèle de prévision longitudinale ne sont pas présentes en grand nombre donc l'enjeu était de maximiser les performances des modèles avec ces données. Il pourrait par la suite être intéressant de tenter d'obtenir plus de données permettant de décrire les évolutions des habitudes de mobilité. Cela pourrait être réalisé par l'intermédiaire de sondage ou de futures enquêtes.

La crise sanitaire du COVID-19 qui a eu lieu pendant la réalisation de cette étude n'est absolument pas considérée dans les prévisions. Elles s'effectuent dans l'hypothèse où la tendance antérieure se conserve. Cela est vraisemblablement une source d'erreur très importante, car l'effet de cette crise a été très prononcé sur l'utilisation des transports en commun. Il faut nécessairement par la suite ajuster les modèles développés pour prendre en compte cet évènement et ainsi ne pas surestimer la demande sur les années suivant 2020.

Enfin, le dernier modèle étudié fait appel à la création d'une typologie de population afin de déterminer des taux de génération de montants. Pour l'instant, ce modèle permet uniquement de modéliser la fréquentation de réseaux de TC, car les prévisions de la population pour les années à venir n'ont pas été disponibles durant la durée du mémoire. Il serait donc intéressant de finaliser ces projections, car le modèle semble prometteur. La modélisation de la pointe du soir via

l'utilisation d'un modèle similaire en utilisant des données sur les lieux d'emplois est également une piste de futurs travaux afin d'avoir une image plus complète d'un jour ouvrable.

RÉFÉRENCES

Abadi, A., Rajabioun, T., & Ioannou, P. A. (2015). Traffic Flow Prediction for Road Transportation Networks With Limited Traffic Data. *Transactions on Intelligent Transportation Systems*.

Arentze, T., Hofman, F., van Mourik, H., & Timmermans, H. (2000). ALBATROSS: Multiagent, Rule-Based Model of Activity Pattern Decisions. *Transportation Research Record*.

Baek, J., & Sohn, K. (2017). Deep-Learning Architectures to Forecast Bus Ridership at the Stop and Stop-To-Stop Levels for Dense and Crowded Bus Networks. *Applied Artificial Intelligence*.

Bonnel, P. (2004). Prévoir la demande de transport.

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*.

Cervero, R. (2006). Alternative Approaches to Modeling the Travel-Demand Impacts of Smart Growth. *Journal of the American Planning Association*.

Cervero, R., Murakami, J., & Miller, M. A. (2009). Direct Ridership Model of Bus Rapid Transit in Los Angeles County. *UC Berkeley: Center for Future Urban Transport: A Volvo Center of Excellence*.

Cheng, L., Lai, X., Chen, X., Yang, S., De Vos, J., & Witlox, F. (2020). Applying an ensemble-based model to travel choice behavior in travel demand forecasting under uncertainties. *Transportation Letters*.

Chiang, W., Russell, R., & Urban, T. (2011). Forecasting ridership for a metropolitan transit authority. *Transportation Research Part A: Policy and Practice*.

Ding, C., Duan, J., Zhang, Y., Wu, X., & Yu., G. (2017). Using an ARIMA-GARCH modelling approach to improve subway short-term ridership forecasting accounting for dynamic volatility. *Transactions on Intelligent Transportation Systems*.

Ding, C., Wang, D., Ma, X., & Li, H. (2016). Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees. *Sustainability*.

Durning, M., & Townsend, C. (2015). Direct Ridership Model of Rail Rapid Transit Systems in Canada. *Transportation Research Record: Journal of the Transportation Research Board*.

Frei, C., & Mahmassani, H. S. (2013). Riding More Frequently: Estimating Disaggregate Ridership Elasticity for a Large Urban Bus Transit Network. *Transportation Research Record*.

Hafezi, M., Liu, L., & Millward, H. (2019). A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation 46*.

He, Y., Zhao, Y., & Tsui, K. L. (2020). An adapted geographically weighted LASSO (Ada-GWL) model for predicting subway ridership. *Transportation*.

Idris, A. O., Habib, K. M., & Shalaby, A. (2015). An investigation on the performances of mode shift models in transit ridership forecasting. *Transportation Research Part A: Policy and Practice*.

Ko, J., Kim, D., & Etezady, A. (2019). Determinants of Bus Rapid Transit Ridership: System-Level Analysis. *Journal of Urban Planning and Development*.

Krakutovski, Z. (2012). Demographic Model 'Age-Cohort' for Modelling of Urban Mobility in Long Term. *2nd International Conference on Road and Rail Infrastructure – CETRA 2012.* Dubrovnik , Croatia.

Lane, C., DiCarlantonio, M., & Usvyat, L. (2006). Sketch Models to Forecast Commuter and Light Rail Ridership: Update to TCRP Report 16. *Transportation Research Record*.

Lawson, C., Muro, A., & Krans, E. (2019). Forecasting bus ridership using a “Blended Approach”. *Transportation*.

Li, S., Yang, H., G. Zhang, Z. L., Xiong, Y., & Li, Y. (2019). What is the best catchment area of a metro station? A study based on station level ridership modeling. *5th International Conference on Transportation Information and Safety (ICTIS)*.

Li, S., Yang, H., Zhang, G., Ling, Z., Xiong, Y., & Y. Li. (2019). What is the best catchment area of a metro station? A study based on station level ridership modeling. *5th International Conference on Transportation Information and Safety (ICTIS).*

Li, X., Yue Liu, A., Gao, Z., & Liu, D. (2016). Decision Tree Based Station-Level Rail Transit Ridership Forecasting. *Journal of Urban Planning and Development* .

Linh, H., Adnan, M., Ectors, W., Kochan, B., Bellemens, T., & Tuan, V. (2019). Exploring the Transferability of FEATHERS – An Activity Based Travel Demand Model – For Ho Chi Minh City, Vietnam. *International Journal of Traffic and Transportation Management*.

Mucci, R. A., & Erhardt, G. D. (2018). Evaluating the Ability of Transit Direct Ridership Models to Forecast Medium-Term Ridership Changes: Evidence from San Francisco. *Transportation Research Record*.

Oort, N., Brands, T., & de Romph, E. (2015). Short-Term Prediction of Ridership on Public Transport with Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*.

Outwater, M. L., Castleberry, S., Shiftan, Y., Ben-Akiva, M., Shuang Zhou, Y., & Kuppam, A. (2003). Attitudinal Market Segmentation Approach to Mode Choice and Ridership Forecasting: Structural Equation Modeling. *Transportation Research Record*.

Pulugurtha, S. S., & Agurla, M. (2012). Assessment of Models to Estimate Bus-Stop Level Transit Ridership using Spatial Modeling Methods. *Journal of public transportation*.

STO. (2017). *Guide d'aménagement et d'accessibilité universelle.* Gatineau.

Sun, L.-S., Wang, S.-W., Yao, L.-Y., Rong, J., & Ma, J.-M. (2016). Estimation of transit ridership based on spatial analysis and precise land use data. *The International Journal of Transportation Research*.

Umlauf, T., Galicia, L. D., & Cheu, R. L. (2015). Ridership estimation procedure for a transit corridor with new bus rapid transit service. *Journal of Advanced Transportation*.

Wegener, M. (1982). Modeling Urban Decline: A Multilevel Economic-Demographic Model for the Dortmund Region. *International Regional Science Review*.

Zhang, D., & Wang, C. (2014). Transit ridership estimation with network Kriging: A case study of Second Avenue Subway, NYC. *Journal of Transport Geography*.

ANNEXE A DÉMONSTRATION DU TAUX DE CROISSANCE MOYEN

Démonstration taux de croissance moyen

Supposons une variable évoluant entre v_i et v_{i+N} respectivement les valeurs de cette variable à l'année i et i+N. On note t_{i+1},..., t_{i+n}, ... et t_{i+N} les taux de croissance (positif ou négatif) des années séparant i et N. t_n est le taux de croissance entre l'année n - 1 et n. n est un entier naturel évoluant entre 1 et N. On a donc la relation suivante :

$$v_{i+N} = v_i * t_{i+1} * t_{i+2} * ... * t_{i+n} * ... * t_{i+N} \tag{1}$$

Entre l'année i et i+N, il y a donc N taux de croissance. Un taux de croissance moyen peut être assimilé à la valeur si le taux était constant entre chaque années. On note donc t_m ce taux moyen. On a donc la relation suivante :

$$v_{i+N} = v_i * t_m * t_m * ... * t_m * ... * t_m \text{ N fois} \tag{2}$$

$$v_{i+N} = v_i * t_m^N \tag{3}$$

Pour obtenir le taux de croissance moyen, on a donc :

$$t_m = (\frac{v_{i+N}}{v_i})^{\frac{1}{N}} \tag{4}$$

t_m est le taux de croissance mais sous la forme d'un coefficient multiplicateur. Pour obtenir ce taux en pourcent on applique la formule suivante :

$$\%_{croissancemoyenne} = ((\frac{v_{i+N}}{v_i})^{\frac{1}{N}} - 1) * 100 \tag{5}$$

ANNEXE B TABLEAUX ANALYSE DE CORRÉLATION

Tableau 0-1 Analyse de corrélation méthode de la courbe de tendance

	Veh-km TC	*Pass-arrêts*	*Validations CAP*	*Véhicules/permis*	*Nombre travailleurs*	*Nombre de personnes*	*Nombre de personnes âgées (65+)*	*Nombre de véhicules*	*Durée moyenne activité travail pour travailleur*	*Durée moyenne activité travail pour tous les depl travail*	*Part modale TC*	*Prix de l'essence*	*Unités de logement*	*Densité d'unités*	*% unifamilales*	*km réseau routier*	*Nombre total de déplacements travail*	*Pourcentage travailleur*	*Pourcentage personnes âgées*	*Validations CAP par personne*
Veh-km TC	1,000																			
Pass-arrêts	0,998	1,000																		
Validations CAP	0,995	0,996	1,000																	
Véhicules/permis	0,857	0,855	0,858	1,000																
Nombre travailleurs	0,997	0,998	0,999	0,860	1,000															
Nombre de personnes	0,992	0,989	0,993	0,844	0,994	1,000														
Nombre de personnes âgées (65+)	0,981	0,981	0,982	0,872	0,984	0,973	1,000													
Nombre de véhicules	0,996	0,997	0,998	0,865	0,999	0,993	0,990	1,000												
Durée moyenne activité travail pour status 1	0,971	0,965	0,972	0,825	0,974	0,991	0,963	0,975	1,000											
Durée moyenne activité travail pour tous les depl travail	0,915	0,905	0,917	0,760	0,918	0,954	0,902	0,919	0,983	1,000										
Part modale TC	-0,371	-0,346	-0,371	-0,222	-0,372	-0,469	-0,323	-0,366	-0,558	-0,699	1,000									
Prix de l'essence	0,721	0,722	0,729	0,540	0,717	0,707	0,686	0,712	0,677	0,627	-0,221	1,000								
Unités de logement	0,991	0,992	0,994	0,825	0,994	0,986	0,970	0,991	0,960	0,900	-0,352	0,744	1,000							
Densité d'unités	0,991	0,992	0,994	0,825	0,994	0,986	0,970	0,991	0,960	0,900	-0,352	0,744	1,000	1,000						
% unifamilales	0,711	0,710	0,716	0,568	0,713	0,704	0,685	0,708	0,677	0,629	-0,227	0,595	0,754	0,754	1,000					
km réseau routier	-0,332	-0,332	-0,342	-0,144	-0,332	-0,320	-0,291	-0,323	-0,285	-0,248	0,039	-0,427	-0,371	-0,371	-0,230	1,000				
Nombre total de déplacements travail	0,984	0,982	0,985	0,862	0,987	0,987	0,995	0,992	0,985	0,940	-0,415	0,686	0,973	0,973	0,686	-0,289	1,000			
Pourcentage travailleur	-0,887	-0,870	-0,889	-0,755	-0,890	-0,927	-0,864	-0,888	-0,954	-0,974	0,700	-0,631	-0,882	-0,882	-0,651	0,291	-0,905	1,000		
Pourcentage personnes âgées	0,917	0,913	0,918	0,866	0,921	0,905	0,967	0,933	0,903	0,838	-0,251	0,638	0,907	0,907	0,660	-0,275	0,953	-0,834	1,000	
Validations CAP par personne	0,965	0,963	0,974	0,877	0,969	0,952	0,964	0,971	0,923	0,849	-0,256	0,736	0,972	0,972	0,735	-0,378	0,956	-0,859	0,948	1,000

Tableau 0-2 Analyse de corrélation méthode interpolation linéaire

	Veh-km TC	*Pass-arrêts*	*Validations CAP*	*Véhicules/permis*	*Nombre travailleurs*	*Nombre de personnes*	*Nombre de personnes âgées (65+)*	*Nombre de véhicules*	*Durée moyenne activité travail pour travailleurs*	*Durée moyenne activité travail pour tous les depl travail*	*Part modale TC*	*Prix de l'essence*	*Unités de logement*	*Densité d'unités*	*% unifamiliales*	*km réseau routier*	*Nombre total de déplacements travail*	*Pourcentage travailleur*	*Pourcentage personnes âgées*	*Validations CAP par personne*
Veh-km TC	1,000																			
Pass-arrêts	0,998	1,000																		
Validations CAP	0,995	0,996	1,000																	
Véhicules/permis	0,917	0,927	0,919	1,000																
Nombre travailleurs	0,997	0,998	0,999	0,921	1,000															
Nombre de personnes	0,987	0,983	0,988	0,906	0,990	1,000														
Nombre de personnes âgées (65+)	0,955	0,953	0,956	0,907	0,959	0,941	1,000													
Nombre de véhicules	0,995	0,995	0,996	0,924	0,998	0,987	0,974	1,000												
Durée moyenne activité travail pour travailleurs	0,941	0,931	0,943	0,870	0,945	0,977	0,928	0,949	1,000											
Durée moyenne activité travail pour tous les depl travail	0,821	0,804	0,823	0,751	0,826	0,894	0,805	0,829	0,963	1,000										
Part modale TC	-0,057	-0,028	-0,058	-0,014	-0,060	-0,200	0,000	-0,050	-0,351	-0,589	1,000									
Prix de l'essence	0,725	0,724	0,729	0,720	0,717	0,714	0,669	0,711	0,678	0,598	-0,072	1,000								
Unités de logement	0,987	0,987	0,991	0,883	0,989	0,973	0,934	0,984	0,913	0,784	-0,016	0,754	1,000							
Densité d'unités	0,987	0,987	0,991	0,883	0,989	0,973	0,934	0,984	0,913	0,784	-0,016	0,754	1,000	1,000						
% unifamilales	0,517	0,508	0,524	0,385	0,519	0,483	0,632	0,547	0,485	0,377	0,192	0,432	0,571	0,571	1,000					
km réseau routier	-0,246	-0,244	-0,257	-0,033	-0,244	-0,187	-0,238	-0,243	-0,106	0,020	-0,375	-0,250	-0,311	-0,311	-0,217	1,000				
Nombre total de déplacements travail	0,962	0,957	0,963	0,906	0,966	0,967	0,991	0,979	0,969	0,876	-0,130	0,679	0,937	0,937	0,593	-0,190	1,000			
Pourcentage travailleur	-0,835	-0,814	-0,836	-0,689	-0,838	-0,899	-0,804	-0,839	-0,951	-0,973	0,537	-0,613	-0,820	-0,820	-0,450	0,129	-0,871	1,000		
Pourcentage personnes âgées	0,820	0,812	0,820	0,755	0,824	0,797	0,944	0,856	0,812	0,690	0,087	0,564	0,804	0,804	0,754	-0,293	0,917	-0,713	1,000	
Validations CAP par personne	0,962	0,962	0,970	0,855	0,966	0,933	0,957	0,970	0,875	0,721	0,107	0,716	0,976	0,976	0,645	-0,395	0,940	-0,771	0,883	1,000

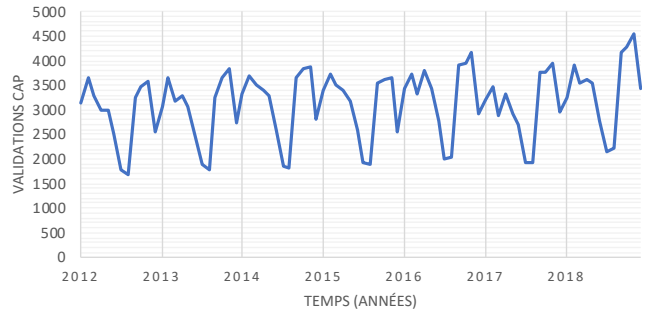
Tableau 0-3 Analyse de corrélation valeurs réelles

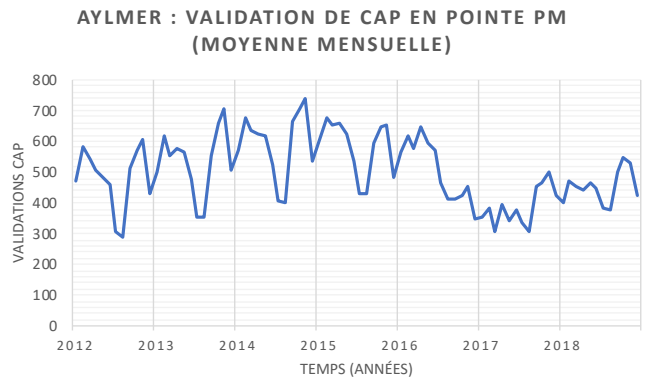
	Véhicules/permis	*Nombre travailleurs*	*Nombre de personnes*	*Nombre de personnes âgées (65+)*	*Nombre de véhicules*	*Durée moyenne activité travail pour travailleurs*	*Durée moyenne activité travail pour tous les depl travail*	*Part modale TC*	*Unités de logement*	*Densité d'unités*	*% unifamiliales*	*Nombre total de déplacements travail*	*Pourcentage travailleur*	*Pourcentage personnes âgées*
Véhicules/permis	1,000													
Nombre travailleurs	0,847	1,000												
Nombre de personnes	0,808	0,998	1,000											
Nombre de personnes âgées (65+)	1,000	0,847	0,808	1,000										
Nombre de véhicules	0,920	0,988	0,974	0,920	1,000									
Durée moyenne activité travail pour travailleurs	0,978	0,939	0,913	0,978	0,981	1,000								
Durée moyenne activité travail pour tous les depl travail	-0,660	-0,958	-0,976	-0,660	-0,901	-0,801	1,000							
Part modale TC	0,972	0,949	0,924	0,972	0,987	1,000	-0,818	1,000						
Unités de logement	0,803	0,997	1,000	0,803	0,972	0,909	-0,978	0,921	1,000					
Densité d'unités	0,803	0,997	1,000	0,803	0,972	0,909	-0,978	0,921	1,000	1,000				
% unifamilales	0,957	0,965	0,944	0,957	0,994	0,996	-0,849	0,998	0,941	0,941	1,000			
Nombre total de déplacements travail	1,000	0,847	0,807	1,000	0,920	0,978	-0,659	0,971	0,802	0,802	0,957	1,000		
Pourcentage travailleur	-0,596	-0,931	-0,954	-0,595	-0,863	-0,749	0,997	-0,768	-0,957	-0,957	-0,802	-0,594	1,000	
Pourcentage personnes âgées	0,980	0,724	0,674	0,980	0,823	0,917	-0,497	0,905	0,668	0,668	0,880	0,980	-0,423	1,000

ANNEXE C FICHES ANALYSES DESCRIPTIVES DES SECTEURS

AYLMER

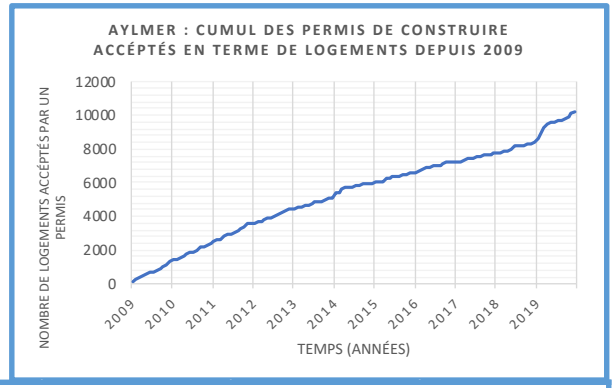
Demande en TC





Descriptif du secteur

- Secteur peu dense qui a récemment pris de décision de densification. Cette zone résidentiel fait preuve d'une utilisation moyenne des TC dont 64% des validations ont lieux pendant la pointe du matin.



Démographie

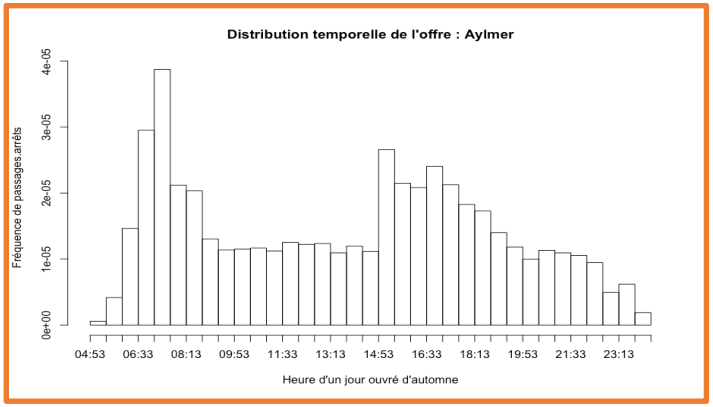
Années	2016	2011	2006	2001
Nombre de ménages privés	25065	20975	15365	13070
Population	64655	54630	41535	36085
Population active	35795 (55%)	31625 (58%)	24105 (58%)	20225 (56%)
Densité de population	6,2	4,71	4,1	7,34

Offre en TC

- Passages.arrets sur un jour ouvré moyen de novembre : 19320 pass.arrêts

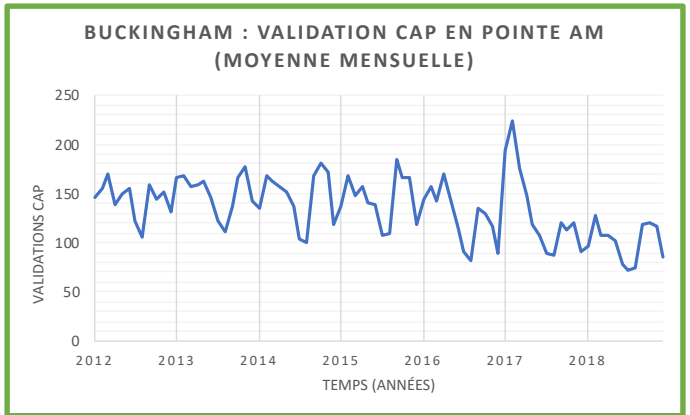
Taux de génération de déplacements TC (jour ouvré) :

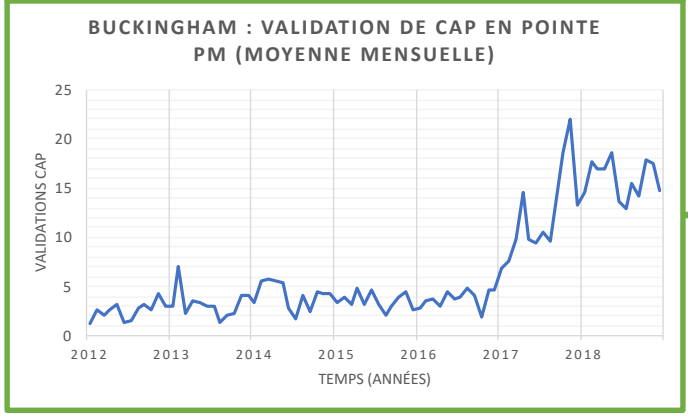
Par logement	0,167
Par personne	0,065



BUCKINGHAM

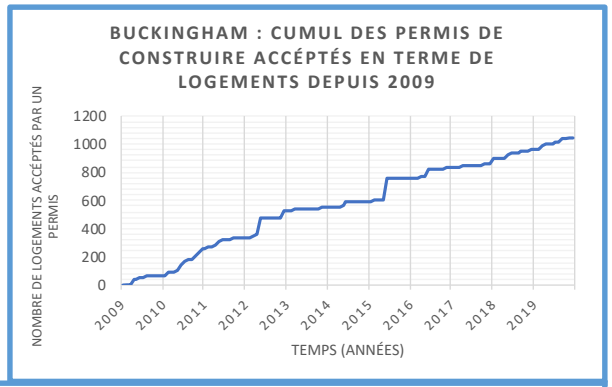
Demande en TC





Descriptif du secteur

- Plus petit secteur de la ville caractérisé par une forte utilisation de la voiture au dépend des TC. Secteur résidentiel expliqué par le déséquilibre de validations entre les pointes.



Démographie

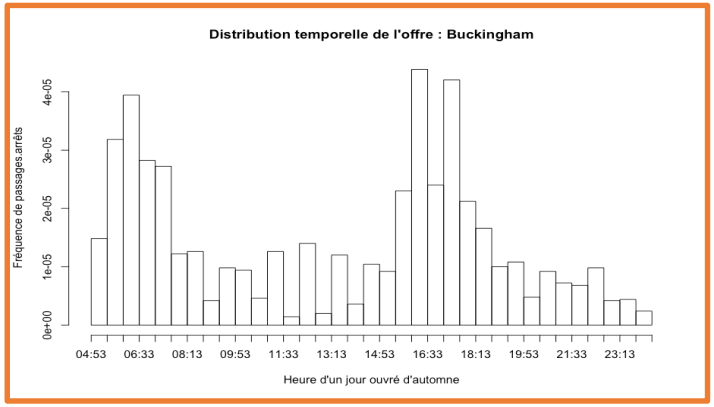
Années	2016	2011	2006	2001
Nombre de ménages privés	5390	5085	4620	4600
Population	12220	12050	11415	11665
Population active	5935 (49%)	6225 (52%)	6010 (53%)	5860 (50%)
Densité de population	7,93	7,51	7,67	8,04

Offre en TC

- Passages.arrets sur un jour ouvré moyen de novembre : 2532 pass.arrêts

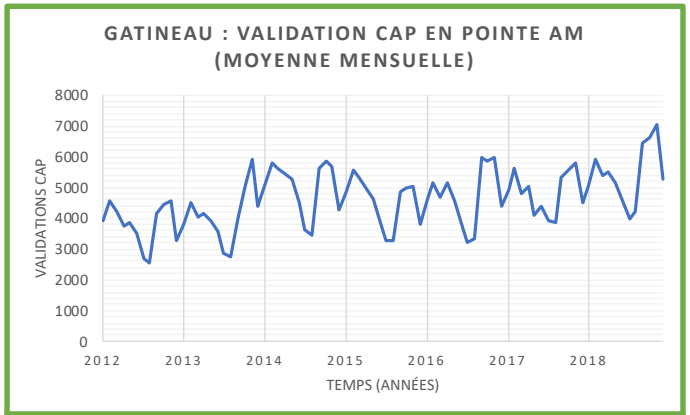
Taux de génération de déplacements TC (jour ouvré) :

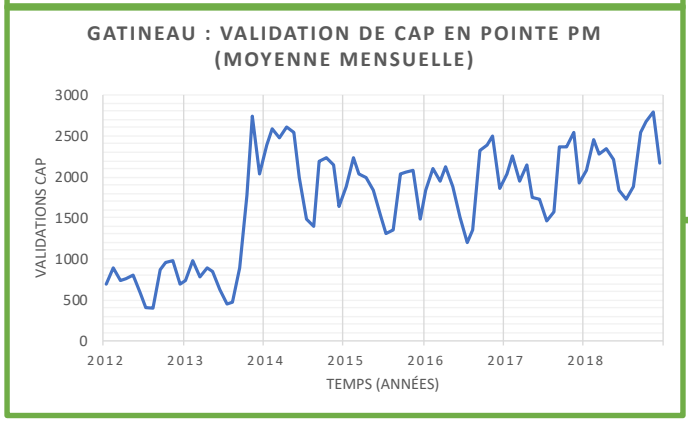
Par logement	0,055
Par personne	0,024



GATINEAU

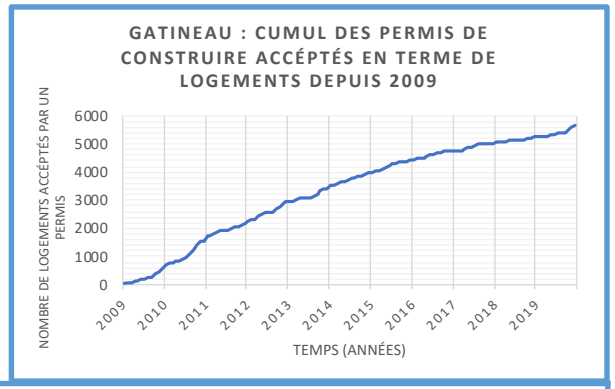
Demande en TC





Descriptif du secteur

- Secteur ancien dont l'aménagement du territoire arrive à maturité. L'utilisation des TC est moyenne et repartie sur toute la journée.



Démographie

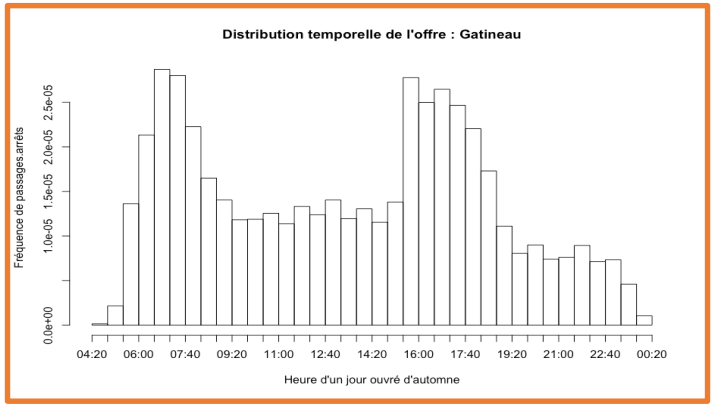
Années	2016	2011	2006	2001
Nombre de ménages privés	49470	48035	43315	39510
Population	117970	116780	108730	102895
Population active	63825 (54%)	65900 (56%)	62270 (57%)	58185 (57%)
Densité de population	7,95	7,4	7	8,03

Offre en TC

- Passages.arrets sur un jour ouvré moyen de novembre : 27043 pass.arrêts

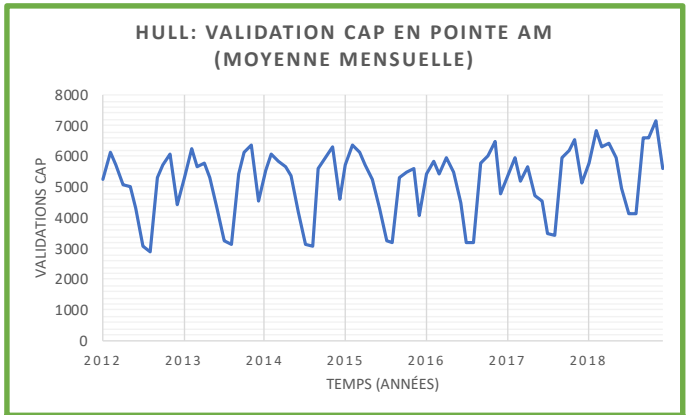
Taux de génération de déplacements TC (jour ouvré) :

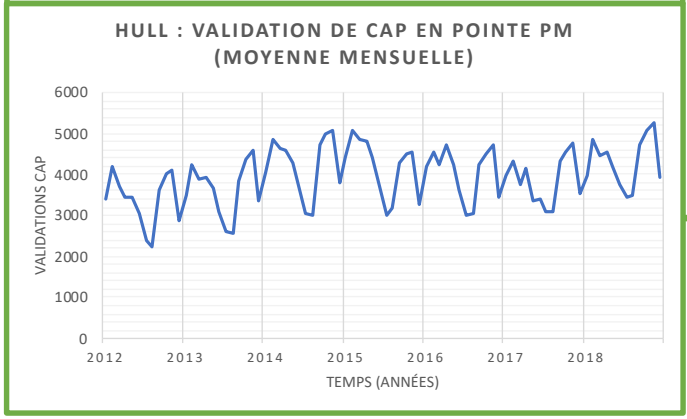
Par logement	0,283
Par personne	0,118



HULL

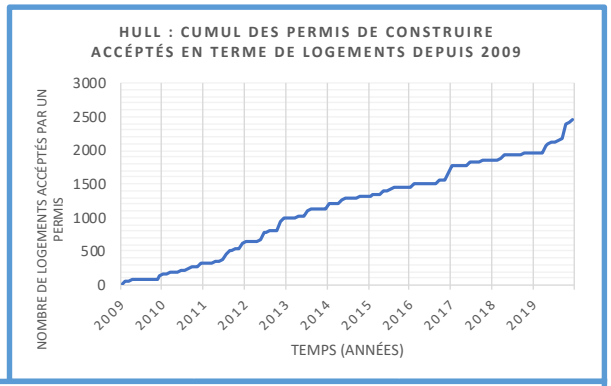
Demande en TC





Descriptif du secteur

- Secteur le plus dense et en pleine expansion. L'utilisation des TC est importante avec une offre répartie tout au long de la journée. Le déséquilibre entre les pointe est faible ce qui indique une présence d'emploie et de logements dans le secteur.



Démographie

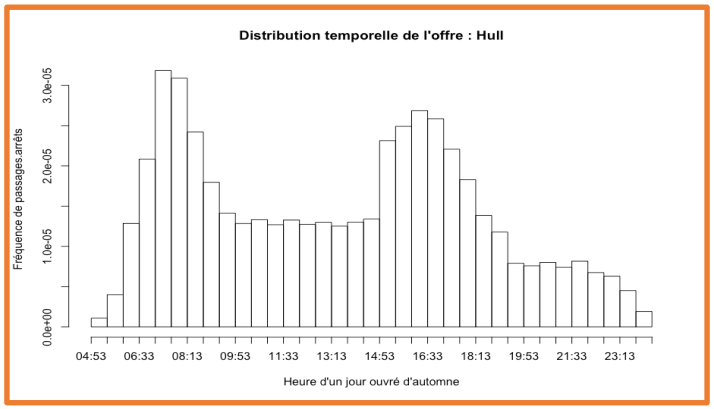
Années	2016	2011	2006	2001
Nombre de ménages privés	33420	33910	32820	31190
Population	68445	69490	69370	66245
Population active	37645 (55%)	40250 (58%)	39700 (57%)	37725 (57%)
Densité de population	18,89	18,86	18,01	18,61

Offre en TC

- Passages.arrets sur un jour ouvré moyen de novembre : 37228 pass.arrêts

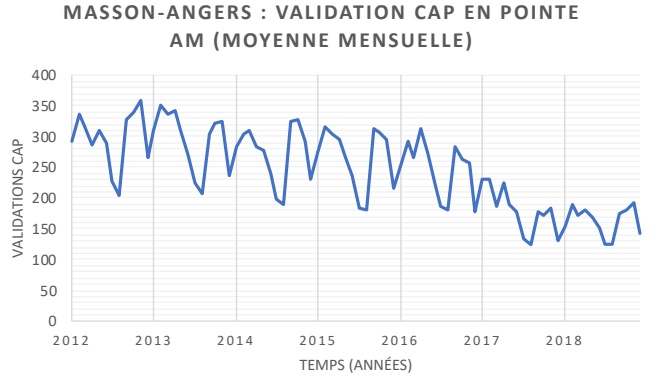
Taux de génération de déplacements TC (jour ouvré) :

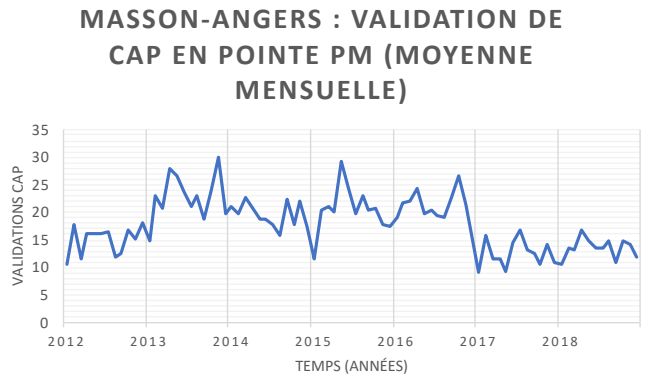
Par logement	0,609
Par personne	0,297



MASSON-ANGERS

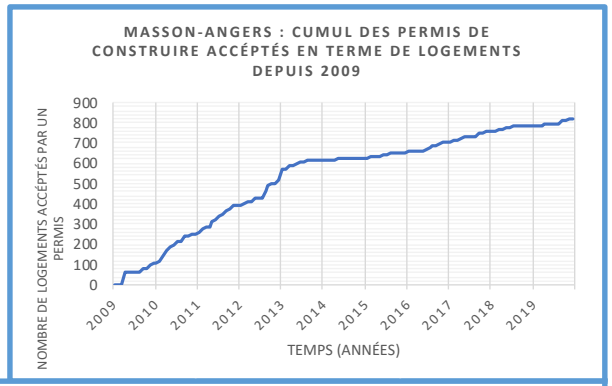
Demande en TC





Descriptif du secteur

- Secteur faiblement peuplé et ne prévoit pas de se densifier immédiatement. L'utilisation des TC est très faible et est en chute ces dernières années.

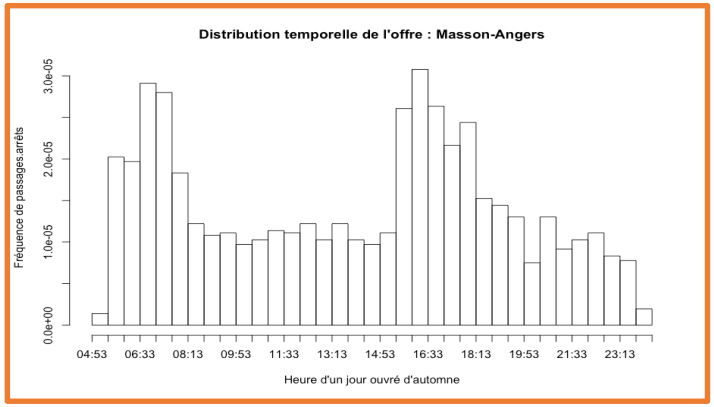


Démographie

Années	2016	2011	2006	2001
Nombre de ménages privés	5040	4750	4140	3560
Population	12945	12400	11070	9800
Population active	7315 (56%)	7070 (57%)	6185 (56%)	5190 (53%)
Densité de population	2,28	2,03	1,8	2,38

Offre en TC

- Passages.arrets sur un jour ouvré moyen de novembre : 1836 pass.arrêts

Taux de génération de déplacements TC (jour ouvré) :

Par logement	0,116
Par personne	0,045