POLYPUBLIE

Fusion des données provenant d'un système de paiement par cartes à puce, d'un système de compte à bord et d'horaire pour l'imputation d'arrêts d'embarquement en transport collectif

Félix Légaré

Document en libre accès dans PolyPublie
Open Access document in PolyPublie

Date	2014
Type	Mémoire ou thèse / Dissertation or Thesis Mémoire de maîtrise/Master's thesis
URL de PolyPublie **PolyPublie URL**	https://publications.polymtl.ca/1628/
Conditions d'utilisation **Terms of use**	© Tous droits réservés / All rights reserved
Référence **Citation**	Légaré, F. (2014). Fusion des données provenant d'un système de paiement par cartes à puce, d'un système de compte à bord et d'horaire pour l'imputation d'arrêts d'embarquement en transport collectif [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/1628/

Détails de la thèse
Thesis details

Directeurs de recherche **Advisors**	Martin Trépanier, & Catherine Morency
Programme **Program**	Génie industriel

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

UNIVERSITÉ DE MONTRÉAL

FUSION DES DONNÉES PROVENANT D'UN SYSTÈME DE PAIEMENT PAR CARTES À PUCE, D'UN SYSTÈME DE COMPTE À BORD ET D'HORAIRE POUR L'IMPUTATION D'ARRÊTS D'EMBARQUEMENT EN TRANSPORT COLLECTIF

FÉLIX LÉGARÉ

DÉPARTEMENT DE MATHÉMATIQUE ET DE GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION

DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(GÉNIE INDUSTRIEL)

DÉCEMBRE 2014

© Félix Légaré, 2014.

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

FUSION DES DONNÉES PROVENANT D'UN SYSTÈME DE PAIEMENT PAR CARTES À PUCE, D'UN SYSTÈME DE COMPTE À BORD ET D'HORAIRE POUR L'IMPUTATION D'ARRÊTS D'EMBARQUEMENT EN TRANSPORT COLLECTIF

présenté par : LÉGARÉ Félix

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

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

M. AGARD Bruno, Doctorat, président

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

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

Mme PELLETIER Marie-Pier, M. Sc. A., membre

REMERCIEMENTS

J’aimerais remercier mon directeur de recherche, Martin Trépanier et ma codirectrice de recherche, Catherine Morency. Leurs expertises m’ont été très précieuses tout au long de la réalisation de cette recherche et ils m’ont grandement aidé à trouver la motivation lorsqu’elle manquait.

Je tiens à remercier le RTL pour nous avoir données accès aux données utilisées pour cette recherche et pour leur collaboration tout au long du projet. J’espère que les réalisations pourront leur être utiles.

Afin, j’aimerais remercier ma famille, mes amis, mon coloc et ma copine pour m’avoir supporté et pour m’avoir encouragé à persévérer.

RÉSUMÉ

Avec l'adoption de systèmes de péage par cartes à puce, une grande quantité de données opérationnelles est maintenant accessible pour la planification des transports. Ces données demandent cependant un traitement préalable comme elles sont souvent récoltées pour répondre à des besoins administratifs comme la collecte de revenus.

Ce projet porte sur l'attribution d'arrêts aux données d'embarquements provenant du système de péage automatisé par cartes à puce du Réseau de Transport de Longueuil (RTL). Cette attribution a été effectuée par un algorithme utilisant des requêtes SQL.

Les données de localisation des véhicules sont déjà utilisées pour évaluer le service et en améliorer la ponctualité. Les données de systèmes de décompte automatisé de passagers sont utilisées pour connaître la demande sur les réseaux de transport et la charge des autobus. Les données de cartes à puce ont plusieurs applications. Une fois enrichies d'arrêts d'embarquement ainsi que d'arrêts de destination, plusieurs analyses peuvent être faites grâce aux données de cartes à puce. Il est possible d'analyser la demande, d'établir les zones du réseau les plus populaires et d'analyser le comportement des usagers. Les données de cartes à puce peuvent aussi compléter et améliorer l'information obtenue grâce aux enquêtes origines-destinations puisqu'elles couvrent un plus grand nombre de déplacements et que l'information sur ceux-ci est plus précise.

La méthodologie utilisée pour cette recherche est la suivante. Tout d'abord, le système d'information du RTL est caractérisé pour pouvoir proposer une structure d'accueil pour les données. Trois tables principales sont utilisées. La table CAP (cartes à puce) comprend les données du système de péage automatisé par cartes à puce où les transactions d'embarquements sont enregistrées. Ces transactions ne contiennent pas de données de localisation. La table SDAP (système de décompte automatique de passagers) comprend les données de localisation des autobus lors des passages aux arrêts et les données de compte à bord. La table GTFS (Google Transit Feed Specification) comprend les données du service planifié. Ensuite, les tables sont comparées entre elles pour vérifier l'intégrité des données. L'algorithme d'attribution d'arrêts est réalisé en trois étapes. La première étape est d'utiliser les données de la table SDAP pour enrichir les embarquements CAP. La deuxième étape est d'utiliser les habitudes des usagers et les embarquements CAP ayant obtenu des arrêts à la première étape pour dériver les arrêts

d’embarquement manquants. La dernière étape est d’utiliser les données GTFS pour finir l’enrichissement des embarquements CAP.

La caractérisation des données a permis de mettre en évidence des variations dans l’intensité des activités semblables pour les trois tables avec des pointes d'appariement maximal le matin et en début de soirée en semaine et beaucoup moins d’activités couplées la fin de semaine. La caractérisation a aussi permis de voir un manque de données au niveau de la table SDAP alors que beaucoup moins d’autobus y sont actifs que dans la table CAP, où il y a significativement moins de passages capturés que de passages prévus. Le manque de données dans la table SDAP rend l’application de l’algorithme beaucoup moins efficace.

L’application de l’algorithme d’attribution d’arrêt aux embarquements CAP a permis d’attribuer un arrêt à 91.7 % des embarquements. 54% ont été attribués lors de l’enrichissement grâce aux données SDAP, 27.2 % lors de l’enrichissement grâce aux habitudes des usagers et 10.5% lors de l’enrichissement grâce aux données GTFS.

ABSTRACT

A huge amount of data is now collected in public transit. With the arrival of automated fare collection system, a lot of operational data is now accessible. This data requires to be processed because it is usually collected for administrative purpose.

This project aims to assign a stop location to the boarding transactions recorded by the RTL's fare collection system using smartcard. The assignment is made using an algorithm running SQL queries.

Vehicle location data is already used to evaluate public transit service and to improve service punctuality. Automated passenger count data is used to know the demand on the transit network and the load on each bus. Automated fare collection system data using smartcard are widely used. Smartcard data enrich with boarding location and alighting location can be analyzed with many purposes. It is possible to use smartcard data to analyze demand on the network, popular boarding and alighting stops and user behavior. Smartcard data can be used to complement and improve origin-destination survey with information on a larger number of trip and more precise information.

The research methodology is the following. First, the RTL's information system is characterized. Three tables are used in this research. The CAP (AFC) table contains the boarding transactions from the fare collection system using smartcard. Those transactions do not contain boarding locations. The SDAP (AVL/ APC) table contains the vehicles locations and the passenger count data. The GTFS table contains data on the planned service. Second, the tables are compared to verify the data integrity. Third, the algorithm is used to assign boarding stops to the CAP data. The algorithm starts by using the SDAP data. Then it uses the commuters' habits to derive boarding stops from results of the first assignment. Finally, the GTFS data is used to assign stops to the remaining boarding transactions.

The data characterization shows similar variations through time between the activities of the three tables with peak periods in the morning and in the evening during week days and a lower level of activities during the week-ends. The characterization also shows missing data in the SDAP table. There are a fewer active buses than in the CAP table and fewer vehicle locations were capture than what was scheduled.

Once the algorithm was run on the CAP data, 91.7 % of the boarding transactions had a boarding stop, 54% were assigned using SDAP data, 27.2% using user habits and 10.5% using GTFS data.

TABLE DES MATIÈRES

REMERCIEMENTS III

RÉSUMÉ IV

ABSTRACT VI

TABLE DES MATIÈRES VIII

LISTE DES TABLEAUX XI

LISTE DES FIGURES XIII

LISTE DES SIGLES ET ABRÉVIATIONS XV

CHAPITRE 1 INTRODUCTION 1

1.1 Contexte de cette recherche 2

1.2 Objectif du mémoire 4

1.3 Structure du mémoire 4

CHAPITRE 2 REVUE DE LITTÉRATURE 5

2.1 Les systèmes de localisation des véhicules et de décompte à bord 5

2.2 Les systèmes de perception tarifaire automatisés par cartes à puce 8

2.2.1 Individualité des cartes 9

2.2.2 Localisation des embarquements 10

2.2.3 Attribution de destination aux embarquements 11

2.2.4 Exploitation des données 13

CHAPITRE 3 MÉTHODOLOGIE 15

3.1 Méthodologie générale 15

3.2 Système d'information 18

3.2.1 Données du système de paiement par cartes à puce (CAP) 18

3.2.2 Données SDAP 21
3.2.3 Données GTFS 24
3.2.4 Relation entre les tables CAP, SDAP et GTFS 27
3.3 Logiciels 29
3.4 Algorithme en SQL Server 30
3.4.1 Indice de temps 30
3.4.2 Ajout du numéro d'autobus à la table SDAP 30
3.4.3 Enrichissement de la table CAP par les arrêts de la table SDAP 31
3.4.4 Enrichissement de CAP par les habitudes des usagers 33
3.4.5 Enrichissement de CAP par la table GTFS 35
CHAPITRE 4 RÉSULTATS 36
4.1 Analyse descriptive 36
4.1.1 Caractérisation des données CAP 36
4.1.2 Caractérisation des données SDAP 39
4.1.3 Caractérisation des données GTFS 44
4.1.4 Comparaison entre les tables CAP et SDAP 47
4.1.5 Comparaison entre la table SDAP et la table GTFS 49
4.2 Résultats de l'algorithme d'attribution des arrêts aux embarquements CAP 52
4.2.1 Enrichissement par les données du système de décompte automatique de passagers 53
4.2.2 Enrichissement de la table CAP selon les habitudes des usagers 55
4.2.3 Enrichissement de CAP à partir des données GTFS 57
4.2.4 Comparaison des enrichissements 58
4.2.5 Caractérisation des résultats 59
4.3 Applicabilité de l'algorithme de destination 63

4.4 Les arrêts d'embarquement pour le mois de mars 2013 64

CONCLUSION 68

BIBLIOGRAPHIE 71

LISTE DES TABLEAUX

Tableau 3.1 : Champs de la table CAP....19

Tableau 3.2 : Champs ajoutés à la table CAP....20

Tableau 3.3 : Champs de la table sdap_courses....21

Tableau 3.4 : Champs de la table sdap_courses_arret....22

Tableau 3.5 : Champs ajoutés à la table SDAP....24

Tableau 3.6 : Champs de la table gtfs_stop....25

Tableau 3.7 : Champs de la table gtfs_trip....25

Tableau 3.8 : Champs de la table gtfs_stop_times....26

Tableau 3.9 : Champs de la table GTFS....27

Tableau 3.10 : Contrainte pour l'ajout du numéro d'autobus....30

Tableau 3.11 : Contraintes pour la première étape de la correspondance entre CAP et SDAP....32

Tableau 3.12 : Contraintes pour la deuxième étape de la correspondance entre CAP et SDAP....32

Tableau 3.13 : Contraintes pour la troisième étape de la correspondance entre CAP et SDAP....32

Tableau 3.14 : Contraintes pour la quatrième étape de la correspondance entre CAP et SDAP....33

Tableau 3.15 : Contraintes pour la première série de l'enrichissement par les habitudes des usagers....34

Tableau 3.16 : Contraintes pour la première série de l'enrichissement par les habitudes des usagers....34

Tableau 3.17 : Contraintes pour l'enrichissement de CAP grâce à GTFS....35

Tableau 4.1 : Éléments de la table CAP....36

Tableau 4.2 : Éléments de la table SDAP....39

Tableau 4.3 : Éléments de la table GTFS....45

Tableau 4.4 : Résultats de l'enrichissement par les données SDAP....53

Tableau 4.5 : Résultats de l'enrichissement de CAP par les habitudes des usagers. 56

Tableau 4.6 : Résultats de l'enrichissement de CAP par les données GTFS.....58

LISTE DES FIGURES

Figure 1-1 : Localisation de Longueuil 3

Figure 1-2 : Carte représentant l'étendue du Réseau de transport de Longueuil 3

Figure 3-1: Représentation schématique de la méthodologie de recherche 15

Figure 3-2 : Liens entre les deux tables SDAP 23

Figure 3-3 : Modèle relationnel entre les tables GTFS 26

Figure 3-4 : Relation entre les tables CAP et SDAP 28

Figure 3-5: Relation entre les tables CAP, SDAP et GTFS 28

Figure 4-1: Nombre d'embarquements moyen par heure pour chaque jour de la semaine 37

Figure 4-2 : Nombre moyen de bus actifs par heure et par jour de la semaine dans la table CAP 38

Figure 4-3 : Nombre moyen de passages d'autobus aux arrêts par heure et par jour de la semaine pour la table SDAP 40

Figure 4-4 : Nombre moyenne de bus actifs par heure et par jour de la semaine pour la table SDAP 41

Figure 4-5 : Carte représentant le nombre de passages aux arrêts enregistrés par SDAP 43

Figure 4-6 : Carte représentant le nombre de passagers embarquant à chaque arrêt selon SDAP. 44

Figure 4-7 : Nombre moyen de passages d'autobus aux arrêts par heure et par jour de la semaine pour la table GTFS 46

Figure 4-8 : Carte représentant le nombre de passages prévus par arrêt selon GTFS 47

Figure 4-9 : Comparaison entre les tables CAP et SDAP du nombre moyen de bus actifs par heure pour l'ensemble des jours de semaine 48

Figure 4-10 : Diagramme de Venn représentant les enregistrements de passages d'autobus prévus dans GTFS communs aux passages capturés dans SDAP 49

Figure 4-11 : Distribution des écarts relatifs entre le nombre de passages d'autobus capturés par SDAP et le nombre de passages prévus dans GTFS par ligne 50

Figure 4-12 : Distribution des écarts relatifs entre le nombre de passages d'autobus capturé par SDAP et le nombre de passages prévus dans GTFS par arrêt.....51

Figure 4-13 : Résumé des résultats de l'algorithme d'attribution d'arrêt 52

Figure 4-14 : Carte représentant les arrêts attribués lors de l'enrichissement par SDAP 54

Figure 4-15 : Carte comparant les montants (en jaune) SDAP et les embarquements CAP pour les bus avec un décompte de passager fonctionnel (en bleu). 55

Figure 4-16 : Carte représentant las arrêts qui ont été attitrés grâce aux habitudes des usagers....57

Figure 4-17 : Représentation des arrêts qui ont été attitrés aux embarquements CAP selon l'étape d'attribution.....59

Figure 4-18 : Taux d'attribution d'arrêt aux embarquements CAP selon les étapes de l'algorithme et selon la ligne.....60

Figure 4-19 : Taux d'attribution d'arrêts aux embarquements CAP selon les étapes de l'algorithme et selon le jour du mois 61

Figure 4-20 : Taux d'attribution d'arrêts aux embarquements CAP selon les étapes de l'algorithme et selon l'heure du jour 62

Figure 4-21 : Carte de l'ensemble des arrêts d'embarquement attribués par l'algorithme.....65

Figure 4-22 : Carte des embarquements pour la pointe du matin 66

Figure 4-23 : Carte des embarquements pour la pointe du soir 67

LISTE DES SIGLES ET ABRÉVIATIONS

CAP	Cartes à puce
SDAP	Système de décompte automatique à bord
GTFS	Google Transit Feed Specification
GPS	Global Positioning System
RTL	Réseau de Transport de Longueuil
SQL	Structured Query Language

CHAPITRE 1 INTRODUCTION

Les nouvelles technologies sont maintenant bien implantées dans les transports en commun. Leur impact se fait ressentir autant au niveau des véhicules, au niveau de l'administration qu'au niveau de l'utilisation. Les nouveaux véhicules, autobus et métro, sont munis de technologies dédiées au confort et à la sécurité des usagers. L'accessibilité pour les personnes à mobilité réduite est devenue un standard pour les nouveaux véhicules. Des autobus électriques et hybrides circulent déjà sur les routes. Les systèmes avancés d'exploitation (SAE) permettent une meilleure gestion des flottes d'autobus. Dans certains cas, ils permettent l'affichage des horaires en temps réel aux arrêts et stations. L'information est maintenant accessible plus facilement. Les usagers possédant des téléphones intelligents peuvent planifier leurs déplacements plus facilement. Les transactions monétaires disparaissent lentement alors que les systèmes de paiement automatisés prennent de l'expansion.

Ces nouvelles technologies apportent aussi un lot d'opportunités. Deux systèmes retiendront notre attention. D'abord, les données des systèmes de comptage automatisé à bord collectent des données de localisation sur les véhicules, en plus de compter les passagers, ce qui permet aux planificateurs de connaître les charges à bord afin de réguler le réseau. Puis, les systèmes de paiement automatisés par cartes à puce collectent des données transactionnelles sur les montées à bord des véhicules, en plus de faciliter la gestion des revenus. Les cartes à puce peuvent servir de réserves de titres qui sont débités à chaque utilisation. Il est aussi possible d'avoir un tarif qui permet l'accès au réseau pour une certaine durée de temps. Dans ces deux cas, la carte est lue à l'entrée de l'autobus ou du métro et le titre est débité, au besoin. Dans d'autres situations, les systèmes de cartes à puce permettent aux réseaux de transport en commun de facturer l'utilisateur selon la distance qui a été parcourue. La carte sert alors à y déposer de l'argent qui est débitée selon l'utilisation. La carte est donc lue deux fois par déplacement, soit à l'entrée et à la sortie du bus ou du métro. Chaque transaction devient un événement qui peut être analysé.

Plusieurs auteurs ont vu le potentiel des données qui sont récoltées grâce à la carte à puce. Leurs recherches portent autant sur les habitudes des usagers que sur l'évaluation de l'efficacité des réseaux. Ces études utilisent principalement les données d'embarquement. Différents auteurs ont été en mesure d'effectuer des analyses supplémentaires en utilisant les données de

localisation des transactions de cartes à puce. Certains auteurs ont été en mesure d'évaluer la charge sur des lignes et sur des autobus individuellement. D'autres ont été capables de créer des matrices origines-destinations pour les usagers sans avoir recours à des enquêtes. Toutes ces analyses permettent ensuite d'ajuster le service offert et de faire une meilleure planification des ressources des sociétés de transport.

1.1 Contexte de cette recherche

Ce mémoire porte sur l'enrichissement des données du système de paiement par cartes à puce du Réseau de Transport de Longueuil (RTL). Les enregistrements du système de cartes à puce du RTL ne contiennent pas de données de localisation. Par contre, il est possible de les coupler avec les données du système de décompte automatique de passagers (SDAP), qui lui, contient des données de localisation de véhicule.

Longueuil est situé au sud-est de Montréal. Les deux villes sont séparées par le fleuve St-Laurent et elles sont reliées par trois ponts, un tunnel et un métro. Beaucoup d'habitants de la ville de Longueuil et du reste de la rive-sud du fleuve Saint-Laurent travaillent à Montréal, le centre économique de la région. Une partie des activités du RTL est de transporter les usagers vers les stations de métro Radisson et Bonaventure. Celles-ci sont situées près des ponts sur l'île de Montréal. Une autre grande partie des activités est de transporter les passagers vers la station de métro Longueuil-Université-de-Sherbrooke. Le métro est opéré par la Société de Transport de Montréal et donne un accès direct au centre-ville de Montréal. Le réseau permet également l'accès au territoire longueuillois, dont entre autres les établissements d'enseignement et les hôpitaux.

La carte de la Figure 1-1 présente la localisation de Longueuil. La carte de la Figure 1-2 illustre le territoire du Réseau de transport de Longueuil, indiquant le tracé des lignes du réseau. Notons que le RTL dessert également la ville de Saint-Bruno-de-Montarville.

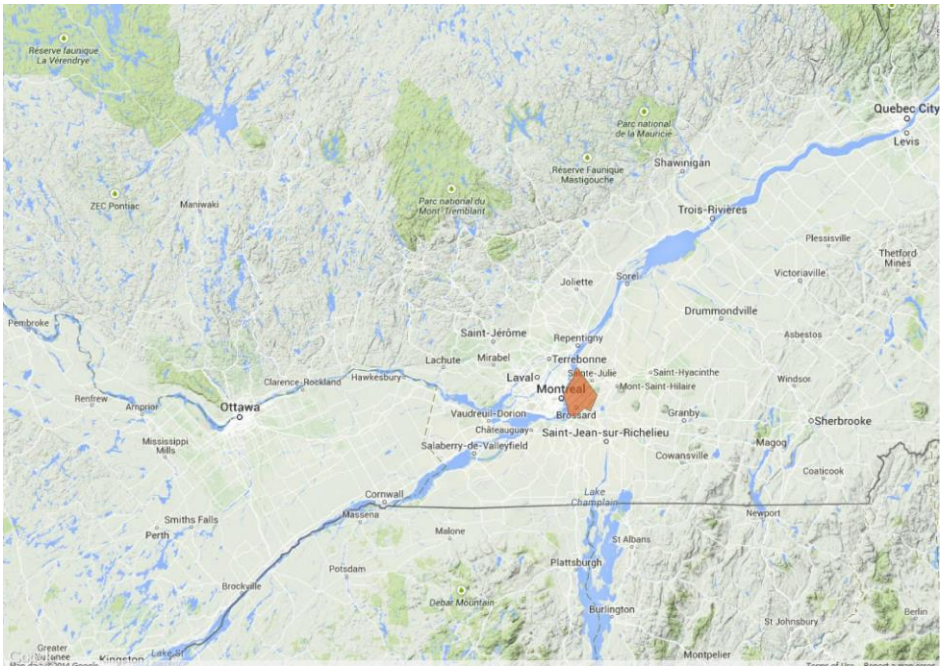


Figure 1-1 : Localisation de Longueuil



Figure 1-2 : Carte représentant l'étendue du Réseau de transport de Longueuil

1.2 Objectif du mémoire

Puisque les données transactionnelles du système de paiement par carte à puce du RTL ne sont pas géolocalisées, l'objectif de la recherche est de proposer une méthode pour attribuer un arrêt (une localisation spatiale) à chacune des transactions.

Pour ce faire, il faudra d'abord élaborer une méthode de traitement préalablement des données. Celles-ci doivent être modifiées pour être manipulables et comparables entre elles.

Ensuite, les différentes données doivent être analysées pour comprendre ce que chacune d'entre elles représentent et les liens qui existent.

Afin, un algorithme est proposé pour attribuer des arrêts aux embarquements. Celui-ci utilise les systèmes de bases de données afin d'enrichir les données de cartes à puce en utilisant les données du système de décompte automatique de passagers (SDAP), couplées aux données sur le réseau d'autobus et les horaires.

Finalement, il sera question des applications supplémentaires possibles grâce aux données de cartes à puce enrichies de localisation d'embarquement.

1.3 Structure du mémoire

Le corps du mémoire est divisé en trois chapitres. Le premier (chapitre 2) est une revue de littérature couvrant l'utilisation de données provenant des systèmes de localisation de véhicule, des systèmes de décompte automatique de passagers et des systèmes de péage automatisés par cartes à puce.

Le chapitre 3 traite de la méthodologie qui a été utilisée pour réaliser cette recherche. Le système d'information du RTL et les outils utilisés sont présentés. Par la suite, le fonctionnement de l'algorithme d'attribution d'arrêt est détaillé.

Le dernier chapitre traite des résultats qui ont été obtenus lors de l'analyse des données du RTL et des résultats obtenus suite à l'application de l'algorithme d'attribution d'arrêt. Une conclusion vient rappeler les contributions de la recherche et énonce quelques perspectives.

CHAPITRE 2 REVUE DE LITTÉRATURE

L'utilisation de données en transport est un sujet qui a grandement évolué avec l'arrivée de systèmes recueillant une abondance de données. Les systèmes de localisation des véhicules ont été les premiers systèmes à générer des données pouvant servir à la planification. Ces systèmes servaient initialement à la gestion des opérations, mais plusieurs utilisations en ont été faites par la suite. L'ajout de systèmes de décompte automatisé des passagers au système de localisation des véhicules a rendu les données encore plus riches et des utilisations supplémentaires leur ont été trouvées. Lorsque les systèmes de perception tarifaire automatisés par cartes à puce ont été implantés pour des raisons principalement administratives, une autre quantité importante de données a commencé à être générée. Les possibilités d'exploitation des données de cartes à puce sont encore plus vastes.

Plusieurs auteurs ont traité des possibilités reliées autant aux systèmes de compte à bord et de localisation des véhicules qu'aux systèmes de perception tarifaire automatisés par cartes à puce. Cette section porte sur ces utilisations et sur les traitements de données devant être préalablement réalisés.

2.1 Les systèmes de localisation des véhicules et de décompte à bord

Les systèmes de localisation des véhicules sont des systèmes qui captent la position GPS du véhicule dans le temps. Certains systèmes captent aussi le chaînage le long de la ligne d'autobus. Les enregistrements se font à différents moments selon le système. Par exemple, certains prennent des enregistrements selon un intervalle de temps régulier et d'autres prennent des enregistrements lorsque les véhicules circulent à un arrêt. La localisation des autobus dans le temps représente une énorme source d'informations et plusieurs utilisations en ont été faites.

Les données de localisation aident à déterminer la vitesse à laquelle se déplaçaient les autobus et le temps qu'ils nécessitent pour parcourir chaque tronçon de ligne. L'ensemble de ces informations permet de connaître la distribution du temps de parcours sur chaque tronçon de ligne. Il est possible d'analyser les délais et les variations de temps de parcours selon les chauffeurs ou la période de la journée et d'apporter des ajustements à la planification (Furth et al. 2006)**.** Les données de localisation de véhicule permettent aussi de connaître le nombre de véhicule-kilomètres parcourus, le nombre de véhicule-heures opérés et la vitesse commerciale

des autobus. La variation de ces résultats peut être décortiquée pour trouver des tendances et améliorer la planification du réseau (Trépanier et al. 2009).

La distribution des temps de parcours permet d'ajuster les temps de passage et d'établir un temps de battement pour s'assurer que les bus soient toujours à l'heure au départ de leur prochaine ligne. Si on ne se sert que du temps de parcours moyen, une partie des autobus seront en retard aux arrêts. La distribution complète des temps de parcours permet d'utiliser les valeurs extrêmes et ainsi de s'assurer que les autobus auront eu le temps de parcourir leur ligne avant le début de la suivante (Furth et al. 2006)**.**

Les données de localisation automatique des véhicules sont utilisées pour améliorer la ponctualité des autobus. La ponctualité du service est un élément important pour l'usager. Lorsqu'un autobus devance son horaire et que l'usager arrive à l'heure, l'usager aura manqué son autobus. Si l'autobus est en retard, il est possible que l'usager manque sa correspondance ou qu'il soit en retard à destination. Ces deux cas peuvent entrainer un dérangement important dans l'horaire de l'usager. Un manque de ponctualité récurrent peut aussi décourager l'usager à utiliser le transport en commun.

À l'aide des données de localisation, il est possible de déterminer la distribution statistique des arrivées des autobus aux arrêts par rapport à l'horaire. Cette distribution suit habituellement une courbe normale. Grâce à cette distribution, les planificateurs peuvent ajuster l'horaire aux arrêts, sans devoir changer les départs des autobus ou le comportement des chauffeurs. Ainsi, il y aura une meilleure adhérence entre l'horaire et l'arrivée des autobus (Cevallos et al. 2011)**.**

Selon la variance de la distribution, différentes options sont possibles pour améliorer la ponctualité. Si la variance est faible, il s'agit de modifier l'heure d'arrivée. Si la variance est élevée, il peut être nécessaire d'améliorer la supervision des activités ou d'avoir un horaire plus fragmenté (Cevallos et al. 2011)**.**

Plusieurs sociétés de transport en commun ont muni leur autobus de systèmes de décompte automatique de passagers. Ceux-ci dénombrent le nombre de passagers embarquant et débarquant à chaque arrêt. Les passagers sont comptés à l'aide de différents capteurs placés aux portes avant et arrière de l'autobus. Les systèmes de décompte sont habituellement combinés aux systèmes de localisation de véhicules. Les données permettent d'analyser le flux de passagers de

plusieurs manières. On peut les utiliser pour connaître la demande sur le réseau. Il est possible de segmenter la demande selon l'heure, la journée, la ligne et la position sur la ligne. Ici, les passagers sont considérés en groupe et il est impossible de connaître la séquence de déplacements d'un usager (Furth et al. 2006).

Du point de vue des arrêts, les données de décompte indiquent à quel moment de la journée et à quel rythme les usagers montent et descendent (Shalaby et Farhan 2004). Il est ainsi possible de connaître les points les plus importants du réseau et les heures auxquelles les déplacements se font à partir ou vers ces arrêts.

Le système de décompte automatique de passagers permet également d'évaluer la charge des autobus. En additionnant le nombre de passagers qui montent et en soustrayant le nombre de passagers qui descendent à chaque arrêt, il est possible de suivre le nombre de passagers à bord des autobus en tout temps. En étudiant la courbe enveloppe, il est possible de voir les segments d'une ligne qui sont le plus achalandés. Il est aussi possible de voir s'il y a un moment de la journée où les autobus sont sur-achalandés ou sous-achalandés et ainsi ajuster la fréquence des passages (Furth et al. 2006).

Une utilisation des données de compte de passagers automatisé a été faite sur le réseau de Toronto par Shalaby et Farhan (2004). Le but était d'afficher l'heure d'arrivée des prochains bus aux arrêts. Avant d'afficher l'heure d'arrivée, il fallait être en mesure de la prédire malgré les variations des activités au cours de la journée. Grâce aux données historiques de localisation des autobus, il est possible de déterminer le temps de trajet entre les arrêts. Par contre, il est plus difficile de calculer la durée des arrêts. Sur la ligne étudiée, la durée de l'arrêt est un facteur très important pour six des arrêts. Ces six arrêts étant situés à des intersections majeures, le nombre d'usagers y embarquant est significatif.

Pour estimer la durée de l'arrêt, ils ont utilisé les données historiques du comptage de passagers. Ces données permettent d'estimer statistiquement le rythme d'arrivée des usagers à chaque arrêt selon le nombre de personnes qui montent dans le bus et le temps entre l'arrivée de ce bus et le départ du bus précédent de cet arrêt. Par ce moyen, ils ont pu estimer le nombre de passagers qui embarqueront. En assumant que chaque passager prend 2,5 secondes à embarquer, ils ont évalué la durée de l'arrêt. Pour cette étude, il est assumé que le temps de descente des passagers n'est pas significatif par rapport au temps d'embarquement (Shalaby et Farhan 2004).

Cette utilisation des comptes de passagers permet aussi de voir l'effet qu'un autobus qui prend du retard ou de l'avance aura sur la ponctualité de l'autobus suivant. S'il passe en avance, plus de passagers arriveront à l'arrêt avant l'arrivée du prochain autobus. Cette autobus aura donc des arrêts plus longs et finira par prendre du retard.

2.2 Les systèmes de perception tarifaire automatisés par cartes à puce

Les cartes à puce en transport en commun avaient initialement comme but de faciliter la perception tarifaire et l'administration. Les cartes sont lues lorsque les usagers les présentent aux lecteurs à l'entrée des autobus. Une transaction est alors effectuée, soit pour débiter un titre ou de l'argent de la carte, ou pour valider le droit de passage. Cette transaction est alors emmagasinée dans une base de données. Plusieurs chercheurs ont vu l'énorme potentiel des données de cartes à puce. Cette section explique les traitements qui ont été faits de ces données et les analyses qu'il est possible d'en faire.

Les enregistrements de transactions cartes à puce contiennent habituellement le jour et l'heure de la transaction. Par contre, il n'est pas commun, notamment dans les systèmes plus anciens, que la localisation de la transaction soit incluse. Il est possible de faire certaines analyses sans la localisation, mais celle-ci devra être imputée pour exploiter tout le potentiel des données de cartes à puce. Plusieurs travaux ont été répertoriés par Pelletier et al. (2011), touchant les aspects stratégiques, tactiques et opérationnels des réseaux de transport collectif. Nous nous limiterons ici aux enjeux reliés plus précisément aux travaux de ce mémoire.

Il est possible d'observer des tendances au niveau de l'achalandage du réseau. Ces tendances s'observent autant au niveau de l'année, de la semaine que de la journée. Il y a une saisonnalité sur le nombre d'embarquements qui sont effectués par semaine au cours de l'année. Des baisses d'achalandage sont généralement remarquées lors des congés scolaires et au cours des vacances estivales (Morency et al. 2007). Au niveau de la semaine, il y a généralement moins d'activités le vendredi que les autres jours de semaines. On enregistre beaucoup moins de transactions les jours de fins de semaine (Trépanier et al. 2009)(Wang et al. 2011). Des tendances très fortes sont observées au niveau du nombre d'embarquements par heure au cours

d'une journée. Il y a habituellement des pointes d'embarquements le matin, lorsque les gens se rendent au travail, et en début de soirée, lors du retour à la maison (Lee et Hickman 2011).

2.2.1 Individualité des cartes

Un des attraits les plus intéressants des cartes à puce est qu'elles ont une identité propre. Les données de compte à bord mentionnées précédemment donnent des informations très intéressantes sur les flux de passagers, par contre les usagers y sont observés de façon agrégée. Chaque numéro de carte à puce dans le système correspondant à un usager (du moins on le suppose), il est possible d'analyser une grande quantité d'information sur les déplacements des usagers de façon désagrégée. Il est donc possible d'analyser l'ensemble des déplacements d'un usager et de créer une chaîne de déplacement. Il est ainsi possible d'analyser le comportement des usagers. Dans une certaine mesure, beaucoup d'information sur l'usager peut être extraite selon ses déplacements en transport en commun. Les différents comportements peuvent être regroupés selon le type de titre (Pelletier et al. 2011). Notons par contre que les données sont anonymes et que les chercheurs n'ont aucunement accès aux données sur les détenteurs.

L'individualité des cartes permet d'établir des journées de déplacements type. En connaissant tous les déplacements d'une carte au cours d'une journée, il est possible d'établir à quels moments l'usager s'est déplacé. Morency et al. (2007) ont regroupé les journées de déplacements d'usagers en grappes. Pour chaque type de carte, quatre grappes de journée d'utilisation typique ont été créées. Les utilisateurs ont des comportements d'utilisation du transport en commun pouvant appartenir à plusieurs grappes, mais une forte régularité a été observée pour la majorité des cartes (Morency et al. 2007).

Lee et Hickman (2011) remarquent que les comportements des usagers ayant une *metro pass* et une *u-pass* sont très différents sur le réseau de Minneanapolis. Les usagers ayant la *metro pass* ont deux périodes de pointe d'utilisation très prononcées entre 6:00 et 9:00 et entre 16:00 et 19:00 et la majorité de ces usagers feront leur deuxième déplacement, le retour à la maison, entre neuf et onze heures après leur déplacement matinal. Les usagers ayant la *u-pass* ont une utilisation distribuée plus également tout au long de la journée avec une très légère pointe entre 8:00 et 9:00. Il est difficile d'observer un intervalle de temps constant entre leurs déplacements. Une plus grande concentration de deuxième déplacement s'effectue quatre heures après le

premier, mais la distribution des déplacements dans le temps est plus égale que dans le cas des *metro pass* (Lee et Hickman 2011).

2.2.2 Localisation des embarquements

Pour pousser l'analyse plus loin, il faut regarder le côté spatial des enregistrements cartes à puce. Dans certains cas, la localisation des transactions est comprise dans l'enregistrement initial (montée). Très peu de réseaux de transport en commun ont des enregistrements correspondant à la destination des déplacements (descente). Les quelques exceptions sont les réseaux qui facturent l'usager selon le trajet qu'il a parcourue. Dans ce cas, la localisation est aussi enregistrée lorsque l'usager descend de l'autobus.

Lorsque la localisation de l'embarquement ne fait pas partie des données de cartes à puce, il faut coupler les données de cartes à puce avec les données de localisation des véhicules pour l'obtenir. Ce couplage de données est possible si les enregistrements de données de cartes à puce contiennent le numéro d'autobus à bord duquel la transaction a été effectuée. Dans ce cas, il suffit d'associer la localisation de l'autobus au moment de la transaction comme étant la localisation de la transaction. L'intégrité des données de localisation des autobus est souvent un problème à cette étape (Shi and Lin 2013) (Munizaga et Palma 2012).

Lorsque la localisation des enregistrements de cartes à puce est connue, il est possible d'observer les arrêts qui sont les plus populaires. Les arrêts d'embarquement habituels des usagers sont aussi analysables.

Morency et al. (2007) ont remarqué que les usagers utilisent peu de nouveaux arrêts. Dans leur recherche, les usagers utilisaient en moyenne 4 arrêts aux cours de la première semaine d'observations et ils utilisaient de nouveaux arrêts au rythme de 0.6 par semaine (Morency et al. 2007).

Dans le cas de Minneapolis, ils ont remarqué que les arrêts les plus fréquentés sont situés dans les deux centres-villes de la région pour tous les types de cartes. Les usagers ayant la *u-pass* ont une forte fréquentation des arrêts autour de l'université, tout en étant très présents dans les centres-villes.

Les arrêts utilisés pour le premier embarquement du matin peuvent être associés avec la localisation de la résidence des usagers. Elle permet de voir que les gens ayant la *metro pass* sont

répartis sur l'ensemble du territoire et en banlieue. De leur côté, les usagers ayant la *u-pas*s sont très concentrés près de l'université (Lee et Hickman 2011).

Foell et al. (2014) ont utilisé les données des cartes à puce pour prédire les arrêts où les usagers embarqueront sur le réseau. Leur objectif est de créer une liste d'arrêts d'embarquement les plus probables pour chaque utilisateur. Ils commencent par utiliser les données d'embarquement de chaque utilisateur. Ils classent les arrêts selon le nombre d'embarquements que l'usager a fait à chaque arrêt.

Ensuite ils regardent la popularité globale de l'ensemble des arrêts du réseau. Selon le nombre d'embarquements qui ont été faits à chaque arrêt, ils trouvent les points d'intérêts de la ville. Ils classent ainsi les arrêts n'ayant pas été utilisées par les usager avec un ordre logique. Il devient aussi plus facile de faire des prédictions pour les usagers ayant un petit historique d'embarquement, et les nouveaux usagers.

Par la suite, la géographie du réseau est analysée. L'hypothèse est que plus un arrêt est loin ou difficile à atteindre, moins il est probable qu'un usager s'y rende. Ainsi, il est moins probable qu'un usager monte à un arrêt qui est loin des arrêts qu'il utilise habituellement. En ayant cette connaissance du réseau, l'historique de l'utilisateur et la popularité globale des arrêts, les auteurs ont établi un coefficient pour établir la probabilité qu'un usager utilise les arrêts selon la localisation de ses arrêts habituels.

Ils tentent ensuite de trouver des usagers ayant des habitudes de déplacement similaires entre eux et des arrêts qui peuvent être attribués à ces types d'habitudes. Ils assument donc que la probabilité qu'un usager se rende à un arrêt où il n'est jamais embarqué, mais qui est utilisé par des gens ayant des habitudes semblables est plus forte. La combinaison de leur modèle leur a permis d'observer des embarquements à des arrêts qui correspondaient, en moyenne, au 97^{e} percentile des listes d'arrêts des usagers. (Foell et al. 2014)

2.2.3 Attribution de destination aux embarquements

Une des lacunes principales des données provenant des systèmes de péages automatisés par cartes à puce est le fait que beaucoup de ces systèmes n'enregistrent que les montées. Il n'y donc aucune indication sur la destination des usagers. Or, pour vraiment être en mesure de

connaître la demande sur le réseau, il est important de connaître les trajets complets qui ont été effectués.

Certains chercheurs se sont penchés sur ce problème. Ils ont créé un algorithme recréant le trajet d'un usager pour trouver les arrêts de destination pour chacun de ses déplacements. Ils utilisent l'identifiant unique des usagers, la séquence des arrêts sur les lignes et la séquence de déplacement de l'usager. Pour chaque embarquement de chaque utilisateur, l'algorithme détermine les arrêts de destination possibles sur la ligne. Connaissant l'arrêt d'embarquement suivant de l'usager, l'algorithme recherche l'arrêt de destination possible le plus près du prochain arrêt d'embarquement. Dans le cas où la destination est impossible à déterminer grâce à l'embarquement suivant, l'algorithme utilise les déplacements historiques de l'usager afin de trouver un déplacement similaire et associer la même destination (Trépanier et al. 2007).

Afin d'augmenter l'efficacité de l'algorithme, certaines améliorations lui ont été apportées. Un tri des données est effectué. Il est important de ne pas avoir de rupture dans la chaîne de déplacement, puisque l'algorithme de destination repose sur cela pour fonctionner. Lorsqu'il est établi que l'arrêt d'embarquement est erroné, l'arrêt est déterminé selon les données du service planifié. Par la suite, ce nouvel arrêt est comparé aux arrêts historiques de l'usager. Il est considéré que les données historiques sont plus fiables car les données du service planifié ne comprennent pas les retards d'autobus. Après avoir utilisé l'algorithme de destination sur les données de cartes à puce enrichies, les données historiques sont utilisées de nouveau pour trouver la destination des déplacements n'ayant pas de destination grâce à l'algorithme (Chapleau et al. 2008)**.**

Une fois que la destination a été estimée, il est possible de calculer le temps de déplacement, les distances parcourues et le temps entre les déplacements. Il est aussi possible de calculer la charge de chaque autobus à tous instants. La connaissance de la charge des autobus met en évidence les sections des lignes où les autobus sont le plus achalandés (Wang et al. 2011). Il est possible de calculer le nombre de passagers-kilomètres déplacés par le réseau de transport en commun (Trépanier et al. 2009). En connaissant les points d'embarquements et les destinations des usagers, il est possible de créer une matrice origine-destination pour le réseau (Munizaga et Palma 2012).

L'intégralité des données de localisation des véhicules est très importante. Les données provenant du système de carte à puce de la ville de Shenzhen ont été couplées avec les données de localisation de véhicules. L'objectif était de pouvoir trouver les paires origine-destination des usagers. Comme moins de 10% des autobus du réseau sont munis de système de localisation, seulement 10% des embarquements sont localisés. Sur les 7 millions d'usagers, 13 600 paires O-D ont pu être déterminées. Le manque de données GPS est la plus grosse embûche rencontrée dans cette étude (Shi and Lin 2013)**.**

2.2.4 Exploitation des données

Les cartes à puce représentent un complément aux enquêtes traditionnelles. Les enquêtes sur l'utilisation du transport en commun couvrent généralement une journée ou une semaine d'utilisation pour une fraction des utilisateurs. Elles sont réalisées très peu fréquemment et sont très coûteuses. De leur côté, les transactions enregistrées de cartes à puce représentent l'ensemble de l'utilisation du transport en commun par chacun des usagers. Elles couvrent toutes les journées et représentent "l'univers" des usagers. Les transactions sont associées à des données opérationnelles complètes. L'ensemble des usagers peuvent être associés à des points d'intérêts. Les données cartes à puce sont donc plus précises que les sondages. Dans le cas des sondages, les personnes sondées doivent se rappeler de leurs déplacements. Les cartes à puce enregistrent chaque déplacement automatiquement. En ce qui a trait au temps, les personnes ont tendance à arrondir aux 5 ou 15 minutes près alors que les cartes à puce enregistrent la transaction à la seconde près (Chapleau et al. 2008). Par contre, aucune information socio-démographique n'est disponible pour les données de cartes à puce.

Riegel et Attanucci (2013) ont comparé les résultats d'une enquête origine-destination et les données de cartes à puce. Sur le réseau de Londres, lors du sondage pour l'enquête origine-destination, certains usagers ont accepté de donner leur numéro de carte Oyster, le système de cartes à puce présent sur ce réseau. Les données d'embarquement pour ces usagers particuliers ont été comparées avec les réponses à l'enquête OD.

Sur les 1 700 personnes ayant donné un numéro de carte et ayant utilisé le transport en commun le jour de l'enquête, 39% ont reporté le même nombre de déplacements que ceux enregistrés par leur cartes tandis que 61% ont sous-reporté ou sur-reporté. Au total, 8,7% plus des déplacements ont été reportés dans le sondage qu'enregistrés par les cartes Oyster pour la journée

de l'enquête. Seulement 22% des usagers avaient une correspondance parfaite entre les déplacements rapportés dans le sondage et les déplacements enregistrés. Plusieurs facteurs peuvent expliquer ces résultats, dont le besoin de se souvenir des déplacements effectués pour répondre au sondage. Les estimations de temps par les sondés sont aussi très vagues. Les temps d'embarquements et les durées sont grandement arrondis.

Les données de systèmes de cartes à puce permettent non seulement de "sonder" un plus grand nombre de personnes, elles permettent aussi d'avoir des résultats plus précis sur l'utilisation du transport en commun par chaque usager. Par contre, les données cartes à puce ne capturent pas les autres modes de transport et ne permettent pas de connaître les usagers (lieu d'habitation, lieu de travail, revenu, nombre de voiture par ménage) (Riegel et Attanucci 2013)**.**

Par la suite, il est possible d'utiliser les données cartes à puce enrichies pour déterminer l'occupation des usagers. Le type de carte peut indiquer l'occupation de l'usager, l'heure des déplacements, le temps entre deux déplacements et la localisation de ceux-ci sont de très bons indices pour déterminer l'activité qui a été faite pour un certain déplacement. Avec les données historiques d'un usager, il est possible d'arriver à des conclusions assez complètes. Lorsqu'un usager reste huit heures au même endroit avant de se déplacer et que son patron de déplacement correspond à un horaire de travail, il est possible de déduire que l'usager est un travailleur ainsi que son lieu de travail (Lee et Hickman 2011). Il est possible de déterminer l'école qui est fréquentée par des usagers ayant des passes étudiantes (Chapleau et al. 2008)**.**

Bref, un grand nombre d'utilisations peuvent être faites avec les données de cartes à puce lorsque celles-ci sont enrichies de localisations d'embarquement et de destination.

CHAPITRE 3 MÉTHODOLOGIE

3.1 Méthodologie générale

Dans le but de créer un algorithme capable d'attribuer un arrêt d'embarquement aux enregistrements du système de cartes à puce, plusieurs étapes sont nécessaires. La Figure 3-1 représente de façon schématique les étapes réalisées au cours de ce projet, qui sont décrites dans les sections qui suivent.

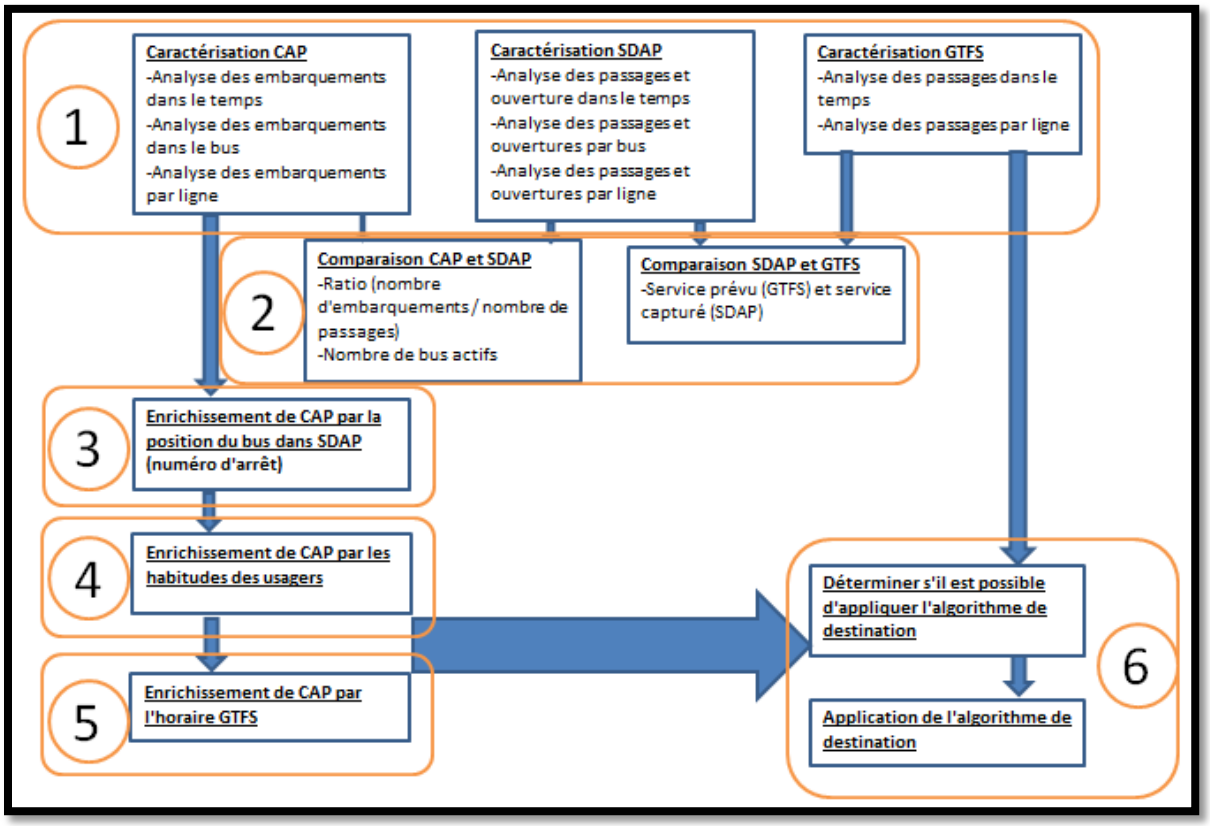


Figure 3-1: Représentation schématique de la méthodologie de recherche

La première étape est de faire un traitement préalable des fichiers de données et de caractériser les tables de la base de données du RTL. Trois tables sont utilisées pour l'étude : une table contenant les données du système de cartes à puce (CAP), une table contenant les données du système de décompte automatique de passagers couplé avec un système de localisation de véhicule (SDAP) et une table contenant le passage aux arrêts planifié des autobus (fichier du standard *General Transit Feed Specification*, GTFS). La caractérisation consiste à voir la

distribution des enregistrements dans le temps, le nombre d'autobus actifs, la localisation des enregistrements.

La deuxième étape est de comparer les tables. L'objectif est d'établir si les données sont complètes, s'il y a des tendances dans les données manquantes et s'il est possible de coupler les tables. Entre les tables CAP et SDAP, la présence des mêmes autobus dans les deux systèmes est un point de comparaison. Entre les tables SDAP et GTFS, la comparaison vise à voir la part du service prévu qui a été capturée par le système SDAP.

La troisième étape est de déterminer quel passage à un arrêt de la table SDAP correspond à chaque embarquement CAP. Cette méthode est la plus précise puisque la localisation de l'autobus est directement disponible. La correspondance entre les deux bases de données est faite via le numéro d'autobus et la date et l'heure d'embarquement. Le moment d'embarquement CAP correspond au moment d'ouverture de porte dans le SDAP. Le succès de cette opération dépend de l'intégrité des données SDAP.

La quatrième étape consiste à trouver un arrêt pour chaque embarquement CAP qui n'a pas trouvé de correspondance dans la table SDAP. La correspondance est établie selon les habitudes des usagers. Sur l'hypothèse qu'un usager utilise régulièrement les mêmes arrêts, les embarquements CAP ayant déjà un arrêt seront utilisés pour combler le manque de données. Le but est de trouver un embarquement par le même usager, pour le même type de jour, à la même heure, sur la même ligne, dans la même direction et qui a déjà une correspondance dans la table SDAP. Plus un usager aura d'embarquements avec un arrêt correspondant, meilleure sera l'estimation des arrêts de ses autres embarquements.

La cinquième étape est de déterminer un arrêt pour les embarquements CAP pour lesquels on n'a toujours pas d'arrêt en fonction des méthodes précédentes. On utilise cette fois-ci la table des passages aux arrêts planifiés (GTFS). La correspondance entre ces deux tables est faite selon la ligne, la direction, la voiture (pièce de travail) et le temps. La problématique est qu'il peut y avoir plusieurs autobus actifs au même moment pour la même ligne. L'arrêt sera accepté seulement s'il est certain que la possibilité de choix est unique.

La sixième étape est de déterminer s'il est possible d'appliquer l'algorithme de destination à cette structure de données et de voir les conclusions qui peuvent être faites à partir des données

cartes à puce enrichies du arrêt d'embarquement et des données du système de décompte automatique de passagers.

3.2 Système d'information

Le système d'information de la RTL est composé de plusieurs tables. Seules les tables détenant les enregistrements de transactions à l'aide de cartes à puce, les enregistrements du système de décompte automatique des passagers et de positionnement des autobus et les données GTFS sont utilisées. Les données utilisées couvrent la période du mois de mars 2013.

3.2.1 Données du système de paiement par cartes à puce (CAP)

Les cartes à puce utilisées en transport en commun ont le format d'une carte de crédit. La carte utilisée au RTL et dans le reste de la région de Montréal se nomme la *carte OPUS*. Elle sert à emmagasiner des titres de transport tel que les titres journaliers, hebdomadaires et mensuels ainsi que des billets individuels. Il existe plusieurs types de titres mensuels qui offrent différentes couvertures du réseau.

Pour valider le titre, l'usager met sa carte sur le lecteur qui est situé à l'entrée de l'autobus ou à l'entrée de la station de métro. La carte est alors activée par le lecteur et elle émet un signal qui est lu par le lecteur. Si la carte est associée à un titre valide, l'usager peut poursuivre sa route.

Même si le réseau de la RTL est connecté au réseau de métro de la ville de Montréal, il est desservi en entier par des autobus. Il y a donc seulement un type d'enregistrement carte à puce. Les enregistrements de transactions carte à puce sont contenues dans une seule table nommée CAP. Cette table contient seulement les transactions réalisées avec une carte à puce et elle ne contient pas les transactions refusées. Pour le mois de mars 2013, 2 481 977 transactions valides ont été enregistrées. La structure de la table est décrite au tableau 3.1.

Tableau 3.1 : Champs de la table CAP

nom du champ	Format ou possibilités	Description
identifica		Identification individuelle de la carte. Il est assumé qu'un usager ne détient qu'une seule carte. Ce numéro peut donc être associé à un usager et être utilisé pour "suivre" celui-ci tout au long du mois. Ce numéro est anonyme et il est impossible de connaître l'identité de l'usager.
date28	yyyymmjj	Date selon une journée finissant à 28 heures (4 :00 le jour suivant). Ainsi, lorsqu'une personne embarque dans l'autobus à 2 :00 AM le 28 mai, l'enregistrement se fait pour le 27 mai à 26:00.
typ_jour	SE, SA, DI, F1	Le type de jour. Jour de semaine, samedi, dimanche ou férié. Pour le mois de mars 2013, il y a une seule journée fériée, le 29 mars. Les horaires sont faits pour chaque type de jour. Il y a donc quatre horaires différents pour le mois de mars.
jour28	1 à 7	Le jour de la semaine, avec le principe du jour finissant à 28 heures. Le lundi est le jour 1.
heure28	hhmm	L'heure à laquelle la transaction a été effectuée selon le principe de jour finissant à 28 heures. Ce champ n'inclut pas les secondes.
sec28	Entier	Le nombre de secondes s'étant écoulées entre minuit et le moment de la transaction. Ce champ permet de connaître le moment exact de la transaction, à la seconde près.
lig_cap	Entier	La ligne sur laquelle la transaction a été effectuée.
Trace	Entier	Plusieurs lignes ont des tracés avec de légères différences. Ce champ indique de quel tracé il s'agit.
Dir	A, R	La direction dans laquelle circulait l'autobus sur la ligne lors de la transaction. Aller ou retour selon les directions établies par le RTL.
no_bus	Entier	Numéro de l'autobus à bord duquel la transaction a été faite.
Voiture	Entier	Pièce de travail, ce qui correspond à l'identification d'une combinaison de lignes empruntées par un autobus. La même voiture peut correspondre à plusieurs lignes à différents moments de la journée.

Afin de faciliter les calculs et les recherches de l'algorithme et pour pouvoir enrichir la table avec la localisation des arrêts, plusieurs champs ont été ajoutés à la table CAP dans le cadre de ce projet. Les champs ajoutés sont présentés dans le tableau suivant 3.2.

Tableau 3.2 : Champs ajoutés à la table CAP

Nom du champ	Format	Description
id_emb	Entier	Ce champ sert d'identification pour les enregistrements d'embarquement. Il devient la clé primaire de la table.
dateheureindex	Entier	Ce champ sert à accélérer la comparaison du temps. Dans la table CAP initiale, pour savoir si deux événements se sont produits en même temps, il faut comparer le champ de la date, puis le champ de l'heure qui est formaté de différente manière selon la table et qui utilise parfois des journées de 24 heures et d'autres fois des journées de 28 heures. Ce champ devient donc le nombre de secondes s'étant écoulées au moment de la transaction depuis minuit le premier mars.
id_arret_sdap	Entier	Servira à identifier l'enregistrement de la table SDAP correspondant à l'arrêt d'embarquement correspondant.
arret_sdap	Entier	Servira à identifier l'arrêt d'embarquement correspondant à l'étape d'enrichissement à partir de la table SDAP.
arret_habitude	Entier	Servira à identifier l'arrêt d'embarquement correspondant à l'étape d'enrichissement à partir des habitudes des usagers.
arret_ligne	Entier	Servira à identifier l'arrêt d'embarquement correspondant à l'étape d'enrichissement à partir des habitudes des usagers.
arret_gtfs	Entier	Servira à identifier l'arrêt d'embarquement correspondant à l'étape d'enrichissement à partir de la table GTFS.
coor_x	Entier	La coordonnée x correspondant soit à l'arrêt SDAP, arrêt habitude ou arrêt ligne avec le système de référencement géographique NAD 27 / MTM zone 8.
coor_y	Entier	La coordonnée y correspondant soit à l'arrêt SDAP, arrêt habitude ou arrêt ligne avec le système de référencement géographique NAD 27 / MTM zone 8.
coor_x_gtfs	Entier	La coordonnée x correspondant à l'arrêt GTFS avec le système de référencement géographique NAD 27 / MTM zone 8.
coor_y_gtfs	Entier	La coordonnée y correspondant à l'arrêt GTFS avec le système de référencement géographique NAD 27 / MTM zone 8.
red_ar_der	Entier	Ce champ servira à identifier le nombre de possibilités lors de l'étape de l'enrichissement par les habitudes de l'usager.
red_ar_der_lig	Entier	Ce champ servira à identifier le nombre de possibilités lors de l'étape de l'enrichissement par les habitudes de l'usager avec des contraintes relaxées.

3.2.2 Données SDAP

Les enregistrements du système de décompte automatique des passagers et de positionnement des autobus sont contenus dans deux tables. La table sdap_courses, qui contient le résumé du compte à bord pour chaque course, et sdap_courses_arret, qui contient le détail du décompte automatique de passagers et les données de localisation du véhicule pour chaque passage à un arrêt, pour chaque course. Une course est le parcours d'un autobus le long d'une ligne, pour une seule direction.

Avant d'être incluses dans ces deux tables, les données ont été validées par le RTL pour éliminer les données erronées. Lorsque les données GPS semblaient erronées, les enregistrements de l'autobus pour le « trip » qu'il effectuait ont été supprimés.

La table sdap_courses compte 75 298 enregistrements et les champs suivants.

Tableau 3.3 : Champs de la table sdap_courses

nom du champ	Format	Description
sdap_date	Aaaammjj	La date de l'enregistrement.
Ligne	Entier	La ligne qui était parcourue par l'autobus au moment de l'enregistrement.
Voiture	Entier	Pièce de travail, soit l'identification d'une combinaison de lignes empruntées par un autobus. La même voiture peut correspondre à plusieurs ligne à différents moments de la journée.
hre_pre_de	hh:mm:ss	L'heure prévue de départ au premier arrêt de la ligne.
hre_reel_d	hh:mm:ss	L'heure réelle de départ au premier arrêt de la ligne.
hre_reel_a	hh:mm:ss	L'heure réelle d'arrivée au dernier arrêt de la ligne.
montants	Entier	Le nombre de passagers qui sont montés dans le bus au cours de la course.
descendant	Entier	Le nombre de passagers qui sont descendus du bus au cours de la course.
charge_max	Entier	La charge maximale au cours de la course.
pos_ch_max	Entier	Le chaînage au point où la charge maximale a été atteinte.
arr_ch_max	Entier	L'arrêt où la charge maximale a été atteinte.
no_bus	Entier	Le numéro de l'autobus.
assignatio		Date du changement du réseau. Pour mars 2013, l'assignation est 20130107 pour tout le mois.

Pour cette recherche, la seule information de cette table qui est pertinente est le numéro d'autobus qui ne se retrouve pas dans la course sdap_courses_arret. Le lien entre les deux tables est décrit ci-dessous.

La table sdap_courses_arret compte 2 193 282 enregistrements et les champs suivants.

Tableau 3.4 : Champs de la table sdap_courses_arret

nom du champ	Format	Description
sdap_date	Aaaammjj	La date de l'enregistrement du passage de l'autobus à l'arrêt.
Ligne	Entier	La ligne qui était parcourue par l'autobus au moment de l'enregistrement.
Voiture	Entier	Pièce de travail, soit une combinaison de lignes empruntées par un autobus. La même voiture peut correspondre à plusieurs lignes à différents moments de la journée.
hre_pre_de	hh:mm:ss	L'heure prévue de départ du premier arrêt de la ligne.
Position	Entier	Le chaînage auquel devrait être rendu l'autobus selon l'arrêt qui est attribué à cette position GPS.
hre_reel_a	hh:mm:ss	L'heure réelle d'arrivée à l'arrêt.
hre_reel_d	hh:mm:ss	L'heure réelle de départ de l'arrêt.
Duree	Entier	La durée de l'arrêt en secondes. Cela correspond à la différence de temps entre les deux champs précédents.
montants	Entier	Le nombre de passagers qui sont montés à l'arrêt.
descendant	Entier	Le nombre de passagers qui sont descendus à l'arrêt.
Charge	Entier	Le nombre de passagers à bord de l'autobus suite à cet arrêt, selon le système de décompte automatique de passagers = charge à l'arrêt précédent + montants – descendants
pos_reel	Entier	Le chaînage selon l'odomètre de l'autobus.
coord_x_re		La coordonnée x selon le système de référencement géographique NAD 27 / MTM zone 8.
coord_y_re		La coordonnées y selon le système de référencement géographique NAD 27 / MTM zone 8.
no_arret	Entier	Le numéro de l'arrêt qui a été associé à cette localisation.
assignatio	Entier	La date du changement du réseau. Pour mars 2013, l'assignation est 20130107 pour tout le mois.

Les relations entre la table sdap_courses_arret et la table sdap_courses sont illustrées à la figure suivante. Il y a un lien très fort entre les enregistrements (champs sdap_date, Ligne, voiture et hre_pre_de) et il est facile d'extraire les données de sdap_courses pour compléter la table sdap_courses_arret.

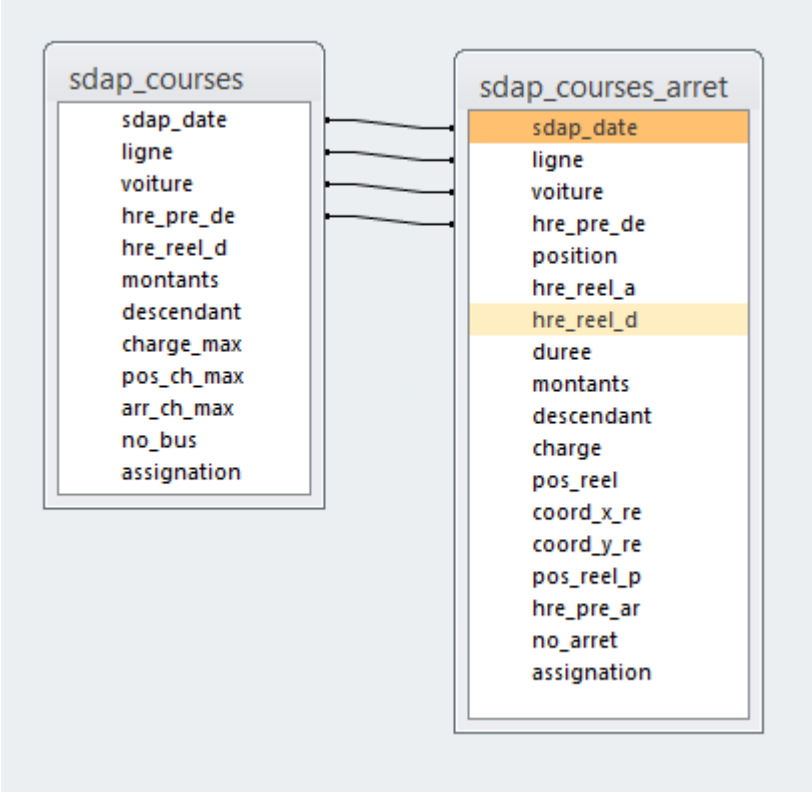


Figure 3-2 : Liens entre les deux tables SDAP

Plusieurs autres champs sont ajoutés à la table sdap_courses_arret pour faciliter les calculs et pour enrichir l'information contenue dans cette table. La table sdap_courses_arret sera appelée SDAP pour le reste de ce document. Les nouveaux champs sont présentés dans le tableau suivant..

Tableau 3.5 : Champs ajoutés à la table SDAP

Nom du champ	Format	Description
id_arret	Entier	Ce champ sert d'identification pour les enregistrements de passage aux arrêts. Il devient la clé primaire de la table.
hre_pre_de_ind	Entier	Ce champ sert à accélérer les calculs en combinant la date et l'heure en un champ et en éliminant les différences de formats. Ce champ devient donc le nombre de secondes s'étant écoulées au moment de l'heure de départ du premier arrêt prévu, depuis minuit le premier mars.
hre_reel_a_ind	Entier	Ce champ sert à accélérer les calculs en combinant la date et l'heure en un champ et en éliminant les différences de formats. Ce champ devient donc le nombre de secondes s'étant écoulées au moment d'ouverture des portes, depuis minuit le premier mars.
hre_reel_d_ind	Entier	Ce champ sert à accélérer les calculs en combinant la date et l'heure en un champ et en éliminant les différences de formats. Ce champ devient donc le nombre de secondes s'étant écoulées au moment de fermeture des portes, depuis minuit le premier mars.
no_bus	Entier	Le numéro d'autobus provenant de la table sdap_courses.
nb_emb_cap	Entier	Ce champ servira à identifier le nombre d'embarquements CAP qui correspondent à chaque passage.

3.2.3 Données GTFS

Les données GTFS (*Google Transit Feed Specification*) sont des données fournies à Google par les sociétés de transport en commun. Elles permettent de représenter le service qui est offert et permettent entre autres au site *Google Transit* de calculer les itinéraires en transport en commun. Dans le cadre de cette recherche, elles sont utilisées pour connaitre la configuration du réseau et les horaires à l'arrêt sur chaque ligne. Les données utilisées se retrouvent dans trois tables : la table gtfs_stop, la table gtfs_stop_times et la table gtfs_trip. Les trois tables sont combinées pour obtenir une table qui se compare plus facilement à la table sdap_courses_arret et qui est plus facile à utiliser dans l'algorithme.

La table gtfs_stop contient l'identification et le positionnement des arrêts. Elle contient 3246 arrêts. Les champs sont les suivants.

Tableau 3.6 : Champs de la table gtfs_stop

nom du champ	Format	Description
stop_id	Entier	Le numéro de l'arrêt.
stop_name		Le nom de l'arrêt (nom des deux rues de l'intersection, nom de la rue et le numéro civique).
stop_lat		La latitude de l'arrêt.
stop_lon		La longitude de l'arrêt.

Les coordonnées géographiques sont difficiles à manipuler. Elles sont donc transformées au système de référencement géographique NAD 27 / MTM zone 8 pour faciliter les calculs. Le logiciel Quantum GIS a été utilisé pour faire cette opération.

La table gtfs_trip contient les détails de chaque course qui sera parcourue par un autobus pour chaque type de jour. Il y a donc quatre horaire complet, soit un pour les jours de semaine, un pour les samedis, un pour les dimanches et un pour le vendredi 29 mars qui est férié. La table contient 11 577 enregistrements de courses différentes.

Tableau 3.7 : Champs de la table gtfs_trip

Nom du champ	Format	Description
route_id	rr ou rrr	L'identification de la ligne.
service_id	Jj	Le type de jour de la semaine (SE, DI, SA, F1).
trip_id	rr_t_d_jj_vvvv_hh_mm	L'identification complète de la course.
direction_id	D	La direction de la course sur la ligne empruntée.
block_id	vvvv_s	La voiture et l'ordre de ses différentes courses.
shape_id	rr_t_d	Le numéro du tracé, de la direction et de la ligne empruntée.

La table gtfs_stop_times contient les moments de passage à chaque arrêt pour chaque course de la table précédente. Elle contient 404 339 enregistrements.

Tableau 3.8 : Champs de la table gtfs_stop_times

nom du champ	Format	Description
trip_id	rr_s_d_jj_vvvv_hh_mm	L'identification de la course.
arrival_time	hh:mm:ss	L'heure d'arrivée à l'arrêt.
departure_time	hh:mm:ss	L'heure de départ de l'arrêt. Presque toujours le même que le temps d'arrivée.
stop_id		L'identification de l'arrêt.

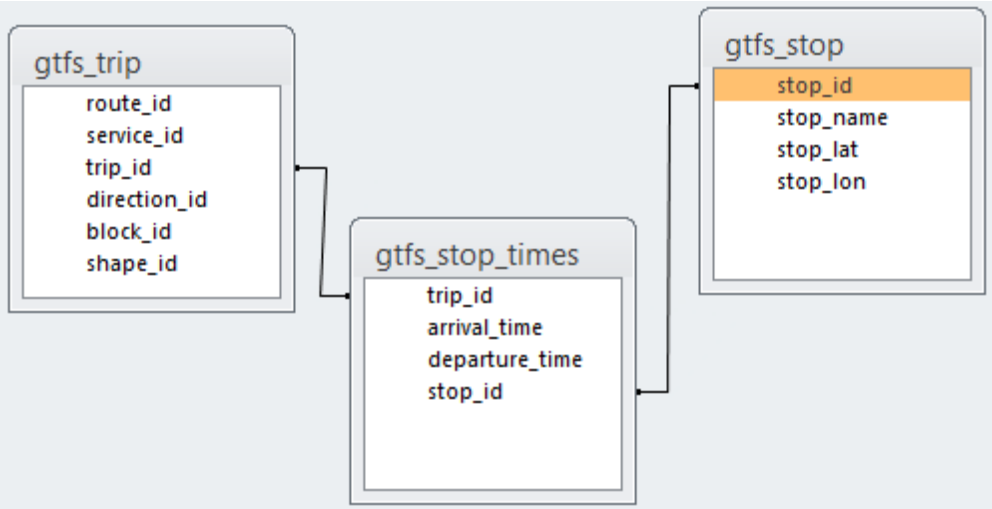


Figure 3-3 : Modèle relationnel entre les tables GTFS

À partir des trois tables, il a été possible de créer la table GTFS, contenant les champs présentés dans le Tableau 3.9 à partir du modèle relationnel de la Figure 3-3. L'information contenue dans les trois tables GTFS est répétée pour chaque jour du mois. Chaque enregistrement correspond au passage prévu d'un autobus à un arrêt sur une ligne à un moment unique du mois de mars 2013. La table contient 4 118 126 enregistrements.

Tableau 3.9 : Champs de la table GTFS

Nom du champ	Format	Description
indextemps	Entier	Ce champ sert à accélérer les calculs en combinant la date et l'heure en un champ et en éliminant les différences de formats. Ce champ devient donc le nombre de secondes au moment d'arrivée prévue de l'autobus à l'arrêt, depuis minuit le premier mars.
arret	Entier	Le numéro de l'arrêt.
ligne	Entier	Le numéro de la ligne.
voiture	Entier	Le numéro de la voiture.
trip	Entier	Le numéro de la course.
date	Entier	La date du passage prévu.
sec	Entier	Le nombre de secondes qui se seront écoulées entre l'heure du passage prévu et minuit la même journée. (calculé à partir du arrival_time)
hre	Entier	L'heure du passage prévu.
trace	Entier	Le tracé de la ligne qui sera emprunté pour cette course.
dir	Entier	La direction de l'autobus sur la ligne pour cette course.

3.2.4 Relation entre les tables CAP, SDAP et GTFS

Cette recherche repose en grande partie sur les relations qui existent entre les tables CAP, SDAP et GTFS. Afin de les comparer entre elles, il est important de connaître quels champs devraient correspondent entre les différentes tables. Il en va de même pour pouvoir enrichir la table CAP grâce aux arrêts qui se retrouvent dans les deux autres tables.

La Figure 3-4 illustre les relations qu'il y a entre les tables CAP et SDAP. Lorsqu'un embarquement CAP est associé à un passage à un arrêt SDAP, tous ces liens existent entre les deux enregistrements. Le moment de l'embarquement CAP est dans l'intervalle entre le moment d'arrivée et de départ à l'arrêt SDAP. Les deux champs dans l'encadré sont les données qui proviendront de la table SDAP.

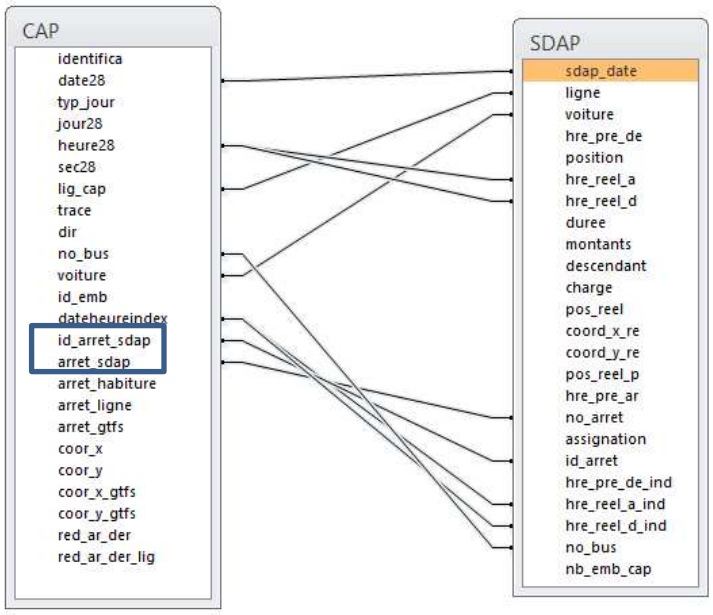


Figure 3-4 : Relation entre les tables CAP et SDAP

La Figure 3-5 illustre les relations qu'il y a entre les tables CAP et GTFS et entre les tables GTFS et SDAP. Lorsqu'un embarquement CAP est associé à un passage prévu GTFS, tous les liens représentent des données qui seront communes entre les deux enregistrements. Il en va de même lorsqu'un passage SDAP et un passage GTFS seront associés au même évènement. Le champ dans l'encadré est pour les données qui seront importées depuis la table GTFS.

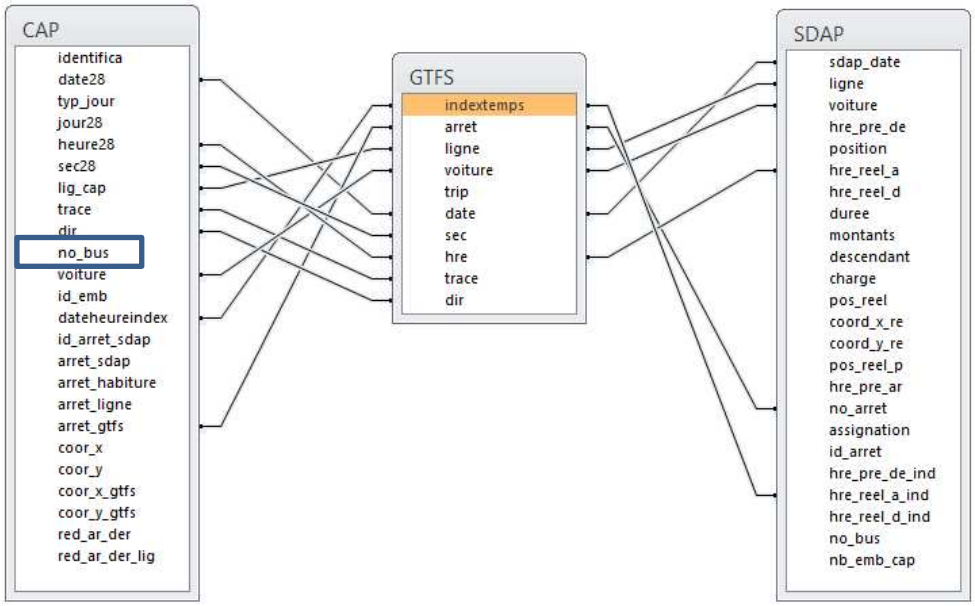


Figure 3-5: Relation entre les tables CAP, SDAP et GTFS

3.3 Logiciels

Trois logiciels sont utilisés dans le cadre de ce projet : SQL Server de Microsoft, Excel de Microsoft et Quantum GIS (QGIS).

Microsoft SQL Server est un logiciel de système de gestion de base de données. Il est l'outil principal utilisé pour faire les manipulations des tables et exécuter l'algorithme. Celui-ci permet de traiter le grand volume de données de l'étude. Le but est de proposer au RTL une série de requêtes SQL Server qui permettront de traiter l'entièreté des données du RTL dans un délai relativement cours et de façon transparente.

Le tableur Microsoft Excel est utilisé pour analyser de petits échantillons des tables et pour tester les résultats. La facilité d'exécution de ce chiffrier accélère les manipulations et permet de voir rapidement les possibilités de traitement sur l'ensemble des données avec SQL Server. Excel permet aussi de faire une analyse descriptive des données et de les illustrer.

Quantum GIS est un logiciel de systèmes d'information géographique (SIG) pour représenter et analyser des données ayant un aspect géographique. Ce logiciel est utilisé pour comprendre le territoire et le réseau du RTL, pour valider les résultats de correspondance entre les embarquements et les arrêts et pour illustrer les utilisations qui peuvent être faites avec les données de la RTL. Quantum GIS est aussi utilisé pour convertir le système de référencement spatial de certaines données.

3.4 Algorithme en SQL Server

3.4.1 Indice de temps

Comme vu dans les tableaux précédents, les formats de temps sont différents entre les trois tables. Afin d'obtenir des indices de temps qui seront facile à comparer, toutes les références au temps sont transformées en nombre de secondes écoulées au moment de l'événement, depuis minuit le premier mars. Cette date du premier mars est une référence pouvant être facilement modifiée au besoin, l'important étant d'utiliser la même référence dans toutes les tables.

Pour la table CAP, la table comprend la donnée du nombre de secondes qui se sont écoulées durant la journée avant l'enregistrement de l'embarquement (sec28) et la date sous le format AAAAMMJJ (exemple 20130317 pour le 17 mars 2013) de l'enregistrement (date28). Il est facile d'obtenir l'indice de temps avec la formule suivante. Le résultat est en seconde.

$$(date28 - 20130301) \times 86400 + sec28$$

Pour l'indice de temps des tables SDAP et GTFS, le format des données sont les mêmes et le calcul se fait de la même manière. Les deux tables possèdent la date en format AAAAMMJJ (sdap_date pour SDAP), (date pour GTFS). Pour sa part, l'heure est en format HH:MM:SS et le champ est en format texte et non en entier (hre_pre_de, hre_reel_a, hre_reel_d pour SDAP), (arrival_time pour GTFS). Il faut donc séparer le texte pour calculer le nombre de secondes s'étant écoulées au cours de la journée. La formule suivante montre la manipulation qui doit être faite pour obtenir les indices de temps pour ces deux tables.

$$(date - 20130301) \times 86400 + left\ (heure, 2) \times 3600 + substring\ (heure, 4{,}2) \times 60 + right(heure, 2)$$

3.4.2 Ajout du numéro d'autobus à la table SDAP

Pour ajouter le numéro du bus à la table SDAP à partir de la table sdap_courses, une correspondance est faite entre les deux tables avec les contraintes du tableau suivant.

Tableau 3.10 : Contrainte pour l'ajout du numéro d'autobus

sdap_courses.sdap_date=sdap.sdap_date
sdap_courses.ligne = sdap.ligne
sdap_courses.voiture = sdap.voiture
sdap_courses.hre_pre_de = sdap.hre_pre_de

3.4.3 Enrichissement de la table CAP par les arrêts de la table SDAP

Cette étape de l'algorithme se fait en quatre sous-étapes. Afin de trouver un arrêt aux embarquements de la table CAP, des correspondances sont effectués entre la table CAP et la table SDAP. Les contraintes varient à chaque sous-étape, mais le principe reste le même. Les contraintes de temps se relaxent d'une sous-étape à l'autre alors que dans certains cas, on accepte des embarquements s'enregistrant 30 secondes avant l'ouverture des portes et 45 secondes après leur fermeture. Ainsi, les lectures de cartes se faisant après la fermeture des portes sont intégrées et on prévoit la possibilité qu'il y ait un léger décalage entre les horloges des deux systèmes. Des contraintes de durée de l'ouverture des portes sont utilisées pour s'assurer que les embarquements correspondent à des passages où l'autobus s'est arrêté. Des contraintes de chaînage sont aussi utilisées pour s'assurer que l'autobus soit situé à un certain endroit sur la ligne. Il y a seulement la contrainte du numéro d'autobus qui reste constante (nous supposons donc que cette information est valide dans les deux tables). Le fait d'avoir plusieurs étapes permet d'éviter les correspondances erronées tout en obtenant un grand éventail de correspondances. Le numéro d'identification de l'enregistrement SDAP et le numéro de l'arrêt sont ajoutés à la table CAP lorsqu'une correspondance est établie.

Pour la première sous-étape, le but est d'aller chercher les embarquements qui se font en début de ligne. Il y a donc une contrainte de chaînage qui favorise les correspondances en début de ligne. Ainsi, on évite qu'un embarquement soit attribué à la fin d'une ligne, qui est souvent très près dans le temps du début de la ligne (dans le cas d'un retour sans battement, par exemple). Pour cette sous-étape, la contrainte de temps est assez large. Des correspondances de 10 secondes avant l'ouverture des portes et de 15 secondes après la fermeture des portes sont acceptées. Comme il arrive que certains passages soient très proches dans le temps lorsque l'autobus ne s'immobilise pas aux arrêts, une contrainte d'ouverture des portes d'une durée plus grande que zéro est aussi ajoutée. Voici les contraintes pour cette première étape.

Tableau 3.11 : Contraintes pour la première étape de la correspondance entre CAP et SDAP

cap.no_bus = sdap.no_bus
cap.dateheureindex +10>= sdap.hre_reel_a_ind
cap.dateheureindex -15 <= sdap.hre_reel_d_ind
sdap.duree > 0
sdap .position < 50

À la deuxième étape, les matchs idéaux sont recherchés. Il est donc assumé que l'enregistrement de l'embarquement se fait après l'ouverture des portes au plus tôt et dans un très court délai (dix seconde) après la fermeture des portes au plus tard. Il n'y a pas de contraintes de chaînage à cette sous-étape. Les contraintes pour cette sous-étape sont les suivantes.

Tableau 3.12 : Contraintes pour la deuxième étape de la correspondance entre CAP et SDAP

cap.no_bus = sdap.no_bus
cap.dateheureindex >= sdap.hre_reel_a_ind
cap.dateheureindex -10 <= sdap.hre_reel_d_ind
sdap.duree > 0

Pour la troisième sous-étape, on recherche les décalages possibles entre les horloges des deux systèmes. Par exemple, si l'horloge du système de cartes à puce est quelques secondes en avance sur l'horloge du système de décompte automatique de passagers, il est fort possible que l'enregistrement CAP soit à l'extérieur de l'intervalle d'ouverture de porte SDAP. Les contraintes sont les mêmes qu'à la sous-étape précédente, mais la contrainte de temps est plus large de 30 secondes.

Tableau 3.13 : Contraintes pour la troisième étape de la correspondance entre CAP et SDAP

cap.no_bus = sdap.no_bus
cap.dateheureindex +15>= sdap.hre_reel_a_ind
cap.dateheureindex -25 <= sdap.hre_reel_d_ind
sdap.duree > 0

Pour la quatrième sous-étape, l'objectif est d'aller chercher des décalages d'horloge plus grands et des erreurs au niveau de la durée de l'ouverture des portes. Il n'y a donc pas de contraintes de durée et la contrainte de temps est encore plus large.

Tableau 3.14 : Contraintes pour la quatrième étape de la correspondance entre CAP et SDAP

cap.no_bus = sdap.no_bus
cap.dateheureindex +30>= sdap.hre_reel_a_ind
cap.dateheureindex -45 <= sdap.hre_reel_d_ind

L'ajustement des contraintes a été réalisé en comparant les résultats de cet enrichissement et les données du système de comptage automatique de passagers. Comme le numéro de l'enregistrement SDAP a été ajouté aux embarquements CAP correspondants, il est facile de faire le décompte des enregistrements CAP associées à chaque passage SDAP. Par la suite, il est possible de comparer le nombre d'embarquements CAP et le nombre de montants SDAP pour valider les contraintes. Cette validation a permis d'ajuster les intervalles de temps et d'établir la nécessité d'une contrainte de chaînage pour les débuts de ligne pour la première sous-étape.

3.4.4 Enrichissement de CAP par les habitudes des usagers

Cette étape de l'enrichissement de la table CAP se fait à partir de la table CAP elle-même. On considère les données historiques de l'usager comme étant plus fiables que l'utilisation des horaires d'autobus pour trouver les arrêts supplémentaires. Les usagers ont tendance à utiliser peu de nouveaux arrêts dans leurs déplacements quotidiens, tandis que les autobus ne respectent pas parfaitement les horaires.

Le but de cette étape est de trouver un arrêt aux embarquements pour lesquels on n'a pas trouvé de correspondance dans les passages SDAP. Pour ce faire, chaque transaction "orpheline" sera comparée avec les autres embarquements du même utilisateur. Si d'autres embarquements sont semblables à celui-ci et qu'il n'y a pas d'ambiguïté (plusieurs choix possibles), le même arrêt sera attitré à la transaction orpheline. L'enrichissement par habitudes se fait avec deux séries de contraintes, en trois sous-étapes.

Dans la première série, les contraintes sont plus fortes. Le but principal est de trouver les arrêts pour les usagers qui font toujours la même utilisation du transport en commun. En plus de la contrainte de l'usager (numéro de carte), il y a les contraintes d'horaire. La première est par rapport au type de jour. Il est estimé que les gens ont un comportement différent les jours de semaine par rapport aux jours de fin de semaine. Par la suite, il y a les contraintes d'heure d'embarquement. Celles-ci visent à avoir des embarquements qui se passent dans la même heure de la journée. Il y a aussi une contrainte sur la ligne et sur la direction, afin qu'elles soient les

mêmes pour les deux embarquements. Les embarquements concordant représenteront donc un usager prenant la même ligne, dans la même direction, le même type de jour et au cours de la même heure. Les contraintes sont présentées au Tableau 3.15.

Tableau 3.15 : Contraintes pour la première série de l'enrichissement par les habitudes des usagers

cap.identifica = c.identifica	
cap.typ_jour = c.typ_jour	
ou	ceiling(cap.sec28/3600) = round(c.sec28/3600,0)
	floor(cap.sec28/3600) = round(c.sec28/3600,0)
cap.lig_cap = c.lig_cap	
cap.dir = c.dir	

Pour la deuxième série, les contraintes sont plus larges. Il y aura beaucoup plus de possibilités de correspondances entre les embarquements. Le risque d'obtenir un arrêt erroné est traité dans les sous-étapes un et deux. Les contraintes pour cette série sont seulement sur l'usager, la ligne et la direction. Les enregistrements concordants seront donc pour un usager embarquant sur la même ligne dans le même sens.

Tableau 3.16 : Contraintes pour la première série de l'enrichissement par les habitudes des usagers

cap.identifica = c.identifica
cap.lig_cap = c.lig_cap
cap.dir = c.dir

La première sous-étape est d'établir, pour chaque embarquement n'ayant pas d'arrêt, le nombre d'arrêts différents pouvant lui être attitrés. Si un seul arrêt peut être attitré à un embarquement selon les contraintes, alors cet arrêt lui est attitré directement à la sous-étape trois. Si plusieurs arrêts correspondent aux contraintes, il y a un risque d'obtenir un arrêt erroné et il faut décider si les arrêts sont valides à la sous-étape deux.

Il n'est pas possible de déterminer quel est le meilleur arrêt parmi les arrêts trouvés précédemment. La deuxième sous-étape consiste à valider si les différents arrêts pour un embarquement sont situés à proximité. L'hypothèse est qu'il est possible que l'usager ait marché en attendant l'autobus ou qu'il ait le choix entre deux différents arrêts sur la même ligne pour son déplacement. La distance entre les différents arrêts est calculée. Si les distances sont trop grandes, aucun arrêt ne sera attiré à cet embarquement. Si toutes les distances entre les arrêts possibles

sont inférieures à 500 mètres, il est alors estimé que tous les arrêts possibles peuvent être associés à l'embarquement sans impacter la validité des résultats. L'arrêt est peut-être erroné, mais il est sur la bonne ligne et il est donc préférable d'être décalé d'un ou deux arrêts que de ne pas avoir d'arrêt pour l'embarquement.

La troisième sous-étape attribue un arrêt à l'embarquement si un seul arrêt correspond aux contraintes ou si la distance entre les arrêts possibles est inférieure à 500 mètres.

3.4.5 Enrichissement de CAP par la table GTFS

Après avoir enrichi la table CAP avec les arrêts provenant de la table SDAP et les habitudes des usagers, la dernière option pour trouver un arrêt à l'embarquement est d'utiliser la table des arrêts planifiés GTFS. Il y a deux inconvénients majeurs à utiliser ces données. Le premier est qu'il peut y avoir plus d'un autobus effectuant le même parcours en même temps et il n'est donc pas possible de savoir dans quel autobus l'enregistrement a été effectué. Le deuxième inconvénient est que l'autobus ne respecte pas nécessairement l'horaire. Un retard de cinq minutes peut se traduire par une différence de quelques kilomètres entre l'arrêt réel et l'arrêt qui sera attribué par cet enrichissement.

L'enrichissement de la table CAP se fait par une correspondance entre les deux tables avec des contraintes de date, de temps (avec un intervalle de 2 minutes), de ligne, de direction et de voiture. Les contraintes sont illustrées dans le Tableau 3.17.

Tableau 3.17 : Contraintes pour l'enrichissement de CAP grâce à GTFS

gtfs.date24 = cap.date28
gtfs.ligne= cap.lig_cap
gtfs.sec > cap.sec28 – 60
gtfs.sec < cap.sec28 + 60
gtfs.dir = cap.dir
gtfs.voiture = cap.voiture

Afin d'éviter qu'il y est deux véhicules correspondent à ces contraintes, un arrêt sera attribué seulement si toutes les possibilités de passages correspondants font partie de la même course (trip). Pour ce qui est des retards, c'est une source d'erreur que l'on ne peut pas éliminer sans les données de localisation de véhicule correspondante (SDAP). Les arrêts trouvés à l'aide de cet enrichissement sont donc moins fiables que les précédents.

CHAPITRE 4 RÉSULTATS

4.1 Analyse descriptive

Afin de comprendre les données utilisées, de trouver les causes d'erreurs possibles et de vérifier l'intégrité des données, celles-ci ont été d'abord caractérisées table par table, puis les tables ont été comparées les unes aux autres.

4.1.1 Caractérisation des données CAP

La table CAP est la table où sont enregistrées les transactions de cartes à puce. Seules les transactions valides sont conservées dans cette table. Chaque ligne correspond à l'embarquement d'un passager et contient l'information de la carte, de l'autobus, de la ligne, de la transaction ainsi que l'heure de la transaction. Les données du mois de mars 2013 contiennent le nombre d'éléments suivant.

Tableau 4.1 : Éléments de la table CAP

Nombre d'enregistrements	**2 481 977**
Nombre de bus différents	**407**
Nombre de lignes	**153**
Nombre de tracés	**617**
Nombre de voitures	**903**
Nombre d'usagers	**103 540**
Nombre de titres différents	**61**

Les enregistrements CAP ont une distribution dans le temps bien caractéristique du transport en commun. Celle-ci est observable dans la Figure 4-1. Il n'y a pratiquement aucun enregistrement la nuit (pas de service). Les activités commencent vers 5:00 le matin. Il y a deux fortes pointes pour les jours de semaine : une le matin, qui correspond aux utilisateurs se rendant au travail et une en début de soirée, qui correspond aux utilisateurs revenant du travail. La pointe

du matin est de 7:00 à 10:00 et celle du soir est de 16:00 à 20:00. Lors de la pointe du matin, la somme des embarquements par heure dépasse 60 000 embarquements pour les jours de la semaine. La pointe du soir est plus étendue. Elle atteint 55 000 embarquements par heure au cours de la 17e heure de la journée, mais elle se prolonge plus longtemps que la pointe du matin. Hors pointe, la somme d'embarquements pour les jours de semaine se situe aux alentours de 15 000 embarquements par heure.

Les jours de fin de semaine sont moins achalandés. Le réseau de la RTL enregistre environ 10 000 embarquements à l'heure pour l'ensemble de la journée. Il n'y a pas de pointe marquée, mais il est notable que l'achalandage augmente en soirée. Le dimanche est légèrement moins achalandé que le samedi.

Dans le graphique ci-dessous, il est notable que le vendredi est moins achalandé que les autres jours de la semaine. Un facteur expliquant ce phénomène est que le vendredi de Pâques était le 29 mars (jour férié).

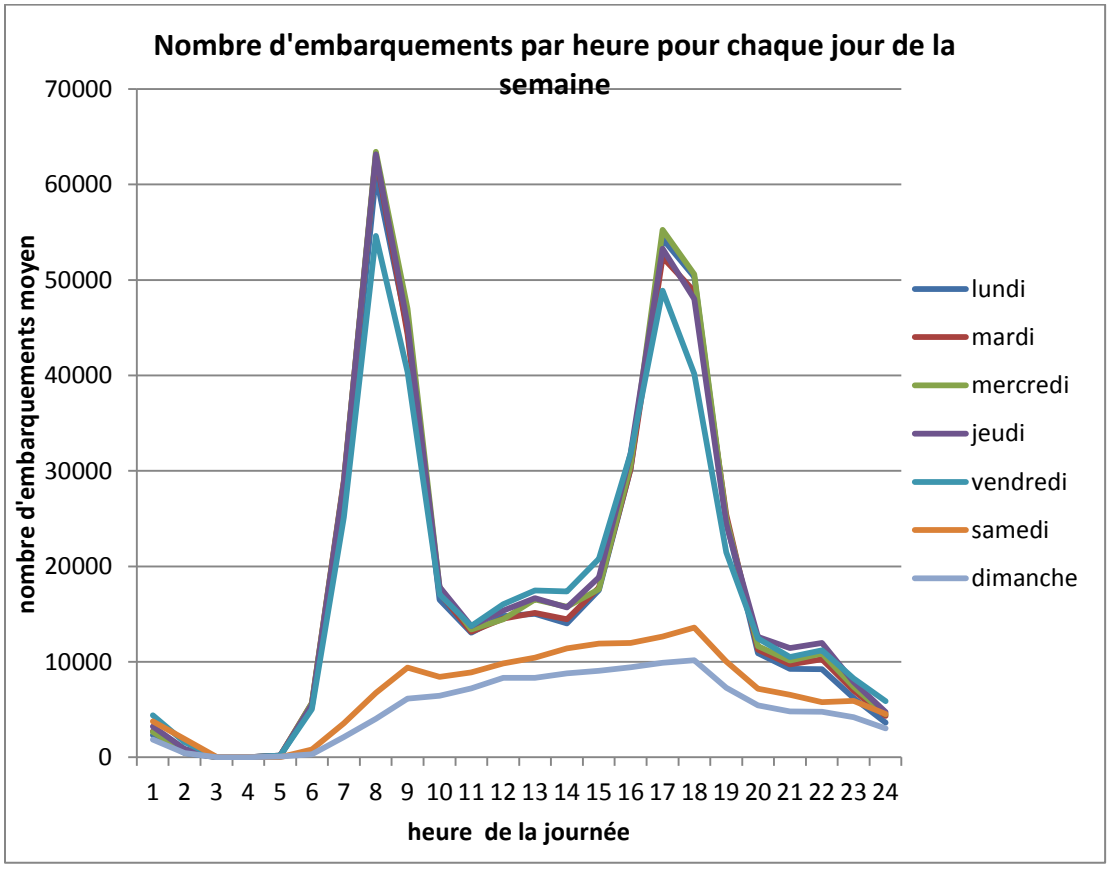


Figure 4-1: Nombre d'embarquements moyen par heure pour chaque jour de la semaine

Grâce aux enregistrements CAP, il est aussi possible de connaître le nombre de bus qui étaient actifs selon le moment de la journée. Pour savoir si un autobus est actif ou non, il suffit d'avoir au moins un enregistrement carte à puce dans ce bus. Par contre, il est possible, mais peu probable, qu'un autobus circule sans qu'aucune transaction de carte à puce ne soit réalisée. Il est donc possible que certains autobus aient été actifs sans qu'ils ne soient représentés dans le graphique de la Figure 4-2.

En regardant ce graphique, il est évidant que la distribution du nombre de bus actifs par heure suit la distribution du nombre d'embarquements par heure. Le nombre de bus capturés est très constant pour les jours de semaine. La seule exception est le vendredi qui voit sa moyenne affectée par la journée du Vendredi Saint. Encore une fois, il y a les deux périodes de pointe du matin et du soir. Lors de la pointe du matin, 350 autobus actifs ont été capturés par le système de cartes à puce. Hors pointe, une moyenne de 100 autobus restent actifs selon le système de cartes à puce. Pour la fin de semaine, il y a beaucoup moins d'autobus actifs. Le samedi, le système cartes à puce a capturé 70 autobus par heure alors qu'il en capturait 55 le dimanche.

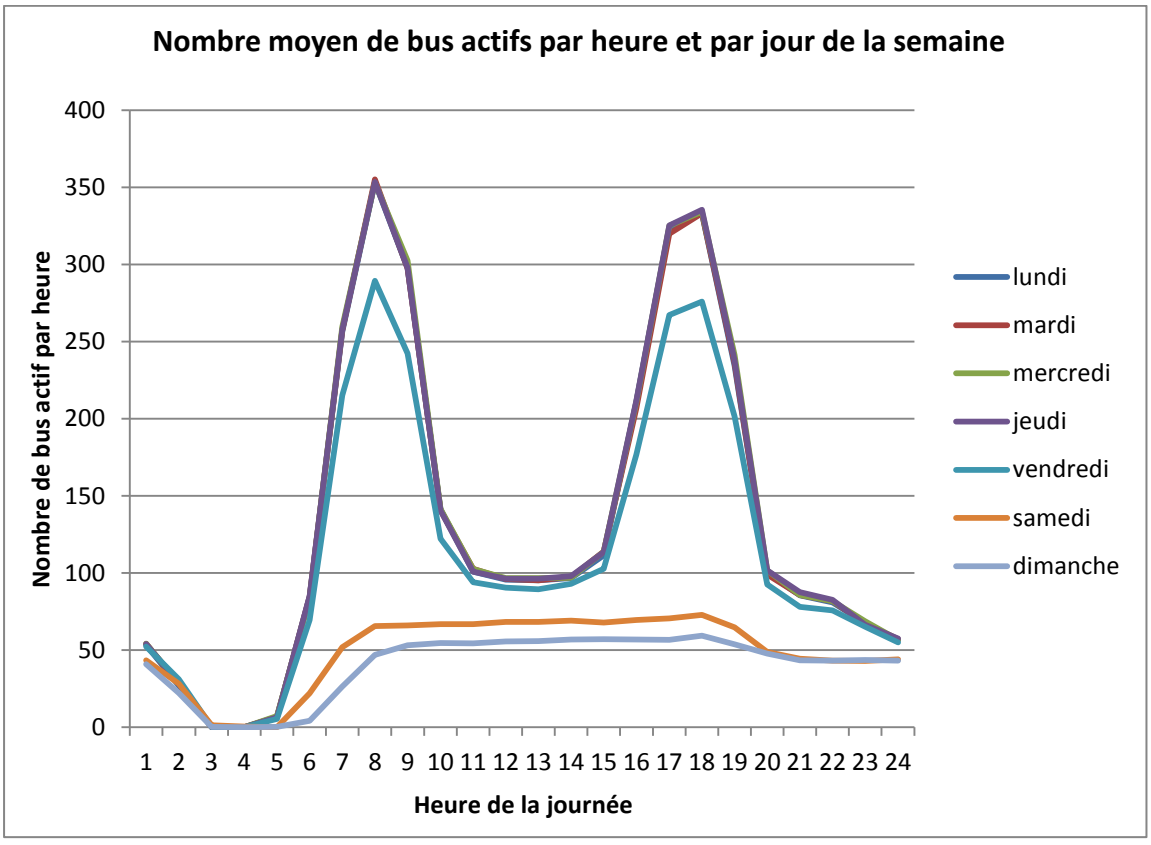


Figure 4-2 : Nombre moyen de bus actifs par heure et par jour de la semaine dans la table CAP

Cette table CAP contient beaucoup d'information sur l'utilisation que les usagers font du transport en commun. Par contre, la partie géographique de l'utilisation n'est pas encore disponible. La table CAP sera enrichie au cours des prochaines étapes.

4.1.2 Caractérisation des données SDAP

La table SDAP contient les informations provenant du système de décompte automatique de passagers. Chaque enregistrement correspond à un passage d'un autobus à un arrêt. Les enregistrements sont effectués lorsque l'autobus passe à un arrêt. Il n'est donc pas obligatoire que l'autobus arrête ou ouvre ses portes. L'enregistrement contient le numéro d'autobus, les détails sur le trajet qui a été effectué par l'autobus, l'heure d'ouverture des portes, le nombre de passagers montants et descendants à l'arrêt, la position GPS, le numéro de l'arrêt et l'heure de fermeture des portes de l'autobus. Seulement une partie des autobus ont un système de compte à bord fonctionnel. Le tableau suivant montre les éléments qui sont contenus dans cette table :

Tableau 4.2 : Éléments de la table SDAP

Nombre d'enregistrements	**2 193 282**
Nombre d'enregistrements avec compte à bord	**1 005 102**
Nombre de bus différents	**351**
Nombre de lignes	**145**
Nombre de voitures	**860**
Nombre d'arrêts (physique)	**3 191**

La Figure 4-3 montre que les enregistrements SDAP ont une distribution dans le temps qui est semblable à la distribution des enregistrements d'embarquement CAP.

Dans le cas des enregistrements de passages SDAP, il y a une activité presque nulle au cours de la nuit. Le nombre de passages par heure au cours des cinq jours de la semaine est distribué de la même manière. Il y a deux périodes de pointe pour chaque journée. Celle du matin

contient de la septième à la neuvième heure et la pointe du soir contient de la 16e à la 19e heure. Lors de la pointe du matin, le système de décompte automatique de passagers a enregistré en moyenne 9 500 passages à des arrêts par heure. Hors pointe, en semaine, le nombre de passages à des arrêts par heure se situe aux alentours de 3 000 passages par heure.

La distribution des passages est différente la fin de semaine. Le nombre de passages enregistrés est constant tout au long des journées. Pour le samedi, le système enregistre 2 400 passages à des arrêts par heure alors qu'il en enregistre 2 000 le dimanche.

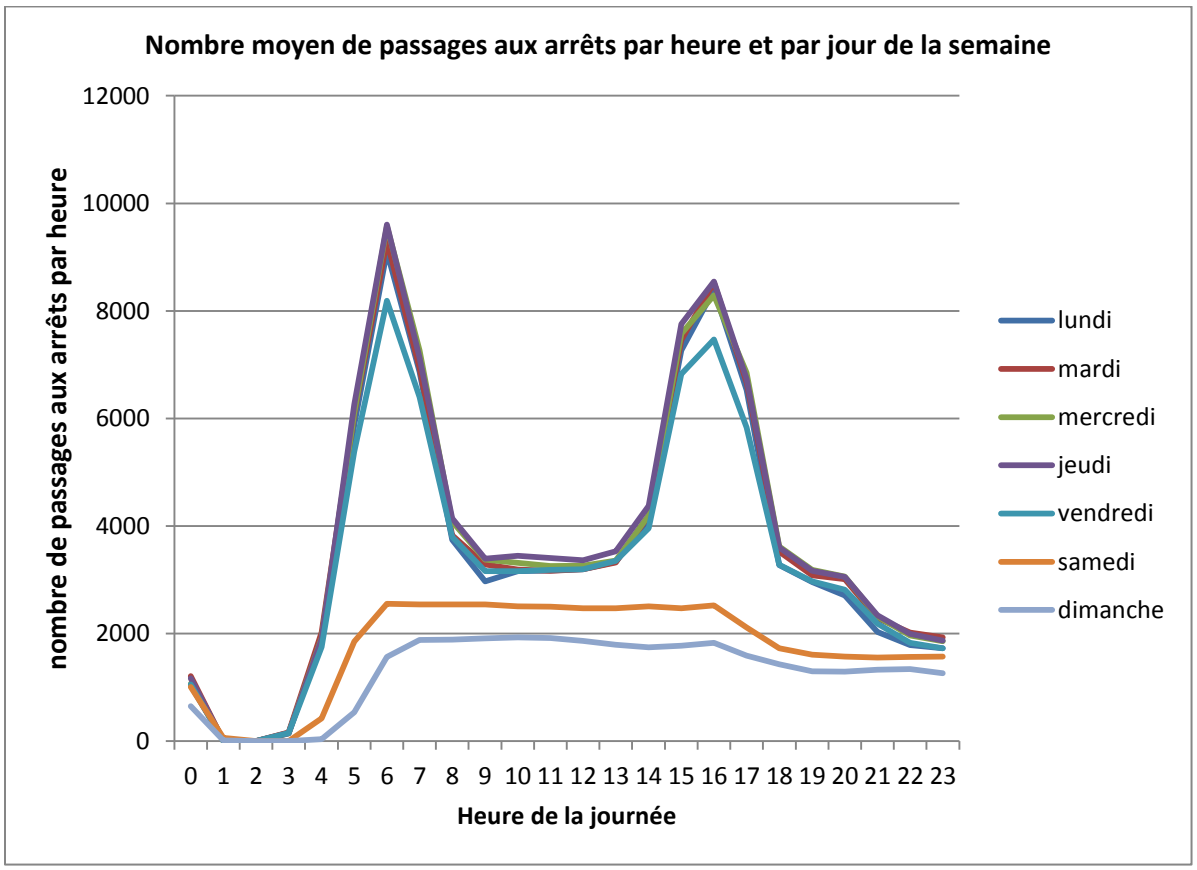


Figure 4-3 : Nombre moyen de passages d'autobus aux arrêts par heure et par jour de la semaine pour la table SDAP

Les enregistrements du système de décompte automatique de passagers permettent aussi de savoir le nombre de bus qui sont actifs selon le moment de la journée. Notons cependant que certains bus du RTL ne sont pas présents dans le système SDAP. La distribution du nombre de bus actifs dans SDAP est présentée par la Figure 4-4. Le nombre de bus actifs par heure selon le système de décompte automatique de passagers est distribué de façon similaire au nombre de

passages aux arrêts. Il y a donc les deux même périodes de pointe le matin et en début de soirée pour les jours de semaine. Lors des périodes de pointe, le nombre d'autobus actifs capturés est de 230 pour le matin et de 210 pour la pointe du soir. Hors pointe, il y a en moyenne 53 autobus actifs par heure pour les jours de la semaine. Il y a moins de bus actifs le vendredi. Cette diminution peut encore une fois être expliquée par le Vendredi Saint. Le samedi et le dimanche ont un nombre de bus actifs par heure constant. En moyenne, 29 autobus actifs ont été capturés les dimanches et 38 les samedis.

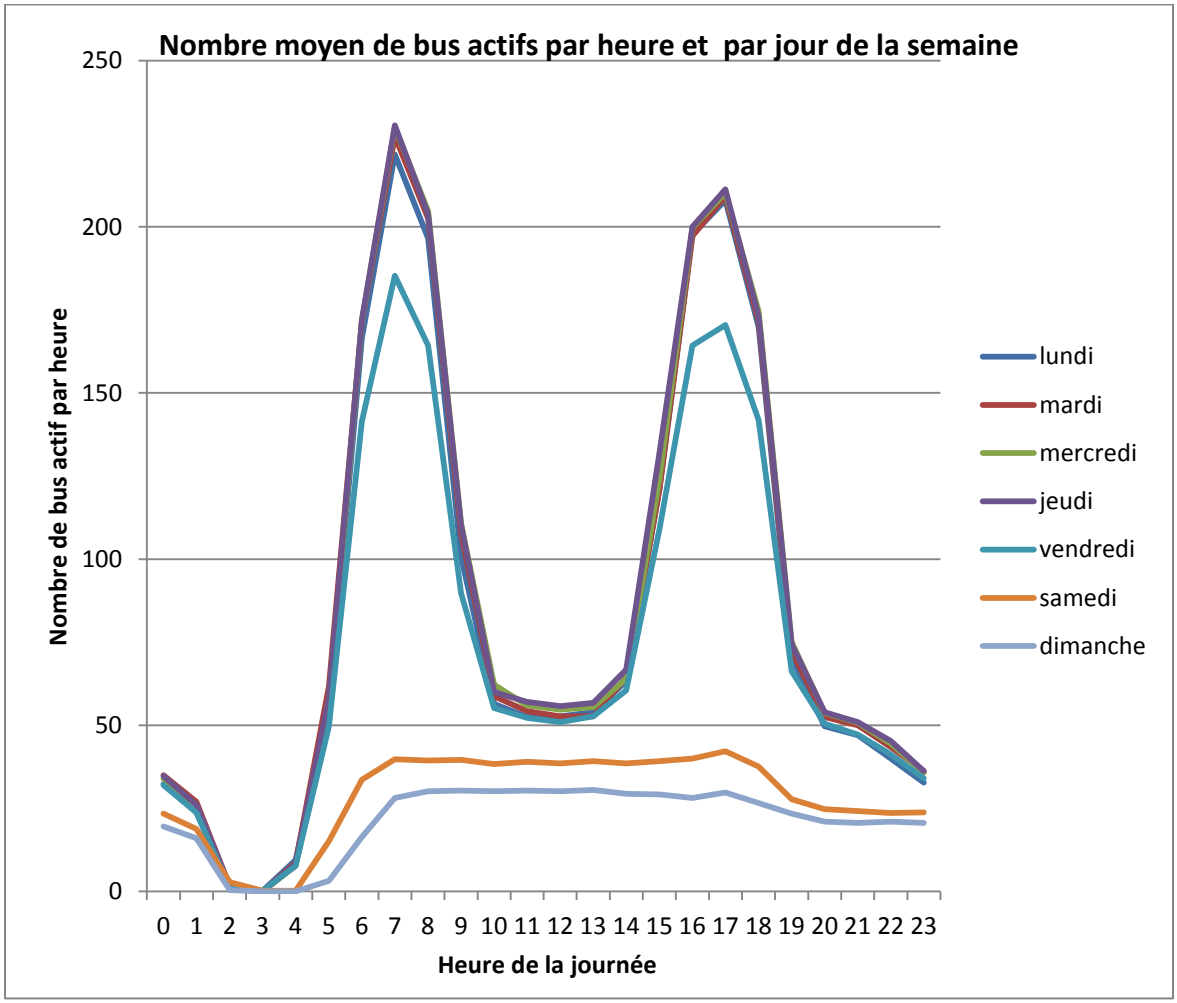


Figure 4-4 : Nombre moyenne de bus actifs par heure et par jour de la semaine pour la table SDAP

Les enregistrements du système de décompte automatique de passagers du RTL est couplé avec un système de localisation des véhicules. Il est donc possible de localiser les enregistrements de la table SDAP. La carte suivante montre où ont été enregistrés les passages des autobus au

cours du mois de mars. La grosseur des points indique le nombre de passages qui ont été faits à chaque arrêt. Les arrêts de la même ligne devraient être de la même grosseur. Il y a une grande concentration de passages près de la station de métro Longueuil-Université-de-Sherbrooke qui donne un accès direct à Montréal. Il y a aussi une grande concentration de passages au terminus Centre-Ville à Montréal (station Bonaventure) et au terminus Panama. Ce sont deux points centraux pour les usagers allant de Longueuil au centre-ville de Montréal où plusieurs lignes passent en plus d'avoir une ligne ne faisant que ces deux arrêts. D'autres arrêts plus éloignés ont enregistré beaucoup moins de passages.

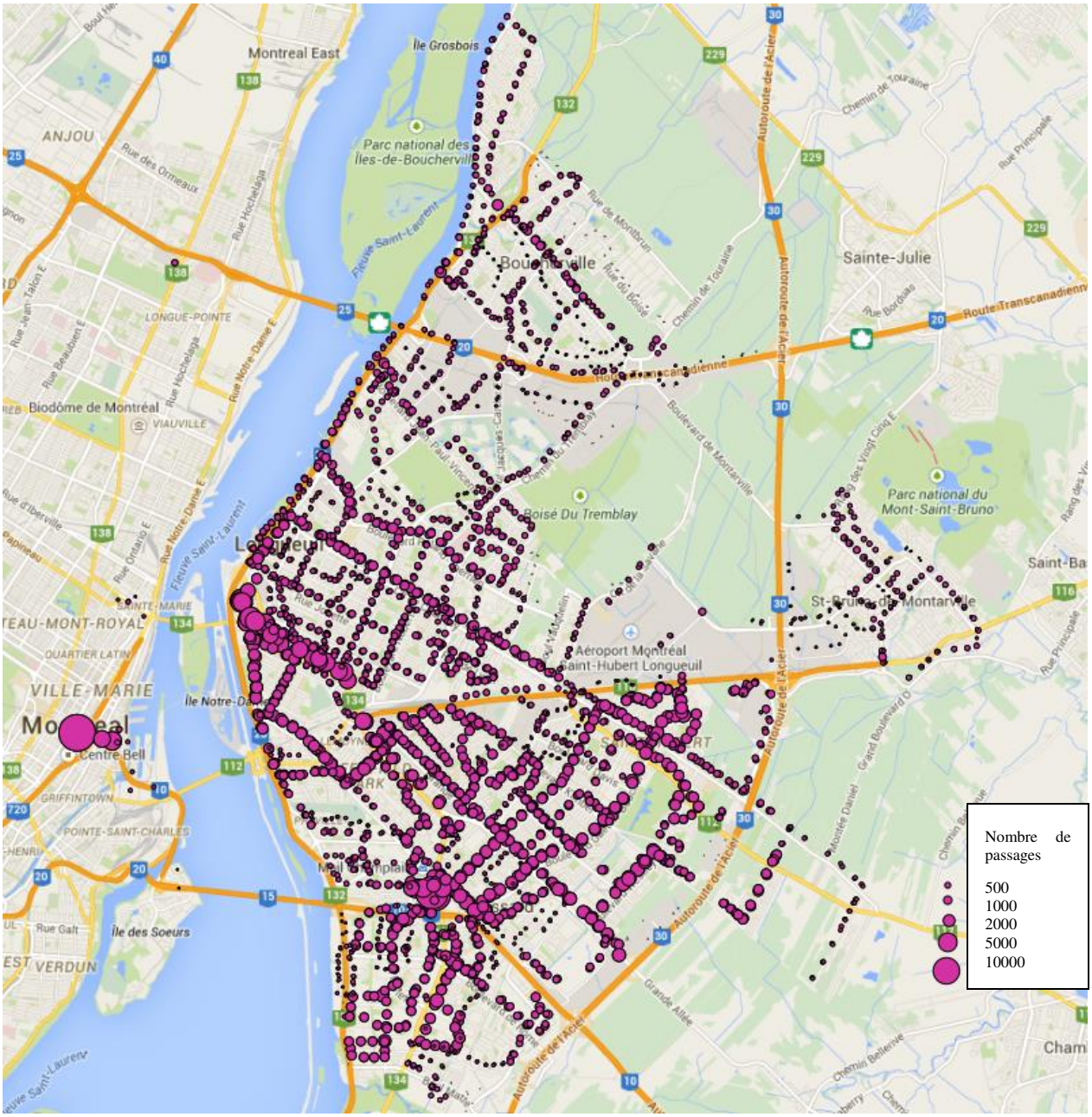


Figure 4-5 : Carte représentant le nombre de passages aux arrêts enregistrés par SDAP

Le comptage de passagers est très représentatif des arrêts importants du réseau. La carte suivante montre le nombre de passagers montants qui ont été enregistrés par le système de décompte automatique de passagers au cours du mois. Il s'agit seulement d'une partie des embarquements réels qui sont représentés car seulement une fraction des autobus sont équipés d'un système de décompte fonctionnel. Il est tout de même possible de remarquer que les embarquements sont distribués sur l'ensemble du réseau et qu'il y a trois points d'embarquement

dominants. Il s'agit du terminus Centre-Ville à Montréal, du terminus Panama et du métro Longueuil-Université-de-Sherbooke. Ces trois points étant des terminus, il est possible d'assumer que la majorité des usagers y embarquant sont des personnes revenant de Montréal en fin de journée et qu'ils débarqueront à l'arrêt prêt de leur domicile.

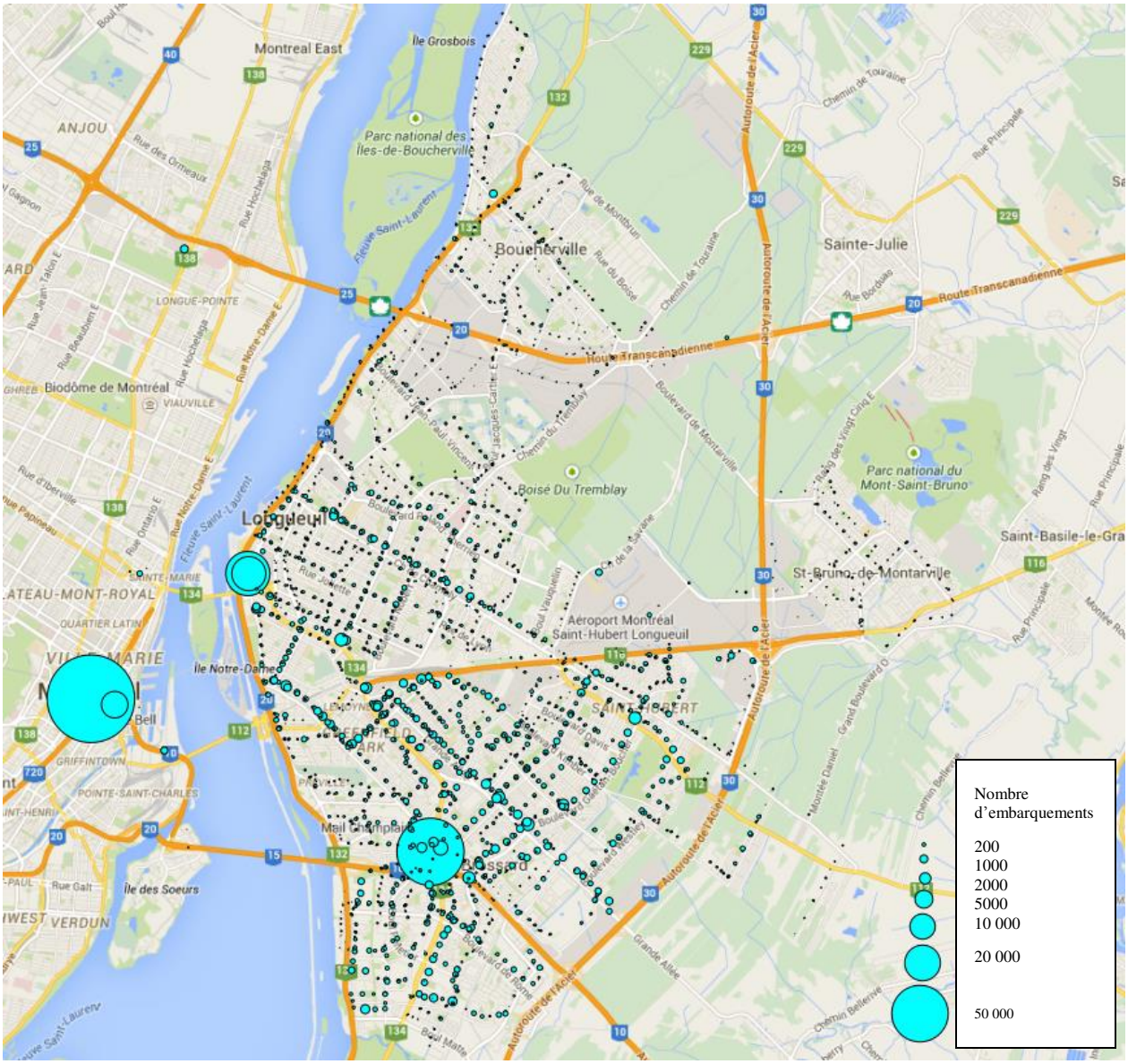


Figure 4-6 : Carte représentant le nombre de passagers embarquant à chaque arrêt selon SDAP

4.1.3 Caractérisation des données GTFS

La table GTFS est une table contenant l'information sur le service offert pour le mois de mars 2013. L'information s'y retrouve sous forme de passages prévus d'autobus aux arrêts. On y

retrouve la position des arrêts, le tracé des lignes, la description des voyages de bus et l'horaire des arrêts. La table contient les éléments suivants.

Tableau 4.3 : Éléments de la table GTFS

Nombre d'enregistrements	**4 118 126**
Nombre de lignes	**92**
Nombre de tracés	**289**
Nombre de voitures	**1 108**
Nombre d'arrêts	**3 245**

La distribution des passages de la table GTFS est présentée à la Figure 4-7. Les quatre types de jours sont illustrés séparément. Pour les jours de semaine, il y a deux périodes de pointe très fortes au cours desquelles plus de 16 000 passages par heures sont prévues. Il y a 6000 passages prévus par heure pour la période hors pointe. Pour la journée fériée, il y a deux petites périodes de pointe à 6200 passages par heure et 4700 passages par heure pour la période hors pointe. Les samedis et dimanches ont des distributions constantes tout au long de la journée avec 4500 passages par heure pour les samedis et 3 600 passages par heure les dimanches.

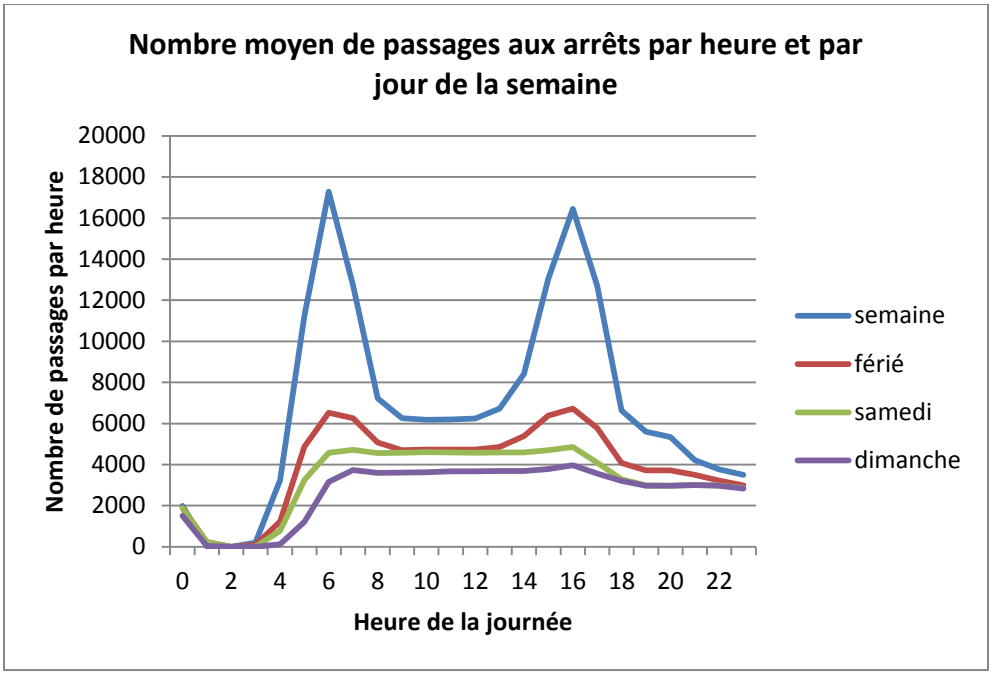


Figure 4-7 : Nombre moyen de passages d'autobus aux arrêts par heure et par jour de la semaine pour la table GTFS

La Figure 4-8 représente le nombre de passages aux arrêts qui est prévu pour le mois de mars 2013. Il est possible de voir les secteurs où le service est plus régulier. Les concentrations d'achalange sont similaires à celles de la table SDAP.

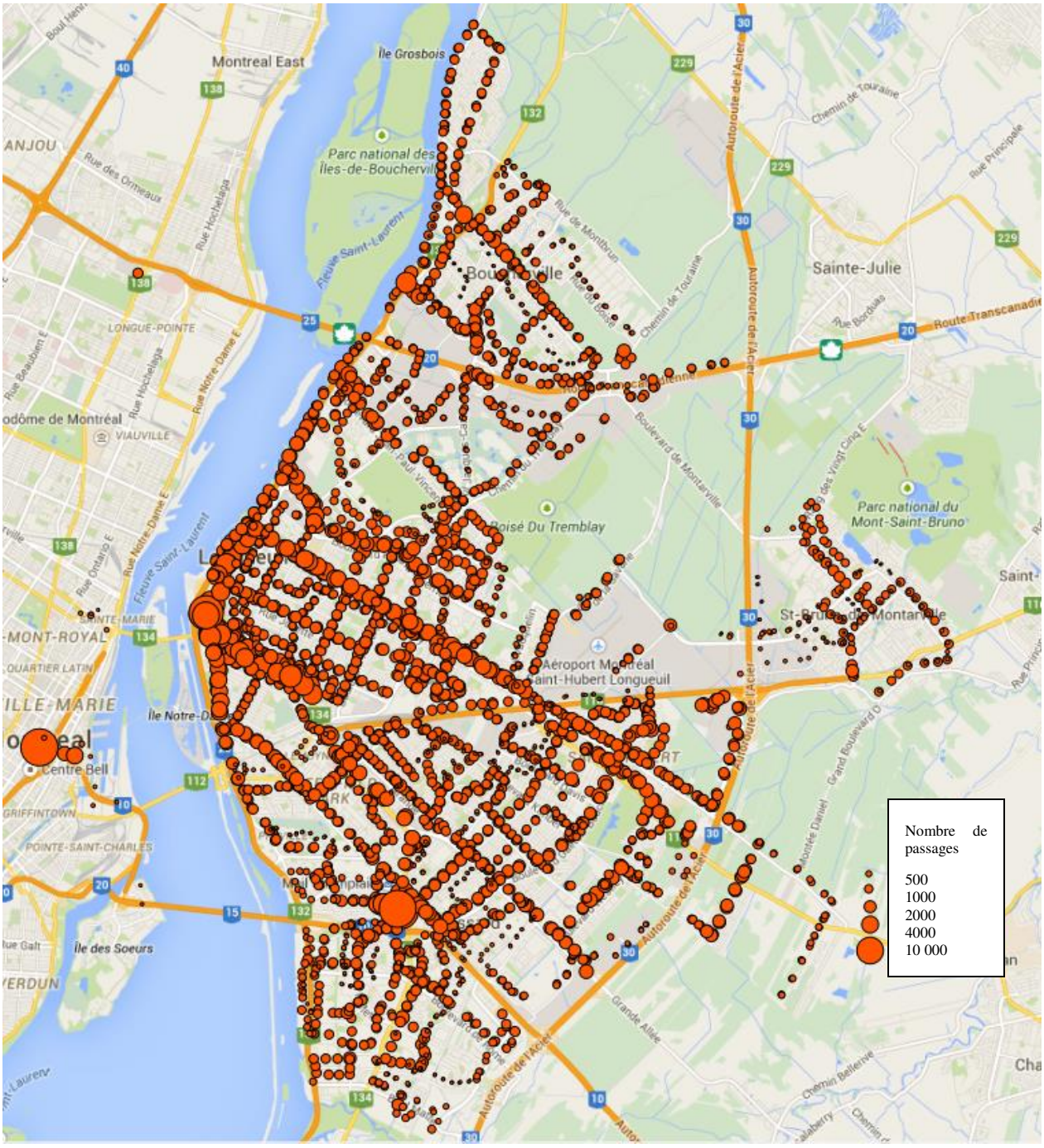


Figure 4-8 : Carte représentant le nombre de passages prévus par arrêt selon GTFS

4.1.4 Comparaison entre les tables CAP et SDAP

Dans cette recherche, les tables CAP et SDAP sont les plus importantes. Il est intéressant d'analyser les éléments qui se recoupent entre les deux pour voir s'il sera possible d'associer un arrêt à chaque embarquement CAP grâce aux passages d'autobus aux arrêts de SDAP.

Comme les deux éléments de jonction principaux sont le temps et les numéros d'autobus, il est intéressant de comparer les autobus actifs de manière agrégée selon le moment de la journée. Comme les enregistrements du système de décompte automatique de passagers devraient contenir chaque passage aux arrêts par les autobus et que le système de cartes à puce nécessite une transaction carte à puce pour valider l'activité d'un autobus, il devrait y avoir plus d'autobus actifs par heure dans la table SDAP que dans la table CAP.

La réalité est tout autre car certains bus du RTL ne sont pas présents dans le SDAP et d'autre bus y sont présents seulement une partie des journées du mois. Le graphique suivant montre le nombre de bus actifs pour CAP et SDAP par heure en moyenne pour les jours de semaine. Il y a une différence très évidente. Dans les périodes de pointe du matin, il y a en moyenne 344 bus actifs dans la table CAP et seulement 227 actifs dans la table SDAP. Pour les périodes hors pointe, 98 autobus sont actifs dans la table CAP tandis qu'il n'y en a seulement 55 dans la table SDAP.

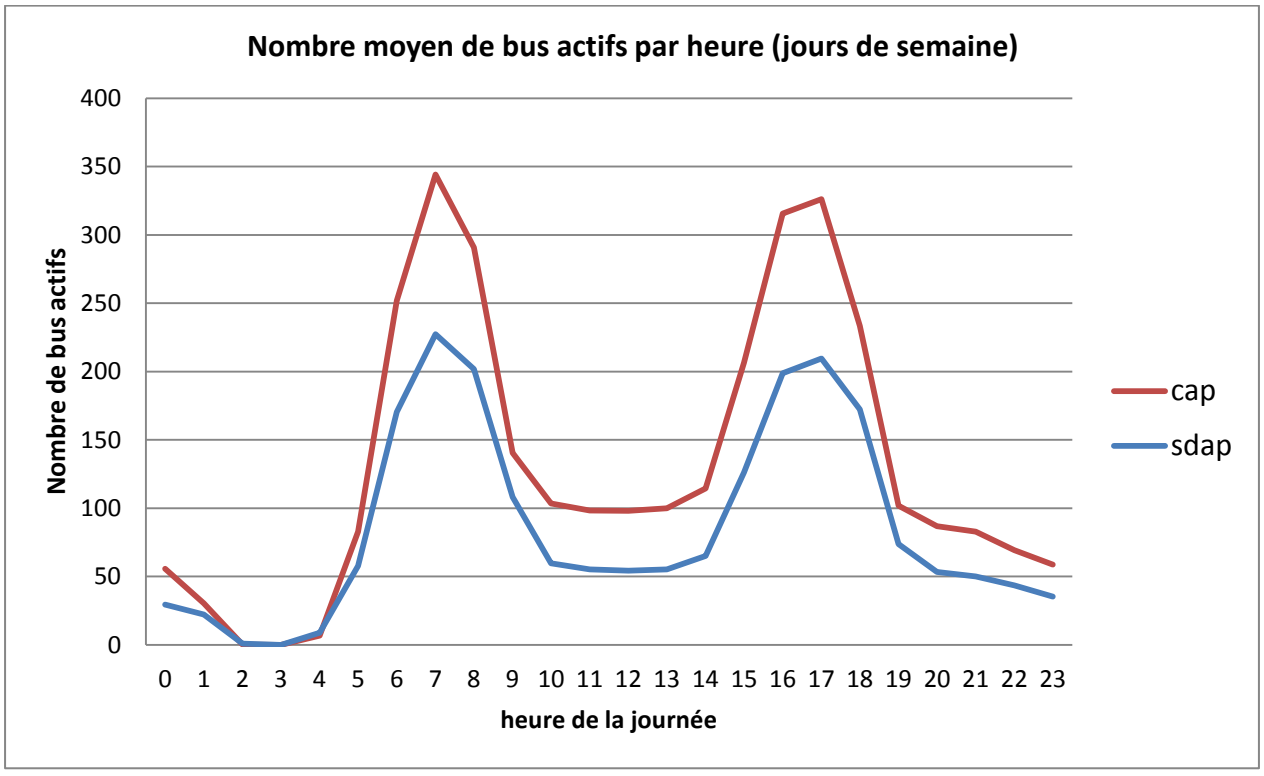


Figure 4-9 : Comparaison entre les tables CAP et SDAP du nombre moyen de bus actifs par heure pour l'ensemble des jours de semaine

De manière plus générale, dans la table SDAP se retrouve 351 numéros de bus différents alors qu'il y en a 407 dans la table CAP. Il manque aussi des journées de données complètes pour certains autres autobus dans la table SDAP.

Cette différence majeure prouve qu'il y a un manque de données importants dans la table SDAP. Il sera donc impossible d'avoir des arrêts pour tous les embarquements CAP à l'aide de la méthode utilisant les passages SDAP seulement.

4.1.5 Comparaison entre la table SDAP et la table GTFS

La table SDAP et la table GTFS contiennent des enregistrements de même nature. Dans les deux cas, ce sont des passages d'autobus aux arrêts. Il est donc facile de voir lesquels sont communs entre les deux tables. Normalement, la table GTFS devrait contenir tous les enregistrements de la table SDAP. Comme la figure suivante le montre, il y a un manque d'information dans les deux cas. Il y a deux fois plus de passages prévus dans GTFS que de passages capturés dans SDAP.

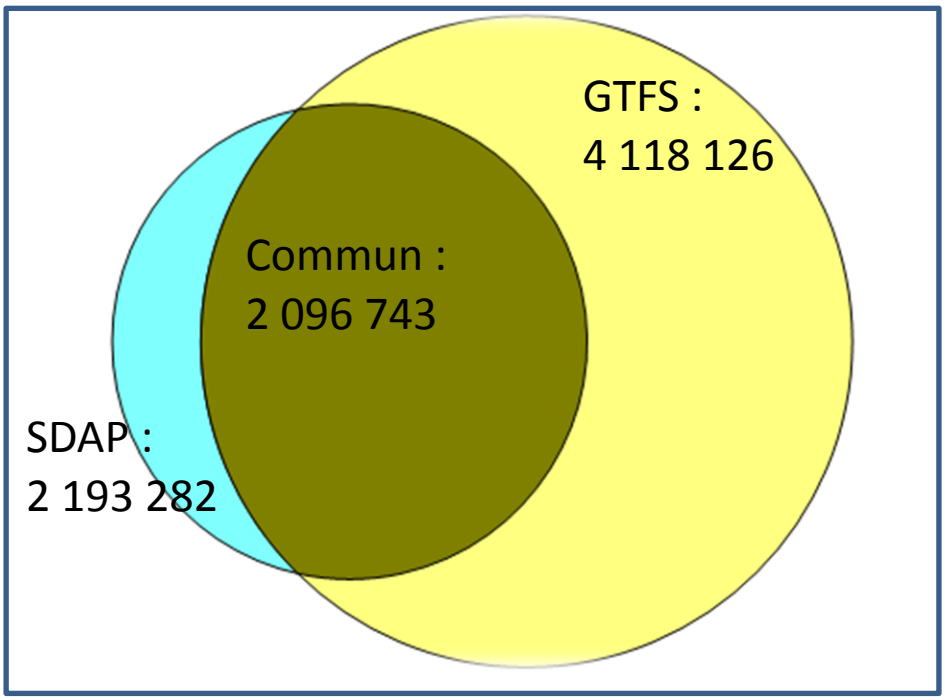


Figure 4-10 : Diagramme de Venn représentant les enregistrements de passages d'autobus prévus dans GTFS communs aux passages capturés dans SDAP

Dans les deux cas, il manque des lignes dans les tables. La table SDAP contient 65 lignes qui ne sont pas dans GTFS. Ces 65 lignes comptent pour 2,4% des passages SDAP. La table GTFS contient 12 lignes qui ne sont pas dans la table SDAP. Ces 12 lignes comptent pour 4,6% des passages prévus dans GTFS. Les lignes manquantes ne sont pas des lignes majeures (ce sont essentiellement des lignes dites scolaires ou des lignes de taxi), mais il y a tout de même un manque de données. Pour les 80 lignes étant communes aux deux tables, il y a en moyenne 44% moins d'enregistrements dans SDAP que dans GTFS. Le graphique de la Figure 4-11 montre la distribution des écarts entre SDAP et GTFS selon l'écart relatif.

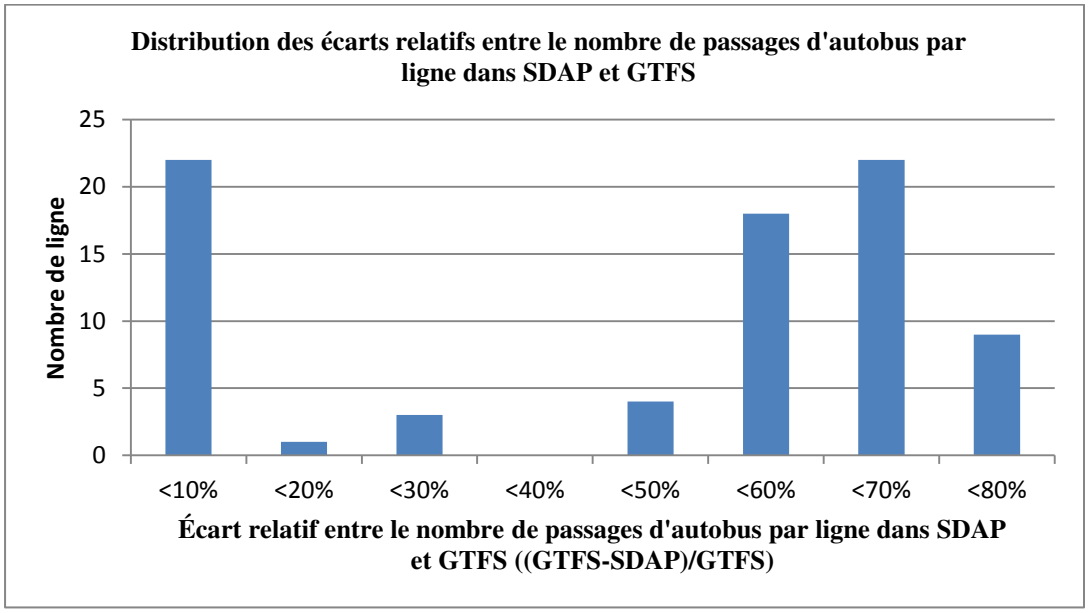


Figure 4-11 : Distribution des écarts relatifs entre le nombre de passages d'autobus capturés par SDAP et le nombre de passages prévus dans GTFS par ligne

Il y a aussi une différence au niveau des arrêts des deux tables. Dans la table GTFS, il y a 3245 arrêts différents où passent les autobus dont 156 qui ne se retrouvent pas dans SDAP. Dans la table SDAP, il y a seulement 3191 arrêts dont 102 sont seulement dans SDAP. Pour les arrêts qui sont compris dans les deux tables, l'écart relatif est de 45,6%. La distribution des écarts est présentée à la Figure 4-12.

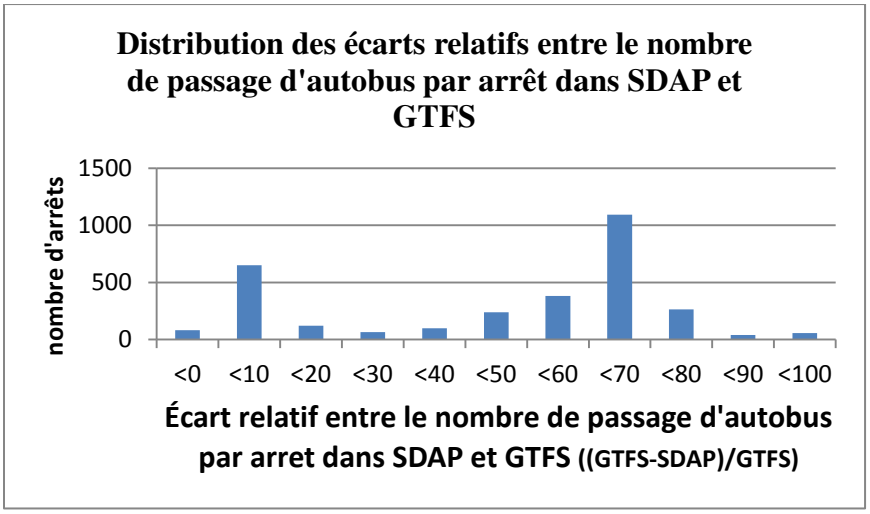


Figure 4-12 : Distribution des écarts relatifs entre le nombre de passages d'autobus capturé par SDAP et le nombre de passages prévus dans GTFS par arrêt

4.2 Résultats de l'algorithme d'attribution des arrêts aux embarquements CAP

Dans cette section, l'objectif est de présenter les résultats de l'algorithme d'attribution d'arrêts. Les résultats sont montrés pour les trois étapes : enrichissement par correspondance aux passages SDAP, enrichissement par correspondance aux habitudes et enrichissement par correspondance aux passages GTFS. Un résumé des résultats est représenté à la Figure 4-13.

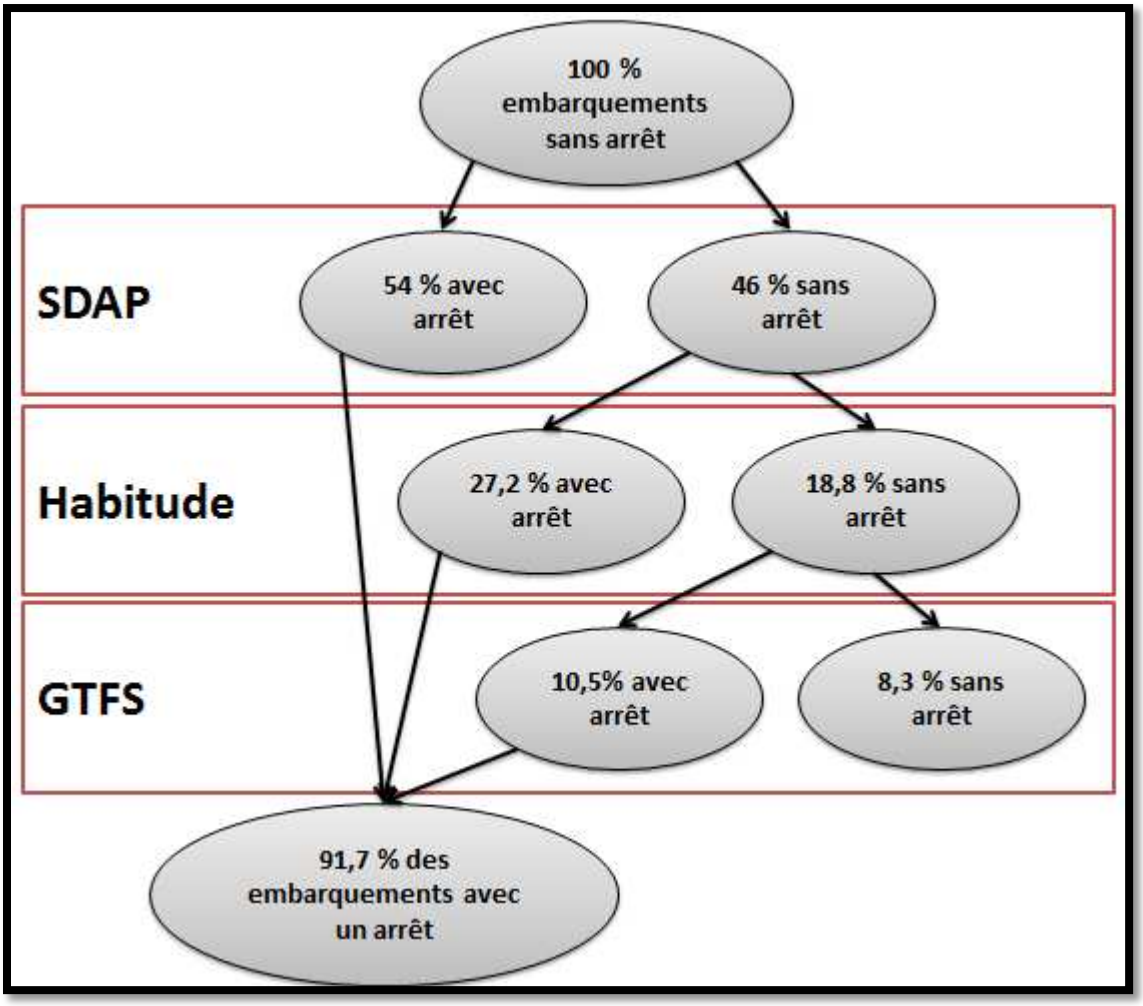


Figure 4-13 : Résumé des résultats de l'algorithme d'attribution d'arrêt

4.2.1 Enrichissement par les données du système de décompte automatique de passagers

L'enrichissement par les données SDAP se fait en quatre sous-étapes. Les résultats détaillés sont présentés dans le Tableau 4.4. La quatrième sous-étape apporte très peu de nouveaux arrêts et il est estimé que de relaxer les contraintes davantage aurait été inutile et aurait pu attribuer un grand nombre d'arrêts erronés.

Tableau 4.4 : Résultats de l'enrichissement par les données SDAP

Contrainte (en plus du numéro d'autobus)	**Embarquements se voyant attribuer un arrêt**	**% des arrêts**	**% cumulatif**
L'enregistrement de l'embarquement s'effectue entre l'ouverture – 10 secondes et la fermeture + 15 secondes des portes de l'autobus et au premier arrêt de la ligne (chainage = 0)	580 615	23.39%	23.3%
L'enregistrement de l'embarquement s'effectue entre l'ouverture et la fermeture + 10 secondes des portes de l'autobus	416 211	16.77%	40.1%
Un léger décalage d'horloge (15 secondes) est accepté. L'enregistrement de l'embarquement s'effectue entre l'ouverture - 15 secondes et la fermeture + 25 secondes des portes de l'autobus	320 463	12.91%	53.0%
Un décalage d'horloge (30 secondes) est accepté. L'enregistrement de l'embarquement s'effectue entre l'ouverture - 30 secondes et la fermeture + 40 secondes des portes de l'autobus	22 967	0.93%	54.0%

Cette première étape d'enrichissement permet d'analyser les données d'embarquement d'un point de vue géographique. La Figure 4-14 montre les arrêts qui ont été attribués au cours de cette étape. Il y a encore une fois une très forte concentration au terminus Centre-Ville à Montréal, à la station de métro Longueuil-Université-de-Sherbrooke et au terminus Panama. Ceci est cohérent avec les données de montants du système de décompte automatique de passagers.

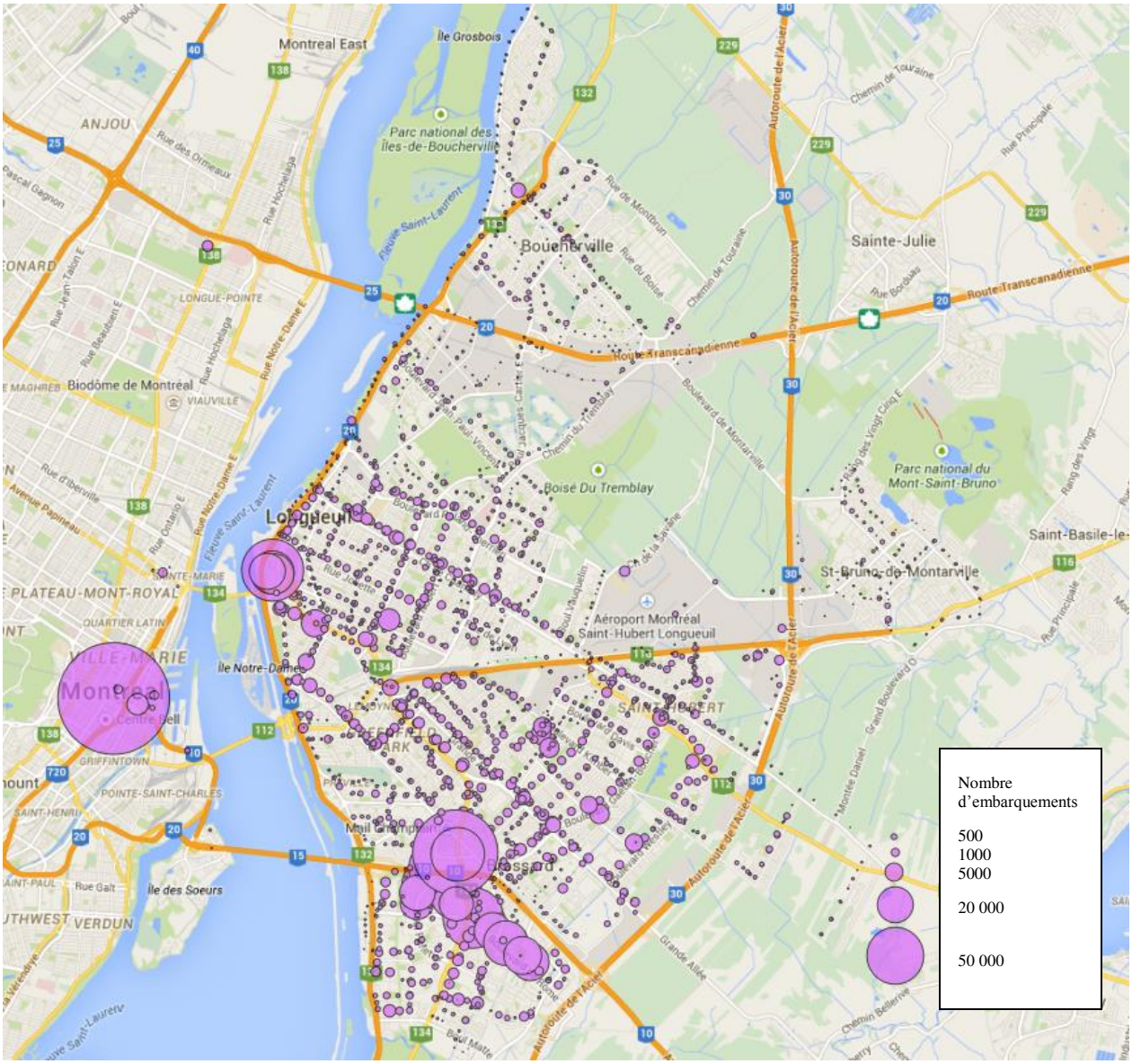


Figure 4-14 : Carte représentant les arrêts attribués lors de l'enrichissement par SDAP

La Figure 4-15 compare les données de montants de la table SDAP et les résultats de l'enrichissement pour les autobus étant munis d'un système de décompte de passagers fonctionnel. Les points bleus représentent le nombre d'embarquements CAP par arrêt. Les points jaunes représentent le nombre de passagers capturés par SDAP par arrêt. La partie grise montre les parts qui sont communes à SDAP et à CAP. On peut voir que le système de compte à bord a détecté moins d'embarquements que les embarquements CAP enrichis.

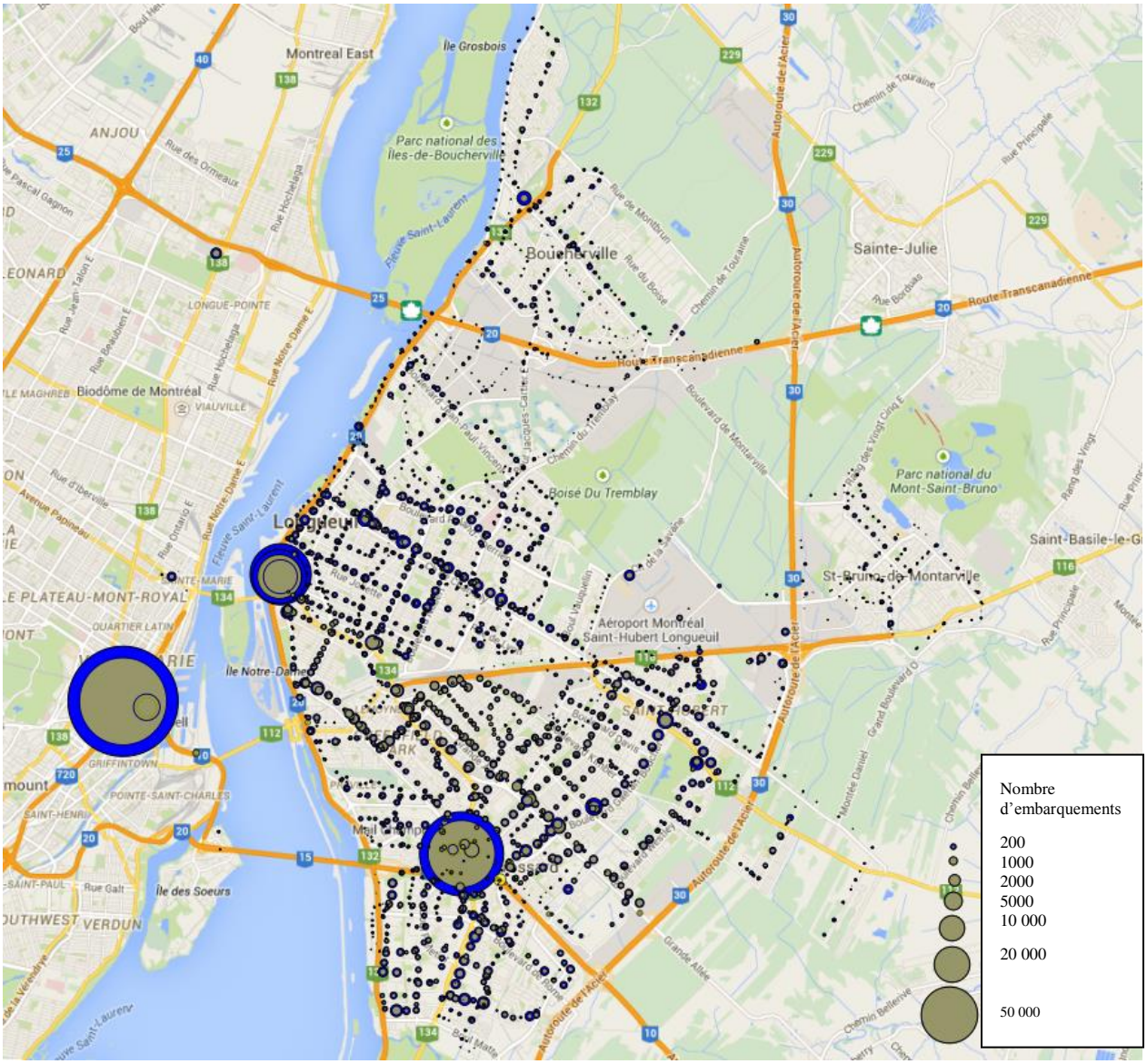


Figure 4-15 : Carte comparant les montants (en jaune) SDAP et les embarquements CAP pour les bus avec un décompte de passager fonctionnel (en bleu).

4.2.2 Enrichissement de la table CAP selon les habitudes des usagers

Les résultats de cette étape dépendent grandement des résultats de l'étape précédente. Plus il y a d'arrêts historiques à comparer, moins il y aura de possibilités d'attitrer un arrêt erroné. Une plus grande quantité d'arrêts connus permet aussi de capturer plus de comportements différents d'un même usager. Les résultats de cet enrichissement sont présentés au Tableau 4.5.

Tableau 4.5 : Résultats de l'enrichissement de CAP par les habitudes des usagers.

Contrainte (en plus du numéro de carte, de la ligne et de la direction)	**Embarquements se voyant attribuer un arrêt**	**% des arrêts**	**Cumulatif**
Les embarquements se sont produits le même type de jour, au cours de la même heure de la journée. Seulement un arrêt correspondait à ces contraintes.	402 216	16.21%	70.21%
Les embarquements se sont produits le même type de jour, au cours de la même heure de la journée. Tous les arrêts correspondants à ces contraintes étaient à moins de 500 mètres les uns des autres.	84 419	3.40%	73.61%
Les embarquements seulement selon les contraintes de lignes. Seulement un arrêt correspondait à ces contraintes.	188 184	7.58%	81.19%

Les résultats de cette étape d'enrichissement peuvent aussi être analysés d'un point de vue géographique. La Figure 4-16 montre les arrêts qui ont été attribués au cours de cette étape. Ceux-ci sont cohérents avec les données de montants de SDAP et avec les arrêts trouvés à l'étape précédente. On retrouve encore les trois mêmes points de concentration d'embarquements.

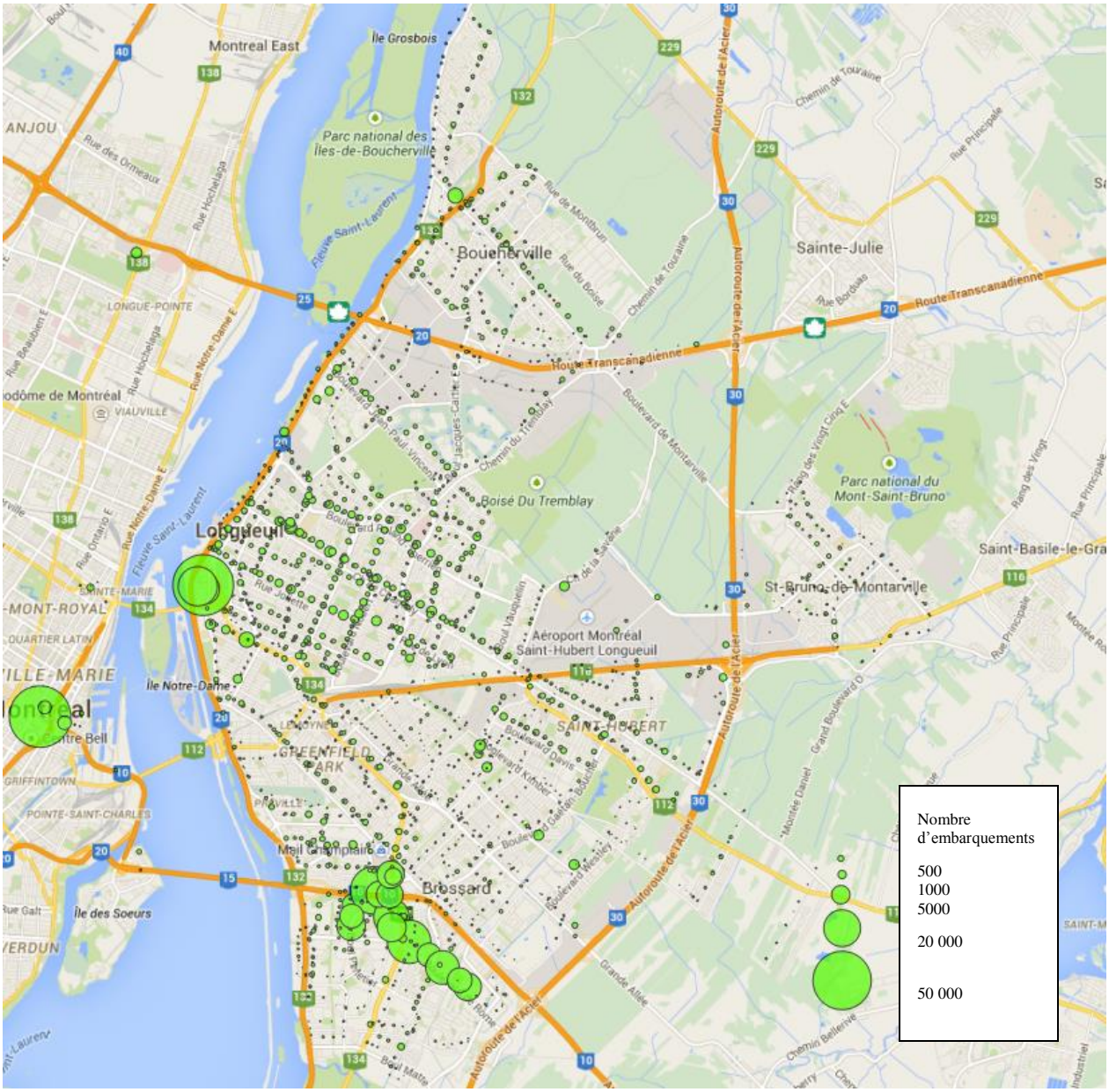


Figure 4-16 : Carte représentant las arrêts qui ont été attitrés grâce aux habitudes des usagers

4.2.3 Enrichissement de CAP à partir des données GTFS

L'enrichissement grâce aux données GTFS se fait en faisant correspondre l'embarquement CAP à un passage d'un autobus à un arrêt GTFS. Comme il n'est pas souhaitable d'avoir des résultats erronés, un filtrage est effectué pour être certain qu'il y a seulement un véhicule à bord duquel l'usager aurait pu embarquer.

Lorsque cette étape est faite sur l'ensemble des embarquements CAP, 53,6 % des embarquements totaux se voient attribuer un arrêt. Les résultats pour l'algorithme d'attribution d'arrêt sont présentés dans le Tableau 4.6.

Tableau 4.6 : Résultats de l'enrichissement de CAP par les données GTFS

Contrainte	**Embarquements se voyant attribués un arrêt**	**% des arrêts**	**Cumulatif**
L'enregistrement CAP à la même date, la même ligne, la même direction et la même voiture que l'enregistrement GTFS. L'enregistrement s'est produit dans l'intervalle allant de 50 secondes avant l'heure du passage prévu à 50 secondes après l'heure du passage prévu.	259 608	10.46%	91.56%

4.2.4 Comparaison des enrichissements

La Figure 4-17 représente l'attribution d'arrêts aux embarquements de la table CAP. Dans cette figure, l'étape d'enrichissement grâce à la table GTFS a été faite sur l'ensemble des embarquements. Il y a donc une partie des embarquements représentés qui ont un arrêt provenant de la table SDAP ou des habitudes des usagers et un arrêt provenant de la table GTFS.

Sur l'ensemble des embarquements, 29% des arrêts ont un arrêt SDAP et GTFS. Pour 89% de ces embarquements, les deux arrêts sont à moins de 500 mètres l'un de l'autre. Pour l'ensemble des embarquements, 8.5% se sont vu attribuer le même arrêt pour les deux enrichissements.

14% de l'ensemble des embarquements ont été enrichis d'un arrêt à l'étape de l'enrichissement par habitude et à l'étape de l'enrichissement par la table GTFS. 94% de ces embarquements ont deux arrêts situés à moins de 500 mètres l'un de l'autre.

Ces résultats montrent qu'il y a une cohérence entre les différentes étapes de l'algorithme d'attribution d'arrêt.

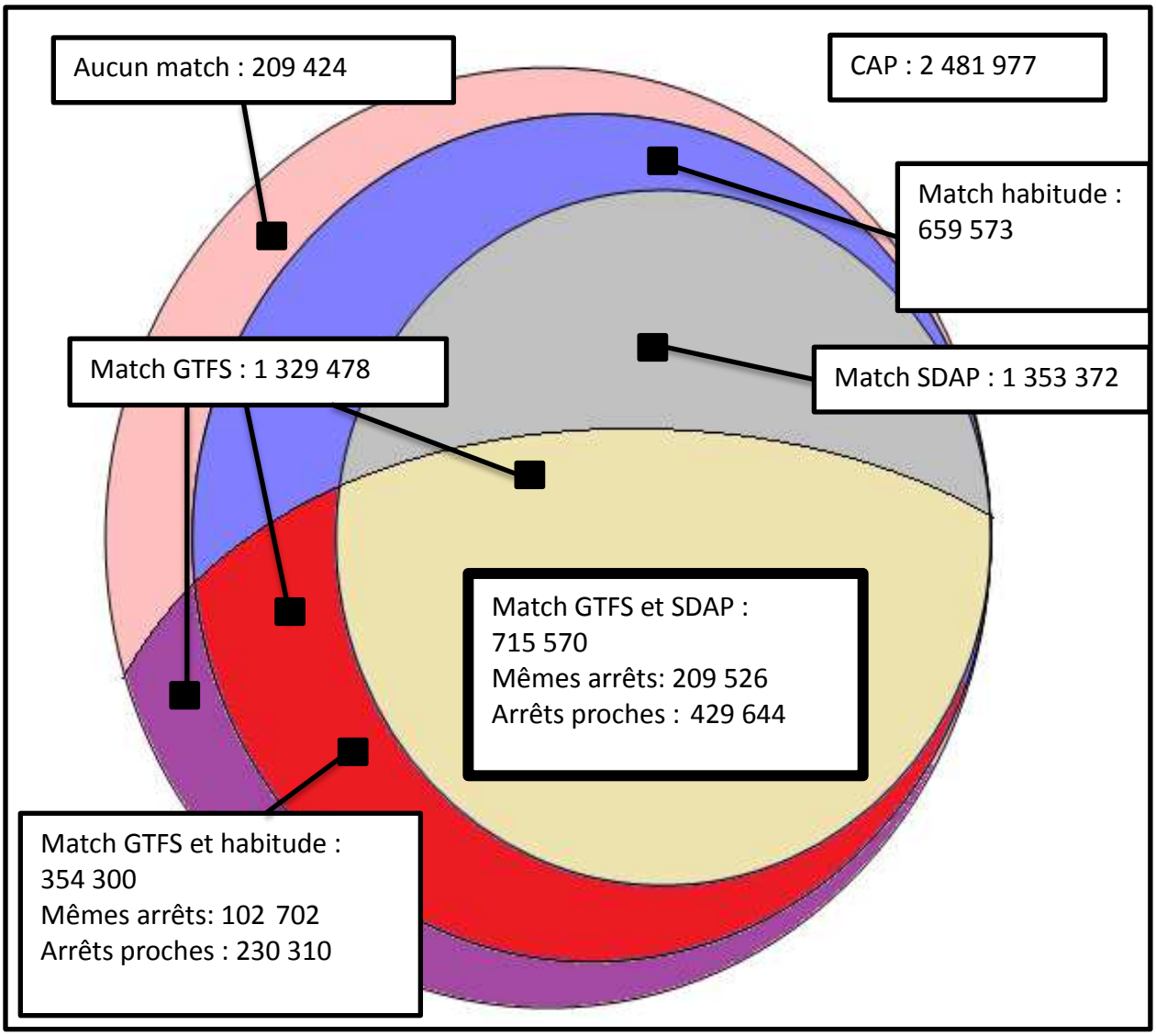


Figure 4-17 : Représentation des arrêts qui ont été attitrés aux embarquements CAP selon l'étape d'attribution

4.2.5 Caractérisation des résultats

Il est intéressant d'analyser les résultats pour voir s'il y a une tendance dans l'attribution d'arrêts et dans les données manquantes. Trois éléments sont analysés : les lignes, la journée et l'heure de la journée.

Le graphique de la Figure 4-18 montre les lignes classées en ordre croissant selon le taux d'attribution d'arrêts. 15 lignes ont un taux d'attribution inférieur à 80%. Les lignes ayant le plus haut taux d'attribution tendent à avoir un meilleur succès à l'étape d'enrichissement SDAP.

L'enrichissement par habitude est aléatoire selon le succès général de l'algorithme. L'enrichissement GTFS viennent combler un manque pour les lignes ayant un moins haut taux d'attribution général.

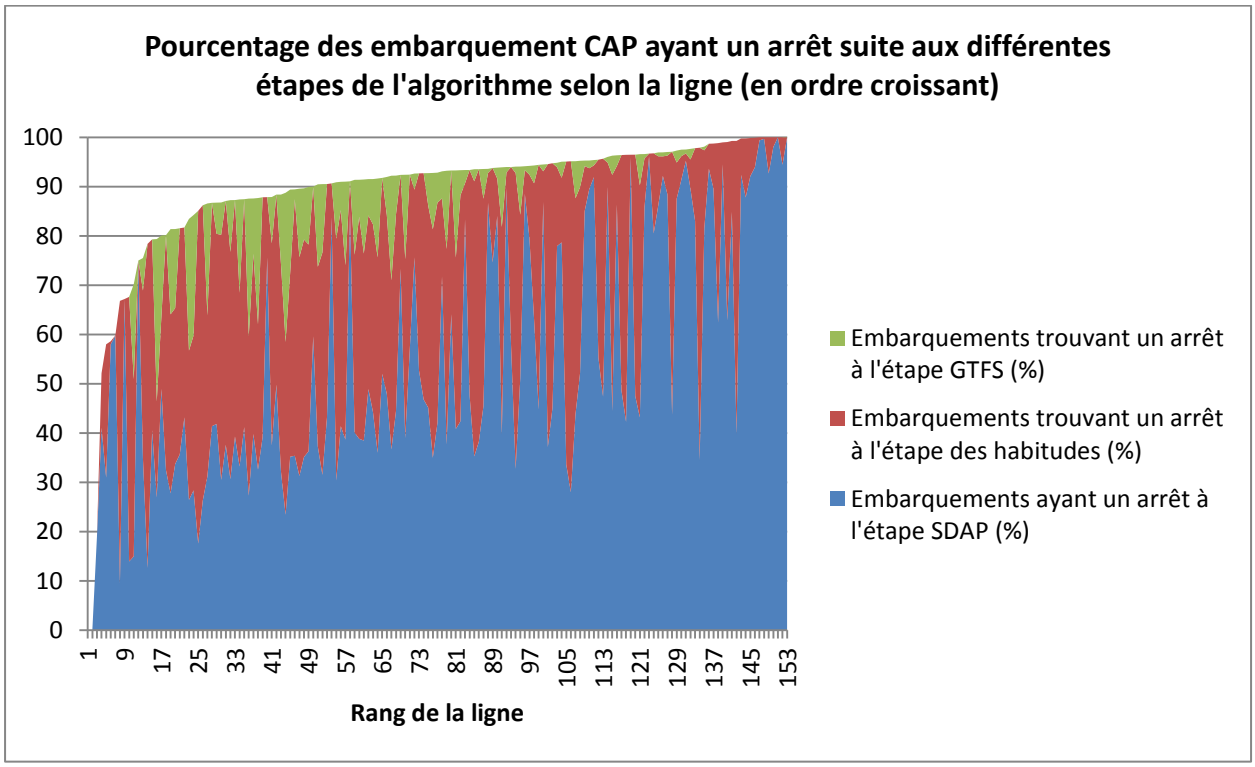


Figure 4-18 : Taux d'attribution d'arrêt aux embarquements CAP selon les étapes de l'algorithme et selon la ligne

Le graphique de la figure suivante montre à quelle étape de l'algorithme les embarquements CAP obtiennent des arrêts. Pour l'étape d'enrichissement grâce aux données SDAP, le taux d'attribution d'arrêt est uniforme tout au long du mois avec un taux légèrement plus bas pour les journées de fin de semaine. L'exception est le 31 mars qui se trouve à obtenir un taux d'attribution d'arrêt SDAP beaucoup plus bas. Pour ce qui est de l'enrichissement par habitudes des usagers, le taux de correspondances est beaucoup plus haut pour les jours de semaine que les jours de fin de semaine. Ce résultat est cohérent alors qu'il y a moins de données historiques pour les jours de fin de semaine et que les usagers sont moins routiniers la fin de semaine. Pour l'étape d'enrichissement grâce aux données GTFS, le taux d'attribution d'arrêts est plus grand lorsque moins d'embarquements ont trouvé un arrêt aux étapes précédentes. Au total, le taux d'attribution est plus haut les jours de semaine.

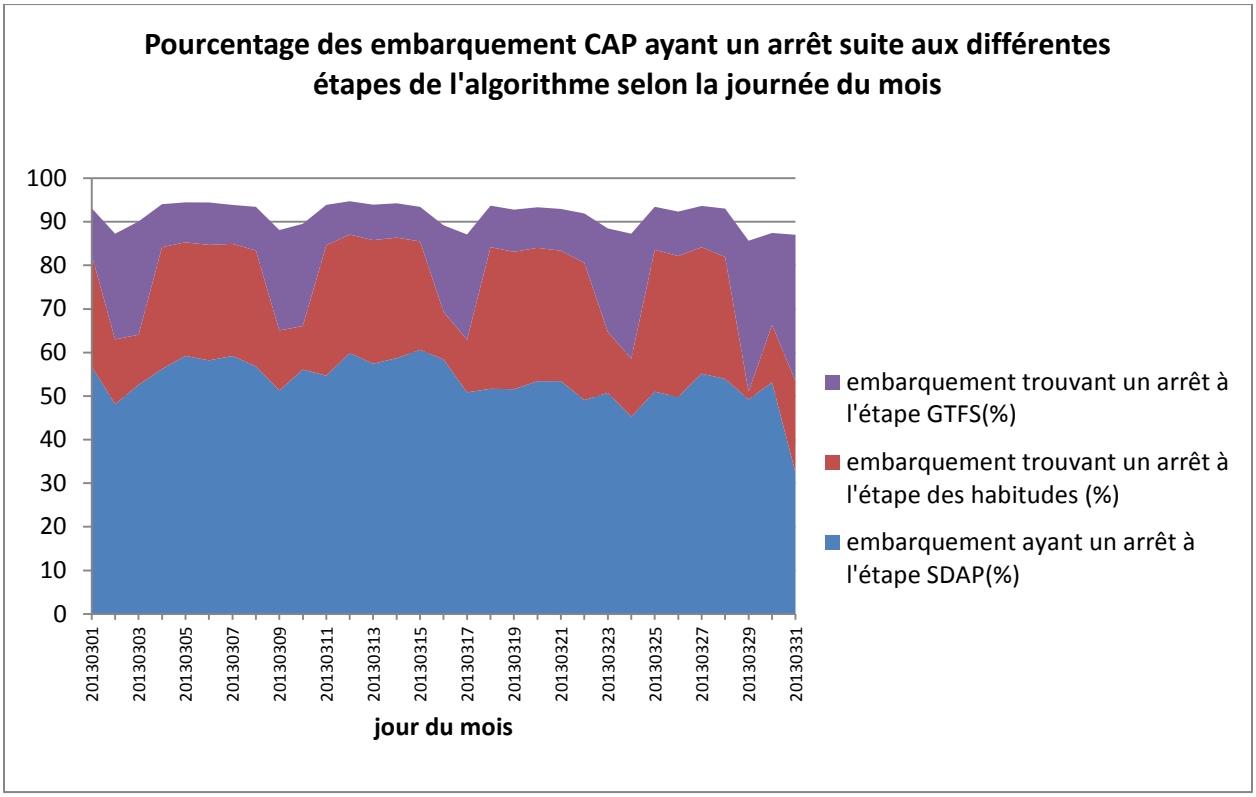


Figure 4-19 : Taux d'attribution d'arrêts aux embarquements CAP selon les étapes de l'algorithme et selon le jour du mois

Le taux d'attribution d'arrêt selon l'heure de la journée est présenté à la table Figure 4-19. Le résultat final de l'algorithme est un taux d'attribution plus élevé le matin que le reste de la journée. Pour l'étape d'enrichissement à l'aide des données SDAP, le taux d'attribution est plus bas au cours de la période de pointe de la soirée. Pour ce qui est de l'enrichissement par habitudes des usagers, le taux de correspondances est beaucoup plus élevé au cours des périodes de pointe que durant le reste de la journée. Ce résultat est cohérent alors que les usagers sont plus routiniers durant les périodes de pointe.

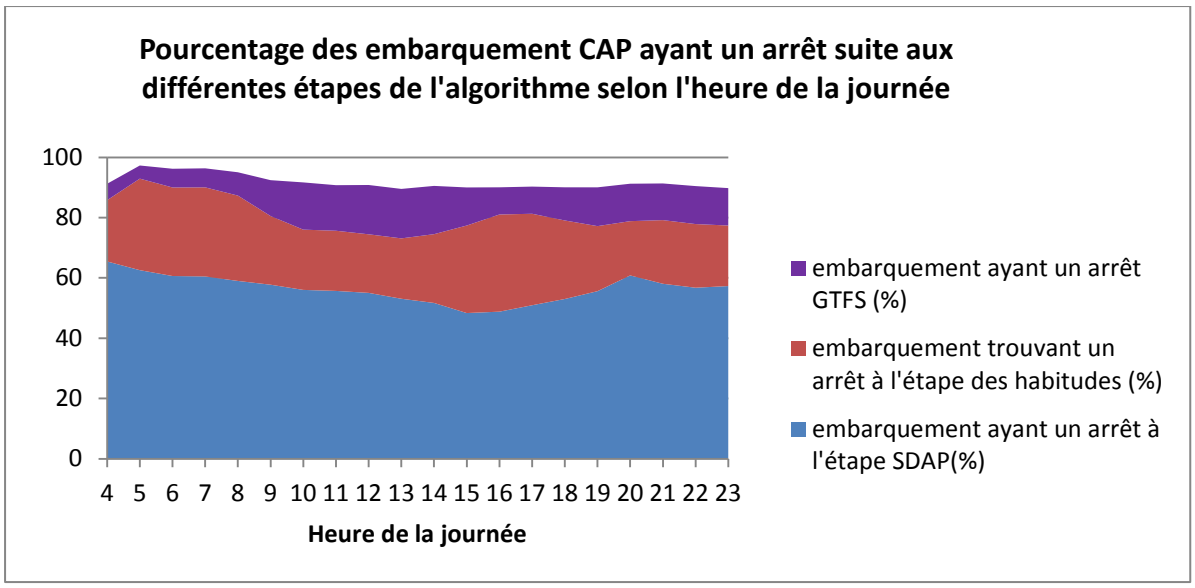


Figure 4-20 : Taux d'attribution d'arrêts aux embarquements CAP selon les étapes de l'algorithme et selon l'heure du jour

4.3 Applicabilité de l’algorithme de destination

L’algorithme de destination pourra être utilisé avec les données de la RTL. Les tables GTFS et la table CAP enrichis par l’algorithme d’attribution d’arrêts ont une structure qui s’y adapte facilement. Il est possible de faire des chaînes d’embarquement pour les usagers et aussi de calculer les distances entre les arrêts.

L’inconvénient principal est l’intégrité des données. Comme 8 % des embarquements n’ont pas d’arrêt, il ne sera pas possible de trouver la destination pour ces embarquements, mais aussi pour les embarquements qui les précèdent. Il y a aussi 38% des embarquements qui ont un arrêt ayant été déduit de manière indirecte. L’application de l’algorithme de destination sur ce genre de données augmente grandement le risque d’avoir des données erronées. Nous n'appliquerons pas cet algorithme dans le cadre de ce mémoire.

4.4 Les arrêts d'embarquement pour le mois de mars 2013

Maintenant que plus de 90 % des embarquements du mois ont un arrêt attitré, il est possible d'utiliser ces données pour aider à la planification du réseau.

En sachant où chaque usager embarque et à quel moment, il est possible d'estimer dans quel secteur celui-ci habite, selon les arrêts utilisés le matin et où il travaille, selon les arrêts utilisés le soir. Il serait donc possible de créer une matrice origine-destination pour une grande partie des usagers réguliers du réseau, et ce sans même connaître les arrêts de destination, vu que la grande majorité des déplacements sont pendulaires.

Il est aussi possible de regarder la popularité des arrêts de façon générale. La carte de la Figure 4-21 montre où sont les arrêts d'embarquement qui ont été attitrés pour l'ensemble du mois de données disponible.

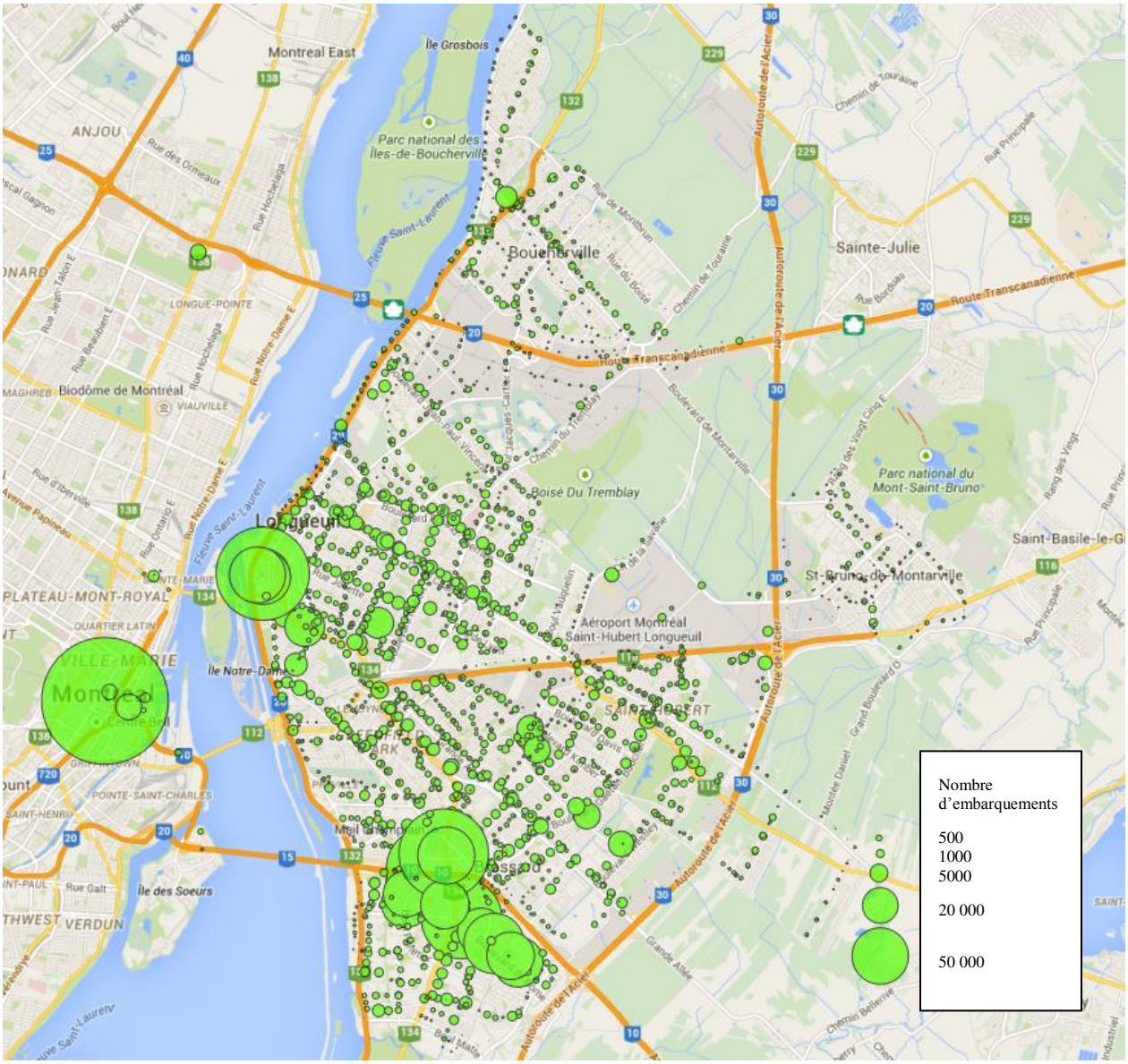


Figure 4-21 : Carte de l'ensemble des arrêts d'embarquement attribués par l'algorithme

La Figure 4-22 montre les arrêts d'embarquement pour la période de pointe du matin. La Figure 4-23, tant qu'à elle, montre les arrêts d'embarquement pour la période de pointe du début de soirée. Il y a une différence très marqué entre les deux pointes. Le matin, les embarquements sont plus réparties sur l'ensemble du territoire, alors que les gens embarquent près de leur domicile et au terminus Panama, alors que beaucoup de bus allant à Montréal passent par cette endroit. Le soir, il y a une très forte concentration d'embarquements à la station Bonaventure à Montréal et au métro de Longueuil et il y a beaucoup moins d'embarquements sur les arrêts plus éloignés. Cela indique une forte affluence de gens revenant du travail qui est situé à Montréal.

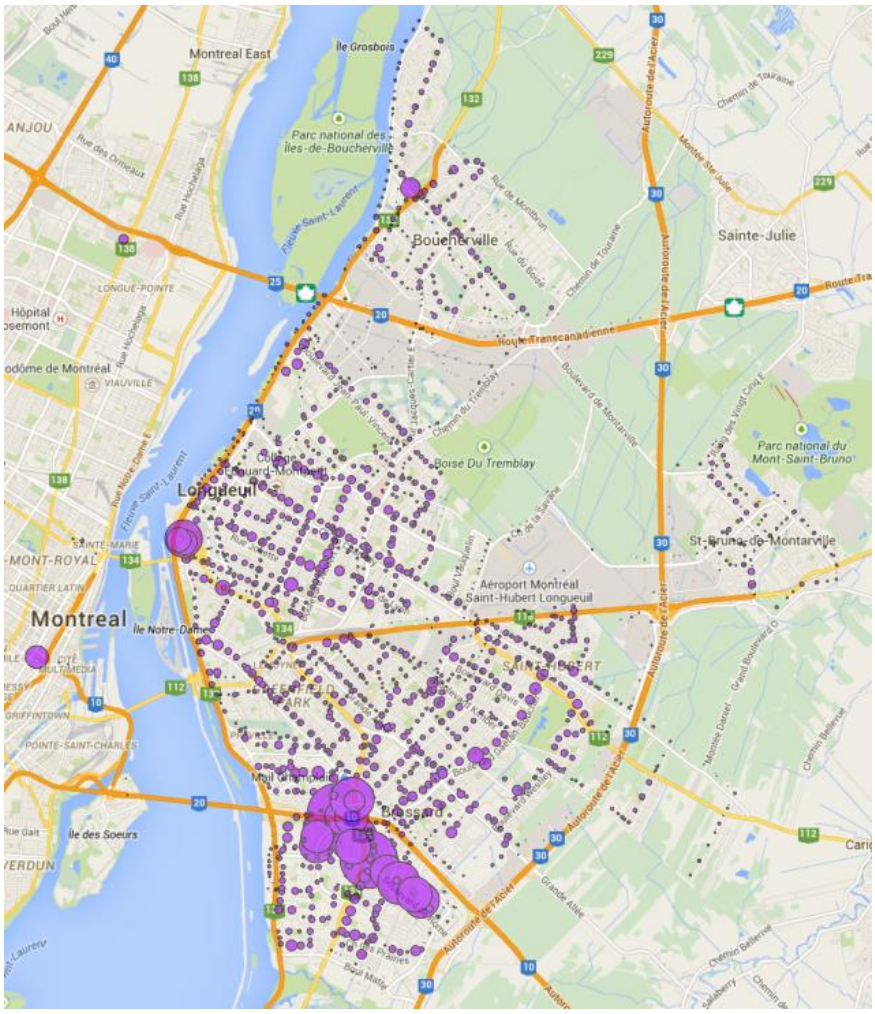


Figure 4-22 : Carte des embarquements pour la pointe du matin

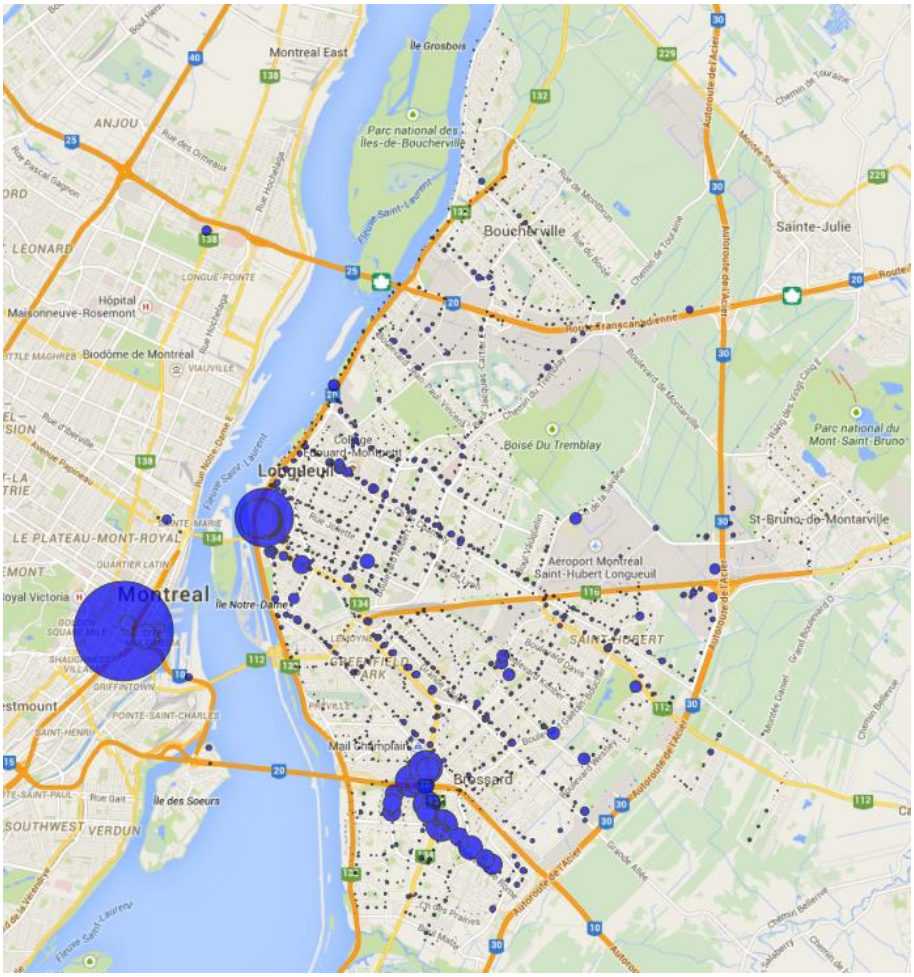


Figure 4-23 : Carte des embarquements pour la pointe du soir

CONCLUSION

Ce projet portait sur l'attribution d'arrêts aux données d'embarquements provenant du système de péage automatisé par cartes à puce du Réseau de Transport de Longueuil. Cet enrichissement des données cartes à puce se fait à partir des données du système de compte à bord et de l'horaire. Après avoir fait un traitement préalable de l'ensemble des données, il a été possible d'en faire l'analyse pour découvrir les relations entre les données. Un algorithme utilisant des requêtes SQL a été élaboré pour attribuer un arrêt aux enregistrements cartes à puce.

Dans un premier temps, une revue de la littérature a été réalisée. L'utilisation des données de localisation des véhicules pour évaluer le service et augmenter la ponctualité du service a été présentée. Par la suite, il a été question de l'utilisation des données de systèmes de décompte automatisé de passagers pour connaître la demande sur les réseaux de transport et la charge des autobus. Enfin, quelques recherches précédentes sur les données de cartes à puce ont été couvertes. Celles-ci parlaient de l'enrichissement des données pour obtenir les arrêts d'embarquement ainsi que de l'algorithme permettant d'obtenir la destination des usagers. Il était aussi question d'analyser la demande grâce à ces données enrichies, d'établir les zones les plus populaires du réseau et d'analyser le comportement des usagers. Elles discutaient aussi de la possibilité d'utiliser les données de cartes à puce pour compléter les enquêtes origines-destinations.

Dans un deuxième temps, la méthodologie de recherche a été présentée. Tout d'abord, le système d'information du RTL a été caractérisé afin de proposer une structure des données. Trois tables principales ont été utilisées. La table CAP comprend les données du système de péage automatisé par cartes à puce où les transactions d'embarquement sont enregistrées. Ces transactions ne contiennent pas de données de localisation. La table SDAP comprend les données de localisation des autobus lors des passages aux arrêts et les données de compte à bord. La table GTFS comprend les données du service planifié. Ensuite, les tables sont comparées entre elles pour vérifier l'intégrité des données. Les étapes de l'algorithme d'attribution d'arrêts ont été détaillées. La première étape a été d'utiliser les données de la table SDAP pour enrichir les embarquements CAP. La deuxième étape a utilisé les habitudes des usagers et les embarquements CAP ayant préalablement obtenu des arrêts pour dériver les arrêts d'embarquement manquants.

La dernière étape a employé les données GTFS pour finir l'enrichissement des embarquements CAP.

Dans un troisième temps, il a été question des résultats de caractérisation des données et de l'application de l'algorithme. Pour l'ensemble des journées de semaine, le nombre d'embarquements par heure enregistrés dans la table CAP et le nombre de passages aux arrêts par heure enregistrés dans la table SDAP montrent des pointes d'activité le matin et en début de soirée. Par contre, beaucoup plus d'autobus sont actifs dans la table CAP (on dénombre 344 autobus actifs en période de pointe dans la table CAP), contre 227 autobus actifs dans la table SDAP. Du côté des passages prévus (GTFS) comparé aux passages enregistrés (SDAP), il y a 44% moins de passages capturés que de passages prévus. Il manque donc une bonne quantité de données dans SDAP.

Pour ce qui est des résultats d'application de l'algorithme d'attribution d'arrêts aux embarquements CAP, les résultats présentés étaient les suivants. Lors de l'étape d'enrichissement grâce aux données SDAP, 54 % des embarquements se sont vus attribuer un arrêt. Lors de l'enrichissement grâce aux habitudes des usagers, 27.2 % des embarquements se sont vus attribuer un arrêt. Lors de l'étape finale, 10.5 % des embarquements se sont vus attribuer un arrêt grâce aux données GTFS. Au final, 91.7 % des embarquements ont été associés à un arrêt grâce à l'algorithme d'attribution.

Suite à ce projet, les perspectives de recherche sont nombreuses. L'amélioration des données de localisation des véhicules (notamment par l'augmentation du nombre de bus équipés) serait une piste pour améliorer les résultats de l'algorithme d'attribution d'arrêts. Les arrêts étant attribués aux données CAP grâce aux données SDAP, cela diminuerait les possibilités d'erreurs. Par la suite, l'enrichissement par habitude serait plus juste puisqu'il y aurait plus d'arrêts dans l'historique de l'usager. Au total, plus d'embarquements se verraient attribuer un arrêt et moins d'erreurs seraient possibles.

Par la suite, appliquer l'algorithme d'attribution d'arrêt sur une plus longue période permettrait aussi d'attribuer un plus grand pourcentage d'arrêts. Une plus longue période permettrait d'avoir accès à plus de données historiques et couvrirait plus d'exception aux habitudes des usagers. Une plus grande quantité d'embarquements historiques permettrait aussi

de modifier l'algorithme pour utiliser des probabilités d'utilisation d'arrêts selon différents paramètres.

Une fois que les données d'embarquements seraient plus complètes, l'application de l'algorithme de destination serait envisageable. La structure de données actuelle permet d'appliquer l'algorithme assez facilement. Grâce aux données de compte à bord présent au RTL, il serait possible d'effectuer une certaine validation de l'algorithme de destination (en comparant les charges estimées aux charges mesurées par comptage). Une fois les destinations connues, beaucoup de nouvelles analyses seront possibles. Il sera possible de connaître la distance que chaque usager parcourt sur le réseau du RTL. Il sera aussi possible de connaître la charge à bord de chacun des autobus, ventilée par type de titre.

En appliquant ces deux algorithmes en continu sur les données, les planificateurs du RTL auront accès à une source d'information formidable. Il sera plus facile de faire des observations sur l'utilisation du réseau et d'ajuster le service pour suivre les variations de la demande. Les ajustements pourront être faits à un niveau plus fin avec ces nouvelles informations. Les données de localisation des embarquements et des destinations permettront aussi de bâtir une matrice des origines et destinations pour l'ensemble des usagers du réseau. Il sera ainsi possible de faire des simulations utilisant des données désagrégées. Connaître les points d'origines et de destination d'une grande portion des usagers permettra aussi d'évaluer les reconfigurations et les ajouts de ligne.

BIBLIOGRAPHIE

Cevallos, F., Xiaobo W., Zhenmin C., et Albert G. (2011). Using AVL Data to Improve Transit on-Time Performance. *Journal of Public Transportation* 14(3): 21–40.

Chapleau, R., Trépanier, M., et Chu, K.K. (2008). The Ultimate Survey For Transit Planning : Complete Information with Smart Card Data and GIS. *Data for Public Transit Planning, Marketing and Model Development*. Annecy, France.

Foell, S., Santi P., Gerd K., Marco V., et Carlos B. (2014). Catch Me If You Can: Predicting Mobility Patterns of Public Transport. http://oro.open.ac.uk/40786/

Furth, P. G., Hemily, B., Muller, T. H. J. et Strathman, J. G. (2006). Using Archived AVL-APC Data to Improve Transit Performance and Management. 113. *Transit Cooperative Research Program*. Washington, D. C.: Transportation research board.

Lee, S. G., et Hickman, M. D. (2011). Travel Pattern Analysis Using Smart Card Data of Regular Users. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*.

Morency, C., Trépanier, M. et Agard, B. (2007). Measuring Transit Use Variability with Smart-Card Data. *Transport Policy* 14(3): 193–203.

Munizaga, M. A. et Palma, C. (2012). Estimation of a Disaggregate Multimodal Public Transport Origin–Destination Matrix from Passive Smartcard Data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies* 24: 9–18.

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

Riegel, L., et Attanucci, J. (2013) Utilizing Automatically Collected Smart Card Data to Enhance Travel Demand Surveys. *TRB 2014 Annual Meeting.*

Shalaby, A., et Farhan, A. (2004) Prediction Model of Bus Arrival and Departure Times Using AVL and APC Data. *Journal of Public Transportation* 7(1): 41–62.

Shi, X, et Hangfei Lin, H. (2013). The Analysis of Bus Commuters Travel Characteristics Using Smart Card Data : The Case of Shenzhen, China. *TRB 2014 Annual Meeting*.

Trépanier, M., Morency, C. et Agard, B. (2009) Calculation of Transit Performance Measures Using Smartcard Data. *Journal of Public Transportation* 12(1): 79–96.

Trépanier, M., Tranchant, N. Chapleau, R. (2007) Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems* 11(1): 1–14.

Wang, W., Attanucci J.P., et Wilson, N.H.M. (2011) Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems. *Journal of Public Transportation* 14(4): 131–150.