Titre: Title:	Développement d'une plateforme de consultation et d'analyse de l'industrie du taxi
Auteur: Author:	Anjeli Narrainen
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Narrainen, A. (2021). Développement d'une plateforme de consultation et d'analyse de l'industrie du taxi [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5545/

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Développement d'une plateforme de consultation et d'analyse de l'industrie du taxi

ANJELI NARRAINEN

Département des génies civil, géologiques et des mines

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

Génie civil

Décembre 2020

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé:

Développement d'une plateforme de consultation et d'analyse de l'industrie du taxi

Présenté par

Anjeli NARRAINEN

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

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

Geneviève BOISJOLY, présidente
Martin TRÉPANIER, membre
Catherine MORENCY, membre et directrice de recherche

DEDICACE

A mes parents

REMERCIEMENTS

Je tiens tout d'abord à remercier ma directrice de recherche Catherine Morency pour m'avoir donné l'opportunité de réaliser cette maitrise dans un domaine qui m'était jusqu'alors inconnu. Ses compétences et sa passion pour le domaine des transports ont éveillé mon intérêt pour ce secteur et m'ont donné envie de m'y consacrer pleinement durant mes deux années de recherche. Elle a partagé ses connaissances et son expertise du milieu, tout en m'accordant sa confiance et une autonomie qui m'ont permis de développer de nombreuses compétences. Je la remercie également pour les moyens qu'elle met à la disposition de ses étudiants tout au long de leur projet de recherche en ce qui concerne les participations aux conférences et la possibilité de réaliser d'autres projets en parallèle. Tout cela ne fait que grandir l'intérêt que l'on porte au domaine et une chose est sûre c'est qu'on ne s'ennuie pas !

Je remercie mes très chers parents, Deva et Sonia, d'avoir tout fait pour assurer mon avenir et mon épanouissement. Merci pour tous leurs sacrifices et efforts sans lesquels je ne se serai pas arrivée jusque-là. Merci à ma sœur Dana et à mon frère Shun, qui malgré les milliers de kilomètres qui nous séparent ont toujours été présents. Merci à ma famille pour leur amour et leur soutien constant.

Je tiens aussi à remercier mes amis et collègues de bureau B-330 à Poly pour tous les bons moments que nous avons passé. Merci particulièrement à Gabriel Lefebvre-Ropars pour son aide précieuse et généreuse avec PostgreSQL, QGIS et pour toutes ses astuces qu'il n'hésite pas à partager. Merci d'avoir toujours pris le temps et eu la patience de répondre à mes innombrables questions. Merci à Elodie Deschaintres pour sa gentillesse et son soutien dans les moments les plus difficiles. Merci aussi d'avoir partagé sa réserve infinie de chocolats qui a su nous donner l'énergie et le réconfort nécessaire pour vivre les longues journées de travail. Merci d'avoir toujours été partante pour les petits déjeuners du premier jeudi du mois à l'AECSP. Et merci à Julien Douville pour sa gentillesse, son amitié et sa bonne humeur inaltérable. Merci d'avoir organisé un super weekend de canot camping pour tout le bureau des étudiants!

Un grand merci également aux associés de recherche, avec une mention particulière pour Pierre-Léo Bourbonnais pour son aide avec la programmation du tableau de bord et Jean-Simon Bourdeau qui, grâce à son expertise et sa patience, a su m'aider à relever les nombreux défis de traitement des données.

Enfin, je tiens à remercier mes amis et colocataires qui ont su faire de ses deux années à Montréal un merveilleux séjour. Merci à Khadija El Iraki, Clément Broggi, Jérémie Mosser et Quentin Panquet avec qui nous avons formé la super Cooloc. Merci d'avoir rendu ce confinement plus surmontable grâce aux séances intensives de sport suivies des préparations tout aussi intensives de macarons et fondants au chocolat. Merci particulièrement à Clément pour ses relectures attentives des chapitres de ce mémoire. Merci à Guillaume Lameynardie et Anthony Cheruel (nos voisins hyperactifs) et à Guillaume David d'avoir motivé les troupes pour l'exploration du Québec.

Merci particulièrement à Khadija et Guillaume D., qui comme moi ont dû rédiger leur mémoire durant le confinement. Merci d'avoir été d'excellents collègues de notre bureau du salon.

Finalement, parce que sans leur soutien le projet n'aurait pas eu lieu, je remercie le Bureau du Taxi de Montréal et Revenu Québec pour le financement de mon projet de recherche.

Je remercie enfin tous ceux qui, d'une manière ou d'une autre, ont contribué à la réussite de ce travail et qui n'ont pas été cités ici.

RÉSUMÉ

Ce projet de recherche s'inscrit dans la continuité des travaux de Lacombe (2016) et Laviolette (2017) portant sur l'élaboration d'indicateurs de performance et de suivi sur l'offre et la demande de déplacements par taxi à Montréal. Cependant, ces études se sont appuyées sur les données de quelques intermédiaires en services de taxi uniquement. Pour le présent projet, ce sont les données de tous les taxis opérant sur l'île de Montréal qui sont rendues disponibles. L'analyse peut donc être élargie à l'ensemble du territoire sans qu'il soit nécessaire de pondérer les objets d'analyse. Dans cette optique, le projet vise à développer une plateforme web de consultation et d'analyse de l'industrie du taxi. Deux principaux objectifs permettent d'atteindre ce but :

1. Automatiser l'extraction des données de ce registre et le calcul des indicateurs.
2. Développer une plateforme de consultation et d'analyse de ces indicateurs qui agira comme un outil d'aide à la planification stratégique.

Dans un premier temps, afin de mettre en évidence les enjeux liés à la conception d'une telle plateforme, une revue de littérature s'intéresse à deux principales thématiques. La première concerne l'analyse du secteur des taxis. Une synthèse des thématiques d'analyses récentes liées à ce secteur y est menée dans le but de dresser un inventaire complet des connaissances sur le secteur des taxis et particulièrement de l'impact de la disponibilité des données GPS de taxis dans ces travaux. Dans la deuxième partie de la revue, l'accent est mis sur la visualisation des données. Les différentes caractéristiques de visualisation essentielles à la conception d'un tableau de bord communiquant les informations de manière efficace et juste y sont mises en évidence.

Les données utilisées représentent plus de 60 millions des points GPS par jour. Afin de faire face aux enjeux de stockage et de temps de calcul que présente ce volume de données, une méthodologie de construction d'une base de données optimisée permettant de répondre à ces problématiques est développée. Une description des données et notamment des quatre statuts disponibles pour caractériser l'activité d'un taxi en service est fournie. Puis, dans l'objectif de pouvoir estimer une variété d'indicateurs, des procédures systématiques et automatisées de traitement des données sont établies. Les principales étapes du processus général de traitement des données sont résumées : de l'extraction des données brutes au traitement de ces dernières jusqu'à la visualisation des indicateurs.

De plus, on s'intéresse particulièrement à la résolution d'un enjeu majeur des données, à savoir l'irrégularité des intervalles de temps entre deux points consécutifs. Des méthodes d'identification des groupes de statut correspondant aux activités des taxis, et plus particulièrement une méthode d'identification des courses régulières se basant sur l'étude d'une population de référence, est développée. Les courses étant identifiées, des critères de validation sont mis en place afin d'identifier les courses valides.

Une fois les méthodes de traitement des données établies, les principaux indicateurs qu'il serait pertinent d'estimer sont identifiés à partir de la littérature et des travaux de Lacombe (2016) et Laviolette (2017). Dans un premier temps, l'estimation des indicateurs caractérisant l'offre et la demande de déplacements par taxi est mise à jour à la lumière des données disponibles dans le Registre des taxis. Ces indicateurs sont classifiés selon six objets d'analyse, à savoir la course, le véhicule de taxi, le chauffeur, l'intermédiaire en service, le poste d'attente et les origines et destinations des courses. Puis, les déclinaisons possibles de ces principaux indicateurs ainsi que les diverses formes de visualisation réalisables sont détaillées. Les échelles spatiales et temporelles retenues pour l'analyse sont également présentées. Enfin, les défis méthodologiques liés au calcul de ces indicateurs sont mis en évidence.

Finalement, le tableau de bord assurant la visualisation et l'analyse des indicateurs d'offre et de demande en déplacements de taxi est présenté. La représentation des indicateurs pouvant être visualisés dans le tableau de bord est explicitée. Cette dernière se base sur les règles de visualisation mises en évidence dans la revue de littérature dont l'objectif est d'assurer une consultation intelligible des données. La structure générale du tableau de bord est donc dévoilée : de la page d'accueil de faits saillants à des distributions fréquentielles et temporelles plus détaillées. Cette structure en entonnoir du tableau de bord assure à l'utilisateur une exploration flexible des données afin de passer des résumés agrégés à des analyses plus désagrégées. Il lui en effet possible d'affiner ses requêtes grâce à la sélection de filtres spatio-temporels. Afin d'illustrer les différents niveaux de visualisation des données du tableau de bord, l'objet course est analysé ainsi que les origines et destinations. Enfin, le volet comparatif permettant de comparer, sur la même fenêtre de visualisation, un ou plusieurs indicateurs selon deux périodes d'analyses distinctes est explicité. Les faits saillants des mois d'avril 2019 et 2020 y sont comparés, révélant notamment l'impact de la pandémie et du confinement sur l'industrie du taxi.

Le document se conclut par des perspectives de recherche qui permettraient d'améliorer le tableau de bord conçu en ce qui concerne la méthodologie d'estimation des indicateurs, les autres sources de données pertinentes à considérer, l'optimisation des processus développées ainsi que les analyses à effectuer afin de poursuivre les études sur le transport par taxi. L'utilisation du tableau de bord permettra de brosser un portrait plus représentatif que ceux établis précédemment des taxis opérant sur l'île Montréal et favorisera donc la prise de décisions quant à l'opération et à la planification de ce service.

ABSTRACT

This research project is a continuation of the work conducted by Lacombe (2016) and Laviolette (2017) on developing performance indicators and monitoring the supply and demand for taxi travel in Montreal. However, these studies relied solely on the data of a few taxi compagnies. For the current project, data from all the taxis operating on the island of Montreal are made available. The analysis can therefore be extended to the entire territory without the need to weight the objects of analysis. To that end, this project aims to develop a dashboard for consultation and analysis of the taxi industry. There are two main objectives to achieve this goal:

1. To automate the extraction of data from the Register and the calculation of the indicators.
2. To develop a dashboard for consultation and analysis of these indicators that will act as a tool to support strategic planning.

In order to highlight the issues involved in the design of such a platform, a literature review first focuses on two main themes. The first one focuses on the analysis of the taxi sector. A review of recent analyses related to this sector is carried out in order to draw up a complete inventory of knowledge on the taxi sector and particularly the impact of the availability of taxi GPS data in these studies. In the second part of the literature review the emphasis is on data visualization. The various visualization features essential to the design of a dashboard that communicates information efficiently and accurately are highlighted.

The data collected represents more than 60 million GPS points per day. In order to meet the challenges of storage and calculation time raised by this volume of data, a methodology for building an optimized database to address these issues has been developed. A description of the data and particularly the four statuses available to characterize the activity of an operating taxi is provided. Then, in order to estimate a variety of indicators, systematic and automated data processing procedures are established. The main steps in the general data processing operation are summarized: from the extraction of the raw data to the processing of the latter up to the visualization of the indicators.

In addition, emphasis is placed on solving a major issue related to the data, namely the irregularity of the time intervals between two consecutive points. Methods for identifying status groups corresponding to taxi activities, and especially a procedure to identify regular trips based on the

study of a reference population, is established. Once the trips have been identified, validation criteria are developed in order to identify valid trips.

Once the data processing methods have been established, the main indicators, whose estimation would be relevant, are identified from the literature and the work of Lacombe (2016) and Laviolette (2017). First the estimation of the indicators which characterize the supply and demand for taxi travel is brought up to date in the light of the data available in the Taxi Register. These indicators are classified according to six objects of analysis, namely the trip, the taxi vehicle, the driver, the taxi company, the taxi stand and the origins and destinations of the trips. Then, the possible variations of these main indicators and the various forms of visualization that can be implemented are detailed. The spatial and temporal scales used for the analysis are also presented. Finally, the methodological challenges related to the calculation of these indicators are highlighted.

To conclude, the dashboard ensuring the visualization and analysis of the indicators of supply and demand for taxi journeys is presented. The representation of the indicators that can be visualized in the dashboard is explained. This representation is based on the visualization rules highlighted in the literature review, the aim of which is to ensure that the data can be consulted in an intelligible manner. The general structure of the dashboard is thus revealed: from the highlights to more detailed frequency and time distributions. This funnel structure of the dashboard ensures the user a flexible exploration of the data in order to move from aggregate summaries to more disaggregated analyses. Queries can be refined through the selection of spatio-temporal filters. In order to illustrate the different levels of visualization of the dashboard data, the trip object is analyzed as well as the origins and destinations. At last, a comparative component allowing to confront on the same viewing window one or several indicators regarding two distinct periods of analysis is explained. The highlights of April 2019 and 2020 are compared, revealing the impact of the pandemic and containment on the taxi industry.

The document concludes with research perspectives that would improve the designed dashboard regarding the methodology for estimating the indicators, the additional relevant data sources to be considered, the optimization of the processes developed as well as the analyses to be carried out in order to pursue the studies on transportation by taxi. The use of the dashboard will provide a more representative picture than those previously established of the taxis operating on the Island of

Montreal and will therefore support decision-making with respect to the operation and planning of this service.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABREVIATIONS

BTM Bureau du Taxi de Montréal

CSV Comma-separated Value (type de fichier où les données sont séparées par des virgules)

CTQ Commission des Transports du Québec

GPS Global Positioning System

JSON JavaScript Object Notation

MaaS Mobility as a Service

PDF Portable Document Format

SQL Structured Query Language

TLC New York City Taxi and Limousine Commission

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

1.1 Problématique

Le secteur des taxis est en pleine mutation. En effet, on constate depuis quelques années une diversification des services de transport, notamment avec le développement des services de transport sur demande assurés par des entreprises de transport technologiques (« Transportation Network Companies » ou « TNCs ») telles que Uber (Xu et al., 2018). Outre cette diversification de l'offre, on ne dispose que d'une compréhension limitée de la demande de déplacements en taxi, ce qui a conduit à remettre en question un certain nombre de points. On s'interroge à la fois sur les modalités opératoires, les réglementations et l'état de l'offre de service ainsi que sur la demande en déplacements. Si des villes comme New York, Beijing ou Singapour disposent de nombreuses études et analyses de leur industrie du taxi (Yuan, Zheng, Zhang, & Xie, 2013), à Montréal il est manifeste que les connaissances scientifiques sur le sujet restent limitées (Laviolette, 2017). Le rôle du taxi dans la mobilité est peu défini, les facteurs qui en déterminent le choix sont peu connus et l'évolution de ses opérations insuffisamment documentée. Tout cela accroît considérablement la difficulté de développer, ou même d'optimiser cette industrie, d'évaluer son cadre législatif et réglementaire de même que son positionnement dans le mouvement de la mobilité intégrée. Dans le contexte actuel où l'on cherche à réduire la dépendance à l'automobile privée et à mettre en place une mobilité plus durable, la mobilité en tant que service (MaaS, Mobility as a Service) offre un véritable changement de paradigme. La mobilité en tant que service correspond à l'intégration de différents types de services de transport en un service unique de mobilité accessible à la demande. L'objectif est d'offrir la meilleure combinaison de services permettant de répondre au besoin de mobilité du client. Le service de mobilité est donc centré sur l'utilisateur. Dans ce concept de MaaS, une seule application est utilisée pour accéder et payer pour différents modes de transport, qu'ils soient publics ou privés, pour un déplacement urbain ou interurbain (Kamargianni & Matyas, 2017; Maas Alliance, 2020). Or, dans les discussions sur la mobilité en tant que service le taxi apparait comme un acteur incontournable dans le cocktail des services sur lesquels peuvent compter les voyageurs. Il est en effet souvent utilisé lorsqu'aucune autre option n'est disponible et constitue donc un élément clé pour répondre à tous les besoins en matière de voyage.

Dans le cas d'une future implémentation d'un système de mobilité en tant que service, afin d'évaluer les contributions d'un tel système, il est essentiel de pouvoir analyser la performance actuelle du taxi à Montréal.

Ainsi, bien que les compagnies de taxi aient choisi d'intégrer des dispositifs de collecte de données GPS (« Global Positioning System ») dans leurs flottes depuis quelques années, à Montréal, ces données sont encore peu exploitées et elles ne reçoivent pas de la part des acteurs du secteur l'attention qu'elles méritent (Laviolette, 2017; Liao, Yu, & Chen, 2010). En effet, les données générées par les appareils GPS sont remarquablement riches et peuvent être mises à profit pour l'analyse, la planification et l'amélioration des services de taxi (Alisoltani, Zargayouna, & Leclercq, 2020; Andrienko, Andrienko, & Wrobel, 2007; Laviolette, 2017; Wang, H., Zou, Yue, & Li, 2009).

Ces importants volumes de données offrent une chance de transformer le modèle commercial actuel des services de taxi. En analysant ces ensembles de données, il est désormais possible de connaître les itinéraires de conduite optimaux, la répartition de l'offre et de la demande, de réduire le temps d'attente ou encore de fournir des suggestions aux chauffeurs sur la manière d'améliorer leurs revenus. En effet, une plus grande efficacité dans les opérations de taxi se traduit par une réduction du temps et de la distance parcourus à vide et donc par une diminution de la consommation de carburant et des émissions polluantes et de gaz à effet de serre (Ge, Liu, Xiong, & Chen, 2011; Zhou et al., 2020). En outre, la circulation des taxis à vide génère un trafic supplémentaire et participe à la saturation du réseau routier. Une exploitation plus efficace des taxis est donc indispensable (Billhardt, Fernández, Ossowski, Palanca, & Bajo, 2019; Kamimura et al., 2013; Yuan et al., 2013).

Ainsi selon Ferreira (2013), le défi n'est plus aujourd'hui dans l'obtention des données nécessaires mais il consiste désormais à les exploiter (Ferreira, Poco, Vo, Freire, & Silva, 2013).

1.2 Objectifs

Dans cette optique d'analyse de l'offre et de la demande de déplacements par taxi, le Bureau du Taxi de Montréal (BTM) s'est doté d'un outil pertinent : le Registre des taxis. Tous les véhicules associés à un permis de propriétaire de taxi doivent désormais être reliés à ce Registre et fournir des données spatiotemporelles dans un format normalisé. Lancé le 16 mai 2018, le Registre des taxis est un point d'échange de taxis (TXP, Taxi Exchange Point) dont le but est d'améliorer la

liaison entre les taxis et leurs clients (Bureau du Taxi de Montréal, 2020a). À Montréal depuis septembre 2017, un règlement du BTM exige que tous les véhicules soient équipés d'un système de positionnement global (GPS, Global Positioning System) afin de "localiser la position et suivre la route du taxi en temps réel" (Ville de Montréal, 2017). De plus, le 16 mai 2018, le Comité exécutif de la Ville de Montréal a adopté l'ordonnance concernant l'obligation pour tout détenteur de permis de propriétaire de taxi de se connecter au Registre des taxis avant le 28 novembre 2018 (Bureau du Taxi de Montréal, Automne 2018). Ainsi, il est désormais possible d'avoir une base de données continue de tous les taxis en service à Montréal.

Le projet de recherche vise à fournir une meilleure compréhension du rôle et des performances actuels du taxi dans la mobilité quotidienne afin d'aider les représentants du secteur des taxis dans leur prise de décision ainsi que, plus généralement, ceux impliqués dans la planification des systèmes de transport et de mobilité. Les projets de recherche préalables de Lacombe (2016) et Laviolette (2017) ont permis d'expérimenter différents types de flux de données et d'identifier les mécanismes requis pour les traiter (Lacombe, 2016; Laviolette, 2017). Plusieurs indicateurs ont également été estimés avec les flux de données d'un seul intermédiaire en service opérant sur l'île de Montréal dans le projet de Lacombe (2016) et de trois intermédiaires en service dans celui de Laviolette (2017) (Lacombe, 2016; Laviolette, 2017). Grâce au Registre, il est désormais possible d'analyser les données de l'ensemble des intermédiaires en service opérant sur l'île de Montréal, soit de l'ensemble des taxis en activité et non plus uniquement d'un échantillon de ces derniers.

Le projet de recherche propose de mettre en valeur les contributions analytiques amenées par le Registre en assurant le développement d'un tableau de bord pour soutenir la prise de décision. Avec plus de 4000 taxis branchés, il est désormais possible de développer des procédures de traitement systématique et automatisé des données et d'estimer une diversité d'indicateurs sur l'offre et la demande de déplacements par taxi. Cependant le volume très conséquent de données générés par ces taxis présente de véritables enjeux notamment en termes de traitement de données. En effet, ces données ne peuvent être traitées à l'aide des outils classiques de traitement de données tels qu'EXCEL ou ACCESS (Ferreira et al., 2013; Yang, C. & Gonzales, 2017). De plus, il n'est plus question d'effectuer un nettoyage et un processus de validation unique des données comme c'est le cas pour un ensemble fini de données. Il s'agit désormais d'assurer que les méthodes de traitement et de validation restent valables dans le temps. La définition des indicateurs ainsi que les méthodes d'estimation établies dans les projets de recherche précédents seront donc actualisées

afin d'être adaptées aux données provenant du Registre. Les indicateurs les plus pertinents seront identifiés ainsi que la récurrence souhaitée pour les estimations.

Ces indicateurs pourront être consultés sur une plateforme web dédiée. Le développement de cette plateforme de consultation et d'analyse de type tableau de bord s'inspire des exemples d'interfaces de consultation proposés dans les précédents projets ainsi que des besoins exprimés par les usagers potentiels de l'interface.

Les principaux objectifs du projet de recherche sont récapitulés dans le Tableau 1-1.

Tableau 1-1 Principaux objectifs de recherche

Objectif 1	Automatiser l'extraction des données du Registre et le calcul des indicateurs
Objectif 2	Développer une plateforme de consultation et d'analyse de ces indicateurs qui agira comme un outil d'aide à la planification stratégique

1.3 Structure du mémoire

Le mémoire est organisé comme suit. Le présent chapitre introduit les problématiques, les objectifs de recherche ainsi que la structure du mémoire.

Le Chapitre 2 comporte une revue de littérature qui présente dans un premier temps les enjeux liés à la visualisation et l'analyse de l'offre et de la demande des déplacements en taxi. Les principales thématiques d'analyses et modèles développés à l'aide des données GPS de taxi sont présentés. Puis, les différentes caractéristiques de visualisation essentielles à la conception d'un tableau de bord communiquant les informations de manière efficace et juste sont mis en avant. Le chapitre se conclut sur une synthèse de la revue de littérature.

Le Chapitre 3 présente la méthodologie de traitement des données. Après une mise en contexte des données utilisées, une description de ces dernières est fournie. Puis, dans l'objectif de pouvoir estimer une variété d'indicateurs, des procédures systématiques et automatisées de traitement des données sont établies. Les principales étapes du processus général de traitement des données sont résumées : de l'extraction des données brutes au traitement de ces dernières jusqu'à la visualisation des indicateurs.

Le Chapitre 4 s'intéresse à la résolution de l'enjeu d'irrégularité des intervalles de temps entre deux points consécutifs. Le chapitre traite donc des méthodes d'identification des groupes de statut correspondant aux activités des taxis. Plus particulièrement, une méthode d'identification des courses régulières se basant sur l'étude d'une population de référence est développée.

Le Chapitre 5 concerne les indicateurs. Dans un premier temps, les principaux indicateurs qu'il serait pertinent d'estimer sont identifiés et classifiés selon six objets d'analyse, à savoir la course, le véhicule de taxi, le chauffeur, l'intermédiaire en service, le poste d'attente et les origines et destinations des courses. Puis les déclinaisons possibles de ces principaux indicateurs ainsi que les diverses formes de visualisation réalisables sont détaillées. Enfin, les défis méthodologiques liés au calcul de ces indicateurs sont mis en évidence.

Le Chapitre 6 présente le tableau de bord assurant la visualisation et l'analyse des indicateurs d'offre et de demande en déplacements par taxi. La structure en entonnoir du tableau de bord est présentée : de la page d'accueil de faits saillants à des distributions fréquentielles et temporelles plus détaillées. Afin d'illustrer les différents niveaux de visualisation des données du tableau de bord, l'objet course est analysé ainsi que les origines et destinations.

Le Chapitre 7 conclut ce mémoire en présentant les principaux résultats de la recherche. Les limitations liées aux données et aux méthodologies développées ainsi que les contributions au sujet y sont mises en évidence. Le document s'achève avec quelques éléments de discussion et des perspectives de recherche futures sur l'industrie du taxi.

CHAPITRE 2 REVUE DE LITTERATURE

Ce chapitre propose une revue de la littérature sur le taxi en définissant tout d'abord ce service ainsi que son rôle dans la mobilité. Puis un résumé des travaux réalisés sur le taxi et notamment des différentes modélisations est présenté. Une synthèse des indicateurs pertinents identifiés dans la littérature est proposée. Finalement, une revue sur les tableaux de bord est réalisée. Les caractéristiques essentielles de visualisation à prendre en compte lors de la conception d'une interface de visualisation sont énoncées.

2.1 Présentation de l'industrie du taxi

Une courte présentation des faits majeurs de l'histoire du taxi en Europe et en Amérique de Nord est réalisée avant de définir le service de taxi ainsi que son rôle dans la mobilité.

2.1.1 Histoire et définition

2.1.1.1 Historique

Le taxi est la plus ancienne forme de transport public conventionné au monde, les premières réglementations sur les taxis remontant au XVIIe siècle (Cooper, Mundy, & Nelson, 2010), avec par exemple au Royaume-Uni l'ordonnance de 1654 pour la réglementation des cochers de Hackney (« the 1654 Ordinance for the Regulation of Hackney Coachmen »). Ce règlement vise alors à garantir un niveau de service pour la prestation de taxis tirés par des chevaux. Les premiers taxis motorisés remontent quant à eux aux années 1890, d'abord sous la forme de véhicules électriques alimentés par batterie. Ainsi, la Figure 2-1 présente le « London Electrical Cab », également connu sous le nom de "Hummingbird" (Colibri) en raison du bruit qu'il génère ou le "Bersey Taxi" d'après son créateur Walter Charles Bersey, qui a fait son apparition dans les rues de Londres le 19 août 1897 (Cooper et al., 2010; Wade, 2018).

Figure 2-1 Premier taxi motorisé (Wade, 2018)

Ces véhicules électriques sont rapidement remplacés par des véhicules à essence et diesel qui représentent encore aujourd'hui la grande majorité de la flotte de l'industrie. Le premier taxi à essence a ainsi été réalisé par le constructeur français Prunel et est introduit dans les flottes de Paris et de Londres à partir de 1903. Ce-dernier a été suivi, au Royaume-Uni, par un grand nombre de modèles de taxis d'autres fabricants. Cette grande diversité des modèles a conduit, à Londres, à l'introduction de règles, telles que les « Metropolitan Conditions of Fitness » en 1906, afin de limiter le service de taxi à certains modèles. De plus, en 1907, l'installation de taximètres dans la flotte de taxis Londoniens est rendue obligatoire. Cette exigence aurait donné naissance au nom de "taxi" que nous utilisons encore aujourd'hui. Le terme "taxi" proviendrait en effet de l'allemand "Taxemeter" qui désigne le compteur conçu pour mesurer la taxe (soit le prix du trajet), inventé en 1891 par Wilhelm Bruhn (Cooper et al., 2010).

L'histoire du taxi en Amérique du Nord est plutôt similaire à celle de Londres. En effet, en Amérique du Nord, les chariots à chevaux, « Hackney Carriages » étaient également un moyen de transport populaire au début du XIXe siècle. De plus, les véhicules électriques font une brève apparition, avec notamment une flotte de taxis électriques introduite en 1897 par l'Electric Carriage and Wagon Company. De la même manière, ceux-ci sont rapidement remplacés par des véhicules à essence. En 1907, l'entrepreneur new-yorkais, Harry N. Allen introduit une flotte de 65 taxis à essence importés du constructeur français Darracq (Hodges, 2020). Il met alors en place un nouveau modèle de gestion en employant des chauffeurs sous contrat, en contrôlant l'offre de taxis dans la ville et en assurant un contrôle des tarifs basés sur la distance parcourue, déterminés à l'aide

de taximètres préconfigurés réduisant ainsi les possibilités d'abus de prix. Les taxis d'Allen seraient également les premiers taxis à être peints en jaune, puisque le jaune serait la couleur la plus visible de loin (Cooper et al., 2010). Il s'en suit une standardisation des véhicules de taxis avec le développement à partir de 1910 du « Yellowcab », un modèle conçu par John Hertz qui à la fois fabrique le véhicule et exploite une compagnie de taxis (Cooper et al., 2010; Hodges, 2020).

Le premier service de taxi connu au Canada date quant à lui de 1837, dans la ville de Toronto (Lacombe, 2016). Dans leurs mémoires, Lacombe (2016) et Laviolette (2017) retracent l'historique du taxi et notamment l'évolution de la régulation de cette industrie (Lacombe, 2016; Laviolette, 2017). Mathieu (2020) retrace quant à lui l'historique de la compétitivité des déplacements par taxi (Mathieu, 2020).

2.1.1.2 Définitions du taxi

Si les services de taxi peuvent varier selon les villes en englobant différents véhicules, plusieurs formes de services ou structures organisationnelles, une définition générale peut tout de même être établie (Lacombe, 2016). Le Tableau 2-1 présente deux définitions générales du taxi.

Tableau 2-1 Définitions générales du taxi

Auteurs	Définition
Cooper, Mundy, & Nelson (2010)	Le taxi propose, sous différentes formes, un transport privé pour un individu ou un petit groupe d'individus, en échange d'une rémunération (Cooper et al., 2010).
Salanova, Estrada, Aifadopoulou, & Mitsakis (2011)	Les taxis sont des véhicules privés utilisés pour les services de transport public assurant un transport individuel porte à porte (Salanova, Estrada, Aifadopoulou, & Mitsakis, 2011).

La définition du taxi la plus appropriée pour cette étude est celle du gouvernement du Québec, qui décrit le taxi comme une automobile qualifiée utilisée pour offrir du transport rémunéré de

personnes, le prix de la course étant réglementé et devant être calculé selon les tarifs établis par la Commission des transports du Québec (CTQ) (Gouvernement du Québec, 2020).

Les taxis sont généralement affiliés à des intermédiaires en service qui coordonnent les appels et les répartissent aux taxis (Billhardt et al., 2019). Mais ils peuvent également être indépendants.

Les services de taxi peuvent être regroupés en trois principaux modes opératoires (Salanova et al., 2011) :

- **Poste d'attente** : un client peut se rendre à un poste d'attente afin de commencer un déplacement en taxi.
- **Héler** : un client peut héler un taxi. Un avantage de ce mode opératoire est que le client n'a pas à se déplacer jusqu'à un poste d'attente. Cependant, le temps d'attente peut être très variable.
- **Centre de répartition** : un client peut commander un taxi en appelant un centre de répartition (ou via une application mobile). Ce mode présente l'avantage pour le client de pouvoir choisir le prestataire de service ainsi que le choix du lieu d'embarquement.

Enfin, certains auteurs estiment que les services de taxi au niveau de l'aéroport constituent un mode opératoire à part (Salanova et al., 2011).

2.1.2 Rôle du taxi

Selon Cooper et al. (2010), le taxi est un mode de transport urbain important mais peu exploité. Ce dernier jouerait en effet un rôle significatif dans le secteur du transport dans le monde entier en offrant un service constant et directement identifiable. Il contribue de plus de manière significative à l'économie d'un lieu en permettant notamment l'accès à des activités essentielles, en soutenant le tourisme et en offrant une accessibilité en cas d'urgence aux individus ne possédant pas de voiture ou ne pouvant accéder aux transports en commun (Cooper et al., 2010). Les sections 2.1.2.1 et 2.1.2.2 mettent en évidence le rôle du taxi dans la mobilité urbaine.

2.1.2.1 Dans la mobilité quotidienne

Un des principaux acteurs impliqués dans la circulation quotidienne dans les zones urbaines sont les flottes de taxis. Elles se composent de plusieurs milliers de véhicules dans les grandes villes avec par exemple, environ 15 000 taxis à Madrid, en Espagne (Billhardt et al., 2019) et plus de 13

500 taxis à New York (New York City Taxi & Limousine Commission, De Blasio, & Joshi, 2018). Certes le taxi est présent en ville mais qu'en est-il de son rôle dans la mobilité quotidienne?

De nombreux auteurs affirment que le taxi joue un rôle essentiel dans le réseau de transport des villes (Billhardt et al., 2019; Kamimura et al., 2013). Phiboonbanakit et Horanont (2016) déclarent ainsi que le taxi a été reconnu comme un moyen de transport de base essentiel dans la plupart des capitales du monde (Phiboonbanakit & Horanont, 2016). En effet, selon Darbéra (2010) les besoins d'utiliser une automobile ne sont qu'occasionnels en ville et le taxi pourrait satisfaire à ces besoins (Darbéra, 2010).

Selon les circonstances, le taxi peut être un mode complémentaire ou un substitut du transport collectif (Salanova et al., 2011; Wang, F. & Ross, 2019; Yang, C. & Gonzales, 2017).

Wang et Ross (2019) identifient ainsi trois catégories de courses dans l'agglomération de New York: celles qui sont en compétition avec le transport en commun, celles qui sont en substitution avec ce dernier et celles qui sont complémentaires au transport en commun (Wang, F. & Ross, 2019).

La première catégorie correspond aux déplacements qui auraient pu s'effectuer en transport en commun. Les taxis sont alors en concurrence avec les bus, métros, trains de banlieue et autres systèmes de transport public (Yang, C. & Gonzales, 2017).

La deuxième catégorie correspond aux déplacements qui, d'un point de vue spatial ou temporel, ne peuvent pas s'effectuer par le transport en commun. Les taxis peuvent également être utilisés pour offrir une mobilité à certaines catégories de la population, notamment là où les transports publics sont limités, tels que pour le ramassage scolaire ou pour les personnes à mobilité réduite (Austin Jr, 2011; Darbéra, 2010).

Enfin, la troisième catégorie constitue les déplacements ayant pour origine ou pour destination des stations de transport en commun. Les taxis complètent alors le transport en commun en transportant des passagers de leur origine à une station de transport en commun ou inversement d'une gare de transit à leur destination finale, en desservant les premiers et derniers kilomètres (Hartikainen et al., 2019; Wang, F. & Ross, 2019; Yang, C. & Gonzales, 2017).

Le taxi, en raison de son service porte-à-porte, de sa flexibilité, de son confort et intimité ainsi que de sa plage de durée de fonctionnement, est un mode qui peut s'avérer plus pratique que les

transports en commun qui peuvent parfois manquer de flexibilité et ne disposer que d'une couverture spatiale et temporelle limitée. Aussi, l'absence de frais de stationnement présente un avantage en ville face à l'automobile individuelle (Salanova et al., 2011). Cependant, en absence de gestion et de contrôle de la flotte des taxis, le nombre important de véhicules de taxis circulant en ville présente des enjeux, notamment lorsque ces derniers circulent à vide. Yang et al. (2000) révèlent ainsi que les véhicules de taxis représentent jusqu'à 60% du flux de circulation global à Hong Kong durant les heures de pointe, la plupart des véhicules étant vides (Yang, H., Lau, Wong, & Lo, 2000). En plus de présenter des enjeux pour les chauffeurs de taxis qui réalisent moins de bénéfices lorsque la distance à vide parcourue est élevée, cela pose également des enjeux en termes de congestion et de pollution (Lee & Sohn, 2017; Salanova et al., 2011; Yuan et al., 2013; Zhou et al., 2020).

2.1.2.2 Dans un contexte de Mobilité en tant que service

Dans le contexte actuel où l'on cherche à réduire la dépendance à l'automobile privée et à mettre en place une mobilité plus durable, la mobilité en tant que service (MaaS, Mobility as a Service) offre un véritable changement de paradigme (Li & Voege, 2017; Little, 2018; Maas Alliance, 2020; Smith, G., 2020; Sochor, Arby, Karlsson, & Sarasini, 2018). La MaaS Alliance (2020) propose la définition suivante de la mobilité en tant que service : « La mobilité en tant que service (MaaS) est l'intégration de diverses formes de services de transport en un seul service de mobilité accessible à la demande » (Maas Alliance, 2020). L'objectif est d'offrir la meilleure combinaison de services permettant de répondre au besoin de mobilité du client. Le service de mobilité est donc centré sur l'utilisateur. Dans ce concept de MaaS, une seule application est donc utilisée pour accéder et payer pour différents modes de transport, qu'ils soient publics ou privés, pour un déplacement urbain ou interurbain. La planification, la réservation, l'accès aux informations en temps réel, la billetterie et ou encore le paiement, tous seront intégrés à la même interface (Kamargianni & Matyas, 2017). Ainsi, avec un forfait de MaaS, un utilisateur peut par exemple bénéficier d'une utilisation illimitée des transports publics, d'un certain nombre de courses en taxi et de plusieurs jours de location de voiture (Li & Voege, 2017).

Dans les discussions sur la mobilité en tant que service, le taxi apparait comme un acteur incontournable de l'ensemble des services sur lesquels peuvent compter les voyageurs (Lyons, Hammond, & Mackay, 2019; Vij, Ryan, Sampson, & Harris, 2020). En effet, la combinaison de

différents modes de transport tels que le covoiturage, le taxi ou encore le vélo, peut permettre de compléter les lignes fixes et les itinéraires classiques des transports publics (Hartikainen et al., 2019; Li & Voege, 2017).

A titre d'exemple, le premier rapport de l'application Whim indique que les taxis détiennent un rôle important dans l'écosystème de mobilité en tant que service puisqu'ils assurent un segment de mobilité que les transports publics ne couvrent pas nécessairement (Hartikainen et al., 2019). Cette application de mobilité en tant que service permet d'intégrer le paiement des transports publics, des taxis, de la location de voitures, du covoiturage et des déplacements à vélo. Le rapport révèle également que les utilisateurs de l'application combinent l'utilisation du taxi aux transports en commun trois fois plus souvent que le résident typique d'Helsinki. En outre, il est suggéré que le taxi aide à résoudre le problème du premier/dernier kilomètre (Hartikainen et al., 2019).

2.1.3 Synthèse des travaux sur le taxi

Les sections suivantes présentent les différentes modélisations réalisées dans l'optique d'une meilleure gestion de la flotte des taxis. De nombreux modèles sur la répartition des courses ou encore la prévision de la demande des déplacements en taxis ont été réalisés (Alisoltani et al., 2020). Salanova et al. (2011) réalisent ainsi une revue détaillée des différentes modélisations de l'industrie du taxi depuis les années 1970, notamment des premières modélisations sur la régulation de cette industrie (Salanova et al., 2011). Cependant, grâce aux récentes avancées technologiques et à la présence des appareils de relevé GPS dans les taxis, il a été possible d'améliorer ces modélisations. Dans les sections suivantes l'attention est donc particulièrement portée sur l'impact de la disponibilité et de l'utilisations des données GPS de taxis dans les différentes modélisations.

Les méthodes de collecte des données GPS peuvent différer selon les villes mais la structure générale des données recueillies est semblable. En effet, les principaux attributs des données GPS de taxi relevées sont (Kuang, An, & Jiang, 2015; Liao et al., 2010; Pan, Qi, Wu, Zhang, & Li, 2012; Phiboonbanakit & Horanont, 2016; Wang, Y., Zhu, He, Yue, & Li, 2011; Zheng, Rasouli, & Timmermans, 2014):

- **l'identifiant du taxi**
- **la position du taxi**
- **la vitesse instantanée du véhicule**

- **l'état du taxi** (cet attribut peut par exemple être un booléen indiquant si un client est à bord ou pas (Liao et al., 2010) ou dans le cadre du Registre une chaine de caractères indiquant le statut du taxi (Beaudoin, David, 2017))
- **l'horodatage** (« *timestamp* ») (soit la date et l'heure correspondant au moment où l'information a été recueillie)

Les appareils GPS des taxis prélèvent les données en fonction du temps ou de la distance parcourue, et le taux d'échantillonnage diffère selon les systèmes. Il peut également varier selon que le statut est vacant ou occupé (Wang, Y. et al., 2011). Liao et al. (2010) utilisent par exemple des données GPS de centaines de taxi afin de visualiser les traces GPS et d'identifier des comportements anormaux de conduite. Les données GPS sont récupérées à chaque 10 à 20 secondes d'intervalle. Dans le cadre du Registre des taxis, les taxis en exploitation sont tenus d'envoyer leurs données au serveur aux 5 secondes, conformément à l'ordonnance du BTM (Beaudoin, David 2017).

Cependant, on ne dispose pas toujours de l'ensemble des points parcourus par un taxi. Parfois, seules les origines et destinations des courses sont disponibles. Le prix de la course peut dans ce cas également être renseigné (Ferreira et al., 2013; Yang, C. & Gonzales, 2017; Zhou et al., 2020). Dans le cas où l'on dispose de plusieurs points GPS, les courses de taxi sont alors reconstituées à partir des données GPS brutes. Les débuts et fins de courses sont identifiés grâce au changement de statut.

Cette similarité dans les attributs principaux relevés permet notamment de faciliter le transfert d'un modèle d'une ville à l'autre.

2.1.3.1 Modèles de répartition des courses commandées (« dispatching »)

Aujourd'hui, la popularisation des appareils mobiles et le développement des systèmes GPS rendent possible l'adaptation dynamique de l'offre de transport à la demande de déplacements pour les opérateurs de transport. Dans le cas des services de taxi traditionnels, la demande est effectuée par l'intermédiaire d'un centre d'appel et les chauffeurs doivent s'appuyer sur leur expérience pour tenter d'optimiser leurs déplacements. De nos jours en revanche, notamment grâce aux plateformes basées sur des applications, les demandes sont centralisées et des techniques d'optimisation plus avancées peuvent être mises en œuvre pour affecter les passagers aux véhicules de taxis (Darbéra, 2017).

En général, les intermédiaires en service appliquent la stratégie du premier arrivé premier servi pour affecter les taxis aux clients. Les chauffeurs de taxi étant généralement des acteurs autonomes, ils peuvent librement choisir d'accepter ou de rejeter une affectation proposée par le service de médiation (Billhardt et al., 2019). Cependant, une fois la course acceptée par le taxi, la répartition est irréversible. Cette méthode se révèle être très inefficace (Egbelu & Tanchoco, 1984). Ainsi, il est essentiel de mettre en place des méthodes de répartition des courses commandées afin de réduire les distances parcourues à vide et par conséquent diminuer le flux de circulation ou encore les émissions de gaz à effet de serre (Billhardt et al., 2019).

Le problème d'affectation des taxis présente deux objectifs principaux. Le premier est la réduction du temps de réponse ou temps d'attente du client, soit le temps entre l'appel d'un client et le moment où le taxi arrive à l'endroit de prise en charge du client. Le second est la réduction des coûts de déplacements à vide, soit les coûts liés au déplacement que le taxi doit effectuer pour se rendre au lieu de prise en charge du client (Alisoltani et al., 2020; Billhardt et al., 2019).

Il est également essentiel de prévoir les temps de trajet avec précision afin de déterminer la disponibilité du taxi et les heures de prise et de dépose du client afin d'assurer une compatibilité avec l'heure de prise en charge souhaitée du client et/ou de l'heure d'arrivée à destination.

Xu et al. (2018), Billhardt et al. (2019) et Alisoltani et al. (2020) réalisent une revue des différents modèles de répartition (Alisoltani et al., 2020; Billhardt et al., 2019; Xu et al., 2018).

Xu et al. (2018) proposent d'optimiser la répartition des commandes sur une longue période (par exemple plusieurs heures ou une journée), en répondant à la fois à la demande actuelle des passagers et en optimisant la demande future estimée (Xu et al., 2018). Toutefois, la méthode proposée est coûteuse en termes de temps de calcul et ne peut être facilement transférée à une autre ville.

Parmi les modèles les plus récents on peut retrouver celui de Billhardt et al. (2019). Les auteurs proposent un modèle d'affectation dynamique des taxis aux clients dans le but de minimiser le temps d'attente global des passagers. Leur algorithme d'affectation heuristique considère des réaffectations de taxis si cela peut conduire à des solutions plus performantes au niveau global. Ainsi des taxis qui ont été affectés à une course mais qui sont toujours en route peuvent être réaffectés à un autre client. De plus, si le changement d'affectation améliore l'efficacité globale de la flotte mais est désavantageux pour le chauffeur (celui-ci doit par exemple parcourir une distance

plus importante pour rejoindre le client de la nouvelle affectation), les auteurs proposent un système de compensation afin que la nouvelle affectation conviennent également au chauffeur (Billhardt et al., 2019).

Alisoltani et al. (2020) estiment que l'enjeu principal de l'affectation dynamique est de pouvoir résoudre très rapidement le processus d'optimisation, en particulier pour les réseaux à grande échelle qui peuvent comporter des milliers de demandes, tout en prenant en compte la situation du trafic sur le réseau au moment de l'optimisation. Ils proposent donc un algorithme intitulé DTaD («Dynamic Taxi Dispatching Algorithm») qui examine toutes les affectations possibles en tenant compte d'un seuil maximum de temps d'attente des passagers et, enfin, choisit l'itinéraire optimal pour chaque taxi afin de minimiser le temps de déplacement du taxi. La méthode réalise également des groupes des trajets ayant le plus grand potentiel d'être servis en séquence par le même taxi afin de minimiser la distance de déplacement totale et d'optimiser le nombre de véhicules (Alisoltani et al., 2020).

2.1.3.2 Modèles de prévision de la demande

Afin de répartir efficacement la flotte de taxis, il est essentiel de comprendre les facteurs qui déterminent la demande. En effet, l'offre de service proposée s'effectue en fonction de la demande, notamment si l'on souhaite atteindre un équilibre (Lacombe, 2016). Il est donc nécessaire de déterminer les facteurs qui influencent le choix de prendre un taxi et de pouvoir prédire la demande dans l'espace et dans le temps afin de planifier et gérer efficacement la flotte de taxis (Yang, C. & Gonzales, 2017).

La disponibilité des points GPS a permis d'améliorer les modèles tels que les modèles de génération. Ainsi des premiers modèles se basant sur des régressions linéaires multiples permettent d'identifier les facteurs ayant le plus d'impact sur la demande de déplacements en taxi. Yang et Gonzales (2014) identifient ainsi six variables qui ont un fort impact sur la demande à partir de traces GPS de courses de taxi à New York. La population, le temps d'accès au transport en commun, l'âge, le niveau d'éducation, le revenu et le nombre d'emplois ont un impact significatif sur la génération des déplacements (Yang, C. & Gonzales, 2014). Les deux derniers facteurs sont identifiés comme ayant le plus fort impact. Ce modèle est repris par Lacombe (2016) et adapté aux déplacements par taxi à Montréal. Les résultats confirment l'impact des facteurs identifiés par Yang et Gonzales (2014). Dans le cas de Montréal, le pourcentage des ménages dont le revenu est

supérieur à 80 000$ et le pourcentage de la population âgée entre 20 et 29 ans sont alors identifiés comme facteurs ayant le plus d'impact (Lacombe, 2016).

Yang et Gonzales (2017) proposent une amélioration de leur premier modèle de régression. En effet, les auteurs considèrent que le nombre de déplacements en taxi générés par secteur de recensement est une variable de comptage. Les variables de comptage sont un type de variables statistiques prenant des valeurs entières positives. Ces valeurs proviennent d'un processus de comptage (ou processus de dénombrement) qui permet de modéliser un nombre entier aléatoire évoluant dans le temps (Winkelmann, 2008). Le modèle de génération doit donc être basé sur un processus de comptage or les modèles de régression linéaire sont inadéquats pour les données de comptage. Ils proposent donc de modéliser le nombre d'origines de courses dans un secteur de recensement par une loi de Poisson ou par une régression binomiale négative (Yang, C. & Gonzales, 2017). Dans leur précédent modèle basé sur la régression linéaire, le revenu par habitant avait été identifié comme une variable explicative significative (Yang, C. & Gonzales, 2014). Or les auteurs ont établi que le revenu est fortement corrélé avec la mesure du niveau d'éducation à New York. Par conséquent, afin d'éviter les problèmes liés à l'autocorrélation de ces deux variables explicatives, seul le niveau d'éducation est conservé dans leur plus récent modèle (Yang, C. & Gonzales, 2017).

Dans son mémoire sur la caractérisation et l'étude de compétitivité des déplacements par taxi, Mathieu (2020) réalise une revue détaillée des facteurs déterminants de la demande de déplacements par taxi ainsi que des modèles de prévision de la demande. Parmi les facteurs identifiés comme ayant un impact sur le choix du taxi, l'auteur identifie la longueur de déplacement, le coût du déplacement, la disponibilité et présence des autres modes, le jour de la semaine ou encore les points d'intérêts tels que l'aéroport ou le centre-ville (Mathieu, 2020). En ce qui concerne les modèles de prévision de la demande, l'auteur les classe selon si la modélisation est à court ou à long terme. Dans le cas de la prévision de la demande à long terme, une autre catégorisation est faite selon si le modèle génère une quantité de déplacements ou la probabilité de choisir un mode en particulier. La première catégorie concerne les modèles de génération et de distribution et la seconde les modèles de choix modal (Mathieu, 2020).

2.1.3.3 Modèle de recherche du prochain client

La recherche de passagers par le chauffeur est au cœur de l'activité de taxi. Il est donc dans l'intérêt du chauffeur du taxi de connaître les lieux et trajets où il est à même de trouver des clients afin de réduire sa distance et durée parcourues à vide et de maximiser son revenu (Laviolette, 2017).

L'utilisation des données GPS de taxi permet le développement de modèles de recherche de lieux ou d'itinéraires pour les conducteurs afin notamment d'améliorer l'efficacité énergétique, le profit ou encore de réduire la distance et la durée de déplacement lors de la recherche du prochain client. Ainsi la plupart des approches essaient soit de maximiser les profits de la prochaine course et la probabilité de trouver le prochain client et/ou de minimiser la consommation d'énergie lors de la recherche du prochain client (Ge, Liu, et al., 2011; Hu, Yang, & Huang, 2015; Kamimura et al., 2013; Qu, Zhu, Liu, Liu, & Xiong, 2014; Yuan et al., 2013; Zhang, M., Liu, Liu, Hu, & Yi, 2012). Cependant, ces travaux se concentrent sur l'optimisation des mesures pour la prochaine course uniquement et non sur une optimisation plus globale, sur toute une période. Face à cette limite, Zhou et al. (2020) proposent un modèle de recommandation d'itinéraire dont l'objectif est la maximisation du profit (soit les revenus moins le coût du carburant) sur une période temporelle et non uniquement sur la prochaine course (Zhou et al., 2020).

2.1.3.4 Détection des anomalies GPS

Si les données GPS de taxi sont utilisées pour de nombreux modèles, la qualité de ces données doit tout d'abord être vérifiée. Avant d'être utilisées pour des modèles, les données GPS des taxis sont traitées et nettoyées afin que l'analyse qui en découle soit fiable. Les données erronées peuvent être dues à une mauvaise concordance entre les données GPS et les coordonnées cartographiques, à la faible précision des appareils de navigation GPS, à la faible fréquence d'échantillonnage ou encore induites par les systèmes d'information tels que les satellites. Ce type de données est également sujet à des erreurs induites par des manipulations humaines, notamment lorsque l'envoi de l'information sur le début et la fin d'une course dépend de la manipulation d'un appareil par le chauffeur (Zhang, J., 2012; Zheng et al., 2014). Si les études sur les sources d'erreurs des données GPS sont nombreuses, l'étude des méthodes de détection systématique des données erronées a cependant reçu moins d'attention (Zheng et al., 2014).

Les approches existantes de détection des valeurs aberrantes et de nettoyage des données utilisent souvent des règles simples telles que le fait que les déplacements doivent se faire à l'intérieur des

limites de la ville et/ou que les distances géométriques doivent respecter certains seuils (Zhang, J., 2012). Veloso et al. (2011) considèrent par exemple que les courses d'une distance inférieure à 200 mètres ou supérieure à 30 kilomètres sont erronées. Aucune justification n'est donnée pour la limite inférieure. La limite supérieure de 30 kilomètres est déterminée en fonction du trajet le plus long reliant les deux extrémités de la ville de Lisbonne qu'ils estiment être d'environ 22 kilomètres (Veloso, Phithakkitnukoon, & Bento, 2011).

La disponibilité des données GPS permet désormais d'examiner l'ensemble de la trace GPS afin d'identifier d'éventuelles anomalies. Cependant, dans certains cas, pour des raisons de protection de la vie privée, de volume de données ou des raisons politiques ou de gestion, les traces GPS complètes ne sont pas disponibles (Zhang, J., 2012). Lorsque les données sont disponibles, au lieu de comparer à des valeurs seuils, il est par exemple possible de comparer la trace GPS obtenue à partir des données à la trace du trajet réseau le plus court. Cependant ce calcul du chemin le plus court peut présenter un enjeu en termes de temps d'exécution selon les algorithmes utilisés et selon la taille du réseau étudié (Zhang, J., 2012).

Parmi les algorithmes classiques on peut retrouver l'algorithme de Dijkstra et les algorithmes A* (Wagner & Willhalm, 2007). De nos jours, plusieurs algorithmes de chemin le plus court, conçus spécifiquement pour les réseaux routiers, atteignent désormais des rendements nettement supérieurs à ceux des algorithmes génériques. Parmi eux, on peut citer l'algorithme contraction hiérarchiques (« Contraction Hierarchies », CH) développé par un groupe de l'Institut de technologie de Karlsruhe en Allemagne (Geisberger, Sanders, Schultes, & Delling, 2008). Deux logiciels libres basés sur l'algorithme CH, à savoir MoNav et OSRM, sont actuellement disponibles (Geisberger et al., 2008; OSRM, 2020). L'Open Source Routing Machine ou OSRM est un routeur open-source conçu pour être utilisé avec les données du projet OpenStreetMap (OSRM, 2020).

Différentes approches de validation des points GPS sont identifiées par Zhang (2012) (Zhang, J., 2012) :

- Lorsque l'ensemble de données ne fournit que les deux extrémités (Origine/Destination) des trajectoires GPS :
    - Utilisation de valeurs seuils pour éliminer des courses trop courtes ou trop longues (Liu, Y., Kang, Gao, Xiao, & Tian, 2012; Veloso et al., 2011)

  - Utilisation de distribution d'une mesure telles que la distance ou la durée de course afin d'éliminer les valeurs qui ne se situent pas dans une certaine fourchette (Wang, Y. et al., 2011)
  - Utilisation d'une analyse géospatiale afin d'identifier si les points d'origine ou de destination se trouvent à l'extérieure d'une zone ou dans des endroits aberrants tels que les lacs ou les rivières.

- Lorsque l'ensemble de données fournit les traces GPS complètes :
  - Association des trajectoires GPS aux réseaux de rues et intégration de la topologie des réseaux de rues afin de détecter les valeurs aberrantes (Liu, W., Zheng, Chawla, Yuan, & Xing, 2011).
  - Discrétisation des traces GPS en unités de maillage et utilisation de ces dernières comme unités de base pour la détection des valeurs aberrantes (Ge, Xiong, Liu, & Zhou, 2011; Zhang, D. et al., 2011).
  - Validation des traces et points GPS sur le réseau routier (« map-matching ») (Zhang, J., 2012).

Lacombe (2016) utilise par exemple une analyse géospatiale puisqu'elle vérifie si les points GPS se trouvent dans une zone appelée « zone de Montréal » s'étendant de Toronto à Sept-Îles en longitude et de New York à Sept-Îles en latitude, afin d'éliminer les points GPS aberrants (Lacombe, 2016).

Les différentes approches mentionnées peuvent également être combinées. Ainsi, en plus d'appliquer les techniques de détection des valeurs aberrantes basées sur les seuils, la distribution et l'analyse spatiale, Zhang (2012) utilise le logiciel MoNav de chemin le plus court et compare les distances enregistrées avec les distances les plus courtes calculées afin de détecter les valeurs aberrantes (Zhang, J., 2012).

Zheng et al. (2014) combinent quant à eux les quatre critères suivants dans leur processus de validation des courses : la précision du signal des données GPS, la comparaison de la vitesse instantanée du véhicule et du mouvement de ce-dernier, l'utilisation des valeurs maximales de vitesse autorisée sur les routes et la comparaison des distances calculées à partir des données GPS à celles calculées à partir des cartes.

Laviolette (2017) combine également plusieurs critères puisqu'il utilise les distributions des distances, durées et vitesses de course afin de déterminer des fourchettes de validité ainsi que des valeurs seuils telles qu'un critère de vitesse maximale correspondant à la limite de vitesse autorisée sur les autoroutes au Québec (Laviolette, 2017).

Enfin, Djenouri et al. (2019) proposent une revue exhaustive des différentes approches et méthodes de détection des valeurs aberrantes à partir des données GPS et on peut y retrouver des méthodes impliquant des données de taxis (Djenouri, Belhadi, Lin, Djenouri, & Cano, 2019).

2.1.3.5 Analyses des données montréalaises

En 2014, une première étude de la demande en déplacements de taxi à Montréal a été effectuée par Pele et Morency (2014) à partir d'un mois de données GPS de 968 véhicules de taxi de l'intermédiaire Taxi Diamond (Pele & Morency, 2014). L'étude révèle notamment que 95% des courses sont effectuées en semaine et que lors de la période de pointe du matin (de 6h à 9h) 32% des origines des courses sont concentrées dans une zone de 12.3 km^2. Les auteurs mettent également en évidence l'incidence des conditions météo ou des jours fériés sur le nombre de courses (Pele & Morency, 2014). Cette première analyse descriptive a permis de mettre en évidence la nécessité de définir des indicateurs d'analyse et de suivi de l'industrie du taxi afin de déterminer le rôle du taxi dans la mobilité quotidienne à Montréal. Lacombe (2016) à nouveau à partir des données de Taxi Diamond, poursuit la caractérisation des déplacements en taxi à Montréal en définissant notamment une liste d'indicateurs de performance et de suivi des activités de taxi identifiés comme pertinents (Lacombe, 2016). Ces derniers sont catégorisés selon l'objet d'étude, à savoir la course, le chauffeur, le véhicule de taxi, la zone ou le client. Elle propose également un modèle de génération des déplacements en taxi se basant sur une régression linéaire des points d'origines et de destination. Laviolette (2017) poursuit le travail entrepris par Lacombe (2016) en intégrant les données GPS des taxis de deux autres intermédiaires en service et en proposant de nouveaux indicateurs (Laviolette, 2017). Il s'intéresse notamment à la régulation de cette industrie ainsi qu'aux enjeux que présentent les nouveaux services de *ridesourcing*. Ce mode de transport alternatif propose un service de transport sur demande par des entreprises de transports technologiques (« Transportation Network Companies » ou « TNCs »), telles que Uber (Laviolette, 2017). L'auteur parvient à mettre en évidence des variations saisonnières de la demande ainsi que la répartition de la demande et de l'offre sur le territoire. Enfin, il analyse également l'utilisation

des permis et estime qu'en 2016 le taux d'utilisation moyen d'un permis est de 46,6% (Laviolette, 2017). Plus récemment, Mathieu (2020) réalise une caractérisation et étude de compétitivité des déplacements en taxi à Montréal. Il identifie les facteurs influençant le choix du taxi et réalise une analyse de disponibilité unimodale et multimodale des alternatives. Enfin, l'auteur propose une méthodologie de caractérisation des origines et destinations de déplacements en taxi pour l'analyse de la disponibilité et de la compétitivité des alternatives. Un modèle de prédiction de la compétitivité du taxi pour des courses commandées n'ayant que le transport en commun comme alternative est également réalisé (Mathieu, 2020).

2.1.3.6 Synthèse des indicateurs d'offre et de demande

Une première méthode de classification de ces indicateurs d'offre et de la demande basée sur un seul intermédiaire en service a déjà été élaborée par Lacombe (Lacombe, 2016). Cette classification a été complétée par Laviolette qui s'est basé sur trois intermédiaires en service (Laviolette, 2017). Lacombe (2016) et Laviolette (2017) classifient les indicateurs retrouvés dans la littérature selon l'objet analysé, soit :

- la **course** : soit le trajet en taxi, de l'origine (embarquement) à la destination du client (débarquement). C'est l'objet le plus étudié (Lacombe, 2016).
- le **véhicule** : soit le véhicule utilisé pour assurer le service de taxi. Un véhicule de taxi se partage souvent parmi plusieurs chauffeurs (Mathieu, 2020).
- le **chauffeur** : soit la personne assurant le service de taxi. Elle doit détenir un permis de chauffeur (Austin Jr, 2011; Geneste, 2017).
- le **poste d'attente** : soit un endroit où les taxis peuvent se garer en attendant d'être engagés dans une course. Un passager peut se rendre à un poste d'attente afin de commencer une course de taxi. En règle générale, la réglementation contraint les passagers à prendre le premier taxi de la file (« First In, First Out ») (Moreira-Matias et al., 2012; Salanova et al., 2011).
- le **client** : soit la personne bénéficiant du service de taxi. Cet objet est peu étudié car les informations sur le client sont rarement relevées et/ou disponibles. Dans le Registre des taxis par exemple aucune information n'est disponible sur les clients.

Trois objets sont ajoutés à cette liste :

- le **permis** : soit la licence ou l'autorisation délivrée par les autorités pour assurer le service de taxi. Le propriétaire du véhicule doit se procurer un permis de propriétaire de taxi (Austin, 2011; Geneste, 2017) .
- le **paiement** : soit tout ce qui est associé au prix de la course et au mode de paiement.
- la **zone** : soit l'étude de zones en particulier telles que l'aéroport.

Le Tableau 2-2 résume la liste des indicateurs d'offre et de demande en déplacements de taxi identifiés comme pertinents dans la littérature. Tel que mentionné précédemment, l'objet course est celui le plus étudié dans la littérature, comme en témoignent les nombreux indicateurs qui y sont liés.

Tableau 2-2 Liste des indicateurs d'offre et de demande en déplacements de taxi inspirée de Lacombe (2016)

OBJET	INDICATEUR	REFERENCES	
COURSE	Nombre de courses	• (Ferreira et al., 2013) • (Savage & Vo, 2013) • (Pele & Morency, 2014) • (Bischoff, Maciejewski, & Sohr, 2015)	• (Lacombe, 2016) • (New York City Taxi & Limousine Commission, De Blasio, & Joshi, 2016; New York City Taxi & Limousine Commission et al., 2018) • (Laviolette, 2017) • (Zhou et al., 2020)
	Distance et durée de course	• (Veloso et al., 2011) • (Zhang, J., 2012) • (Savage & Vo, 2013) • (Pele & Morency, 2014) • (Bischoff et al., 2015)	• (Lacombe, 2016) • (Laviolette, 2017) • (New York City Taxi & Limousine Commission et al., 2018) • (Mathieu, 2020)
	Vitesse de course	• (Zhang, J., 2012) • (Savage & Vo, 2013)	• (Pele & Morency, 2014)
	Nombre de courses commandées (via une application mobile ou via un centre de répartition des appels)	• (Lacombe, 2016) • (New York City Taxi & Limousine Commission et al., 2016, 2018)	• (Mathieu, 2020)
	Nombre de courses hélées	• (New York City Taxi & Limousine Commission et al., 2016) • (Lacombe, 2016)	• (Mathieu, 2020)
	Nombre de courses partagées	• (New York City Taxi & Limousine Commission et al., 2018)	
	Nombre de courses de transport adapté	• (New York City Taxi & Limousine Commission et al., 2016, 2018)	

Tableau 2-2 (Suite) Liste des indicateurs d'offre et de demande en déplacements de taxi inspirée de Lacombe (2016)

OBJET	INDICATEUR	REFERENCES
COURSE	Nombre d'origines et de destinations	• (King, Peters, & Daus, 2012) • (Ferreira et al., 2013) • (Savage & Vo, 2013) • (New York City Taxi & Limousine Commission, Bloomberg, & Yassky, 2014) • (Zhang, Y., 2014) • (Lacombe, 2016) • (New York City Taxi & Limousine Commission et al., 2016, 2018) • (Riascos & Mateos, 2020)
	Carte de chaleur des Origines et Destinations	• (Zhang, J., 2012) • (Liu, Y. et al., 2012) • (Ferreira et al., 2013) • (Pele & Morency, 2014) • (Zhang, Y., 2014) • (Liu, X., Gong, Gong, & Liu, 2015) • (Lacombe, 2016) • (Phiboonbanakit & Horanont, 2016) • (Zhou et al., 2020) • (Riascos & Mateos, 2020)
	Visualisation des traces GPS et détection de comportements anormaux de conduite	• (Liao et al., 2010)
	Indicateurs de disponibilité d'alternatives et de compétitivité	• (Mathieu, 2020)
VEHICULE	Nombre de véhicules actifs	• (Veloso et al., 2011) • (Lacombe, 2016) • (New York City Taxi & Limousine Commission et al., 2016, 2018) • (Laviolette, 2017)
	Nombre de véhicules actifs accessibles aux fauteuils roulants	• (New York City Taxi & Limousine Commission et al., 2016, 2018)

Tableau 2-2 (Suite) Liste des indicateurs d'offre et de demande en déplacements de taxi inspirée de Lacombe (2016)

OBJET	INDICATEUR	REFERENCES	
VEHICULE	Age moyen des véhicules	• (New York City Taxi & Limousine Commission et al., 2018)	
	Nombre d'accidents impliquant un véhicule de taxi	• (New York City Taxi & Limousine Commission et al., 2016, 2018)	
	Distance parcourue par le véhicule	• (Lacombe, 2016)	• (Laviolette, 2017)
	Distance et durée parcourue à vide	• (Pele & Morency, 2014) • (Zhang, D.-Z., Peng, & Sun, 2014) • (Bischoff et al., 2015)	• (Lacombe, 2016) • (Laviolette, 2017)
	Taux d'occupation	• (Veloso et al., 2011) • (Ge, Liu, et al., 2011) • (Yuan et al., 2013)	• (Zhang, Y., 2014) • (Lacombe, 2016) • (New York City Taxi & Limousine Commission et al., 2016)
	Distance et durée parcourue entre deux courses	• (Veloso et al., 2011)	• (Zhou et al., 2020)
	Heures de service	• (Ge, Liu, et al., 2011) • (Pele & Morency, 2014) • (Lacombe, 2016)	• (Laviolette, 2017) • (Zhou et al., 2020)
	Consommation du véhicule	• (Zhou et al., 2020)	
CHAUFFEUR	Nombre de chauffeurs actifs	• (Lacombe, 2016) • (New York City Taxi & Limousine Commission et al., 2016, 2018)	• (Laviolette, 2017)
	Début et fin des quarts de travail	• (Lacombe, 2016) • (New York City Taxi & Limousine Commission et al., 2016, 2018)	• (Laviolette, 2017)

Tableau 2-2 (Suite) Liste des indicateurs d'offre et de demande en déplacements de taxi inspirée de Lacombe (2016)

OBJET	INDICATEUR	REFERENCES	
CHAUFFEUR	Caractéristiques des chauffeurs (âge moyen, sexe, pays d'origine)	• (New York City Taxi & Limousine Commission et al., 2016, 2018)	
	Revenu	• (Veloso et al., 2011) • (Yuan et al., 2013) • (Zhang, Y., 2014) • (Zhang, D.-Z. et al., 2014)	• (Phiboonbanakit & Horanont, 2016) • (New York City Taxi & Limousine Commission et al., 2016) • (Laviolette, 2017) • (Zhou et al., 2020)
	Rentabilité	• (Zhou et al., 2020)	
PAIEMENT	Prix des courses	• (Zhang, J., 2012) • (Ferreira et al., 2013) • (Phiboonbanakit & Horanont, 2016)	• (Laviolette, 2017) • (New York City Taxi & Limousine Commission et al., 2018) • (Mathieu, 2020)
	Moyen de paiement utilisé	• (New York City Taxi & Limousine Commission et al., 2016, 2018)	
PERMIS	Taux d'utilisation	• (Laviolette, 2017)	
ZONE	Classification de l'usage du sol	• (Pan et al., 2012)	• (Mathieu, 2020)
	Attractivité des zones	• (Wang, H. et al., 2009)	• (Pan et al., 2012)
	Étude d'une zone spécifique (ex : aéroport)	• (New York City Taxi & Limousine Commission et al., 2016, 2018)	• (Laviolette, 2017)
	Relations entre deux zones	• (Veloso et al., 2011) • (Zhang, L., Ahmadi, Pan, & Chang, 2012)	• (Laviolette, 2017)

Tableau 2-2 (Suite) Liste des indicateurs d'offre et de demande en déplacements de taxi inspirée de Lacombe (2016)

OBJET	INDICATEUR	REFERENCES
CLIENT	Distance de marche entre l'origine et le centre du lieu de concentration	• (Zhang, D.-Z. et al., 2014)
	Temps d'attente	• (Phiboonbanakit & Horanont, 2016) • (Mathieu, 2020)
	Nombre de déplacements hebdomadaires/mensuels effectués par taxi	• (Laviolette, 2017)
	Caractéristiques des clients (âge, sexe, etc.)	• (New York City Taxi & Limousine Commission et al., 2014) • (New York City Taxi & Limousine Commission et al., 2016, 2018)
	Objets perdus dans les taxis	• (New York City Taxi & Limousine Commission et al., 2018)
POSTE D'ATTENTE	Proportion de courses s'amorçant aux postes d'attente	• (Lacombe, 2016)
	Nombre de taxis visitant un poste d'attente	• (Laviolette, 2017)

2.2 Visualisation des données

Dans cette deuxième partie de la revue de littérature, l'accent est mis sur la visualisation des données ainsi que sur les caractéristiques essentielles à la conception d'un outil de visualisation. En particulier, le tableau de bord en tant qu'outil d'aide à la prise de décision est explicité.

2.2.1 Tableau de bord

Les dernières années ont été caractérisées par une croissance rapide et une influence de plus en plus importante des technologies de l'information et de la communication (Few, 2006; Kitchin, 2014). Mais qui dit technologies de l'information dit également données. En effet, ces technologies permettent de récupérer une grande quantité de données. Et nous avons pris conscience du grand potentiel qu'offrent des données aussi massives. Ces dernières contribuent en effet à la naissance de nouvelles opportunités en termes d'analyse, pouvant conduire à l'amélioration de la vie des citoyens grâce à des décisions, politiques et mesures correctives fondées sur leur analyse (Ferreira et al., 2013; Vila, Estevez, & Fillottrani, 2018).

Mais les données brutes, telles qu'elles se présentent, ne fournissent pas d'informations compréhensibles. Des lignes consécutives de données brutes ne nous permettent pas d'identifier une tendance ou un problème. Elles doivent être traitées et structurées avant que nous puissions en tirer un sens. Mais surtout, elles doivent être présentées de manière à rendre l'analyse possible. En effet, la visualisation des données nous permet d'exploiter la capacité de notre système visuel à identifier les relations et les tendances (Wexler, Shaffer, & Cotgreave, 2017).

2.2.1.1 Le tableau de bord pour la prise de décisions

Bien que plusieurs outils aient été développés pour permettre la visualisation de données, les tableaux de bord ou *dashboard*, en particulier, sont devenus très populaires ces dernières années, notamment en ce qui concerne l'amélioration des processus de prise de décision au sein du gouvernement et des organisations (Vila et al., 2018) .

L'utilisation du tableau de bord dans le milieu de l'entreprise privée n'est pas récente. En effet, ces dernières intègrent des tableaux de bord dans leur processus internes afin de contrôler et vérifier l'atteinte des objectifs stratégiques. Ils permettent par exemple d'aider à mieux comprendre les modèles de vente, à gérer les ressources humaines ou encore l'efficacité de la production. Ce sont

donc des outils de gestion importants pour la gestion et la prise de décisions au sein d'une organisation donnée (Vila et al., 2018).

Le développement des tableaux de bord dans le secteur public est quant à lui plus récent et est lié à la disponibilité croissante d'importants volumes de données. En effet, pour produire des visualisations, les tableaux de bord requièrent des données en entrée, pouvant être fournies par des initiatives de données ouvertes et par les gouvernements. Selon Kourtit et Nijkamp (2018), les villes sont aujourd'hui de véritables « entrepôts de données » (Kourtit & Nijkamp, 2018). Et les gouvernements utilisent de plus en plus les données dans toutes leurs activités (Matheus, Janssen, & Maheshwari, 2020).

La conception des tableaux de bord dans le secteur public présente toutefois de nombreux enjeux tels que l'anonymisation des données, le respect des réglementations et lois concernant les données ainsi qu'assurer la qualité des données et leur exactitude (Vila et al., 2018). Les organismes publics peuvent utiliser les tableaux de bord à des fins diverses, notamment la transparence, la planification, le suivi des performances ainsi que pour l'élaboration et l'évaluation de politiques en se basant sur des données probantes. Si les tableaux de bord peuvent être utilisés pour soutenir les processus décisionnels, ils peuvent également l'être pour communiquer et interagir avec le public (Allio Michael, 2012; Matheus et al., 2020). Ainsi, dans leur étude sur la ville de Rio de Janeiro au Brésil, Matheus et al. (2020) démontrent que les tableaux de bord améliorent la transparence et la responsabilisation des citoyens (Matheus et al., 2020).

2.2.1.2 Origine

Si traditionnellement les visualisations d'informations étaient des présentations visuelles statiques, avec les progrès technologiques, on assiste aujourd'hui à une utilisation croissante de représentations graphiques interactives et dynamiques de l'information (Meirelles, 2013).

Ainsi, la visualisation de données et la visualisation d'informations font référence à l'utilisation de représentations visuelles interactives et assistées par ordinateur de données abstraites pour renforcer la cognition selon une traduction libre de Mackinlay Card dans *Readings in Information Visualization : Using Vision to Think* (Card, 1999).

Le tableau de bord, en tant qu'outil de communication, provient des Executive Information Systems (EIS), également connus sous le nom de Executive support system (ESS). Développé pour la

première fois dans les années 1980, l'objectif d'un EIS était de simplifier et de soutenir les besoins d'information et de prise de décision des cadres supérieurs en affichant les principales mesures financières par le biais d'une interface facile à utiliser (Few, 2006; Smith, V. S., 2013).

Tout d'abord, bien que la ligne de démarcation entre les tableaux de bord et les rapports soit souvent floue, il est important de différencier ces deux éléments. Un rapport peut être défini comme un document qui fournit des données à des fins de visualisation et d'analyse. Il peut être aussi basique qu'une table de données. La plupart du temps, les rapports exigent du lecteur qu'il exerce son propre jugement et analyse des données. Au contraire, le tableau de bord guide l'analyse de l'utilisateur puisque le tableau de bord a un objectif spécifique (Alexander & Walkenbach, 2010).

2.2.1.3 Définition

Diverses définitions d'un tableau de bord ont été proposées :

- Few (2006) (Few, 2006): "Un tableau de bord est une représentation visuelle des informations les plus importantes nécessaires pour atteindre un ou plusieurs objectifs ; consolidées et disposées sur un seul écran afin que les informations puissent être évaluées en un coup d'œil".

- Alexander et Walkenbach (2010) (Alexander & Walkenbach, 2010): "Un tableau de bord est une interface visuelle qui fournit des vues d'ensemble sur les mesures clés pertinentes pour un objectif ou un processus opérationnel particulier".

- Smith (2013) (Smith, V. S., 2013): "Les tableaux de bord sont des affichages visuels qui présentent sur un seul écran les informations les plus importantes nécessaires pour atteindre des objectifs spécifiques".

- Wexler, Shaffer et Cotgreave (2017) (Wexler et al., 2017): "Un tableau de bord est un affichage visuel de données utilisé pour suivre des conditions et/ou faciliter la compréhension".

Il existe plusieurs critères pour définir un tableau de bord :

- Les tableaux de bord sont des affichages visuels. En effet, pour qu'un tableau de bord soit considéré comme tel, il doit avant tout être doté de mécanismes d'affichage graphique. Un

tableau de bord est un moyen spécifique de représentation ou d'affichage visuel, il ne s'agit pas d'un type de technologie donné (Few, 2006; Wexler et al., 2017).

- Les tableaux de bord permettent de suivre les informations de manière synoptique (Few, 2006; Smith, V. S., 2013).

- Un tableau de bord doit tenir sur un seul écran (Eckerson, 2011; Few, 2006; Hartikainen et al., 2019).

Il n'y a aucune restriction quant à la fréquence des mises à jour des données. Les informations ne doivent pas nécessairement être mises à jour en temps réel. Cela dépend uniquement de la finalité du tableau de bord (Few, 2006). En effet, il existe plusieurs types de tableaux de bord selon la finalité visée. Il existe des tableaux de bord à des fins stratégiques, à des fins analytiques ou à des fins opérationnelles (Eckerson, 2011; Few, 2006; Smith, V. S., 2013). Les tableaux de bord les plus utilisés aujourd'hui sont ceux à vocation stratégique. Le "tableau de bord exécutif" qui affiche les performances de l'entreprise et les tableaux de bord qui sont destinés aux gestionnaires de tous les départements sont par exemple des tableaux de bord stratégiques. Ces tableaux de bord sont centrés sur des indicateurs de performance clés et donnent un aperçu de ce qui est nécessaire à la prise de décision (Few, 2006). Comme ils sont destinés à des objectifs à long terme et non à une intervention d'urgence immédiate, ces tableaux de bord ne nécessitent pas de données en temps réel. Quant aux tableaux de bord analytiques, ils offrent des informations plus détaillées afin de pouvoir aller jusqu'à identifier les causes de ce qui peut être observé.

2.2.2 Le pouvoir de la perception visuelle

Lors de la conception d’un outil de visualisation, l’objectif est de répondre à la question suivante : quelle serait la meilleure façon de montrer cette information? Pourtant, la plupart des tableaux de bord ne parviennent pas à communiquer de manière efficace et efficiente en raison d’une mauvaise conception. En effet, il est important de noter qu’un objectif de la visualisation est de présenter les données aux observateurs d'une manière qui soit informative et compréhensible, d'une part, mais aussi intuitive et sans efforts, d'autre part (Healey, Booth, & Enns, 1996). Aussi bonne que soit la technologie, le succès d'un tableau de bord comme moyen de communication est selon une traduction libre de Stephen Few dans *Information Dashboard Design* un produit de la conception, le résultat d'un affichage qui parle clairement et immédiatement (Few, 2006). Les tableaux de bord

peuvent exploiter la puissance de la perception visuelle pour communiquer des quantités considérables de données, mais seulement si leurs concepteurs appliquent les bons principes et bonnes pratiques de visualisation (Few, 2006; Smith, V. S., 2013).

En effet, notre sens le plus puissant est la vue. Et la vue et la pensée sont inextricablement liées. Par conséquent, pour afficher efficacement des données, il est crucial de comprendre les limites et les capacités de la cognition et de la perception visuelles (Few, 2006; Meirelles, 2013). En suivant des règles axées sur la perception, les données peuvent être affichées de manière à révéler des tendances pertinentes et instructives. Si les règles ne sont pas suivies, les données risquent d'être mal comprises ou susceptibles de nous induire en erreur (Ware, 2004). Pour aider les utilisateurs à reconnaître visuellement les tendances et les anomalies et les amener à prendre des décisions plus éclairées, les tableaux de bord doivent exploiter les capacités visuelles des personnes. La conception graphique de l'information est un enjeu fondamental (Brath & Peters, 2004).

2.2.3 Les défis de conception en matière de visualisation

Malheureusement, l'accent est souvent mis sur des dispositifs d'affichage clinquants qui nuisent aux objectifs d'une communication claire. Un tableau de bord efficace devrait être le résultat d'une conception intelligente : "plus de science que d'art, plus de simplicité que d'éblouissement" selon Stephen Few dans *Information Dashboard Design* (Few, 2006). Il devrait s'agir avant tout d'une question de communication. Les différents auteurs s'accordent sur une règle importante : privilégier la simplicité (Alexander & Walkenbach, 2010; Card, 1999; Few, 2006; Tufte, 2001; Wexler et al., 2017). Par exemple, il ne faut pas tomber dans le piège de la diversité par peur d'ennuyer l'utilisateur. Il faut toujours choisir la méthode d'affichage qui convient le mieux, même si cela se traduit par un tableau de bord qui n'est constitué que de plusieurs instances d'un même type de graphique. Il y a donc d’importants défis à relever, en ce qui concerne non seulement le choix des bons indicateurs mais aussi le choix du support de visualisation approprié pour chaque indicateur. Par exemple, les graphiques à secteurs (ou « camemberts ») ne présentent pas les données quantitatives de manière adéquate car les aires des différents secteurs sont difficilement comparables (Few, 2006; Smith, V. S., 2013; Wexler et al., 2017).

Cependant, outre le choix de l'indicateur et du type de représentation, il faut également, au sein de la représentation elle-même, prêter attention à l'utilisation des couleurs, par exemple. En effet, des couleurs vives doivent par exemple être utilisées pour des données spécifiques qui sont censées se

démarquer du reste. Il ne doit pas y avoir de décorations inutiles et distrayantes comme la visualisation en 3D qui occulte souvent les graphiques (Few, 2006; Tufte, 2001; Wexler et al., 2017). Ainsi, dans *The Visual Display of Quantitative Information*, Edward R. Tufte suggère de maximiser le " ratio données/encre " (Tufte, 2001). Chaque tache d'encre sur un graphique requiert une raison qui devrait presque toujours être que l'encre présente de nouvelles informations (Tufte, 2001). Un pourcentage important de l'encre du graphique et donc du tableau de bord devrait être consacré aux données et non à ce qu'il appelle « chartjunk » que l'on choisit de traduire par "déchets graphiques", à savoir : les bordures, les lignes de quadrillage, les lignes de tendance, les étiquettes, les arrière-plans, etc. L'expression "déchets graphiques" désigne tous les éléments visuels des schémas et des graphiques qui ne sont pas essentiels pour comprendre les informations qui y sont présentées et qui détournent l'attention de l'observateur (Harris, 2000; Tufte, 2001, 2006). Dans le cas des tableaux de bord, nous pourrions minimiser ce que Stephen Few, dans *Information Dashboard Design*, appelle les "pixels non liés aux données" (Few, 2006).

Retenons que la conception graphique et en particulier la conception visuelle de l'information repose sur des processus cognitifs et sur la perception visuelle tant pour sa création que pour son utilisation, ce qu'Isabel Meirelles nomme dans *Design for Information* respectivement "encodage" et "décodage". Un échec dans le processus de décodage implique que la visualisation était également un échec (Meirelles, 2013).

2.2.4 Les attributs pré-attentifs

Enfin, divers auteurs insistent particulièrement sur les attributs de pré-attention et les principes de la Gestalt (Few, 2006; Ware, 2004; Wexler et al., 2017). Ces derniers sont des principes de design destinés à améliorer efficacement la détection des motifs et des inférences perceptives (Few, 2006; Meirelles, 2013). Quant à l'analyse pré-attentive, elle a pour but l'extraction rapide de caractéristiques visuelles de base qui sont détectées par le système visuel humain sans que l'attention soit portée sur une zone spécifique d'une image ou d'un graphique (Healey, C. G., Booth, & Enns, 1996; Treisman, 1985). Le traitement pré-attentif ne nécessite qu'un simple coup d'œil à ce qui est affiché, il est effectué rapidement, sans effort et sans qu'aucune attention ne soit portée sur l'affichage (Treisman, 1985). L'analyse pré-attentive se fait généralement en moins de quelques centaines de millisecondes, avant l'analyse consciente. Ainsi, une analyse ayant lieu en moins de 200 à 250 ms est considérée comme pré attentive (Healey, C. & Enns, 2011; Healey, C.

G. et al., 1996; Kosara, Miksch, & Hauser, 2002; Wexler et al., 2017). En effet, les mouvements des yeux prennent au moins 200 millisecondes pour se déclencher. Toute perception possible avant ce laps de temps implique donc uniquement les informations disponibles en un seul coup d'œil (Healey, C. G. et al., 1996).

Les Figure 2-2 et Figure 2-3 illustrent le traitement pré-attentif. Un exemple de tâche pré-attentive est la détection d'une cible dans un groupe, par exemple la détection d'un triangle orange dans un groupe de triangles bleus (Figure 2-2 (b)). L'objet cible a une caractéristique visuelle « orange » que les objets bleus n'ont pas. Tous les objets non cibles, dans le cas présent les triangles bleus, sont considérés comme des « distracteurs » (Healey, C. & Enns, 2011). Un observateur peut dire d'un simple coup d'œil si la cible est présente ou absente car notre système visuel identifie la cible par une différence de teinte ou couleur.

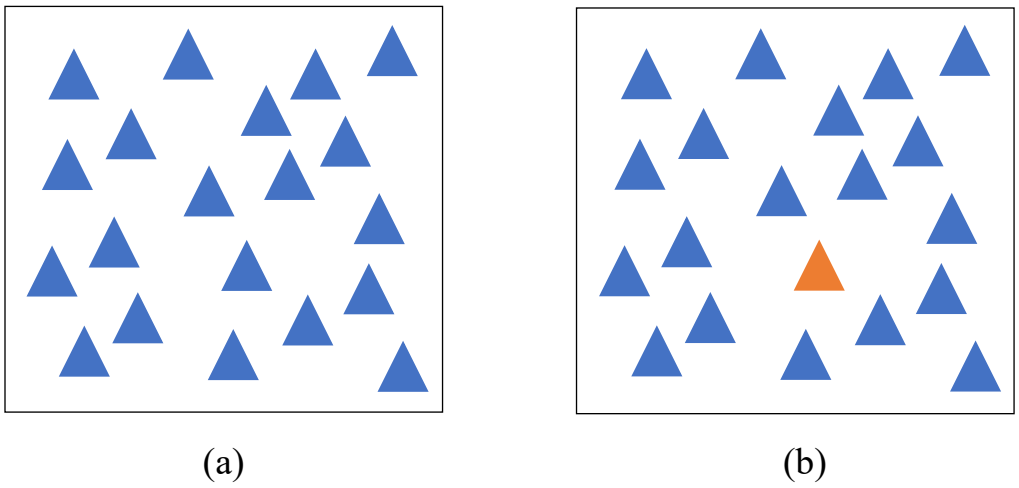
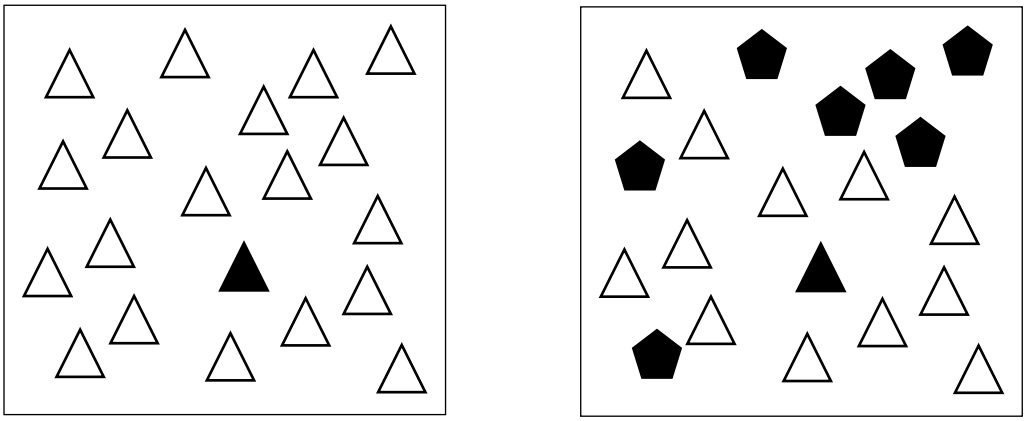


Figure 2-2 Exemple de traitement pré-attentif : (a) population monochrome d'éléments (b) détection d'une cible à l'aide d'un attribut pré-attentif de couleur

Aussi, la présence ou l'absence de teinte ou de remplissage peut également permettre le traitement pré-attentif. Ainsi, dans Figure 2-3 (a), la cible peut être détectée de manière préventive car elle possède la caractéristique "remplie" (« filled »). Mais dans la Figure 2-3 (b), la cible ne peut pas être détectée d'un seul coup d'œil car elle ne présente pas de caractéristique visuelle qui soit unique par rapport aux autres objets.



(a) (b)

Figure 2-3 Exemple de traitement pré-attentif : (a) détection d'une cible à l'aide d'un attribut pré-attentif de « remplissage » (b) absence de détection pré-attentive

C'est également en raison des caractéristiques de pré-attention que l'utilisation des diagrammes en barres (« bar chart ») est recommandée pour les visualisations. En effet, la longueur est l'un des attributs de pré-attention les plus efficaces à traiter (Wexler et al., 2017). Ainsi, les différences de hauteur ou longueur entre les différentes barres d'un tel diagramme sont traitées et analysées rapidement par l'utilisateur.

Il est donc suggéré d'utiliser des caractéristiques de pré-attention pour améliorer la détection d'informations pertinentes dans les visualisations (Meirelles, 2013; Ware, 2004). Stephen Few, dans *Information Dashboard Design*, fait référence à Colin Ware dans *Information Visualization* : *Perception for Design*, lorsqu'il suggère que ces attributs de pré-attention peuvent être classés dans les catégories suivantes (Few, 2006; Ware, 2004) :

- Couleur
- Forme
- Position spatiale
- Mouvement

La Figure 2-4 présente des exemples d'attribut de pré-attention selon les catégories définies précédemment.

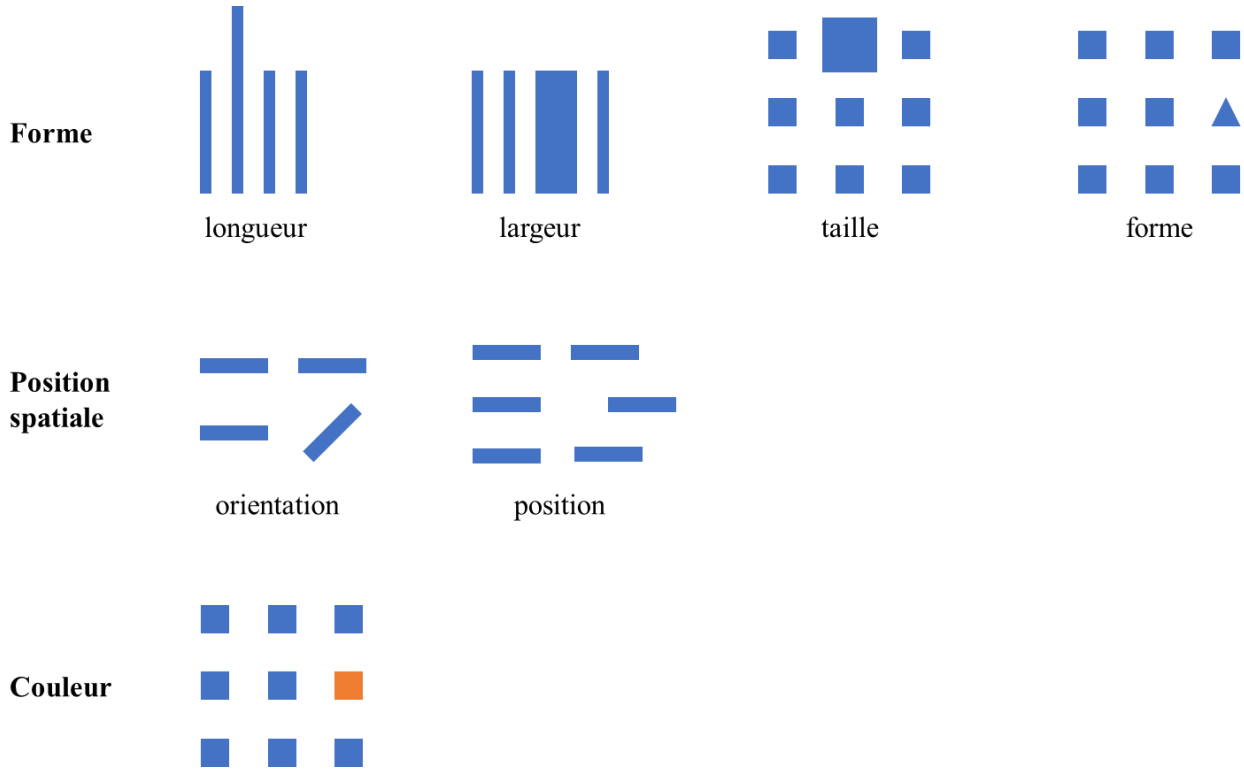


Figure 2-4 Exemples d'attributs pré-attentifs (Few, 2006; Wexler et al., 2017)

Cependant, des limites s'appliquent à ces attributs. Ainsi, lors de l'organisation des données en groupes distincts utilisant différentes déclinaisons d'un même attribut de pré-attention, il est déconseillé d'utiliser plus de cinq expressions distinctes (Few, 2006). Par exemple si l'on utilise plus de cinq nuances d'une certaine couleur il devient compliqué de distinguer les différentes nuances sans fournir un effort ou sans examiner avec minutie ce qui est contraire à l'objectif d'un attribut de pré-attention.

2.2.5 Importance des couleurs

Enfin, il est essentiel que l'utilisation des couleurs dans la visualisation soit pertinente. Les couleurs ne doivent pas être utilisées uniquement dans une optique d'embellissement de la visualisation. Les couleurs doivent être utilisées pour des raisons ou des objectifs précis. Dans *The Big Book of Dashboard*, les auteurs indiquent que la couleur doit être utilisée dans la visualisation des données de trois manières principales : séquentielle, divergente et catégorielle. La couleur peut également être utilisée pour attirer l'attention du lecteur sur un point important ou l'alerter ou pour mettre en évidence une partie des données (Wexler et al., 2017). Ces utilisations sont décrites dans la Figure 2-5. L'utilisation séquentielle d'une couleur correspond à l'utilisation de plusieurs nuances d'une

même couleur. Les couleurs divergentes peuvent être utilisées pour montrer une plage divergente à partir d'un point médian afin de représenter deux catégories ou des variations positives et négatives d'une mesure. Enfin, des couleurs catégorielles peuvent être utilisées pour représenter différentes catégories. Aussi, on peut souhaiter mettre en évidence certaines données mais sans vouloir alarmer le lecteur. Une couleur peut être choisie pour mettre en évidence et faire ressortir certaines données. Cependant, si l'on souhaite alerter le lecteur, une couleur vive et alarmante peut être utilisée (Wexler et al., 2017).

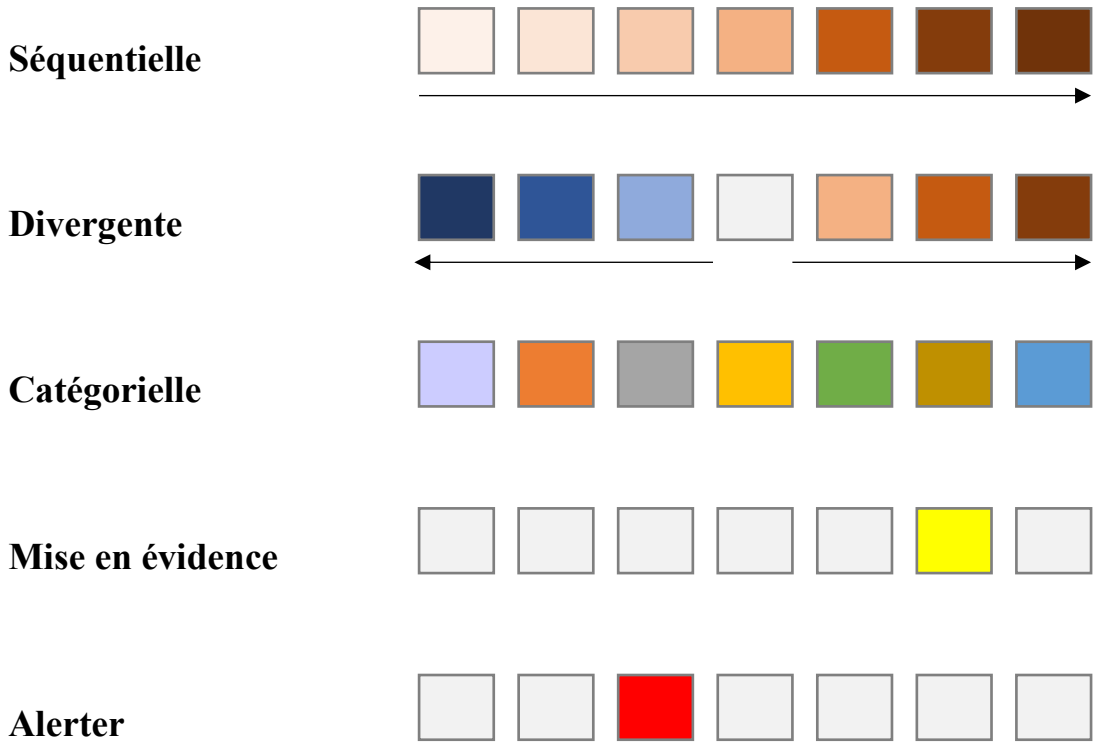


Figure 2-5 Utilisation des couleurs dans la visualisation des données - reproduit de *The Big Book of Dashboards* (Wexler et al., 2017)

Il faut toutefois faire attention au fait que selon le public visé, la couleur choisie pour alarmer n'apparaît pas forcément comme telle. Par exemple, le rouge ne signifie pas forcément qu'une information est importante ou alarmante dans toutes les cultures. En Chine par exemple, le rouge évoque le bonheur (Few, 2006).

De plus, les couleurs ne seront pas perçues de la même manière par tous les lecteurs. En effet, l'utilisateur peut être atteint d'une anomalie ou trouble de la vision des couleurs. Selon le type d'anomalie, il peut être très difficile pour les personnes atteintes de distinguer certaines couleurs en raison de leur perception du spectre des couleurs (Wexler et al., 2017). Il est important de prendre en compte ces enjeux lors de la conception des visualisations puisque 8% des hommes et 1% des femmes souffriraient de ces troubles (Wexler et al., 2017). Si la couleur est utilisée pour représenter les données et qu'il est nécessaire pour les lecteurs de distinguer entre les couleurs afin de comprendre la visualisation, il est alors indispensable d'utiliser des palettes compatibles avec les troubles de la vision des couleurs (Few, 2006; Wexler et al., 2017). Ainsi, il est préférable de choisir une couleur et d'en faire varier son intensité pour indiquer des degrés variables d'importance ou des variations d'une mesure plutôt que d'utiliser des couleurs différentes, car même ceux atteints de troubles de la vision des couleurs peuvent détecter des intensités distinctes de la même couleur (Few, 2006). Il est par exemple peu recommandé d'utiliser du rouge, du vert, du marron et de l'orange dans un même graphique puisque le rouge, le vert et l'orange apparaissent tous marron pour une personne atteinte d'un fort trouble de la vision des couleurs (Wexler et al., 2017).

Si l'on souhaite mettre en évidence certaines informations, il peut donc être préférable d'utiliser d'autres attributs que la couleur. Le flou par exemple (« blur ») présente l'avantage de fonctionner de façon indépendante de la couleur en utilisant une caractéristique visuelle inhérente à l'œil humain (Kosara et al., 2002). Une zone nette sera ainsi traitée pré-attentivement si le reste de la visualisation est flou. Cela fonctionne pour les images en noir et blanc ainsi que pour les utilisateurs atteints de trouble de la vision des couleurs. Mais si l'usage de la couleur est à contrôler, il est toutefois également important que le tableau de bord soit agréable à consulter (Wexler et al., 2017).

Les tableaux de bord devraient donc pouvoir condenser beaucoup de données sur un seul écran et les présenter sous forme de vue d'ensemble sans sacrifier la moindre information importante ni compromettre la clarté (Few, 2006).

Lors de la création d'un tableau de bord, il est donc nécessaire de s'assurer à la fois de la qualité des processus d'arrière-plan ou *back-end* (côté serveur) et de celle des processus frontaux ou *front-end* (côté interface client). En ce qui concerne l'arrière-plan, il est important de privilégier les outils qui assurent une grande interactivité et un chargement rapide (Andreeva et al., 2012). Une interactivité élevée est ce qui permet à l'utilisateur de personnaliser facilement la visualisation des

données. Et le chargement rapide est lié à la capacité de générer de nouvelles vues des mêmes données côté client tout en évitant un appel côté serveur qui peut être coûteux en temps.

2.2.6 Exemple de tableau de bord

Les opportunités d'analyse reposent essentiellement sur la disponibilité des données. Aussi certaines villes reviennent très souvent dans la littérature. C'est le cas de la ville de New York, très étudiée en raison de sa grande flotte de taxis et de l'importance de ces derniers dans la mobilité des New Yorkais mais surtout en raison de la disponibilité des données. En effet, depuis 2009, les données de l'ensemble des taxis opérant dans la ville sont disponibles sur le site de la New York Taxi Limousine Commission (New York City Taxi & Limousine Commission, 2018a). Les principales évolutions et tendances sont présentées dans leur *Fact Book* (New York City Taxi & Limousine Commission et al., 2014; New York City Taxi & Limousine Commission et al., 2016, 2018) telles que des statistiques agrégées sur l'offre et la disponibilité des taxis, sur le nombre de courses et leur répartition spatiotemporelle ainsi que différentes caractéristiques des chauffeurs et passagers. Les données étant disponibles, elles sont à la base de nombreuses études externes à la TLC mais aussi internes (Laviolette, 2017). Fausto Lopez, responsable des données et des analyses pour la New York City Taxi and Limousine Commission, met ainsi à disposition du public un tableau de bord fonctionnel et open source : le *TLC Fast Dash* (New York City Taxi & Limousine Commission, 2018a, 2018b).

La Figure 2-6 présente une capture d'écran du *TLC Fast Dash.* Ce dernier est réalisé à partir du package *Shiny* du logiciel libre de statistiques et sciences des données *R* (The R Foundation, 2020). On y retrouve des statistiques sur le nombre de courses, la durée de ces dernières, le nombre de chauffeurs et de véhicules en opération ainsi que sur les heures en service de ces derniers. L'utilisateur peut choisir la période d'analyse temporelle souhaitée ainsi que filtrer selon le jour, le mois ou l'année (New York City Taxi & Limousine Commission, 2018b).

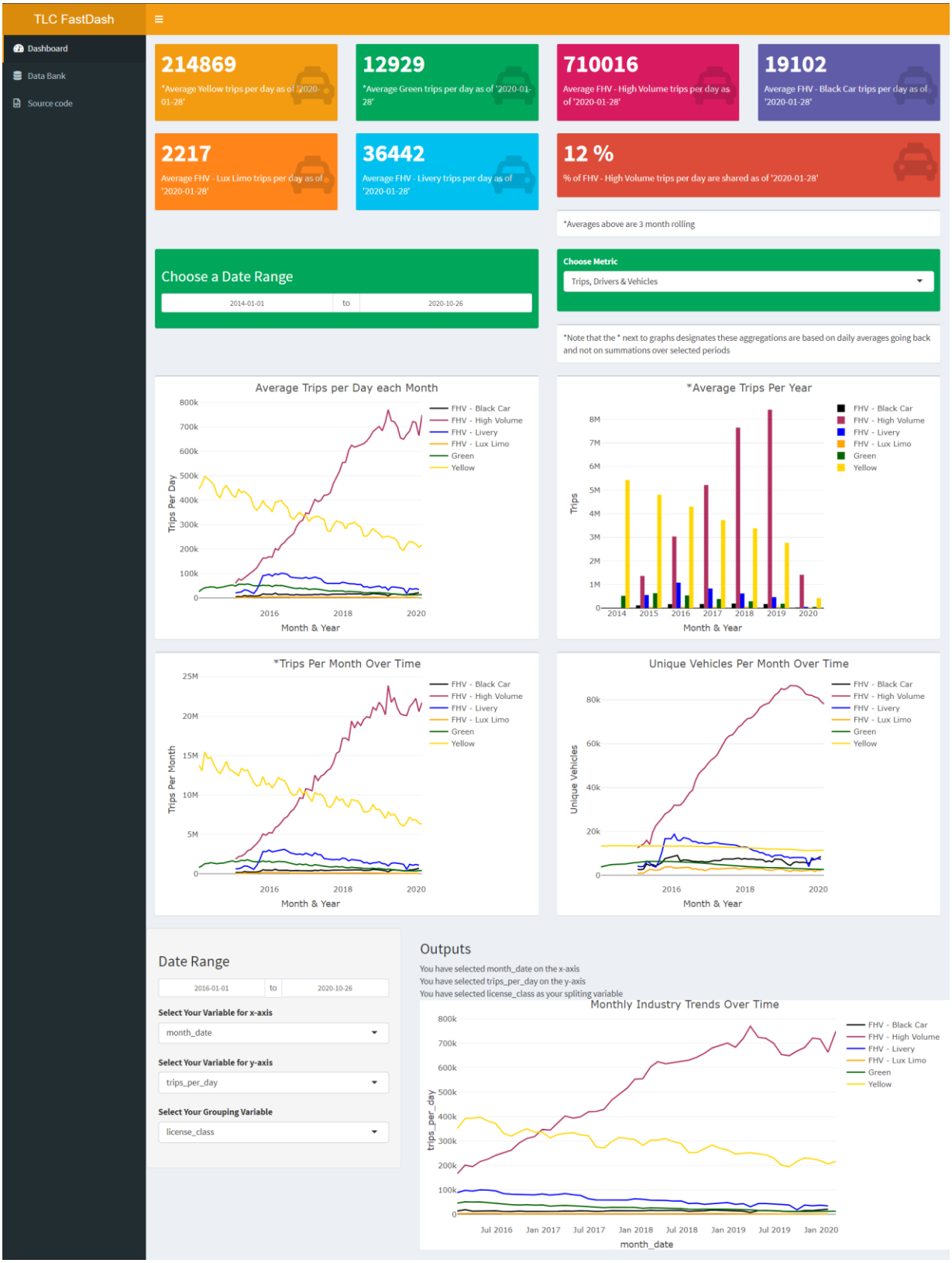


Figure 2-6 Capture d'écran du TLC FastDash (New York City Taxi & Limousine Commission, 2018b)

2.3 Synthèse de la revue de littérature

La première partie de la revue traite du secteur des taxi. Un rappel des évènements marquants de l'histoire du taxi ainsi que la définition du service sont d'abord énoncés, suivi du rôle du taxi dans la mobilité quotidienne et notamment de son interaction avec le transport en commun. Enfin une synthèse des thématiques d'analyses récentes liées à ce secteur est menée. Les récentes améliorations des différents modèles, tels que ceux de répartition des courses, de prévision de la demande ou de recherche du prochain client, permises par la disponibilité des données GPS de taxis sont mises en exergue. Les modèles de détection des anomalies sont également présentés car si les données GPS sont utilisées dans la construction de modèles, il est essentiel d'en éliminer les éventuelles erreurs au préalable. Enfin une synthèse des indicateurs pertinents permettant de caractériser l'offre et la demande de déplacements en taxi est présentée dans le Tableau 2-2.
Dans la deuxième partie de la revue l'accent est mis sur la visualisation des données. Le tableau de bord en tant qu'outil de prise de décision est défini. Enfin les différentes caractéristiques de visualisation essentielles à la conception d'un tableau de bord communiquant les informations de manière efficace et juste sont mises en avant. L'importance des attributs pré-attentifs et d'une bonne utilisation des couleurs est soulignée. Enfin, la revue se conclut sur un exemple de tableau de bord présentant les activités des taxis de la ville de New York.

La revue de littérature présentée dans ce chapitre met en évidence la nécessité d'une plateforme de visualisation pour l'analyse de l'industrie du taxi à Montréal. En effet, le secteur des taxis est aujourd'hui en pleine mutation car il fait face à l'émergence de nouveaux modes de transport notamment avec le développement des services de transport sur demande. Outre cette diversification de l'offre, à Montréal, on ne dispose que d'une compréhension limitée de la demande de déplacements en taxi ce qui rend difficile le développement et l'optimisation de cette industrie. On dispose pourtant de données GPS mais elles sont encore peu exploitées. Or en analysant ces ensembles de données, il est possible de connaître la répartition de la demande et donc d'optimiser l'offre en conséquence, de réduire le temps d'attente en identifiant les itinéraires de conduite optimaux mais aussi de diminuer le temps et la distance parcourus à vide et donc également la consommation de carburant, les émissions polluantes et gaz à effet de serre.

De plus, l’efficacité des tableaux de bord en tant qu’outil de visualisation pour faciliter la prise de décisions a été démontrée. Il est donc nécessaire de poursuivre les recherches permettant d’analyser l’offre et la demande des déplacements en taxi à Montréal. La conception d’un tableau de bord en tant qu’outil de visualisation pour faciliter la prise de décisions relatives à l’industrie du taxi à Montréal est donc justifiée. Cette plateforme intègrera donc à la fois des fonctionnalités stratégiques et analytiques.

CHAPITRE 3 METHODOLOGIE GENERALE

3.1 Description des données

Lacombe (2016) et Laviolette (2017) présentent dans leurs travaux respectifs une mise en contexte du taxi sur l'île de Montréal. La situation réglementaire du taxi au Québec et en particulier à Montréal y est détaillée. De plus, une présentation du réseau de transport de la ville ainsi que les habitudes de déplacements des montréalais est disponible (Lacombe, 2016; Laviolette, 2017).

Seuls les principaux points de la structure organisationnelle du taxi à Montréal seront ici détaillés afin de fournir un contexte aux données utilisées dans le cadre de ce projet. Puis la description des données et la méthodologie de traitement de ces dernières seront définies.

3.1.1 Provenance des données

3.1.1.1 Intervenants directs de l'industrie du taxi

Les principaux acteurs du secteur des taxis sont l'intermédiaire en service, le propriétaire du véhicule et le chauffeur. Le propriétaire du véhicule est celui qui détient un permis de propriétaire de taxi. Il peut conduire son propre véhicule s'il détient un permis de chauffeur de taxi, ou alors autoriser une tierce personne possédant un permis de chauffeur de taxi à conduire son véhicule. L'intermédiaire en service est la dénomination québécoise d'une entreprise de taxi. Cependant au Québec, les intermédiaires en service ne possèdent pas de véhicules de taxi, contrairement à d'autres villes. Ils offrent uniquement des services de répartition des appels ou de publicité aux détenteurs d'un permis de propriétaire de taxi ou d'un permis de chauffeur de taxi. Ces derniers ne sont pas dans l'obligation de s'affilier à un intermédiaire en service et sont alors considérés comme indépendants. Selon Laviolette (2017), environ 1400 détenteurs de permis de propriétaire de taxi seraient indépendants (Laviolette, 2017).

3.1.1.2 Bureau du taxi de Montréal

Le Bureau du taxi de Montréal (BTM) est un organisme paramunicipal de la Ville de Montréal. En vigueur sous sa forme actuelle depuis le 1er janvier 2014, il a pour principale mission le développement de l'industrie du transport par taxi, tout en veillant au respect de l'application de la

loi et de la règlementation sur l'ensemble du territoire de l'île de Montréal (Ville de Montréal, 2020a).

Plusieurs mesures ont été prises par le BTM dans un objectif de développement de l'industrie en misant particulièrement sur l'innovation technologique. Ainsi, depuis septembre 2017, un règlement du Bureau du taxi de Montréal (BTM) exige que tous les véhicules de taxis soient équipés d'un système de positionnement global (GPS, Global Positioning System) afin de "localiser la position et suivre la route du taxi en temps réel" (Ville de Montréal, 2017). Et afin de tirer profit des données recueillies et dans une optique d'analyse de l'offre et de la demande de déplacements par taxi, la Ville de Montréal et le Bureau du Taxi de Montréal (BTM) se sont dotés d'un outil pertinent : le Registre des taxis.

3.1.1.3 Le Registre des taxis

Lancé le 16 mai 2018, le Registre des taxis est une plateforme de données ouvertes, appelée également point d'échange de taxis (TXP, Taxi Exchange Point) dont le but est d'améliorer la liaison entre les taxis et leurs clients (Bureau du Taxi de Montréal, 2020a). De plus, le 16 mai 2018, le Comité exécutif de la Ville de Montréal a adopté l'ordonnance concernant l'obligation pour tout détenteur de permis de propriétaire de taxi de se connecter au Registre des taxis avant le 28 novembre 2018 (Bureau du Taxi de Montréal, Automne 2018). Tous les véhicules associés à un permis de propriétaire de taxi doivent désormais être reliés à ce Registre et fournir des données spatiotemporelles dans un format normalisé. Les taxis en exploitation sont tenus d'envoyer leurs données au serveur du Taxi Exchange Point toutes les 5 secondes, conformément à l'ordonnance (Beaudoin, David 2017). En outre, on peut noter que les différents intermédiaires en service de taxi n'ont pas accès aux données des autres intermédiaires.

Un des objectifs est que les données fournies par le Registre facilitent la prise de décision en aidant à la compréhension du fonctionnement de l'industrie du taxi. La disponibilité des données peut également permettre de favoriser la création et le développement de nouveaux services en lien avec le secteur. Ainsi, si le Registre des taxis n'est pas une application mobile et n'effectue pas de répartition de demandes de déplacement, il peut toutefois permettre à des applications mobiles pour clients de relayer des demandes de transport. A terme, le BTM souhaite que cette plateforme puisse indiquer en temps réel où se trouve chacun des taxis en service sur le territoire de l'île de Montréal,

et s'il est disponible pour recevoir une demande de transport, afin notamment de réduire les voyages à vide (Bureau du Taxi de Montréal, Automne 2018).

Le Tableau 3-1 présente les différents intermédiaires en service opérant sur l'île de Montréal et qui doivent désormais être affiliés au Registre des taxis (Bureau du Taxi de Montréal, 2020b).

Tableau 3-1 Liste des intermédiaires en service opérant sur l'île de Montréal et partenaires du Registre des taxis

• Amical	• Angrignon	• Atlas
• Boisjoly	• Champlain	• Coop de l'est
• Coop de l'ouest	• Coop de Montréal	• Diamond
• Diamond accessible	• Hochelaga	• Hochelaga adapté
• Hypra intermédiaire	• Ici taxi	• McGill taxi
• Netlift intermédiaire	• Pontiac	• Rosemont plus
• Rosemont van adapté	• Taxi expert	• Téo
• Van Médic		

Ainsi, il est désormais possible d'avoir un flux continu de données de tous les taxis en service à Montréal. Ces données GPS collectées passivement par les taxis indépendants ou associés à des intermédiaires en service à Montréal sont utilisées dans le présent mémoire.

Les sections suivantes fournissent une description détaillée des données contenues dans le Registre.

3.1.2 Description des tables du Registre des taxis

Le Registre des taxis fournit cinq ensembles de données :

- Un ensemble de données sur les **permis** (ADS : "Autorisation de Stationnement")

- Un ensemble de données sur les **véhicules**

- Un ensemble de données sur les **chauffeurs**

- Un ensemble de données sur les **taxis**

- Un ensemble de données sur les **positions GPS** des taxis

Les relations entre ces cinq ensembles de données ainsi que les principaux attributs de chacun sont présentés dans la Figure 3-1. Des tableaux en Annexe A présentent les définitions de chaque attribut des tables des permis, des véhicules, des chauffeurs et des taxis.

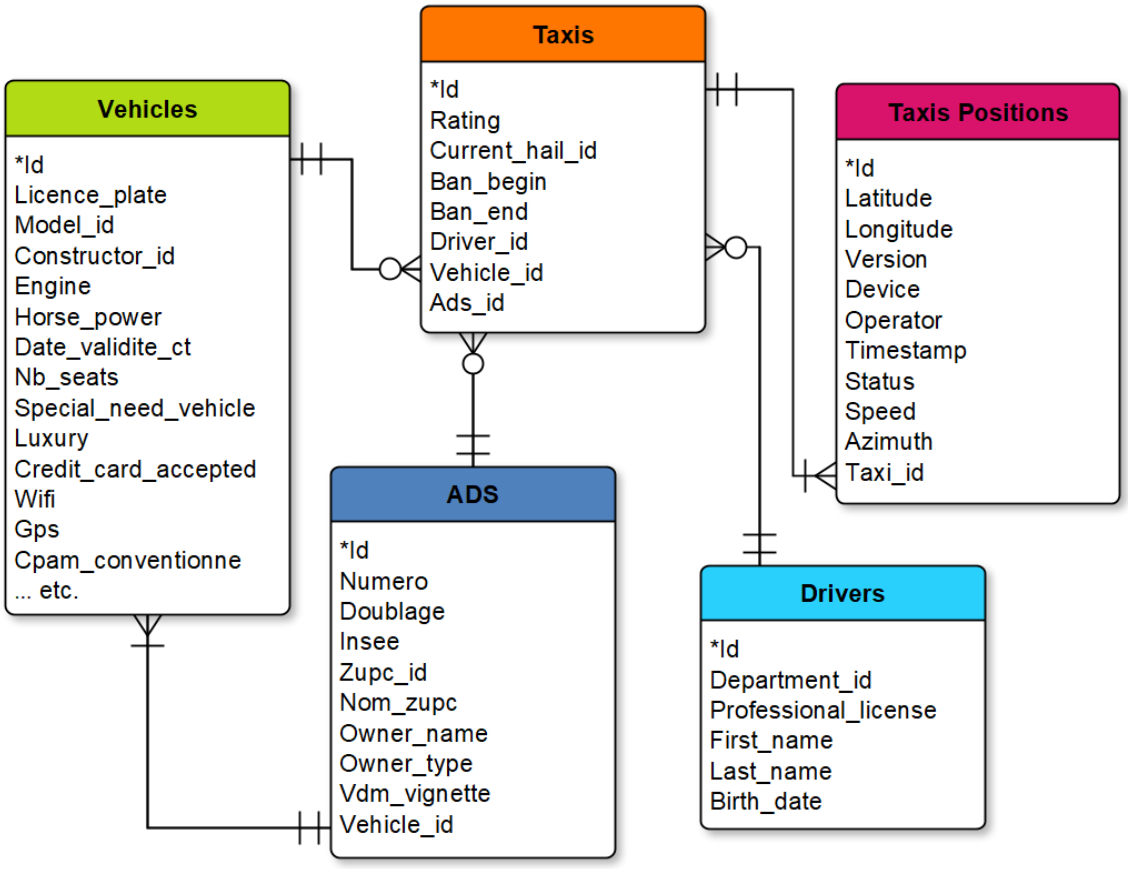


Figure 3-1 Structure des données du Registre

La table de données des permis (ADS) fournit des informations sur le permis de taxi telles que le numéro de permis, le nom du propriétaire du permis, la zone de permis ainsi que le véhicule associé au permis.

La table de données des véhicules fournit des informations sur le véhicule utilisé pour le service de taxi telles que le numéro de plaque d'immatriculation, le modèle du véhicule, le nombre de sièges ainsi que de nombreuses informations sur les différentes options présentes dans le véhicule (ex : disponibilité de données wifi dans le véhicule pour le client).

La table des données des chauffeurs fournir des informations sur les chauffeurs telles que leur numéro de permis de conduire, leurs nom, prénom et date de naissance.

Dans le cadre des données du Registre, un taxi est défini comme la combinaison d'un chauffeur, d'un véhicule et d'un permis (cf. Figure 3-2). Chaque combinaison différente d'un véhicule, d'un chauffeur et d'un permis correspond donc à un identifiant unique de taxi généré par le Registre. (TXP : Taxi Exchange Point) (Beaudoin, David 2017).

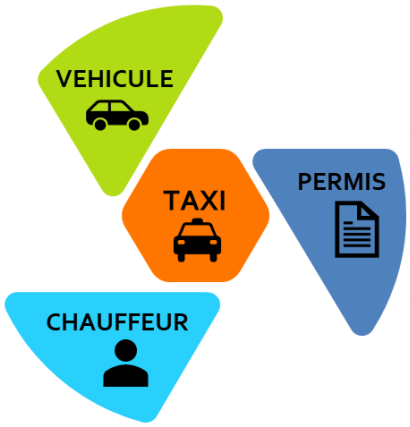


Figure 3-2 Définition d'un taxi dans le cadre du Registre

Un taxi étant la combinaison d'un véhicule, d'un chauffeur et d'un permis, **la table des taxis** indique donc quel est le chauffeur associé à l'identifiant de taxi ainsi que le véhicule et permis associés.

3.1.3 Description des statuts

Lorsqu'un taxi est en service, il doit fournir au Registre des informations sur son statut ou état. Il existe quatre statuts possibles pour les taxis en opération. Ces derniers sont spécifiés dans la Figure 3-3.

Figure 3-3 Description des quatre statuts possibles

Les différents statuts des taxis sont uniquement définis pour des taxis en service. Lorsque le taxi n'est pas en service, par exemple si le véhicule est utilisé à des fins personnelles, il ne doit pas envoyer ses données au Registre. Un taxi est *free* lorsque le taxi est à la recherche d'un client. Il peut être stationné dans un poste d'attente ou circulant en ville. Lorsqu'un client commence une course, qu'elle soit hélée ou commandée, le taxi doit alors changer son statut à *occupied* indiquant qu'il est en course et qu'il n'est donc plus disponible pour répondre à une autre demande de transport. Lorsqu'un taxi accepte une course commandée, son statut doit être changé à *oncoming* pendant toute la période où le taxi est en route pour récupérer le client. Il ne peut donc pas répondre à une autre demande ou accepter une course hélée puisqu'une course lui a déjà été attribuée. Enfin, le statut *unavailable* indique que le taxi est tout de même en service mais qu'il effectue un type particulier d'activité.

Ce statut *unavailable* présente d'ailleurs un enjeu. En effet, lorsque le taxi est *unavailable* (indisponible), il peut soit être en opération pour un service de taxi collectif pour une société de transport en commun, soit fournir des services de taxi adapté, ou encore effectuer des contrats corporatifs ou hospitalier. Toutefois, les informations actuellement disponibles dans le Registre ne permettent pas de différencier et d'identifier quel type d'activité été réalisé. Ainsi, seule une partie des trajets effectués par les taxis peut être identifiée avec les données disponibles dans le Registre. Seuls les trajets réguliers peuvent être comptabilisés pour l'instant, c'est-à-dire lorsque le statut

indique "occupied" soit occupé. Une des limites du Registre des taxis est ainsi mise en évidence puisqu'on ne peut identifier l'ensemble des activités réalisées par les taxis en exploitant uniquement les données du Registre. Les flux de données du transport adapté et du taxi collectif seraient nécessaires pour croiser les données et ainsi identifier les différents types de d'activités.

3.1.4 Les traces GPS

Les données de localisation des taxis par GPS (table Taxis Positions) sont l'ensemble de données qui nous permet de calculer la plupart des indicateurs. Un point GPS correspond à la localisation spatiale et temporelle d'un véhicule. Les points GPS relevés permettent d'obtenir la trace GPS d'un taxi, soit le chemin effectué par le taxi dans le temps et l'espace. D'autres informations sont recueillies telles que l'identifiant du taxi, le nom de l'intermédiaire auquel le taxi est affilié ou encore la vitesse instantanée du véhicule. Les différents attributs sont présentés dans le Tableau 3-2, avec un exemple de données pour chaque attribut.

Tableau 3-2 Description de la table des positions GPS

Attribut	Description	Exemple
Taxi	L'identifiant du taxi	-
Latitude	La latitude du taxi à l'instant où l'information est récupérée	45.53892
Longitude	La longitude du taxi à l'instant où l'information est récupérée	-73.55624
Device	Le dispositif à partir duquel la position est relevée (ex: phone, tablet, taximeter ou otherdevice)	phone
Operator	Nom de l'intermédiaire en service	-
Status	Le statut du taxi à l'instant où l'information est récupérée	Free
Version	La version actuelle de l'API	2
Speed	La vitesse du taxi à l'instant où l'information est récupérée (km/h)	15
Azimuth	L'orientation du taxi à l'instant où l'information est récupérée	234
Timestamp	Le moment exact où la position a été déterminée par le taxi, formaté à l'heure Unix	1500580203
TimestampUTC	L'horodatage au format UTC à l'instant où l'information est récupérée	2017-07-20T19:50:03.000Z

C’est également l’ensemble de données qui présentent le moins d’erreurs. En effet, mis à part pour le changement de statut, les entrées de données ne sont pas manuelles. Cela limite les erreurs humaines et seules les éventuelles erreurs liées aux appareils de mesure demeurent présentes. Au contraire, les autres tables de données présentent de nombreuses erreurs humaines car grand nombre des entrées sont effectuées manuellement et sans validation des champs. Des exemples d’entrées erronées liées à une absence de validation sont présentés dans le paragraphe suivant.

3.1.5 Absence de règles de validation

Comme mentionné précédemment, les tables des véhicules, des chauffeurs et des permis présentent de nombreuses entrées de données erronées liées à une erreur humaine. Seule la table des taxis n’en présente pas puisqu’elle est générée automatiquement par le Registre à partir de la combinaison des tables des chauffeurs, des véhicules et des permis.

Ainsi les numéros de permis sont par exemple souvent mal renseignés dans la plateforme car aucune validation n’est faite sur les caractères à remplir. Pourtant puisque les numéros de permis respectent un même format, l’utilisation d’une expression régulière correspondant au format de numéro de permis délivré par les autorités permettrait de contrôler le remplissage de ce champ. L’entrée d’un caractère ou d’une expression ne respectant pas le format autorisé serait donc refusée et l’utilisateur serait avisé de son erreur. Puisque cette étape de validation n’est pas présente dans le Registre de nombreux numéros de permis renseignés sont erronés tels que ceux présentés dans le Tableau 3-3. Si l’on souhaite utiliser cette information, il est donc nécessaire de croiser les données du Registre avec une autre source de données, à savoir la liste des permis valides dont dispose le BTM. Cela rajoute des étapes et complexifie l’analyse de données puisqu’en plus de devoir croiser les sources de données, il faut également s’assurer de la mise à jour de la source externe de données et de sa validité.

Tableau 3-3 Exemples de numéros de permis erronés extraits de la table des permis (ADS)

Numéro de permis
0
411584
12121212
2006340001
3.8209E+11
000a00a00a
01M00001200
0C2000140001
0C200022001A

Le Tableau 3-4 présente un autre exemple d'entrées erronées liées à une erreur humaine. Quinze écritures différentes du constructeur automobile Hyundai peuvent être relevées dans la table des véhicules. S'il est simple d'identifier qu'un même mot est écrit en lettres majuscules ou minuscules (non sensible aux caractères), il serait cependant très complexe d'automatiser la reconnaissance d'un nom de constructeur automobile. Cela aurait pu être évité en imposant le choix d'un nom de constructeur parmi une liste déroulante des noms de constructeurs mise à jour lorsque nécessaire.

Tableau 3-4 Exemples d'entrées erronées extraits du champ "constructor name" de la table des véhicules

Constructor name	Occurrences
HUINDAY	1
hundai	1
HUNDAY	1
huyandai	1
HUYDAI	1
Huyndai	17
HUYNDAY	1
Hyndai	1
HYNDAY	2
Hyudai	1
HYUNDAI	821
Hyundai	5
Hyunday	28
Hyundi	6
hYUNDY	11

De plus, certains champs des tables du Registre ne s'appliquent pas au Québec mais aux réglementations françaises, tel que l'attribut booléen « cpam_conventionne » de la table des véhicules indiquant que le véhicule a une convention avec la sécurité sociale pour le transport des patients. Cela génère une colonne uniquement renseignée de valeurs « faux » ou vides. Si cela ne présente pas d'enjeux en termes d'analyse puisque ces champs peuvent être ignorés, ces derniers occupent toutefois de l'espace mémoire inutilement.

Enfin, un manque de cohérence dans le nom des champs est observé. Certains noms de champs utilisent des tirets bas (« underscores ») afin de séparer les différents mots tels que « vehicle_id » ou « birth_date » alors que d'autres non (« constructorname » au lieu de « constructor_name »). Cela complique l'automatisation des processus.

Ces exemples mettent en évidence l'importance des processus de validation des données lorsque l'entrée de ces dernières dépend d'un utilisateur. Il est essentiel que la base de données soit rigoureusement construite afin de limiter le besoin ultérieur d'opérations de traitement et de nettoyage des données.

3.1.6 Autres sources de données

Des données complémentaires sont nécessaires pour le calcul des indicateurs d'offre et de demande en déplacements de taxi qui seront disponibles sur le tableau de bord et qui sont présentés dans le Chapitre 5. Il s'agit surtout de données géographiques de limites territoriales permettant de réaliser des analyses spatiales. Ces données sont au format de *shapefile* ou « fichier de formes » qui est un format de fichier pour les systèmes d'informations géographiques. Un tel fichier contient toute l'information liée à la géométrie des objets décrits qui peuvent être des points, des lignes ou des polygones (Contributeurs de Wikipédia, 2020).

Les principaux découpages du territoire utilisés sont :

- la couche des limites administratives des villes liées et des arrondissements de l'île de Montréal;
- la couche des secteurs de recensement (SR) de 2016.

Des données liées à l'utilisation du taxi sur le territoire sont également incluses telles que la couche des postes d'attente de taxis fournie par le BTM. Cette dernière est disponible sur le portail de

données ouvertes de la Ville de Montréal. L'information sur les postes d'attente de taxis comprend l'identifiant du poste, ses coordonnées spatiales, le nombre de places disponibles, son type (privé, public ou commun), l'état du poste (actif, temporaire ou fermé) ainsi que la période d'opération (ex : 24H / 7J, 00h00-06h00, etc.). La dernière mise à jour date du mois d'octobre 2020 (Ville de Montréal, 2020c).

Selon les besoins des utilisateurs de la plateforme, d'autres découpages territoriaux peuvent être ajoutés tels que les secteurs municipaux de l'enquête Origine-Destination. Cependant, la plus petite zone d'étude est fixée aux secteurs de recensement.

Enfin, les autres sources de données mentionnées par Lacombe (2016) et Laviolette (2017) sont également incluses telles que la couche délimitant les trois agglomérations de taxi de Montréal fournie par le BTM (Lacombe, 2016; Laviolette, 2017).

3.2 Méthodologie générale de traitement des données

Afin de concevoir un tableau de bord permettant la visualisation et l'analyse d'indicateurs, il est d'abord nécessaire de collecter et de traiter les données utilisées pour le calcul de ces indicateurs. Les principales étapes du processus général de traitement des données sont résumées dans la Figure 3-4, des données brutes à la visualisation. Trois principales étapes sont identifiées : l'extraction des données, le traitement de ces dernières et la visualisation des indicateurs. Ces étapes sont détaillées dans les paragraphes suivants.

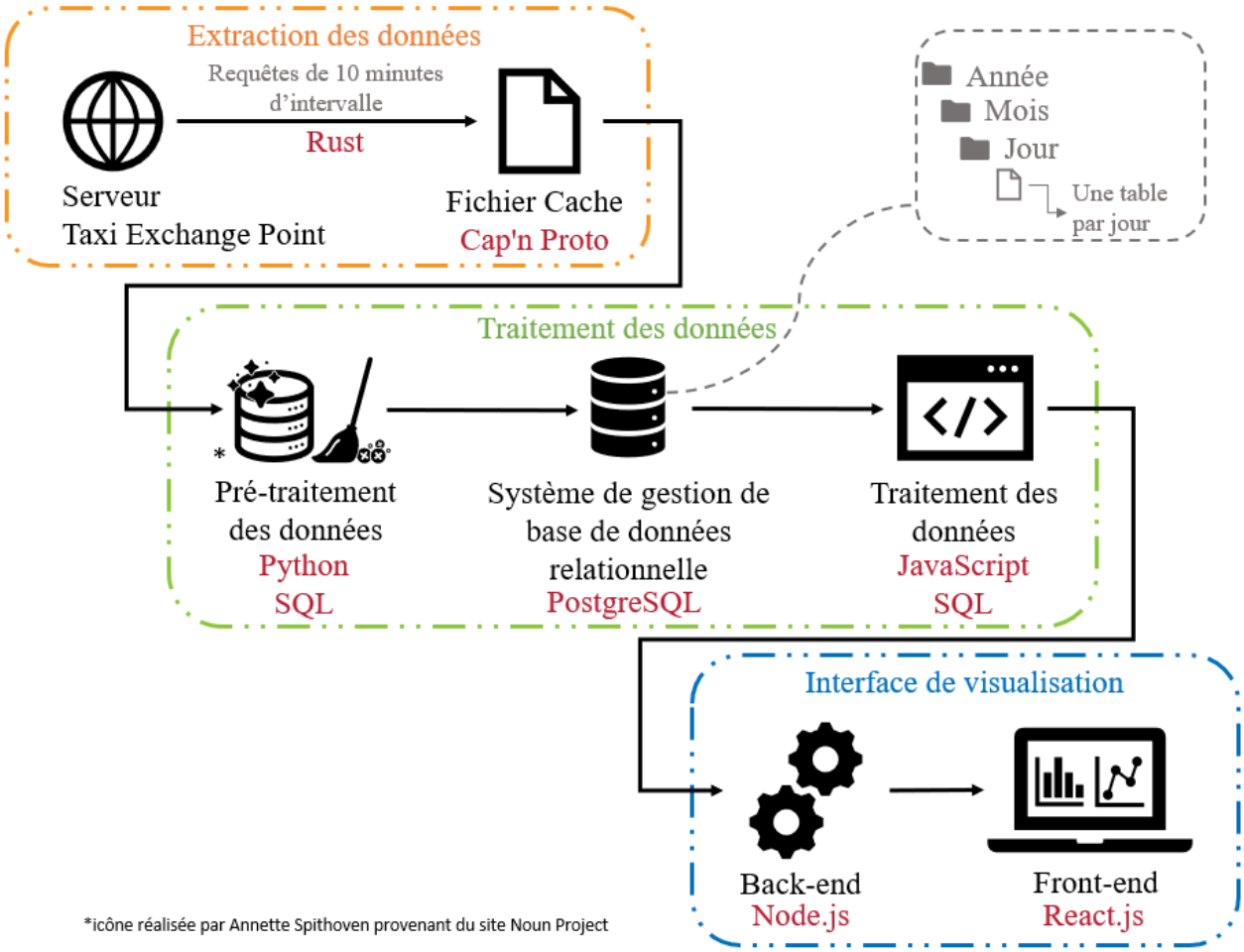


Figure 3-4 Schéma d'architecture de données

3.2.1 Importation des données

Comme mentionné précédemment, les taxis en exploitation sont tenus d'envoyer leurs données au serveur du Taxi Exchange Point toutes les 5 secondes, conformément à l'ordonnance (Beaudoin, David 2017). En raison du volume conséquent de données que cela génère, l'importation des données brutes ne peut se faire que par plage de dix minutes. Des requêtes sont donc lancées à l'interface de programmation d'application du BTM afin de récupérer les données.

Une interface de programmation d'application ou *API*, de l'anglais « Application Programming Interface », est un moyen de communication entre deux logiciels. Une API facilite la communication entre le client et le serveur en servant d'intermédiaire, tel qu'illustré dans la Figure 3-5. Le client demande à l'API une information, en formulant une requête, et l'API va chercher

cette information dans la base de données du serveur puis la renvoyer au client sous forme de réponse (Pedro, 2020).

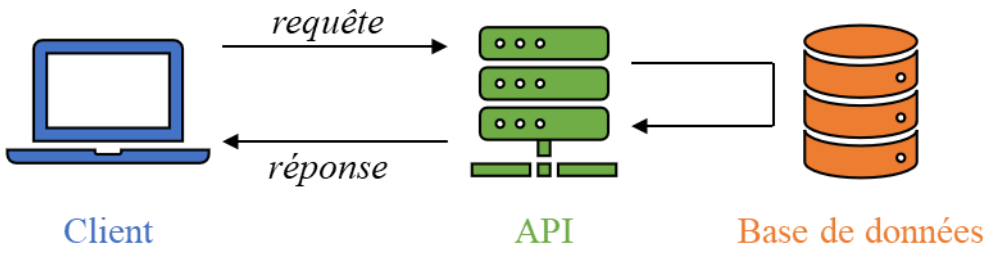


Figure 3-5 Communication à l'aide d'une API

L'API utilisée pour récupérer les données du Registre est privée. En effet, une clé API fournie par le BTM doit être renseignée dans la requête. Cela garantit un contrôle des utilisateurs ayant accès aux données.

La réponse envoyée par l'API est en format JSON (« JavaScript Object Notation »). JSON est une syntaxe pour le stockage et l'échange de données texte. Un des principaux avantages de ce format est que le texte peut être lu et utilisé facilement par n'importe quel langage de programmation. La simplicité de sa structure le rend également aisément compréhensible par un humain. Ce format est donc simple à mettre en œuvre tout en étant complet puisqu'il permet de gérer de nombreux types de données (W3Schools, 2020).

Une fois reçu, ce fichier de réponse est dans un premier temps traité à l'aide des langages Python et SQL. La Figure 3-6 présente le pseudo-code du processus de récupération des données GPS de taxis. Les données étant imbriquées sur plusieurs niveaux, plusieurs boucles sont requises pour aller les récupérer. Seuls dix attributs sont conservés parmi les onze relevés. Les attributs récupérés sont énoncés dans la section 3.4.1. Des modifications sont également réalisées à l'import des données telles que la troncature des décimales dans le cas des données spatiales.

L'objectif du tableau de bord conçu pour le projet de recherche n'étant pas l'analyse en temps réel, les données seront importées à une certaine fréquence (ex : quotidiennement ou de manière hebdomadaire) et non en temps réel.

Algorithme 1 Importation des données GPS des taxis

Début

1: Jour_à_récupérer ← date du jour que l'on souhaite récupérer
► *ex* : 2020-01-31

2: Liste_dates_locales ← liste de dates incrémentées par heure commençant à 23h10 le jour précédent le jour que l'on souhaite récupérer et se terminant à 00h10 le jour suivant
► *ex* : [2020-01-30 23h10 ; 2020-01-31 00h10 ; 2020-01-31 01h10 ... 2020-02-01 00h10]

3: Liste_dates_UTC ← conversion de la liste Liste_dates_locales en liste de dates UTC

► les données ne peuvent être récupérées que par plage de 10 minutes
►on doit donc créer des listes de dates avec un pas de 10 minutes

4: **Pour** *x* allant **de** 0 **à** longueur(Liste_dates_locales) -1 :

5: Liste_heures_locales ← création d'une liste de 6 valeurs afin de couvrir une heure à partir de l'élément d'indice *x* de la liste Liste_dates_locales
► *ex* : si *x* = 2020-01-31 00h10 alors Liste_heures_locales = [2020-01-31 00h10 ; 2020-01-31 00h20 ; 2020-01-31 00h30 ; 2020-01-31 00h40 ; 2020-01-31 00h50 ; 2020-01-31 01h00]

6: Liste_heures_UTC ← conversion de la liste Liste_heures_locales en liste de dates UTC

7: Liste_données ← création d'une liste vide

8: **Pour** *i* allant **de** 0 **à** longueur(Liste_heures_locales) - 1 :

9: Taxis_positions ← requête API au Registre des données GPS de taxis en indiquant dans les options de la requête l'élément d'indice *i* de la liste Liste_heures_UTC comme plage temporelle à récupérer
► la requête requiert une date UTC et non locale
► la requête requiert également une clé API

10: Taxis_positions_json ← récupération du fichier JSON à partir de Taxis_positions

11: Taxis_positions_items ← récupération des attributs de "items" dans Taxis_positions_json

► les données étant imbriquées sur plusieurs niveaux, on rentre dans différentes boucles pour pouvoir les récupérer : Taxi_positions_json.get("items")

```
12:     Pour k dans Taxi_positions_items :

13:         Taxi_positions_items_items  ←  récupérer les attributs de "items" dans
                                           Taxis_positions_items
            ► 2ème niveau d'imbrication "items" : Taxi_positions_items.get("items")

14:         Pour l dans Taxi_positions_items_items :

15:             a1   ← récupérer l'item "taxi"
16:             a2   ← récupérer l'item "lat"      ► Ne conserver que 5 décimales
17:             a3   ← récupérer l'item "lon"      ► Ne conserver que 5 décimales
18:             a4   ← récupérer l'item "version"
19:             a5   ← récupérer l'item "device"
20:             a6   ← récupérer l'item "operator"
21:             a7   ← récupérer l'item "status"
22:             a8   ← récupérer l'item "speed"
23:             a9   ← récupérer l'item "azimuth"
24:             a10  ← récupérer l'item "timestampUTC"

25:             ligne ← [a1; a2; a3; a4; a5; a6; a7; a8; a9; a10]

26:             ajouter ligne à la liste Liste_données

27:     Données_heures  ←  conversion de la liste Liste_données en table ou dataframe et la
                           nommer en précisant l'heure et la date locales
                           ► ex : raw_data_2020_01_31_00h
                           ► chaque table correspond à une heure de données pour la
                             journée renseignée
                           ► l'heure renseignée dans le nom de la table est l'heure locale
                             car il est plus intuitif pour l'utilisateur de travailler en heure
                             locale
                           ► chaque table correspond à une heure de données, cependant
                             les courses peuvent s'étaler sur plusieurs heures
                           ► on crée donc une table qui regroupe toutes les heures (26
                             heures)

28: Données_journée  ←  concaténation des 26 tables d'heures en une table de journée
                        ► On prend en compte 26 heures et non 24 heures car une
                          course peut commencer à 23h et se terminer après minuit. On
                          ne veut donc pas considérer les points à partir de minuit
                          comme une autre course mais bien comme une course ayant
                          débuté dans l'heure précédente.

29: Suppression des doublons (lignes identiques) dans Données_journée

30: Suppression des doublons de statut (attributs "taxi" et "timestampUTC" identiques) en ne
    conservant que la première occurrence dans Données_journée
    ► En effet, il arrive que lors d'un changement de statut deux informations sont envoyées :
      il existe donc deux points pour un même taxi au même moment mais de statuts différents.
      On décide de conserver le premier statut.

    Fin
```

Figure 3-6 Pseudocode d'importation des données GPS de taxis

3.2.2 Traitement des données

Une fois les données brutes importées, ces dernières sont traitées avant de pouvoir servir de base à l'analyse. Le processus de traitement englobe plusieurs étapes.

3.2.2.1 Prétraitement des données brutes

Une première étape de prétraitement des données est nécessaire afin de nettoyer les données et de notamment éliminer les doublons de données. Ces données, une fois nettoyées, peuvent ensuite être conservées dans des bases de données afin de ne pas avoir à effectuer directement des requêtes dès lors que l'on a besoin de données. Le temps de chargement serait considérable si pour chaque calcul d'indicateur et chaque choix de période spatiale et/ou temporelle par l'utilisateur de la plateforme, les données devaient être préalablement récupérées directement du Registre et ensuite traitées. De plus, en raison de l'important volume des données, le format de ces dernières est optimisé afin de ne pas stocker les données brutes nettoyées mais des données optimisées dans un objectif de réduction de l'espace mémoire utilisé et d'amélioration du temps de calcul des indicateurs. Les processus de nettoyage des données et de création de la base de données optimisée sont explicités dans les sections 3.3 et 3.4. De plus, la méthodologie utilisée pour le regroupement des points GPS en groupes de points consécutifs de même statut est détaillée dans le Chapitre 4.

3.2.2.2 Stockage des données optimisées

L'espace requis pour stocker les données optimisées est nettement inférieur à celui requis pour le stockage des données brute. Mais il n'en reste pas moins considérable. Les données optimisées sont donc stockées dans un serveur de base de données de la Chaire Mobilité à Polytechnique à l'aide de PostgreSQL. PostgreSQL est un système de gestion de base de données relationnelle et objet, utilisant le langage SQL («« Structured Query Language »). C'est un outil développé sous une licence libre, il est donc gratuit et libre de droits. Il fonctionne de plus sur tous les principaux systèmes d'exploitation. Cet outil permet la gestion de bases de données de très grandes tailles. Il possède de plus un module d'extension de prise en charge de données géospatiales, PostGis, rendant possible la connexion au logiciel de système d'information géographique QGIS. Ce dernier, également libre d'utilisation, permet notamment le traitement, l'analyse et la visualisation des données spatiales (PostgreSQL, 2020a). Cet outil est donc adapté pour la gestion d'une base de données GPS des taxis (Laviolette, 2017).

3.2.2.3 Calcul des indicateurs

Le calcul des indicateurs se fait par la suite à partir des données optimisées. En effet, ces dernières ont certes été optimisées de sorte à diminuer l'espace mémoire requis pour le stockage mais également de sorte à faciliter le calcul des indicateurs. Les données GPS optimisées des taxis sont en effet regroupées par identifiant de taxi et par statut. Ainsi, les points GPS consécutifs d'un taxi donné et présentant le même statut sont regroupés dans un même élément. De plus, les coordonnées spatiales et données temporelles des limites du groupe sont mises en évidence. A titre d'exemple, la durée totale peut donc facilement être déduite puisqu'il suffit de réaliser la différence entre le l'horodatage correspondant au premier point et celui correspondant au dernier point du groupe. Sans la disponibilité directe des horodatages de début et fin des groupes dans la base optimisée, afin d'obtenir la durée, il aurait fallu à chaque fois réaliser toutes les opérations de tri et de regroupement des points décrites dans le Chapitre 4. Le temps de calcul requis aurait donc été considérable. Dans ce même objectif, d'autres informations comme la distance totale parcourue ou les intersections avec les secteurs de recensement sont également calculées lors du pré-traitement des données et disponibles dans la base optimisée. Le détail des étapes de construction de cette dernière est fourni dans la section 3.4.

Les données optimisées doivent toutefois encore être traitées lors du calcul des indicateurs. En effet, des règles de validation peuvent être appliquées, telles que celles pour les courses, détaillées dans la section 4.4. Ces règles ne sont pas appliquées préalablement au stockage des informations car elles peuvent être amenées à évoluer. Il est donc plus judicieux de les appliquer lors du calcul des indicateurs. La méthodologie de calcul de ces derniers est détaillée au Chapitre 5.

3.2.3 Visualisation

Enfin, concernant l'interface de visualisation, cette dernière est codée en JavaScript avec la librairie ReactJs. Cette dernière, développée par Facebook depuis 2013, est une librairie libre qui permet de réaliser des interfaces utilisateurs interactives (Facebook Inc, 2020). React est utilisé côté serveur avec NodeJs. Ce dernier est un environnement bas niveau supportant l'exécution de JavaScript côté serveur (Node.js, 2020).

L'ensemble des processus de traitement de données et calculs des indicateurs sont intégrés à la plateforme afin de permettre la visualisation des indicateurs identifiés dans le Chapitre 5.

La structure du tableau de bord développé dans le cadre du projet de recherche est quant à elle détaillée dans le Chapitre 6. Elle est en entonnoir : soit du plus global au plus précis. Ainsi, une page de faits saillants constitue la vue d'accueil de l'interface. Ces derniers permettent d'obtenir un aperçu général des caractéristiques d'offre et de demande en déplacements de taxis. Si l'utilisateur souhaite obtenir plus d'informations sur un de ces faits saillants, il peut alors cliquer sur celui-ci et naviguer entre les différents niveaux de détails disponibles. L'interface étant interactive, plusieurs choix de filtres temporels (tels que la période d'analyse), spatiaux (tels que la zone d'étude) mais aussi de l'objet étudié (tels que les chauffeurs, véhicules ou permis) sont intégrés. Ces derniers sont détaillés dans la section 6.3.2.

Les processus énoncés précédemment de traitement de données et de construction d'une base de données optimisée sont détaillés dans les paragraphes qui suivent.

3.3 Prétraitement des données

Afin de pouvoir développer des procédures systématiques et automatisées de traitement des données et pour estimer une variété d'indicateurs, il faut d'abord se familiariser avec les données. En effet, les bases de données ont leur part de "bad data", c'est-à-dire des données inexactes, incomplètes ou susceptibles d'être mal interprétées, qui doivent être rectifiées (Berman, 2018; Loshin, 2001). La détection des valeurs aberrantes et le nettoyage des données est une exigence indispensable à une analyse fiable (Zhang, J., 2012). Une analyse de la qualité des données constitue donc la première étape du projet. Il est essentiel de pouvoir identifier les données erronées ainsi que les sources potentielles d'erreur, soit des erreurs qui ne sont pas actuellement observées mais qui pourraient être rencontrées. En effet dans le Registre, comme mentionné précédemment, de nombreuses saisies de données sont effectuées manuellement par les différents utilisateurs, que ce soient les chauffeurs, les intermédiaires de services de transport en taxi ou encore les titulaires de permis. De plus, la plupart des champs d'entrées ne sont pas soumis à validation et des mesures correctives doivent donc être prises en aval.

3.3.1 Suppression des doublons

Il existe par exemple de nombreux cas de doublons dans les données. Ces derniers sont de deux types. Les doublons « exacts », c'est-à-dire lorsque toutes les colonnes sont identiques, peuvent être facilement traités à l'aide de procédures d'élimination des doublons intégrées dans les langages

de programmation. Le Tableau 3-5 fournit un exemple de doublon exact. Les valeurs des différents attributs sont identiques. La première occurrence, soit la première ligne, est conservée.

Tableau 3-5 Illustration d'un doublon exact

id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC
-	-	45.50401	-73.63797	Phone	-	Free	2	15	217	1575546690
-	-	45.50401	-73.63797	Phone	-	Free	2	15	217	1575546690

Cependant, il existe également des doublons « de statut ». Cela signifie que pour le même taxi au même moment, deux statuts différents sont fournis en même temps. Ce type de doublon est sans doute lié au niveau de résolution qui n'est pas à la seconde. On obtient donc deux lignes de données qui diffèrent uniquement par le statut. Ce type d'erreur est particulièrement rencontré lors d'un changement de statut. Des règles de validation sont donc établies pour définir les informations à conserver. La première occurrence est conservée. Le Tableau 3-6 fournit un exemple de doublons de statut. Les valeurs des différents attributs sont identiques mis à part pour le statut. Le premier statut *Free* serait, dans l'exemple présenté, celui conservé.

Tableau 3-6 Illustration d'un doublon de statut

id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC
-	-	45.50401	-73.63797	Phone	-	Free	2	15	217	1575546695
-	-	45.50401	-73.63797	Phone	-	Oncoming	2	15	217	1575546695

3.3.2 Défis du traitement d'un flux de données continu

Lorsque l'on travaille avec des données, le nettoyage de celles-ci fait inévitablement partie du défi. Cependant, lorsque les procédures de traitement doivent être automatisées, les défis sont encore plus grands. En effet, il ne s'agit plus de nettoyer les données une seule fois. Dans les projets de recherche précédents, Lacombe (2016) et Laviolette (2017) disposaient de données GPS sur une période finie et pour un échantillon des taxis. Ils ont donc effectué un nettoyage et un traitement de l'ensemble de leur échantillon, puis ont établi des règles de validations en se basant sur leur

échantillon (Lacombe, 2016; Laviolette, 2017). Pour le projet actuel, il s'agit désormais de traiter les données GPS de l'ensemble des taxis de l'île de Montréal et ce traitement doit s'effectuer régulièrement puisque le flux de données est continu. La plateforme n'étant pas une plateforme de visualisation en temps réel, le traitement ne sera pas effectué en temps réel. Mais les données seront importées régulièrement et le traitement de ces dernières doit donc s'effectuer à la même fréquence.

Les méthodes de nettoyage doivent donc rester valables dans le temps ou pouvoir s'adapter à d'éventuelles nouvelles sources d'erreur qui n'avaient pas été observées auparavant. Cela est d'autant plus nécessaire puisque certaines entrées de données sont effectuées manuellement et qu'il y a peu de validation des champs dans le Registre des taxis. Dans le cadre de l'automatisation des procédures de traitement, il est donc crucial de veiller à ce que ces méthodes de traitement restent viables dans le temps et puissent, par exemple, résister à un changement de dispositif d'acquisition de données GPS. En effet, les méthodes doivent assurer la comparabilité des données dans le temps.

Grâce au Registre, il n'y a pas de problème de fusion des données GPS. En effet, même si l'on a accès aux données des taxis des différents opérateurs, ces données proviennent d'une seule structure, à savoir le Registre des taxis, et non des différents opérateurs individuellement. Elles respectent donc un même format.

3.4 Construction d'une base de données optimisée

Avec plus de 60 millions de points GPS collectés par jour, le stockage d'une telle quantité de données, même après les processus de nettoyage des données, est un enjeu fondamental du projet. L'espace disque requis pour le stockage d'un seul mois de données s'élève à plus de 48000 méga-octets, soit 48 giga-octets. Il n'est pas nécessaire de stocker toutes les données brutes pour les procédures de calcul des indicateurs. En effet, les indicateurs qui ont été proposés pour caractériser l'offre et la demande de courses en taxi nécessitent l'utilisation de certaines données qui sont mises en évidence. Ainsi, une base de données optimisée est créée afin de répondre à la problématique du stockage des données et d'accélérer le calcul des indicateurs en regroupant le plus d'informations possible et en supprimant toute information redondante des données brutes.

3.4.1 Schéma

La Figure 3-7 présente les différentes étapes du traitement des données brutes et de la construction d'une table optimisée. De nombreuses opérations sont nécessaires telles que : la suppression des doublons, la correspondance des traces GPS avec le réseau routier (« map matching »), l'intersection spatiale avec les zones de recensement jusqu'à la compilation de données optimisées. Ces étapes sont détaillées dans les paragraphes suivants.

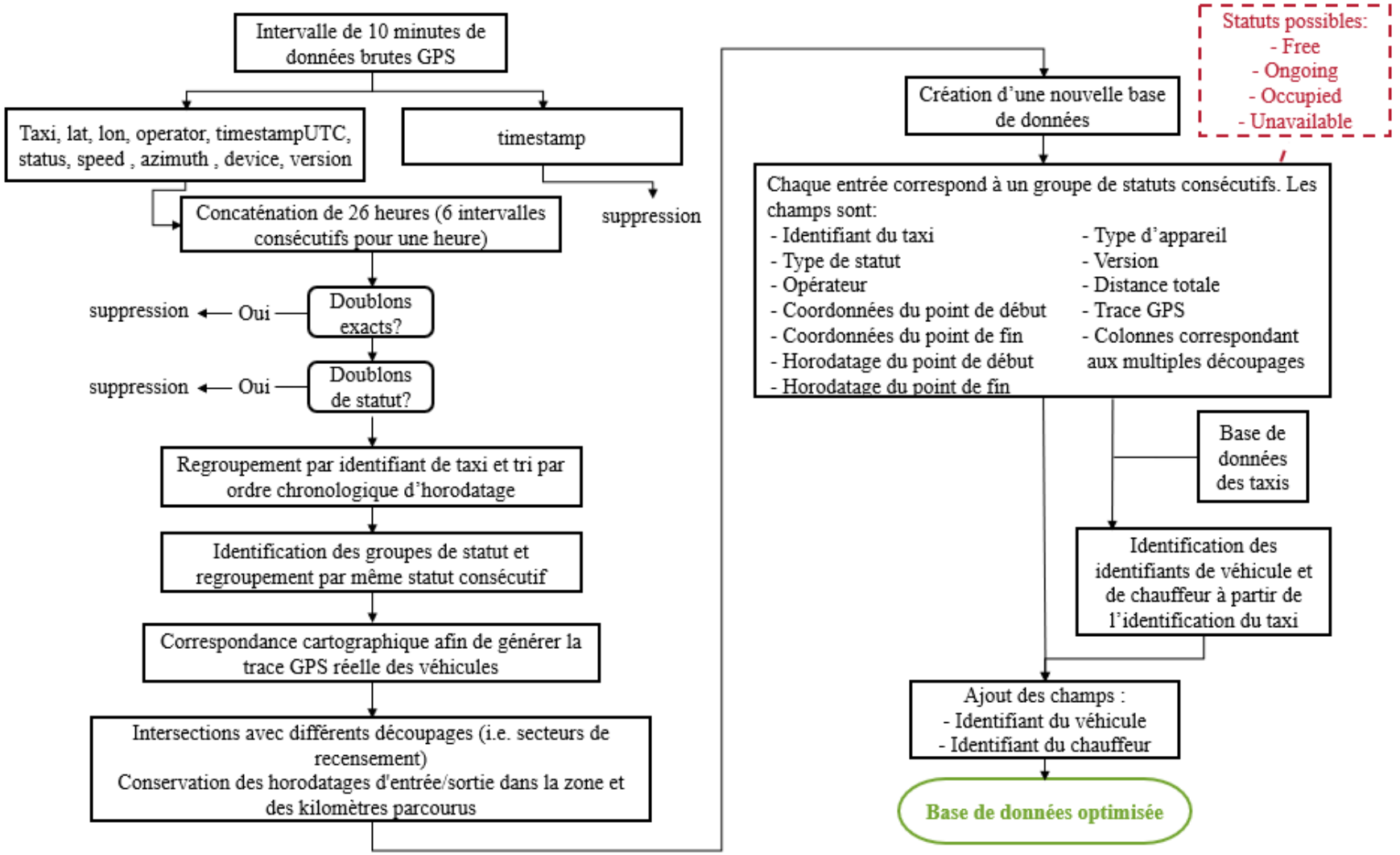


Figure 3-7 Processus de création d'une base de données optimisée

Requête

Tout d'abord, une requête vers une API est lancée afin de récupérer les données du Registre. En raison du grand nombre de données il est possible de récupérer les données de la table Taxis Positions uniquement par plage de dix minutes (Bureau du Taxi de Montréal, 2020a). Par exemple si l'on souhaite récupérer les données d'un jour donné de 10h00 à 11h00 du matin, il faut effectuer 6 requêtes : de 10h00 à 10h10, de 10h10 à 10h20, de 10h20 à 10h30, de 10h30 à 10h40, de 10h40 à 10h50 et de 10h50 à 11h00.

Attributs

Enfin dans la requête, les attributs que l'on souhaite récupérer sont précisés. Ces derniers ont été décrits dans le Tableau 3-2. Tous les attributs sont sélectionnés mis à part l'attribut Timestamp. En effet deux attributs sont disponibles pour indiquer le moment où l'information a été envoyée : Timestamp et TimestampUTC. Le Tableau 3-7 illustre ces deux formats d'horodatage. Cependant l'information fournie par les deux attributs étant la même mais sous deux formats différents, il n'est pas nécessaire de conserver les deux, un seul suffit.

Tableau 3-7 Exemple des deux formats d'horodatage

Attributs	Format
timestamp	1500580203
timestampUTC	2017-07-20T19:50:03.000Z

Aussi, les données brutes de longitude et latitude possèdent de nombreuses décimales, comme l'illustre le Tableau 3-8. Ces dernières si toutes conservées occupent beaucoup d'espace mémoire. Ainsi, seules 5 décimales sont conservées puisque cela permet d'avoir une précision suffisante (au mètre près environ).

Tableau 3-8 Illustration des décimales des coordonnées spatiales

	Données brutes	Données optimisées
Latitude	45.665154348799	45.66515
Longitude	-73.516392380129	-73.51639

Plage de récupération

Dans un premier temps les données sont récupérées par « journée ». Pour l'étude d'une dite « journée », on considère 26 heures de données. On récupère en effet la dernière heure du jour précédent et la première heure du jour suivant en plus des 24 heures de la journée considérée, tel qu'illustré dans la Figure 3-8. On considère en effet qu'une course commencée à 23h00 peut se terminer après minuit. Si pour l'étude d'une journée on ne considère que les données entre 00h00 et 24h00, on déterminerait des bornes erronées pour les courses commençant le jour précédent mais se terminant après 00h00 ou pour les courses commençant avant 24h00 mais se terminant le jour suivant. On émet également l'hypothèse qu'une course de nuit ne dure pas plus de 2 heures. Enfin, on considère que les courses d'une période donnée correspondent aux courses commençant pendant cette période, soit les courses dont l'horodatage du point d'origine est compris dans la période d'analyse souhaitée. Les courses d'une journée correspondent donc à celles commencées entre 00h00 et 24h00. Ainsi une course commencée à 23h00 le jour précédent mais se terminant après minuit sera considérée comme étant une course du jour précédent et non de la journée étudiée.

Figure 3-8 Plage de récupération des données pour l'étude d'une journée

Suppression des doublons

Une fois ces données récupérées on effectue alors diverses opérations. La première est la suppression des doublons. Comme mentionné dans le paragraphe sur le Prétraitement des données à la section 3.3, il existe deux types de doublons : les doublons « exacts » et les doublons « de statut ». On supprime donc ces deux types de doublons en conservant la première occurrence.

Regroupement

Lorsque les données brutes sont récupérées, elles sont par ordre chronologique d'horodatage uniquement, comme illustré dans le Tableau 3-9. Ainsi chaque ligne correspond à un point GPS d'un taxi et la ligne suivante à un point GPS d'un autre taxi. Or, il est nécessaire de regrouper tous les points appartenant à un même taxi afin notamment de pouvoir identifier les courses réalisées par ce taxi.

Tableau 3-9 Illustration de l'ordre des données brutes

id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC
0	taxi_1	45.57323	-73.73004	Phone	-	Free	2	59	146	1575546690
1	taxi_2	45.50574	-73.63573	Phone	-	Occupied	2	3	322	1575546690
2	taxi_3	45.47201	-73.6306	Tablet	-	Free	2	36	212	1575546690
3	taxi_1	45.56199	-73.73696	Phone	-	Free	2	27	156	1575546695
4	taxi_2	45.49754	-73.62688	Phone	-	Occupied	2	9	304	1575546695
5	taxi_3	45.48323	-73.6386	Tablet	-	Free	2	0	106	1575546695
6	taxi_1	45.57007	-73.73016	Phone	-	Free	2	57	142	1575546700
7	taxi_2	45.49398	-73.63871	Phone	-	Occupied	2	9	124	1575546700
8	taxi_3	45.4649	-73.62496	Tablet	-	Unavailable	2	0	70	1575546700

On regroupe donc les données par identifiant de taxi et les données sont triées par ordre chronologique de temps, comme illustré dans le Tableau 3-10.

Tableau 3-10 Regroupement par identifiant de taxi et tri par ordre chronologique de temps

id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC
0	taxi_1	45.57323	-73.73004	Phone	-	Free	2	59	146	1575546690
3	taxi_1	45.56199	-73.73696	Phone	-	Free	2	27	156	1575546695
6	taxi_1	45.57007	-73.73016	Phone	-	Free	2	57	142	1575546700
1	taxi_2	45.50574	-73.63573	Phone	-	Occupied	2	3	322	1575546690
4	taxi_2	45.49754	-73.62688	Phone	-	Occupied	2	9	304	1575546695
7	taxi_2	45.49398	-73.63871	Phone	-	Occupied	2	9	124	1575546700
2	taxi_3	45.47201	-73.6306	Tablet	-	Free	2	36	212	1575546690
5	taxi_3	45.48323	-73.6386	Tablet	-	Free	2	0	106	1575546695
8	taxi_3	45.4649	-73.62496	Tablet	-	Unavailable	2	0	70	1575546700

Fonctions de fenêtrage

Différentes opérations sont réalisées sur les groupes de points grâce aux fonctions de fenêtrage ou « Window function » du langage SQL. En effet, ce type de fonction permet d'appliquer une opération à des groupes de points de tailles différentes sans que l'utilisateur ait à connaître et à renseigner au préalable la taille et les limites des groupes de points. En effet le regroupement est fait selon des critères définis par l'utilisateur, en indiquant par exemple de regrouper selon un même attribut. Mais une fois les regroupements faits, grâce aux fonctions de fenêtrage, diverses opérations peuvent être effectuées au sein d'un regroupement et ce, pour tous les groupes. Il est par exemple possible de calculer la moyenne d'un attribut au sein d'un groupe ou encore d'effectuer des tris au sein des groupes et d'identifier les premières et dernières valeurs du groupe. Les opérations réalisées par les fonctions de fenêtrage peuvent se rapprocher des calculs réalisables par une fonction d'agrégation. Cependant, contrairement à une fonction d'agrégation, l'utilisation d'une fonction de fenêtrage ne résulte pas en un regroupement des enregistrements traités en un seul élément (PostgreSQL, 2020b).

Dans le cadre du traitement des données du Registre, ces dernières sont par exemple regroupées par identifiant de taxi et par statut, comme illustré dans le Tableau 3-11. Au sein de chaque groupe, les données sont ordonnées par ordre chronologique de temps. Des opérations peuvent ensuite être effectuées au sein de ces partitions telles que l'identification des premiers et derniers horodatages de chaque groupe, la somme des distances séparant les points consécutifs d'un groupe, la vitesse moyenne ou encore la valeur maximale de vitesse du groupe.

Tableau 3-11 Illustration du partitionnement par identifiant de taxi et par statut

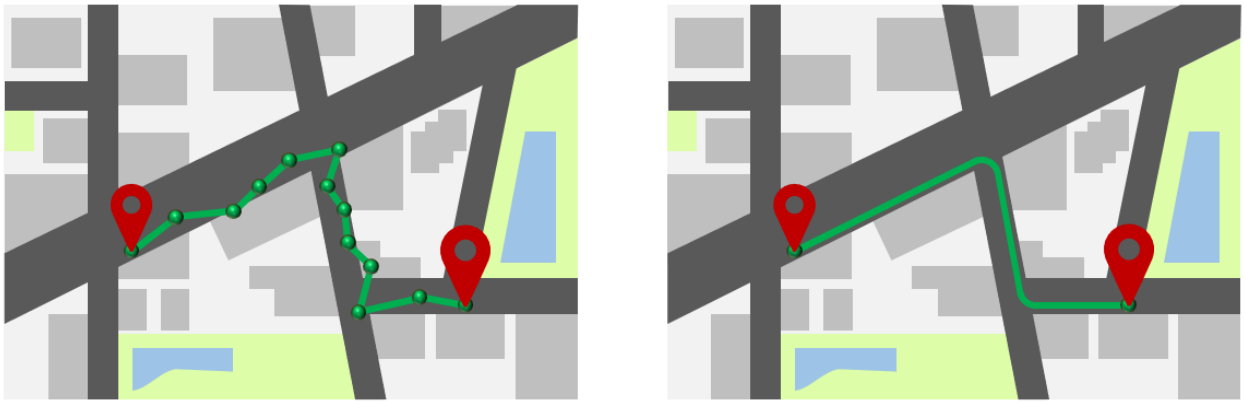
id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC
0	taxi_1	45.57323	-73.73004	Phone	-	Free	2	59	146	1575546690
3	taxi_1	45.56199	-73.73696	Phone	-	Free	2	27	156	1575546695
6	taxi_1	45.57007	-73.73016	Phone	-	Free	2	57	142	1575546700
9	taxi_1	45.54383	-73.71653	Phone	-	Oncoming	2	53	143	1575546705
12	taxi_1	45.55699	-73.73150	Phone	-	Oncoming	2	60	143	1575546710
15	taxi_1	45.55946	-73.73418	Phone	-	Oncoming	2	53	233	1575546715
18	taxi_1	45.56329	-73.73669	Phone	-	Oncoming	2	47	333	1575546720
21	taxi_1	45.56261	-73.73970	Phone	-	Oncoming	2	54	144	1575546725
24	taxi_1	45.54658	-73.71920	Phone	-	Oncoming	2	54	142	1575546730

Enfin, ce format est à nouveau optimisé afin de supprimer les informations redondantes. En effet, il n'est pas nécessaire de conserver toutes les informations de chaque ligne puisque certaines informations se retrouvent dans plusieurs lignes. Par exemple, si le statut d'un taxi reste inchangé pendant une certaine période temporelle, le même statut sera renseigné à chaque ligne.

Les données sont donc regroupées en groupe de points consécutifs de même statut, afin qu'une ligne de données ne corresponde plus à un unique point mais contienne toutes les informations d'un ensemble de points d'un taxi donné. La méthodologie utilisée pour identifier les groupes de points est détaillée dans le Chapitre 4 et le Tableau 3-12 fournit un aperçu de ce regroupement en groupes de points consécutifs de même statut.

3.4.2 Correspondance cartographique

Une étape de correspondance cartographique (« map matching ») est par la suite essentielle pour identifier la trace réelle des véhicules. En effet, la trace établie à l'aide des points GPS transmis peut être très éloignée de la trace réelle des véhicules sur le réseau. Ceci est principalement dû à la précision des outils GPS qui réalisent déjà des interpolations ainsi qu'au type d'environnement dans lequel le véhicule se déplace. En effet, le phénomène des canyons urbains peut être présent, en particulier dans les centres-villes. Il se caractérise par l'occultation d'un satellite en raison du relief tel qu'un bâtiment ou d'un écho du signal contre une surface qui fournira ainsi une fausse localisation (Correia, 2006). La distance calculée à l'aide des points GPS, correspondant à la somme cumulée des distances entre des points consécutifs deux à deux, peut donc être très différente de la véritable distance réseau parcourue, comme l'illustre la Figure 3-9. De plus on ne possède pas les données brutes GPS avec leurs incertitudes. La correspondance cartographique permet donc de reconstruire la trajectoire la plus plausible en utilisant le réseau routier.



(a) (b)

Figure 3-9 Illustration de la différence entre la trace générée par les points GPS (a) et la trace réelle d'un véhicule (b)

L'Open Source Routing Machine ou OSRM est un routeur libre de droits conçu pour être utilisé avec les données du projet OpenStreetMap (OSRM, 2020). Il est utilisé dans le cadre du projet pour établir la correspondance entre les points GPS relevés et le réseau routier.

La correspondance cartographique permet de faire correspondre les points GPS donnés au réseau routier de la manière la plus plausible. OSRM peut renvoyer une, plusieurs ou aucune correspondance selon les données GPS fournies. D'importants intervalles de temps entre les données (supérieurs à 60 secondes) ou des transitions peu probables peuvent résulter en un fractionnement des traces si une correspondance complète n'a pas pu être trouvée. Il est possible que l'algorithme ne soit pas en mesure de faire correspondre tous les points. Les valeurs aberrantes sont alors supprimées si elles ne peuvent pas être comparées avec succès (OSRM, 2020).

3.4.3 Intersection spatiale

Comme mentionné dans la partie 3.1.6, des couches de données géographiques externes au Registre sont incluses afin d'effectuer des analyses spatiales. Après consultation des partenaires du projet, il a été établi que la plus petite résolution spatiale pour l'analyse serait les secteurs de recensements. La Figure 3-10 présente les secteurs de recensement de l'Île de Montréal (identifiés en orange), datant du recensement de 2016.

Les secteurs de recensement (SR) sont de petites régions géographiques dont la population est comprise entre 2 500 et 8 000 habitants. Ils sont créés au sein de régions métropolitaines de

recensement et d'agglomérations de recensement. Les SR sont créés de sorte à être le plus homogène possible en ce qui concerne les caractéristiques socio-économiques et ils doivent respecter les limites des régions métropolitaines de recensement, des agglomérations de recensement ainsi que les limites provinciales (Statistique Canada, 2018).

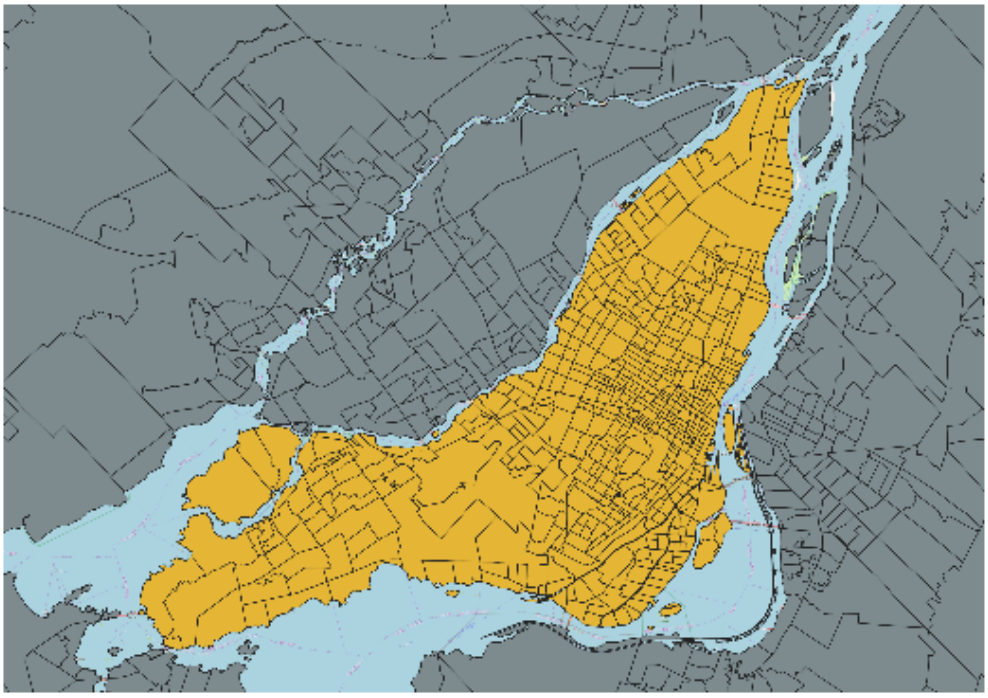


Figure 3-10 Carte des secteurs de recensement de l'île de Montréal (2016)

Une intersection spatiale est donc réalisée entre les traces GPS obtenues et les différents découpages zonaux. Les horodatages lors de l'entrée et de la sortie des zones sont identifiés et conservés ainsi que la distance parcourue dans la zone. La Figure 3-11 illustre le processus d'intersection spatiale.

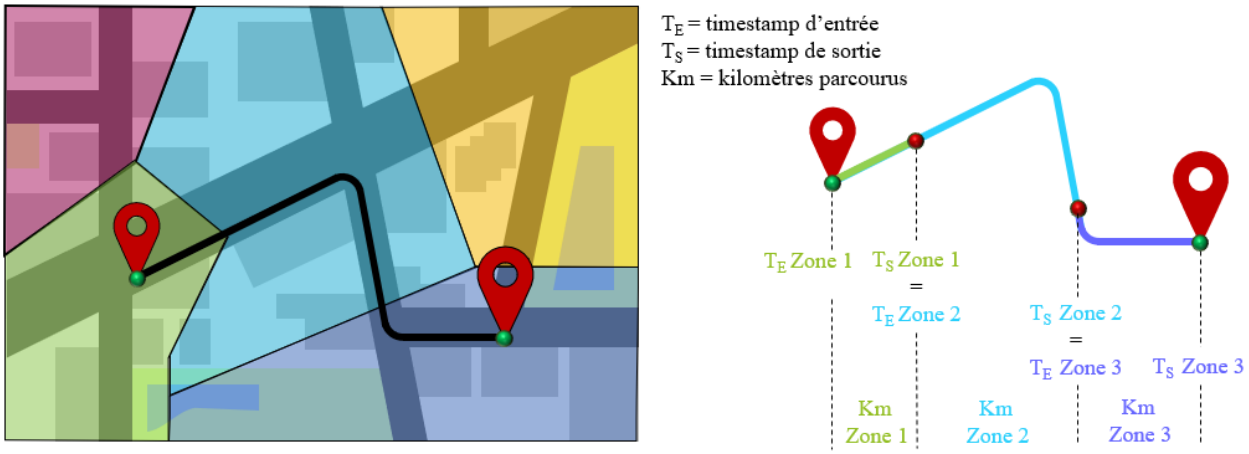


Figure 3-11 Schéma du procédé d'intersection spatiale

Toutes ces informations sont regroupées dans une nouvelle table de données. L'identifiant du véhicule ainsi que l'identifiant du chauffeur sont alors identifiés à partir de l'identifiant de taxi et des tables de données des Véhicules et des Chauffeurs puis ajoutés à la base de données optimisée.

Le Tableau 3-12 présente un aperçu des transformations des données brutes en données optimisées. Les données brutes, dont chaque ligne correspond à un seul point GPS, sont transformées en données optimisées où une ligne correspond à un groupe de points consécutifs ayant le même statut.

Tableau 3-12 Aperçu du traitement des données brutes en données optimisées

Données brutes (un point par ligne)

id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC
-	-	45.50401	-73.63797	Phone	-	Free	2	15	217	1575546695
-	-	45.50382	-73.63819	Phone	-	Free	2	18	221	1575546700
-	-	45.50372	-73.63834	Phone	-	Free	2	15	254	1575546705
-	-	45.50381	-73.63875	Phone	-	Free	2	9	305	1575546710
-	-	45.50414	-73.63949	Phone	-	Free	2	6	302	1575546715
-	-	45.50457	-73.64046	Phone	-	Free	2	5	302	1575546720
-	-	45.50498	-73.64138	Phone	-	Free	2	3	304	1575546725

Données optimisées (un groupe de points consécutifs de même statut par ligne)

id	taxi	latitude_début	longitude_début	horodatage_début	latitude_fin	longitude_fin	horodatage_fin	opérateur	statut
-	-	45.50401	-73.63797	1575546695	45.50498	-73.64138	1575546725	-	Free

appareil	version	distance totale	trace GPS	intersection spatiale	véhicule id	chauffeur id
Phone	2	350	géométrie	géométrie	-	-

CHAPITRE 4 IDENTIFICATION DES COURSES

4.1 Explication du problème

Comme expliqué à la section 3.4, afin de minimiser le temps nécessaire au calcul des indicateurs et l'espace mémoire utilisé, les données brutes sont transformées en une base de données optimisée. Dans cette dernière, chaque ligne correspond à un groupe de statut (regroupement de plusieurs points consécutifs de même statut d'un taxi donné). Afin d'identifier les différents groupes, les données brutes sont triées par ordre chronologique croissant et regroupées par taxi. Ensuite, pour un identifiant de taxi donné, une itération sur les lignes est effectuée : on vérifie si le point suivant (consécutif dans le temps) a le même statut que le précédent.

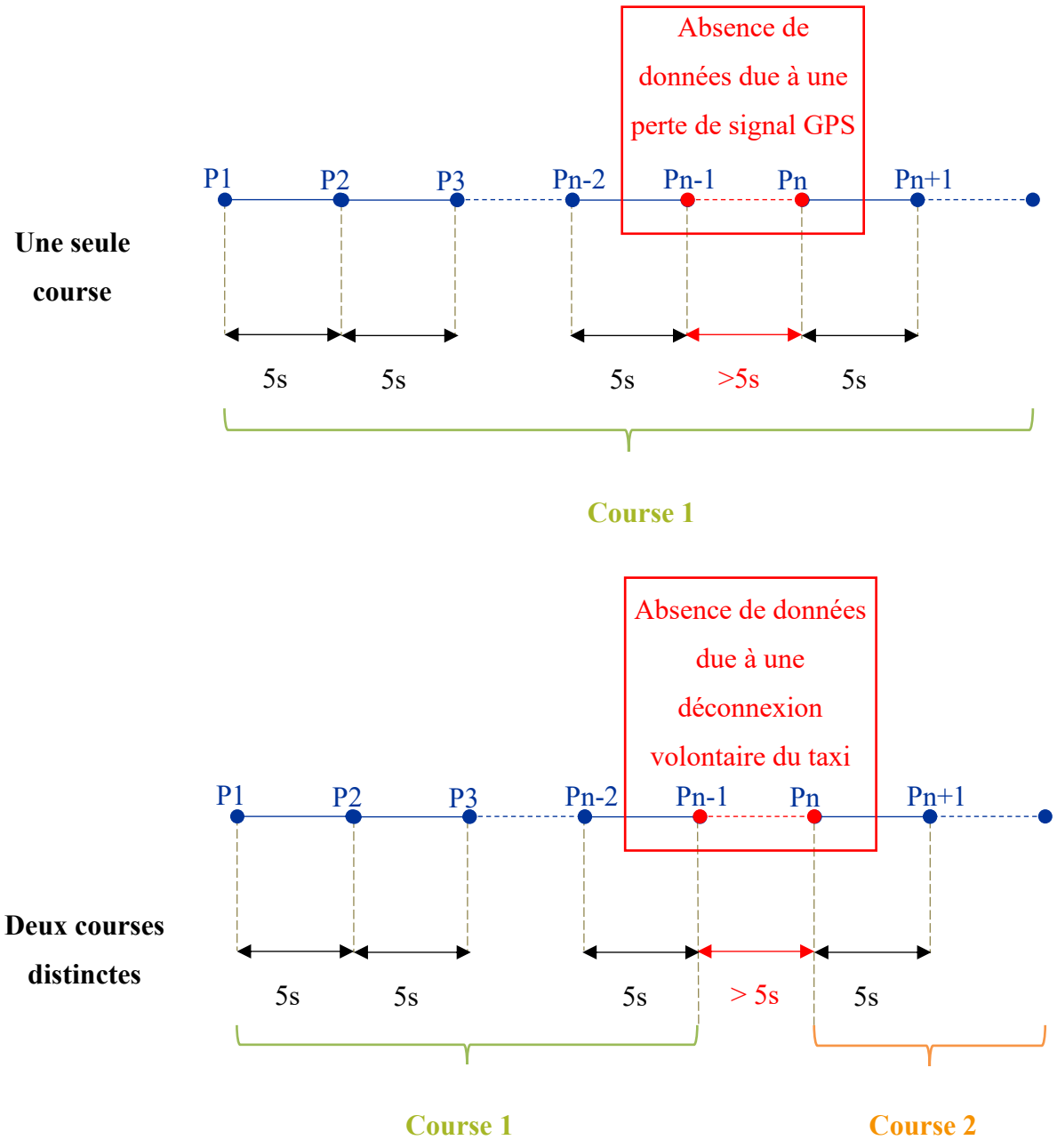
Selon l'ordonnance du BTM, les points GPS doivent être envoyés toutes les 5 secondes (Beaudoin, David 2017). Cependant, les données recueillies présentent des intervalles de temps entre deux points successifs qui excèdent 5 secondes. Lors du regroupement de points consécutifs en groupes de même statut, si l'intervalle de temps entre deux points consécutifs de statuts différents est supérieur à 5 secondes, cela ne pose pas de problème puisqu'il y a changement de statut. Toutefois, dans le cas de deux points consécutifs de même statut, si l'intervalle de temps entre les points dépasse 5 secondes, il est nécessaire de pouvoir déterminer si les deux points appartiennent ou non au même groupe. En effet, on ne peut se contenter d'examiner uniquement s'il y a un changement de statut afin de déterminer s'ils appartiennent au même groupe. Autrement, si le taxi se déconnecte entre deux points (c'est-à-dire qu'il n'est plus en service) sans changement de statut, les points seraient toujours considérés comme appartenant au même groupe. Par exemple, plus de 12 heures peuvent être observées entre deux points consécutifs ayant le même statut *occupied*, comme l'illustre la Figure 4-1. Le *Δtemps* correspond à la différence de temps en secondes avec l'horodatage du point précédent. Même si les deux points consécutifs ont le même identifiant de taxi et ont le même statut *occupied*, il est certain qu'ils n'appartiennent pas à la même course.

taxi	latitude	longitude	opérateur	timestamp	statut	Δ temps (s)
	45.37224	-73.72858		1575542825	occupied	10
	45.37266	-73.72875		1575542835	occupied	10
	45.37359	-73.729		1575542845	occupied	10
	45.37425	-73.72902		1575542855	occupied	10
	45.37513	-73.72883		1575542865	occupied	10
	45.37611	-73.72859		1575542875	occupied	10
	45.37668	-73.72846		1575542885	occupied	10
	45.3776	-73.72823		1575542895	occupied	10
	45.37774	-73.72801		1575542905	occupied	10
	45.37766	-73.72794		1575542915	occupied	10
	45.37761	-73.72795		1575542926	occupied	11
	45.37761	-73.72795		1575542936	occupied	10
	45.37766	-73.728		1575587430	occupied	44494
	45.37766	-73.728		1575587441	occupied	11
	45.37766	-73.728		1575587451	occupied	10
	45.37766	-73.728		1575587461	occupied	10

Figure 4-1 Exemple d'une déconnexion sans changement de statut "occupied"

Si le statut est libre (*free*), cela est moins problématique. En effet, seule la durée pendant laquelle le taxi est libre, soit la somme des durées des groupes de statut libre, est calculée. Ainsi, le nombre de groupes n'a pas d'impact.

Dans le cas où le taxi est en course (le statut indique *occupied*), deux options apparaissent. En effet, lorsque l'intervalle de temps entre deux points consécutifs de même statut *occupied* est, par exemple, de 3 minutes, le second point peut soit être considéré comme appartenant au même groupe que le point précédent (un passage dans un tunnel pourrait expliquer l'absence de données), soit appartenir à une nouvelle course. Si le second point correspond au début d'une nouvelle course, cela suppose que le taxi a terminé sa course précédente, qu'il n'a pas été en service pendant quelques minutes, puis a repris du service en débutant une nouvelle course. La Figure 4-2 illustre ces deux cas de figure.



P_i : Point GPS de statut occupé (« occupied »)

Figure 4-2 Schéma illustrant les deux possibilités de regroupement en courses lorsque des points consécutifs de même statut *occupied* sont séparés d'un intervalle supérieur à 5 secondes

Afin de pouvoir déterminer précisément les indicateurs sur la durée, distance et vitesse moyennes des courses, les limites de ces dernières doivent clairement être identifiées. Il est donc nécessaire d'établir un processus de regroupement des points consécutifs de même statut. Dans un premier temps, l'étude des distributions des intervalles de temps entre deux points consécutifs de même statut est réalisée afin d'identifier des règles de regroupement. Cette étude est d'abord réalisée pour le statut o*ccupied* en vue d'identifier les courses de taxis. En effet, puisque les données du Registre ne fournissent pas une table des données des courses mais uniquement des points GPS, les courses doivent donc être reconstituées à partir des points recueillis.

4.2 Distribution des intervalles de temps

D'après le manuel d'utilisation des données du Registre, la position du taxi peut être relevée par plusieurs types d'appareils (Beaudoin, David 2017) :

- Telephone intelligent « phone »
- Tablette « tablet »
- Taximètre « taximeter »
- Autre appareil « other device »

Selon les opérateurs le type d'appareil utilisé varie comme l'illustre le Tableau 4-1. Seuls deux types d'appareils sont utilisés : la tablette (« tablet ») ou le téléphone intelligent (« phone »).

Tableau 4-1 Appareil utilisé pour relever la position GPS selon les intermédiaires en service

Appareil	Liste des intermédiaires	Appareil	Liste des intermédiaires
Tablet	Intermédiaire1	**Phone**	Intermédiaire11
	Intermédiaire2		Intermédiaire12
	Intermédiaire3		Intermédiaire13
	Intermédiaire4		Intermédiaire14
	Intermédiaire5		Intermédiaire15
	Intermédiaire6		Intermédiaire16
	Intermédiaire7		Intermédiaire17
	Intermédiaire8		
	Intermédiaire9		
	Intermédiaire10		

Pour des raisons de confidentialité des données, les noms des intermédiaires en service sont associés à des identifiants.

On réalise donc la distribution moyenne des intervalles de temps entre deux points consécutifs de même statut *occupied* en regroupant les intermédiaires en service selon le type d'appareil utilisé afin d'identifier si l'appareil a une influence sur la régularité d'envoi des données GPS.

Pour les intermédiaires utilisant une tablette (« tablet ») pour relever la position GPS du véhicule, la Figure 4-3 (a) illustre la distribution cumulée des intervalles de temps entre deux points consécutifs de même statut *occupied*. Pour les utilisateurs de tablette il semble y avoir une distribution semblable des points GPS, mis-à-part pour l'intermédiaire en service d'identifiant 1.

Et pour les intermédiaires utilisant un téléphone intelligent (« phone ») pour relever la position GPS du véhicule, la Figure 4-3 (b) illustre la distribution cumulée des intervalles de temps entre deux points consécutifs de même statut *occupied*.

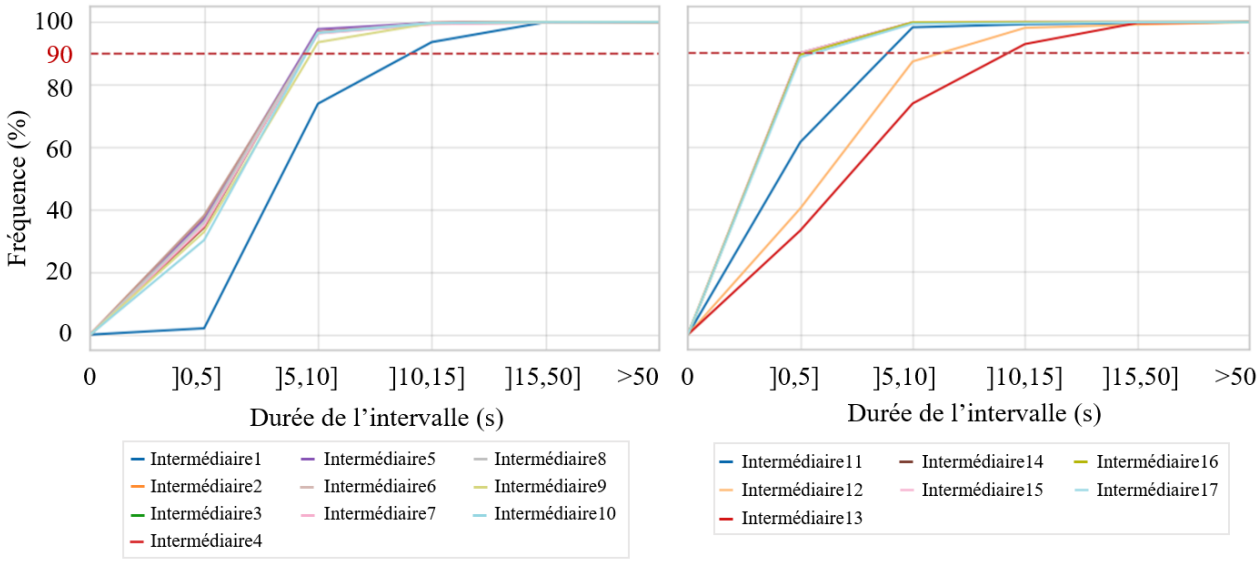


Figure 4-3 Distribution cumulée des intervalles de temps entre deux points consécutifs de même statut *occupied* pour les intermédiaires utilisant comme appareil de relevé des positions GPS : (a) un téléphone intelligent (b) une tablette

Cependant, on peut surtout noter que dans les deux cas, donc pour tous les opérateurs, 90% des intervalles de temps sont inférieurs à 15 secondes.

Tous les opérateurs sont donc regroupés dans la Figure 4-4 (a). Plus de 90% des intervalles de temps sont inférieurs à 15 secondes et ce, pour tous les opérateurs. La Figure 4-4 (b) est un agrandissement de la Figure 4-4 (a). Elle met en évidence que 99% des intervalles de temps entre

deux points consécutifs de même statut *occupied* sont inférieurs à 50 secondes. On traitera donc les données sans différencier selon l'intermédiaire en service ou selon l'appareil de relevé utilisé. Ainsi une seule méthode de traitement des points consécutifs de même statut *occupied* peut donc être établie et appliquée à tous les opérateurs.

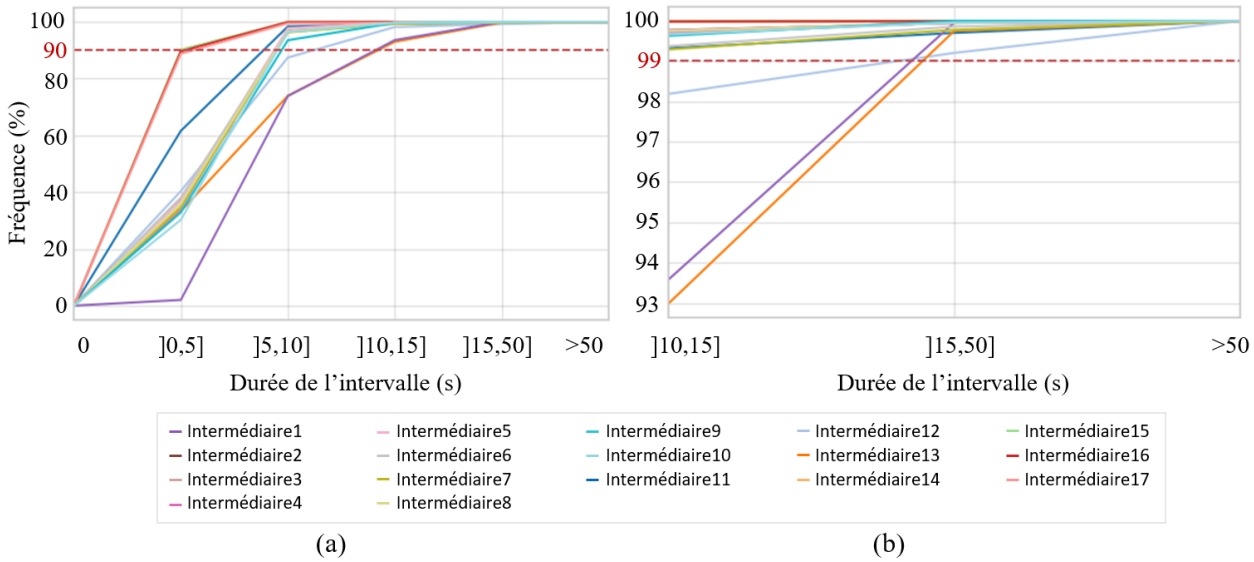


Figure 4-4 Distribution cumulée des intervalles de temps entre deux points consécutifs de même statut *occupied* (a) pour tous les intermédiaires en service (b) agrandissement de (a)

4.3 Hypothèses pour le regroupement des points GPS consécutifs de statut *occupied*

Trois hypothèses ont été émises afin d'identifier les regroupements des points GPS consécutifs de même statut *occupied.* Elles sont présentées dans la Figure 4-5 qui est une schématisation de ce processus de regroupement. Les 2 principaux scénarios qui ont été expérimentés y sont présentés.

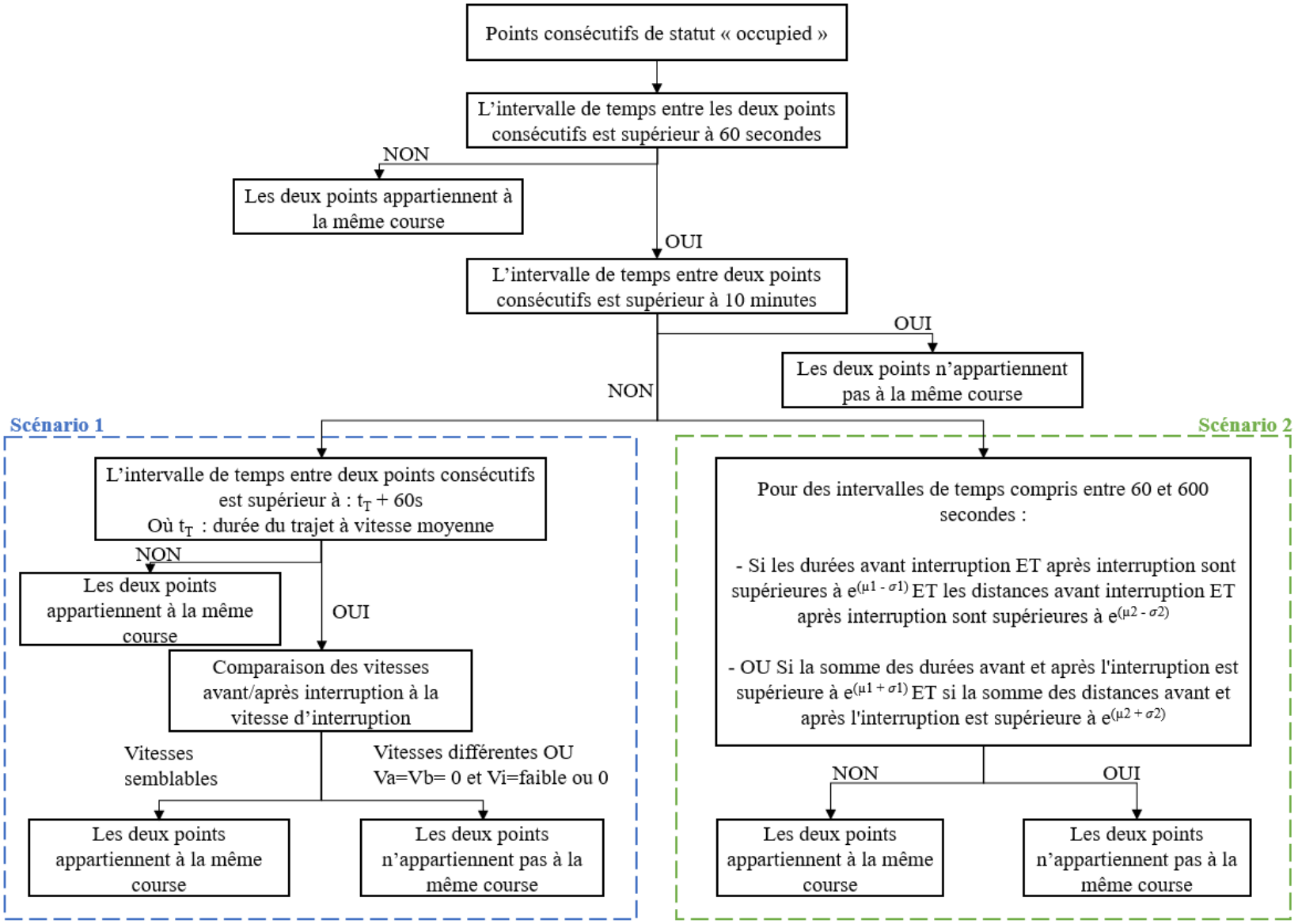


Figure 4-5 Processus de regroupement des points GPS consécutifs de même statut occupied

Le premier scénario repose sur l'étude des vitesses moyennes de déplacements. Le second scénario se base quant à lui sur les distributions des distances et durées des courses d'une population de référence.

Les deux premières hypothèses retenues pour le regroupement des points consécutifs de même statut *occupied* sont communes aux deux scénarios, à savoir :

- **Hypothèse 1** : On considère que le temps minimum pour qu'un client embarque ou débarque est d'une minute. De plus, l'embarquement et le débarquement peuvent se faire simultanément (Dandl, Bracher, & Bogenberger, 2017). Si l'intervalle entre deux points consécutifs est inférieur à 60 secondes, on suppose donc qu'il ne peut y avoir de changement de client et que les deux points font partie de la même course.

- **Hypothèse 2** : On suppose que la forte densité de bâtiments dans le centre-ville peut entraîner une perte de signal GPS en raison du phénomène de canyon urbain (Correia, 2006). La section du centre-ville de Montréal allant de la station Berri-UQAM à la station Atwater est considérée comme la zone où une perte de données peut se produire en raison de la forte densité de gratte-ciel (Skyscraper Source Media, 2020). La Figure 4-6 met en évidence cette forte densité de bâtiments entre les stations Berri-UQAM et Atwater. On se place dans des conditions de circulation défavorables, c'est-à-dire lorsque de la congestion est présente en centre-ville et que les véhicules ne peuvent pas circuler à une vitesse d'écoulement libre. On considère un seuil de congestion de 60%, c'est-à-dire que les véhicules se déplacent à une vitesse moyenne correspondant à 60% de la vitesse d'écoulement libre. La vitesse autorisée (et donc en écoulement libre) est de 40 km/h pour le centre-ville de Montréal. On considère donc une vitesse de déplacement moyenne de 24 km/h (Tessier, 2015; Ville de Montréal, 2020b). Il faut 10 minutes pour traverser le centre-ville (de la station Berri-UQAM à la station Atwater), soit un trajet de 4 km à une vitesse de 24 km/h. Ceci représente la limite supérieure: un intervalle de temps de plus de 10 minutes, soit 600 secondes, est considéré comme une déconnexion du taxi. Les deux points n'appartiennent donc pas à la même course.

C'est aussi l'hypothèse posée par Laviolette (2017) qui estime qu'un intervalle de temps entre deux points supérieur à 10 minutes résulte d'une déconnexion du véhicule (Laviolette, 2017).

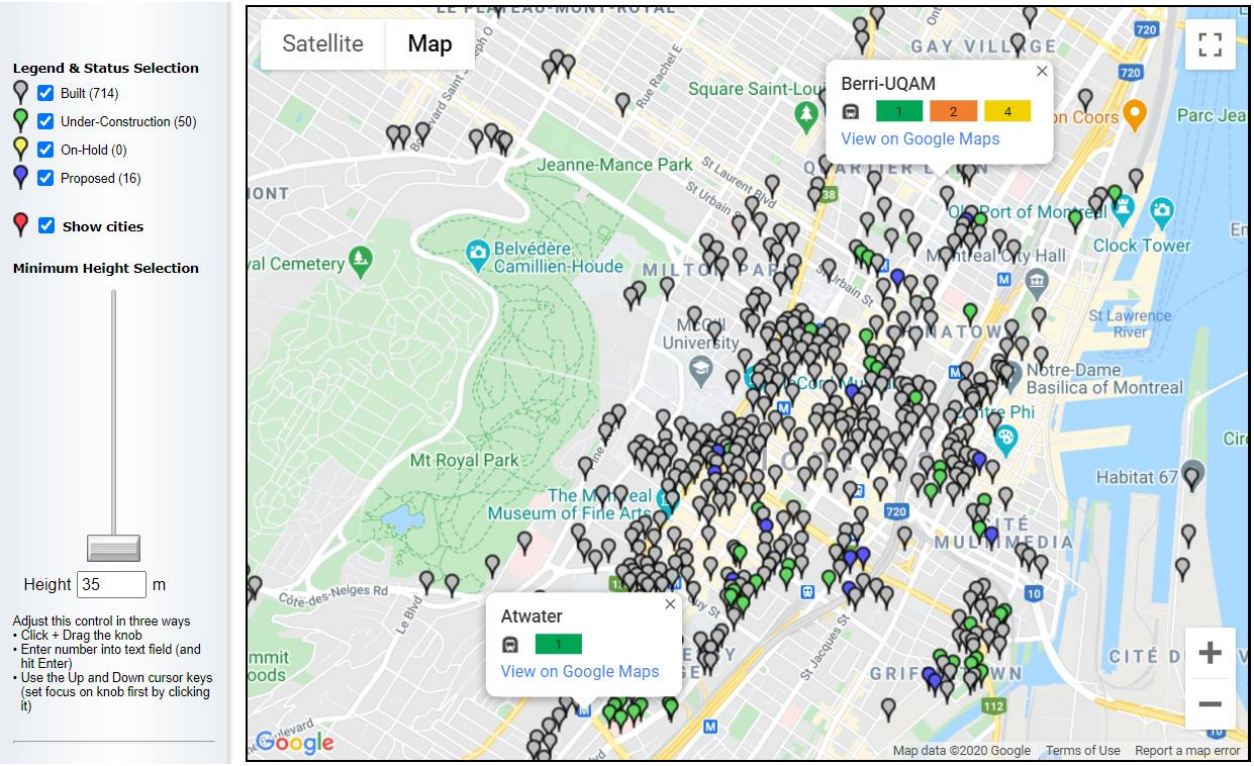


Figure 4-6 Carte montrant la concentration en bâtiments supérieurs à 35 mètres en centre-ville de Montréal (Skyscraper Source Media, 2020)

Les bornes inférieure et supérieure ont donc été définies respectivement à 60 secondes et 600 secondes. Ainsi, si l'intervalle de temps entre les deux points est inférieur à 60 secondes, on estime que les deux points appartiennent au même groupe de points. Et si l'intervalle est supérieur à 600 secondes, soit 10 minutes, on considère qu'ils appartiennent à deux groupes distincts.

Nous allons maintenant nous intéresser au traitement des données comprises entre ces deux bornes. Deux scénarios sont envisagés pour déterminer si l'absence de données résulte d'une déconnexion du taxi ou bien d'une perte de données.

Une analyse préliminaire du scénario n°1 est disponible en annexe B. En raison de la trop grande incertitude quant aux hypothèses prises dans ce scénario, ce dernier n'a pas été retenu pour traiter les intervalles de temps compris entre 60 et 600 secondes.

Le scénario n°2 est celui retenu pour la méthodologie d'identification des regroupements des points GPS consécutifs de même statut *occupied*. Il se base sur les distributions des distances et durées d'une population de référence de courses qualifiées de complètes. L'identification de cette population de référence et de ses caractéristiques est détaillée dans la section suivante.

4.3.1 Détermination de la durée moyenne et distance moyenne de course de la population de référence

Dandl et al. (2017) estiment que le temps correspondant à l'embarquement ou au débarquement d'un client est de 60 secondes (Dandl et al., 2017). On considère donc qu'un quart de ce temps, soit 15 secondes, n'est pas suffisant pour qu'un client débarque et qu'un autre embarque. Ainsi toute course dont les points sont espacés d'au maximum 15 secondes est considérée comme une course complète, c'est-à-dire sans perte de données et sans changement potentiel de client.

On détermine ainsi le nombre de courses complètes et leurs caractéristiques. Les caractéristiques moyenne de ces courses (durée moyenne de course, distance moyenne parcourue) vont nous permettre de déterminer des règles de validation pour les points problématiques. Ces courses représentent 50% des groupes de statut consécutifs (pourcentage obtenu en analysant l'ensemble du mois d'avril 2019). Les courses constituées de points séparés de 5 secondes au maximum représentent environ 5% des groupes de statut consécutifs.

Les courses ne comportant que des points espacés de 15 secondes au maximum peuvent toutefois être des courses erronées. En effet, des regroupements de points de statut *occupied* de moins de 30 secondes peuvent être observés. Ces groupes ne peuvent pas correspondre à une course et sont liés à une erreur de statut (par exemple un oubli de changement de statut par le chauffeur ou une mauvaise entrée de statut). De plus, on observe également que certains chauffeurs ne changent jamais de statut lors de la journée. Ainsi certains taxis indiquent un statut *occupied* tout au long de la journée. Cela induit des groupes de points consécutifs de statut *occupied* de plusieurs heures. On souhaite donc que ces courses erronées ne soient pas prises en compte dans le calcul de la durée moyenne de course et de la distance moyenne de course.

On réalise les distributions des distances et durées des courses constituées de points espacés d'au maximum 15 secondes (qualifiées de courses « complètes »). Ces distributions sont présentées dans les Figure 4-7 et Figure 4-8. Enfin, l'analyse des distributions permet de mettre en évidence que :

- La distribution des durées de courses suit une loi log-normale de paramètres μ_1 et σ_1^2

- La distribution des distances de courses suit une loi log-normale de paramètres μ_2 et σ_2^2

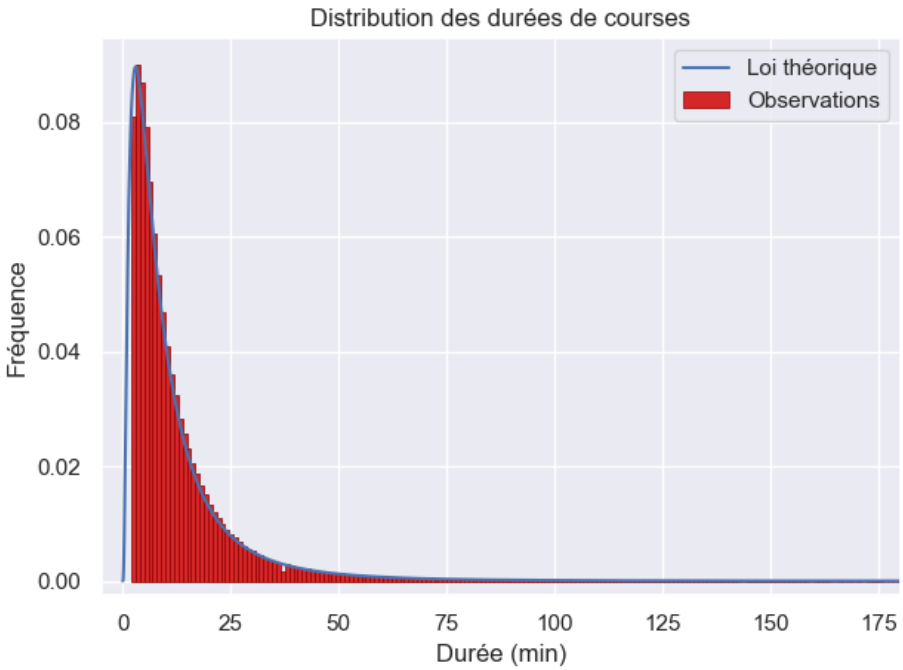


Figure 4-7 Distribution des durées des courses « complètes » du mois d'avril 2019

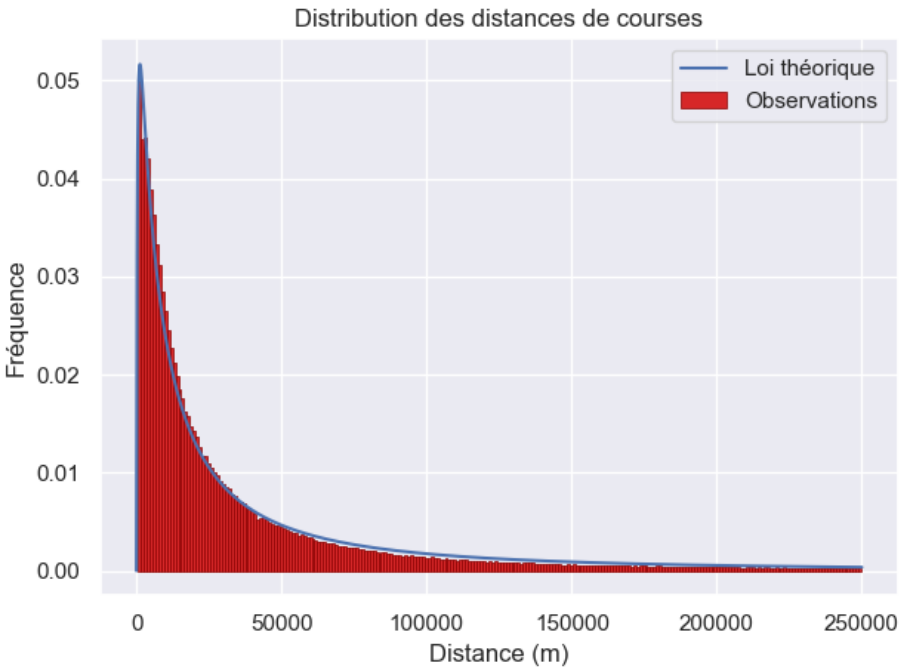


Figure 4-8 Distribution des distances des courses « complètes » du mois d'avril 2019

Afin de ne pas prendre en compte les valeurs aberrantes on considère que les courses correctes sont situées dans l'intervalle [moyenne - écart type ; moyenne + écart-type].

En effet, la variable durée de course X suit une loi log-normale de paramètres μ et σ^2 car la variable Y=ln(X) suit une loi normale d'espérance μ et de variance σ^2. La règle des trois sigmas indique que pour une loi normale 99,73% des valeurs sont situées à plus ou moins 3 écarts-types de la moyenne. Cette règle n'est pas retenue dans notre cas puisque l'on souhaite que les bornes soient suffisamment restrictives. Cependant, si l'on considère l'intervalle [μ - σ ; μ + σ] pour la loi normale, on sélectionne alors 68,27 % des valeurs (Lejeune, 2010). C'est donc cet intervalle qui est choisi. En effectuant une transformation exponentielle afin d'appliquer cette règle à la loi log-normale, on obtient les bornes pour les distributions log-normales, comme l'illustre la Figure 4-9.

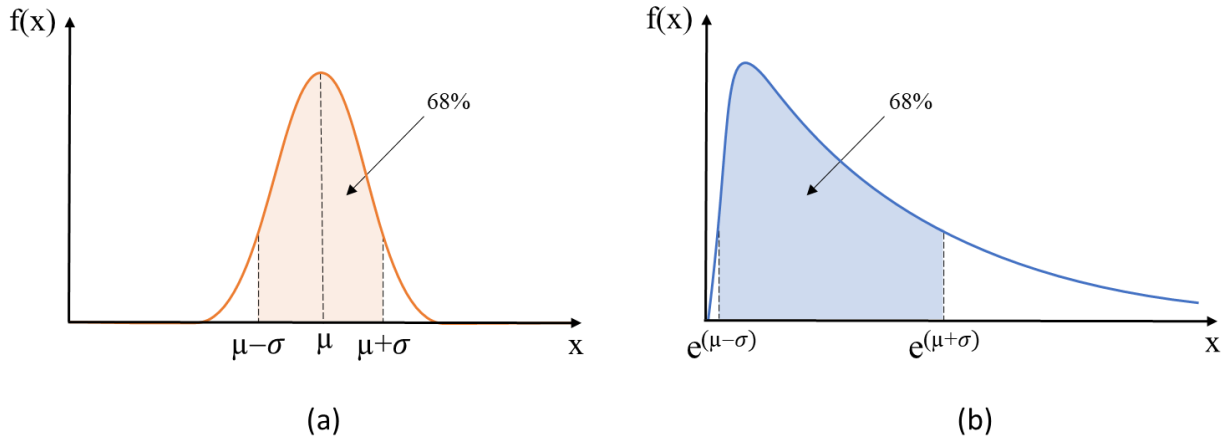


Figure 4-9 Illustration de la règle des sigmas pour : (a) la loi normale où environ 68 % des valeurs se situent à moins d'un écart-type de la moyenne (b) pour la loi log-normale

A titre d'exemple, les moyennes suivantes sont effectuées sur une semaine du mois de décembre 2019.

Durée :

- $e^{(\mu 1 - \sigma 1)}$ = 6,2 minutes = t_{min}
- $e^{(\mu 1 + \sigma 1)}$ = 24,3 minutes =t_{max}

Distance :

- $e^{(\mu 2 - \sigma 2)}$ = 1730 mètres =d_{min}
- $e^{(\mu 2 + \sigma 2)}$ = 8744 mètres =d_{max}

Ces bornes sont utilisées dans le scénario n°2 afin de déterminer les regroupements de points pour les points séparés d'un intervalle de temps compris entre 60 et 600 secondes.

Les bornes calculées pour la semaine du mois de décembre pourraient donc être utilisées dans le processus de regroupement de points consécutifs de même statut *occupied* pour cette même semaine de décembre.
Dans le cadre de l'étude, on choisit d'estimer ces bornes par semaine. Ainsi, l'identification des courses se fait par semaine. Les bornes sont définies pour les courses complètes d'une semaine au complet avant de les appliquer aux données cette même semaine et d'identifier l'ensemble des courses. Par exemple, si l'on souhaite identifier les courses de la première semaine du mois d'avril, les courses complètes de cette même semaine d'avril sont d'abord identifiées et leurs caractéristiques moyennes servent ensuite de valeurs seuils pour déterminer le reste des courses de la même semaine.

4.3.2 Description du scénario n°2

Les Figure 4-10 et Figure 4-11 présentent le processus de traitement par le scénario n°2 des points consécutifs de statut *occupied* séparés d'un intervalle de temps compris entre 60 et 600 secondes. Les premières étapes du pré-traitement des données, détaillées dans la partie 3.2, ont déjà permis d'effectuer le regroupement des points GPS par identifiant de taxi et par statut. Les points GPS de même statut sont donc jusqu'à présent regroupés dans la même partition tant qu'ils sont consécutifs dans le temps et proviennent du même taxi. La Figure 4-10 illustre l'ajout d'un attribut temporaire « repère » permettant de mettre en évidence les intervalles de temps (*Δtemps*) problématiques : soit ceux compris entre 60 et 600 secondes. Cet attribut prend la valeur « 1 » lorsque l'intervalle est entre 60 et 600 secondes, sinon il prend la valeur « 0 ».

id	taxi	latitude	longitude	device	operator	status	version	speed	azimuth	Timestamp UTC	Δtemps (s)	Repère
1	taxi_1	45.493	-73.5758	tablet	-	occupied	2	0	15	1575465071	*null*	0
2	taxi_1	45.4935	-73.5752	tablet	-	occupied	2	31	32	1575465081	10	0
3	taxi_1	45.4943	-73.5746	tablet	-	occupied	2	39	35	1575465091	10	0
4	taxi_1	45.495	-73.5739	tablet	-	occupied	2	29	34	1575465101	10	0
5	taxi_1	45.4955	-73.5733	tablet	-	occupied	2	20	36	1575465112	11	0
6	taxi_1	45.4957	-73.5732	tablet	-	occupied	2	0	43	1575465122	10	0
7	taxi_1	45.4957	-73.5732	tablet	-	occupied	2	0	43	1575465202	**80**	**1**
8	taxi_1	45.4957	-73.573	tablet	-	occupied	2	17	34	1575465212	10	0
9	taxi_1	45.4967	-73.5721	tablet	-	occupied	2	46	34	1575465222	10	0
10	taxi_1	45.4976	-73.5712	tablet	-	occupied	2	40	34	1575465232	10	0
11	taxi_1	45.4984	-73.5706	tablet	-	occupied	2	23	44	1575465322	**90**	**1**
12	taxi_1	45.4993	-73.5697	tablet	-	occupied	2	48	34	1575465332	10	0
13	taxi_1	45.5004	-73.5688	tablet	-	occupied	2	41	34	1575465342	10	0
14	taxi_1	45.5011	-73.5683	tablet	-	occupied	2	35	15	1575465352	10	0
15	taxi_1	45.5012	-73.5684	tablet	-	occupied	2	0	298	1575465362	10	0

Figure 4-10 Détail du scénario 2 de regroupement des points GPS consécutifs séparés d'un intervalle de 60 à 600 secondes : ajout de l'attribut temporaire « repère » permettant d'identifier les intervalles de temps problématiques

Cet attribut « repère » permet de segmenter la partition en plusieurs groupes. La Figure 4-11 présente les différentes étapes permettant d'identifier les groupes qui devraient le plus correspondre à des courses. Dans un premier temps, la partition de points est donc segmentée selon la valeur de l'attribut repère. Dès que la valeur « 1 » est rencontrée, on considère que le point GPS correspondant appartient à un autre groupe. Les points consécutifs à ce point appartiennent à ce même groupe tant que la valeur de l'attribut repère est de « 0 ». Dans la Figure 4-11, trois groupes de points sont ainsi identifiés au sein de la même partition. Il s'agit désormais d'identifier si ces groupes de points correspondent effectivement à des courses distinctes.

- Soit d_i la distance correspondant à la distance réseau parcourue entre le premier point et le dernier point du groupe i.
- Soit t_i la durée correspondant au temps écoulé entre le premier point et le dernier point du groupe i.

- Soit [d_{min} ; d_{max}] et [t_{min} ; t_{max}] les valeurs seuils identifiées à partir des distributions des courses complètes (section 4.3.1).

Un test est effectué sur les valeurs d_i et t_i des différents groupes en les comparant aux valeurs seuils. Les deux premiers groupes sont d'abord testés :

- Si la distance d_1 et la distance d_2 sont toutes deux supérieures à d_{min}, cela signifie que les deux groupes présentent une distance qui peut correspondre à celle d'une course.
- De plus si les durées t_1 et t_2 sont également toutes deux supérieures à t_{min}, alors les groupes présentent aussi une durée qui pourrait correspondre à celle d'une course.
- Si ces deux conditions sont réunies alors les deux groupes peuvent tout deux correspondre à des courses potentielles et ils resteront donc deux groupes distincts.

Si ces deux conditions ne sont pas réunies mais que la somme des distances d_1 et d_2 est supérieure à d_{max} et que la somme des durées est supérieure à t_{max}, alors cela signifie que si on considère de regrouper les deux groupes en un seul, la distance totale et la durée totale du groupe résultant seraient trop élevées pour une course. On ne peut donc pas regrouper les deux groupes et ils doivent de ce fait rester distincts.

Si aucune de ces conditions n'est réunie, alors les deux groupes sont regroupés en un seul.

Le résultat de ce premier test est alors comparé au groupe suivant, soit le groupe 3 :

- Si le résultat du test est que les groupes 1 et 2 sont deux groupes distincts alors le test suivant est effectué entre les groupes 2 et 3. Le groupe 3 étant consécutif dans le temps au groupe 2, il ne peut donc être éventuellement regroupé qu'au groupe 2 et non au groupe 1.
- Si le résultat du premier test indique que les groupes 1 et 2 appartiennent au même groupe, alors le test suivant est réalisé entre le groupe résultant (groupe 1 $\cup$ 2) et le groupe 3.

Ainsi, selon les résultats des différents tests, la partition de points peut être segmentée en différents groupes ou rester telle quelle. On détermine donc les éventuelles courses en se basant sur les caractéristiques de durée et de distance d'une population de référence de courses.

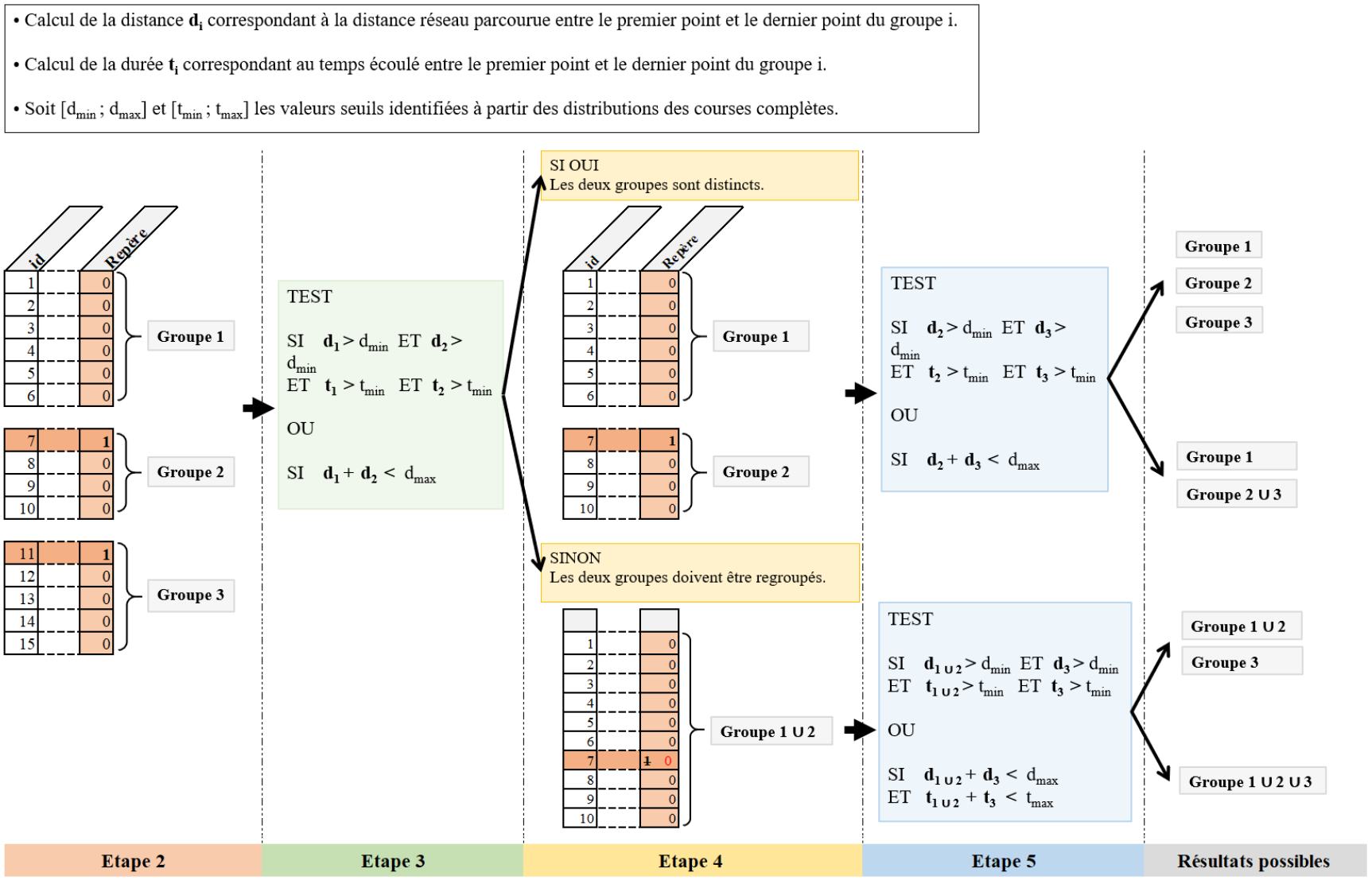


Figure 4-11 Détail du scénario 2 de regroupement des points GPS consécutifs séparés d'un intervalle de 60 à 600 secondes : les différentes étapes pour identifier les courses

La dernière hypothèse retenue pour le regroupement des points consécutifs de même statut *occupied* est donc celle du scénario n°2 :

- **Hypothèse 3** : on suppose que 15 secondes ne suffisent pas pour qu'un client débarque et qu'un autre embarque (Dandl et al., 2017). Par conséquent, toute course dont les points sont séparés par 15 secondes au maximum est considéré comme une course "complète", c'est-à-dire sans perte de données et sans changement potentiel de client. Ces courses "complètes" constituent la population de référence. À partir de cette population de référence, une durée moyenne de course et une distance moyenne de course sont déterminées selon les règles statistiques de la distribution log-normale détaillées précédemment à la section 4.3.1. Ces deux caractéristiques sont ensuite utilisées comme critères pour déterminer les trajets parmi les groupes problématiques.

La Figure 4-12 présente le processus retenu pour le regroupement des points GPS consécutifs de même statut *occupied* en courses.

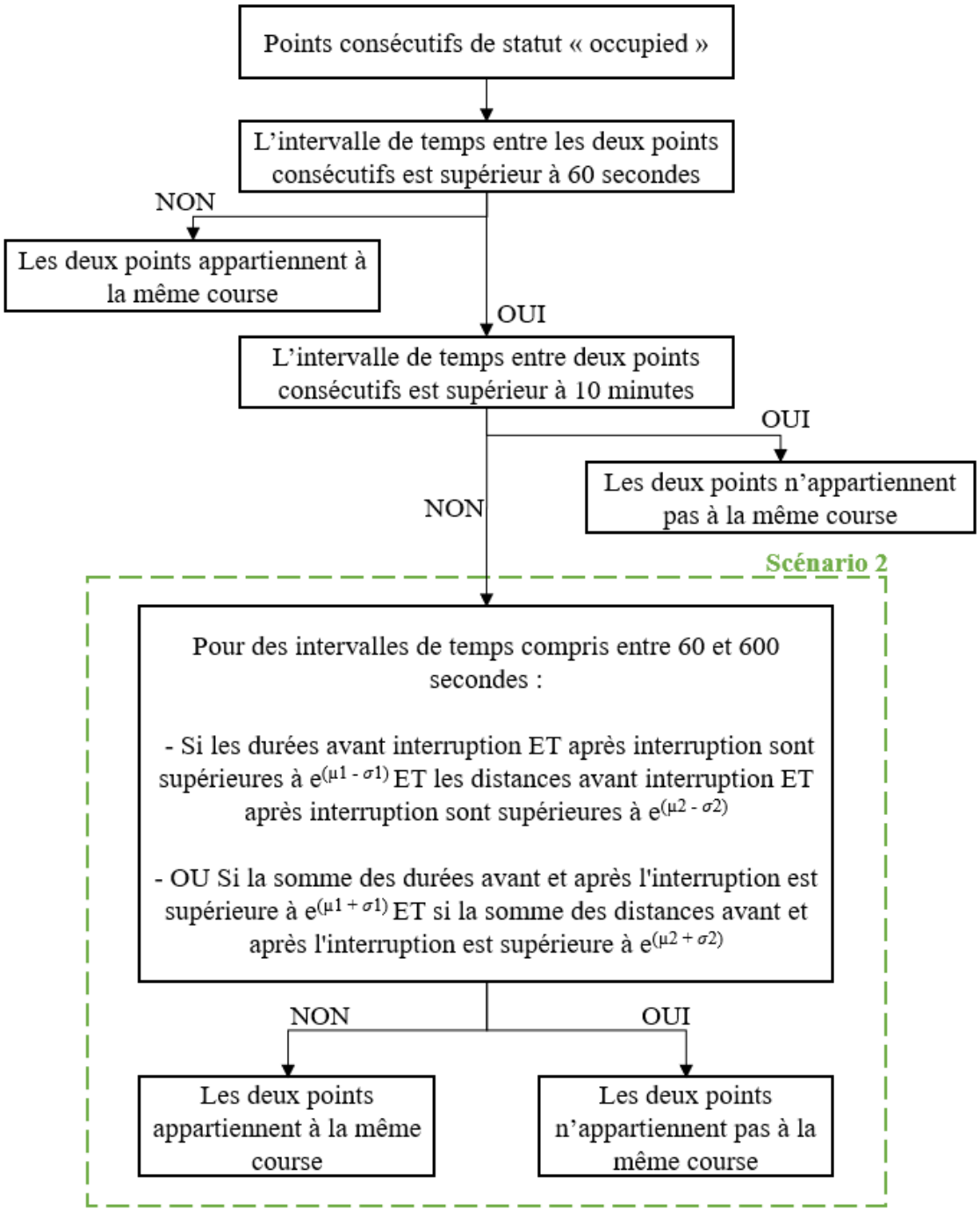


Figure 4-12 Schématisation du processus choisi pour le regroupement des points GPS consécutifs de même statut *occupied*

4.4 Règles de validation des courses

Le processus de regroupement détaillé à la section précédente a permis d'identifier les éventuelles courses de taxi. Cependant les groupes identifiés ne correspondent pas toujours à des courses. En effet, certains groupes présentent des caractéristiques anormales telles que des durées de course de quelques secondes ou des vitesses nulles durant l'ensemble de la course.

Par exemple, des regroupements de points de statut *occupied* de moins de 30 secondes peuvent être observés. Ces groupes ne peuvent pas correspondre à une course et sont probablement liés à une erreur de statut. Il convient en effet de rappeler que les changements de statuts relèvent d'une manipulation manuelle du dispositif présent dans le véhicule par le chauffeur de taxi. Un chauffeur peut donc oublier de changer de statut ou indiquer le mauvais statut. Il peut également indiquer qu'il est en course puis se remettre disponible si le client décide finalement qu'il ne souhaite plus réaliser la course. De la même manière, Zheng et al. (2014) indiquent que des informations peuvent être mal enregistrées si les chauffeurs de taxi utilisent mal leur appareil ou si une interruption du signal survient lorsque le statut de leur taxi change. Il en résulte des trajets extrêmement courts ou longs (Zheng et al., 2014). De nombreuses raisons peuvent donc expliquer ces courses erronées et l'on ne dispose pas de suffisamment d'informations pour toutes les identifier.

Des règles de validation des courses doivent donc être établies. L'enjeu est complexe puisqu'il s'agit d'identifier les vraies courses des fausses. Cependant une durée minimale de validation trop élevée peut certes éliminer une grande partie des courses erronées, mais également éliminer de réelles courses courtes. L'enjeu est d'autant plus complexe qu'il n'existe que très peu d'informations à ce sujet dans la littérature (Laviolette, 2017; Zheng et al., 2014).

Lacombe (2016) considère uniquement que les courses dont la durée ou la distance sont inférieures ou égales à 0 sont erronées. Elle met cependant en évidence les limites de cette méthode en observant que pour un mois de données de l'intermédiaire Taxi Diamond, 9% des courses ont une durée inférieure à 1 minute et que 7% des courses présentent une distance parcourue inférieure à 100 mètres. Elle suggère donc de se baser sur les distances et durées de course pour éliminer les courses non valides (Lacombe, 2016).

Laviolette (2017) combine quant à lui plusieurs critères provenant de l'analyse des distributions des distances, durées et vitesses de course afin de déterminer des fourchettes de validité ainsi que

des valeurs seuils, tels qu'un critère de vitesse maximale correspondant à la limite de vitesse autorisée sur les autoroutes au Québec (Laviolette, 2017).

La méthode de validation suivie dans le présent projet sera similaire à celle proposée par Laviolette (2017). Des valeurs seuils ainsi que des critères provenant des distributions des durées, distances et vitesses moyennes de courses du mois d'avril 2019 seront utilisés pour l'identification et la validation des courses de taxi. Selon le BTM tous les taxis opérant sur l'Île de Montréal ont été connectés au Registre avant le 30 mars 2019. Ainsi, à partir du 1er avril 2019 l'ensemble des taxis peut donc être potentiellement observé et analysé.

4.4.1 Suppression des groupes d'une seule observation

La première étape consiste à supprimer les groupes constitués d'une seule observation, soit d'un seul point GPS. Ces derniers représentent environ 2% des groupes. On peut se demander si ces groupes ne sont pas dus à un mauvais regroupement des points en groupes de points consécutifs de même statut. Or parmi ces groupes, environ 91% sont précédés et suivi d'un groupe de point de statut différent de *occupied*. Dans 91% des cas ce n'est donc pas lié à un mauvais regroupement mais bien à une mauvaise manipulation par le chauffeur du dispositif à bord ou d'une défectuosité de ce dernier. Enfin, dans plus de 60% des cas, cette unique observation de statut *occupied* est précédée et suivie de points de statut *free*.

On s'intéresse aux cas où le point seul est précédé et/ou suivi d'un groupe de points de statut *occupied* (9% des cas). La Figure 4-3 illustre la distribution des écarts de temps entre le point seul de statut occupied et le groupe précédent ou suivant dans le cas où le groupe précédent et/ou suivant est aussi de statut occupied.

Dans les cas où l'unique observation de statut *occupied* est précédée et/ou suivie de groupes de statut *occupied* :

- dans 100% des cas l'intervalle de temps entre le dernier point du groupe précédent (de statut *occupied*) et l'unique observation est supérieur à 10 minutes ;
- dans 100% des cas l'intervalle de temps entre l'unique observation et le premier point du groupe suivant (de statut *occupied*) est supérieur à 10 minutes.

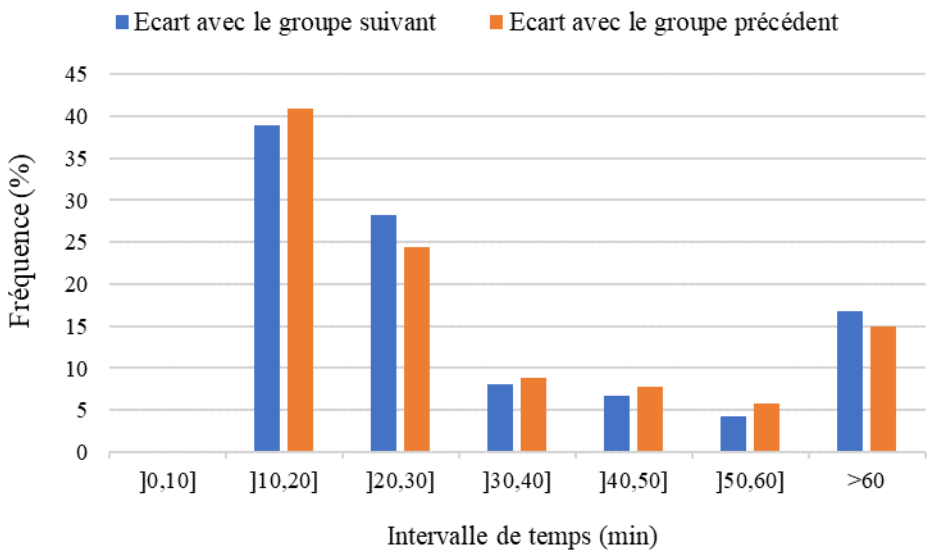


Figure 4-13 Histogramme de distribution des écarts de temps entre le point seul et le groupe précédent ou suivant dans le cas où le groupe précédent et/ou suivant est aussi de statut *occupied*

Cela signifie que dans le cas d'un écart de temps entre deux points consécutifs de statut *occupied* supérieur à 10 minutes, on considère que les deux points n'appartiennent pas à la même course. Cela correspond à l'hypothèse 2 émise dans le cadre de la méthode d'identification des courses détaillée à la section 4.3. Si des intervalles inférieurs à 10 minutes avaient été observés, la méthodologie de regroupement aurait pu être remise en cause.

Les points seuls sont donc supprimés.

4.4.2 Critères de durée

Pour la suite, l'analyse des distributions des durées, distances et vitesses moyennes est réalisée afin de déterminer des critères de validation. Le premier critère concerne la durée minimale pour qu'une course soit considérée comme valide. La Figure 4-14 présente la distribution des durées des courses par intervalle de 30 secondes pour le mois d'avril 2019.

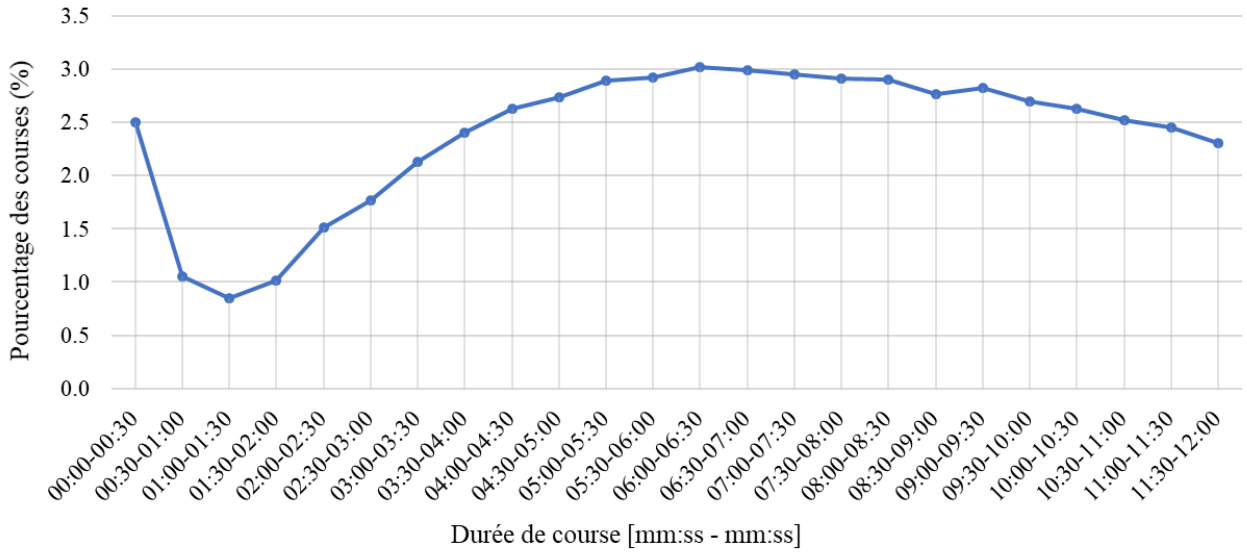


Figure 4-14 Distribution des durées de course par intervalle de trente seconde pour le mois d'avril 2019

Toutefois, jusqu'à quelle durée cette hypothèse peut-elle être étendue sans risquer d'éliminer des courses valides ? Le point d'inflexion de ces trois courbes est situé entre 01:00 et 01:30 et peut être un bon indicateur permettant de choisir la durée seuil. La durée minimale seuil est donc établie à 01:30.

Une proportion de courses de durée inférieure à trente secondes peut être observée, témoignant de mauvaises manipulations par le chauffeur ou de dysfonctionnements du dispositif à bord. Il est en effet très peu probable qu'une course dure moins de trente secondes. Cependant comment déterminer la durée minimale de course afin de ne pas éliminer de vraies courses de courte durée ? Le point d'inflexion de la courbe de distribution des durées, situé entre une minute (01 :00) et deux minutes (02 :00) est considéré comme un bon indicateur pour le choix de cette durée minimale (Laviolette, 2017). Cette dernière est donc établie à une minute trente (01 :30). C'est également la durée minimale identifiée par Laviolette (2017) lorsqu'il analyse la distribution des durées des courses de 2016 provenant de trois intermédiaires : Hochelaga, Coop de l'Ouest et Taxi Diamond (Laviolette, 2017).

Les groupes de course dont la durée est inférieure à une minute trente secondes représentent environ 4.4% des courses (pourcentage calculé une fois les observations seules supprimées).

La durée maximale de course est quant à elle établie à 180 minutes, soient 3 heures. Cela correspond à la durée pour rejoindre la ville de Québec. Les courses supérieures à 180 minutes représentent 0.38% des courses.

4.4.3 Critères de distance

Les critères suivants concernent la distance parcourue. La Figure 4-15 présente la distribution des distances de courses par plage de 100 mètres pour le mois d'avril 2019.

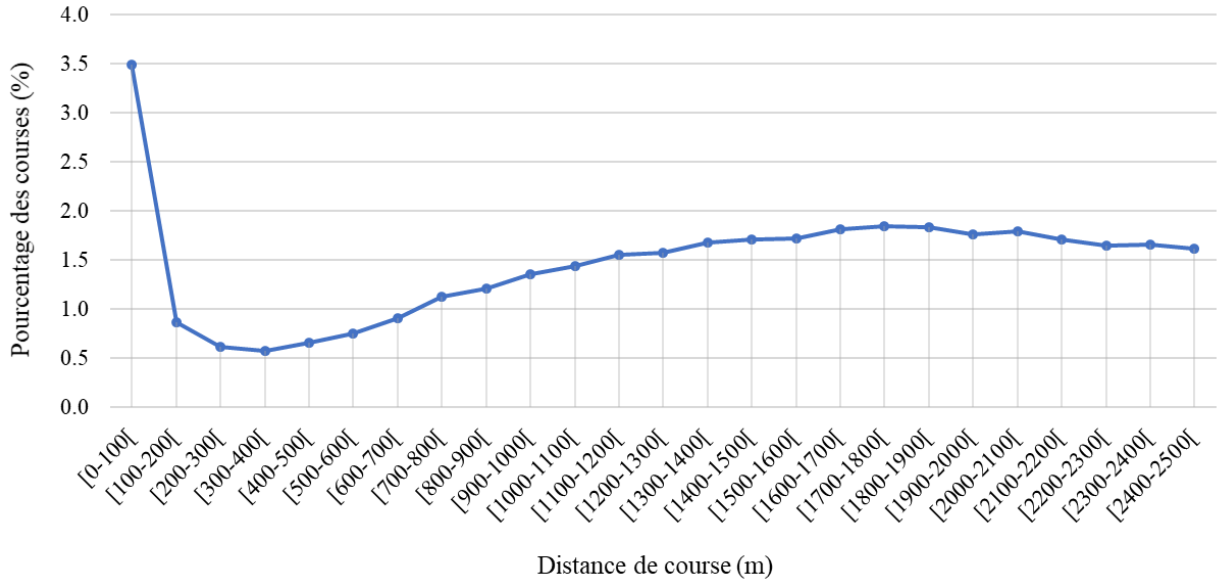


Figure 4-15 Distribution des distances de course par plage de 100 mètres pour le mois d'avril 2019

De manière analogue à la durée, une proportion de courses de distance inférieure à 100 mètres peut être observée. Le point d'inflexion est situé entre 300 et 500 mètres. Le critère de distance minimale de course est donc établi à 400 mètres. Cela correspond à la distance minimale identifiée par Laviolette (2017) lors de son analyse de la distribution des distances des courses de 2016 provenant de trois intermédiaires : Hochelaga, Coop de l'Ouest et Taxi Diamond (Laviolette, 2017).

Les courses de distances inférieures à 400 mètres représentent 5.5% des groupes (pourcentage effectué une fois les observations seules supprimées).

La distance maximale de course est quant à elle établie à 250 kilomètres. Cela correspond à la distance pour rejoindre la ville de Québec. Les courses supérieures à 250 kilomètres représentent 0.01% des courses.

4.4.4 Critères de vitesse moyenne

Enfin, il est nécessaire de valider la correspondance entre la durée d'une course et la distance parcourue. L'estimation de la vitesse moyenne (rapport entre la distance parcourue et la durée de course) permet de mettre en évidence les cas extrêmes correspondant probablement à des courses non valides. La Figure 4-16 présente la distribution des vitesses moyennes de course pour le mois d'avril 2019. Cette distribution est effectuée après nettoyage par les valeurs limites de durée et de distance établies précédemment.

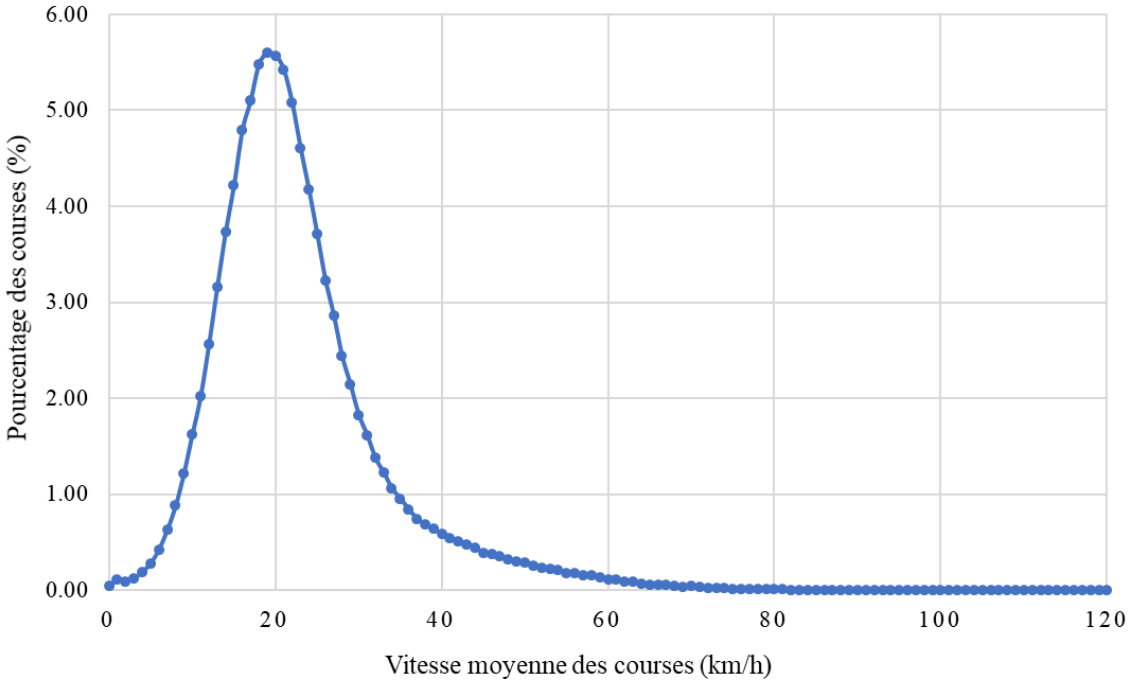


Figure 4-16 Distribution des vitesses moyennes de course pour le mois d'avril 2019

Les courses inférieures à 4 km/h représentent 0.51% des courses. Cette valeur est donc choisie pour le critère de vitesse minimale. La vitesse maximale tolérée sur les autoroutes au Québec étant de

120 km/h (Laviolette, 2017), cette valeur est choisie comme critère de vitesse maximale. Les courses de vitesse supérieure à 120 km/h représentent 0.07% des courses.

Le Tableau 4-2 résume le résultat de ce processus de nettoyage et validation des courses. Le nombre de courses restantes après chaque étape de nettoyage selon chacun des critères mentionnés précédemment y est indiqué. Les courses valides représentent environ 91% des courses brutes et environ 93% des groupes de plus d'une observation.

Tableau 4-2 Règles de validation des courses appliquées aux courses du mois d'avril 2019

Étape	Règles de validation	Nombres de lignes	% par rapport à l'étape 0	% par rapport à l'étape 1
0	Brutes	487780	100.0%	-
1	Nombre d'observations > 1	476874	97.8%	100.0%
2	1,5 min <= durée	455919	93.5%	95.6%
3	durée <= 180 minutes	454086	93.1%	95.2%
4	400 m <= distance	445006	91.2%	93.3%
5	distance <= 250 km	445004	91.2%	93.3%
6	4 km/h <= vitesse moyenne	442735	90.8%	92.8%
7	vitesse moyenne <= 120 km/h	442411	90.7%	92.8%
	COURSES VALIDES avril 2019	**442411**	**90.7%**	**92.8%**

4.5 Règles de validation des autres statuts

Si déterminer les courses est fondamental pour l'analyse des activités de taxis, il est également essentiel de pouvoir identifier les périodes où le taxi n'est pas en course mais est tout de même en service. Par exemple afin de pouvoir calculer l'efficacité d'un taxi, le temps total passé en course (ou la distance totale parcourue) doit être comparé au temps total passé en service (ou distance totale parcourue en service).

Il est donc essentiel de pouvoir distinguer les périodes où le taxi est en service mais est sans client des périodes où le taxi n'est pas en activité. Dès lors qu'un taxi entre en activité, il est tenu d'envoyer ses données au Registre. Ainsi toute donnée envoyée doit théoriquement correspondre aux données du taxi en service. Cependant le même problème soulevé dans les parties précédentes sur l'identification des courses s'applique ici : qu'en est-il des périodes sans données GPS entre

deux points consécutifs de même statut? Est-ce une perte de données ou une déconnexion volontaire du taxi qui serait par exemple en pause ?

Les distributions des intervalles de temps entre deux points consécutifs de même statut sont étudiées pour les trois autres statuts : *free*, *oncoming* et *unavailable*.

4.5.1 Statuts *free, oncoming et unavailable*

La Figure 4-17 présente les distributions cumulées des fréquences des intervalles de temps entre deux points consécutifs de même statut et ce, pour les statuts *free, oncoming* et *unavailable*.

Pour les trois statuts, la Figure 4-17 (a) met en évidence que plus de 90% des points sont séparés d'un intervalle de temps inférieur à 15 secondes, tout opérateur confondu. La Figure 4-17 (b) présente un agrandissement de la Figure 4-17 (a). On y observe que, pour les trois statuts, plus de 99.5% des points sont séparés d'un intervalle inférieur à 60 secondes.

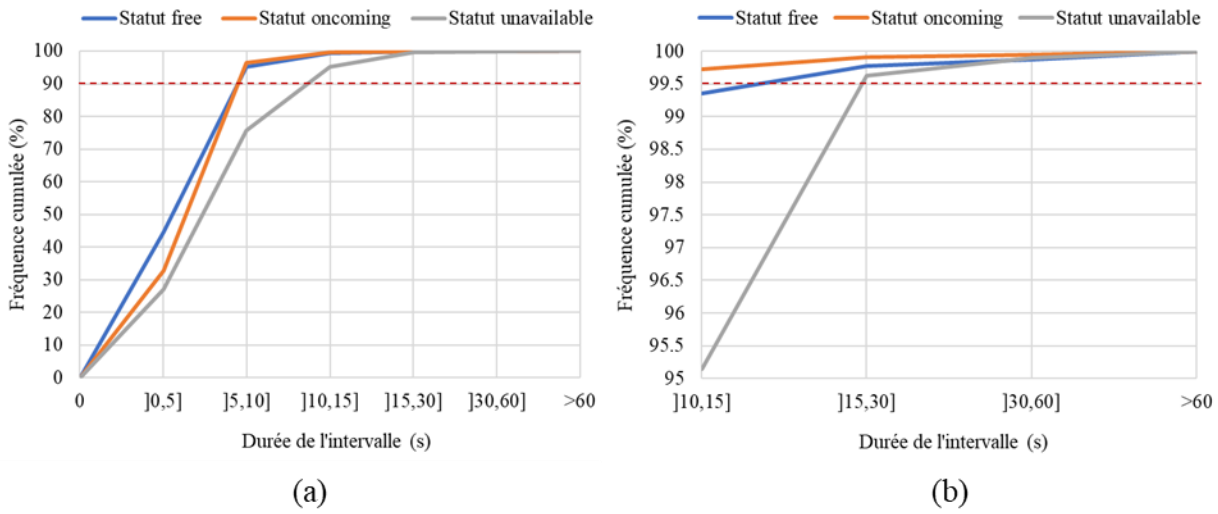


Figure 4-17 Distribution cumulée des intervalles de temps entre deux points consécutifs de même statut (a) pour les statuts free, oncoming et unavailable (b) agrandissement

On considère donc qu'en dessous de 60 secondes d'interruption de données les deux points de statut *free* appartiennent au même groupe de points et qu'il n'y a pas eu d'interruption volontaire du chauffeur entre les deux. Au-delà, on considère que les deux points appartiennent à deux groupes différents. Il est en de même pour les points de statut *oncoming.*

De manière analogue à la détermination du critère de temps pour les statuts *free* et *oncoming*, on pourrait donc considérer 60 secondes comme critère.

Cependant le statut « unavailable » présente un enjeu supplémentaire puisque ce statut peut indiquer que le taxi est en train d'effectuer une course particulière ou un autre type d'activité. En effet, comme mentionné dans la section 3.1.3, lorsque le taxi est *unavailable*, il peut soit être en opération pour un service de taxi collectif pour une société de transport en commun, soit fournir des services de taxi adapté, ou encore effectuer des contrats corporatifs ou hospitaliers. Toutefois, les informations actuellement disponibles dans le Registre ne permettent pas de différencier et d'identifier quel type de course ou activité a été réalisé. Ainsi, seule une partie des trajets effectués par les taxis peut être identifiée avec les données disponibles dans le Registre. Seuls les trajets réguliers, soit lorsque le statut indique *occupied*, ont pour l'instant pu être étudiés (sections 4.1 à 4.4).

Le statut *unavailable* pouvant correspondre à une activité de type course, les mêmes règles de validations que celles établies pour les courses régulières pourraient être déterminées.

Il faudrait pour cela d'abord analyser les groupes de statut *unavailable* afin d'identifier si plusieurs types de courses peuvent être mis en évidence. Les courses de transport adapté ne présentent peut-être pas les mêmes caractéristiques que les courses corporatives. Si différents types sont effectivement mis en évidence, les caractéristiques de durée et distance moyenne pour chaque type doivent être déterminées à partir d'une population de référence à définir. En raison de ces nombreuses incertitudes sur le statut *unavailable*, dans un premier temps, il a été établi que seul un critère de temps serait appliqué aux groupes de statut *unavailable* et qu'il serait préférable d'attendre de disposer des autres flux de courses (transport adapté, taxi collectif, etc.) pour pouvoir établir des règles de validation.

On considère donc qu'en dessous de 60 secondes d'interruption de données les deux points de statut *unavailable* appartiennent au même groupe de points et qu'il n'y a pas eu d'interruption volontaire du chauffeur entre les deux. Au-delà, on considère que les deux points appartiennent à deux groupes différents.

4.5.2 Entre statuts

Les sections précédentes ont permis d'établir les règles de regroupement des points consécutifs de même statut. Cependant il est aussi nécessaire d'en établir pour les intervalles de temps entre deux statuts différents. En effet, l'intervalle de temps entre deux points consécutifs de statuts différents présente également un enjeu. Quel est le statut du taxi pendant cet intervalle de temps ? Est-il le même que le statut précédent ? Ou alors, lorsque cet intervalle est élevé, est-ce que le taxi n'était plus en service entre ces deux groupes de statut ?

La Figure 4-18 illustre cet enjeu ainsi que les trois solutions qui peuvent être envisagées.

La première est de considérer que l'intervalle de temps T_i entre les deux points n'appartient ni au groupe de points le précédant, ni au groupe de points le succédant. On attribuerait donc à cet intervalle un statut « indéterminé », témoignant de l'incertitude sur le fait que le taxi soit en activité ou non.

La seconde méthode proposée serait de séparer cet intervalle de temps en deux période égales. La première période serait rattachée au groupe précédent et la seconde au groupe suivant. Dans ce cas il serait également nécessaire d'effectuer une interpolation entre les points afin d'estimer la demi-distance parcourue. On ne peut en effet rajouter une période de temps sans y ajouter la distance parcourue correspondante.

La troisième méthode consiste à attribuer cet intervalle de temps au groupe le précédant. On considérait dans ce cas que tant qu'un nouveau statut n'a pas été reçu, tout ce qui le précède est du même statut que le précédent. Cependant si le taxi se déconnecte entre le changement de statut, on considèrerait donc que la période hors connexion est une période en activité. Il est donc à nouveau nécessaire d'établir des règles de validation afin de distinguer une déconnexion volontaire du taxi d'une perte de données.

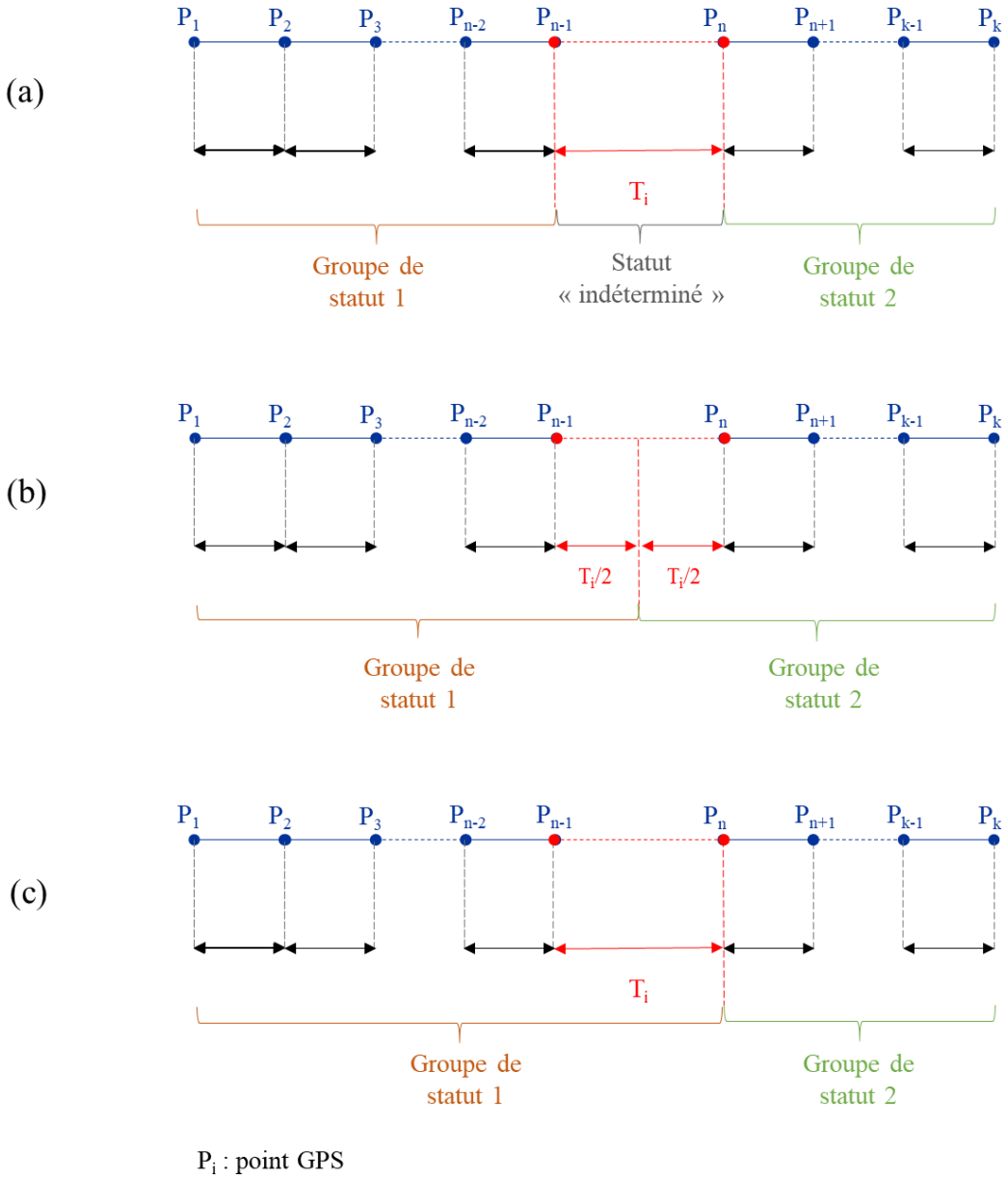


Figure 4-18 Illustration de l'enjeu de l'écart de temps T_i entre deux points consécutifs de statuts différents : (a) on attribue à cet intervalle de temps T_i un statut « indéterminé » (b) l'intervalle de temps T_i est séparé en deux périodes égales (c) l' intervalle de temps T_i est rattaché au groupe le précédant

Les distributions des intervalles de temps entre deux points consécutifs de statuts différents sont réalisées. Comme l'illustre le Tableau 4-3, 12 combinaisons sont possibles puisque 4 statuts existent.

Tableau 4-3 Combinaisons possibles de statut entre deux points consécutifs de différents statuts

Statut du point précédent	Statut du point suivant
free	occupied
	oncoming
	unavailable
occupied	free
	oncoming
	unavailable
oncoming	free
	occupied
	unavailable
unavailable	free
	occupied
	oncoming

Les distributions de ces douze combinaisons sont présentées dans la Figure 4-19. On peut y observer que pour chacune des douze combinaisons plus de 90% des intervalles de temps entre deux points consécutifs de statuts différents sont inférieurs à 60 secondes.

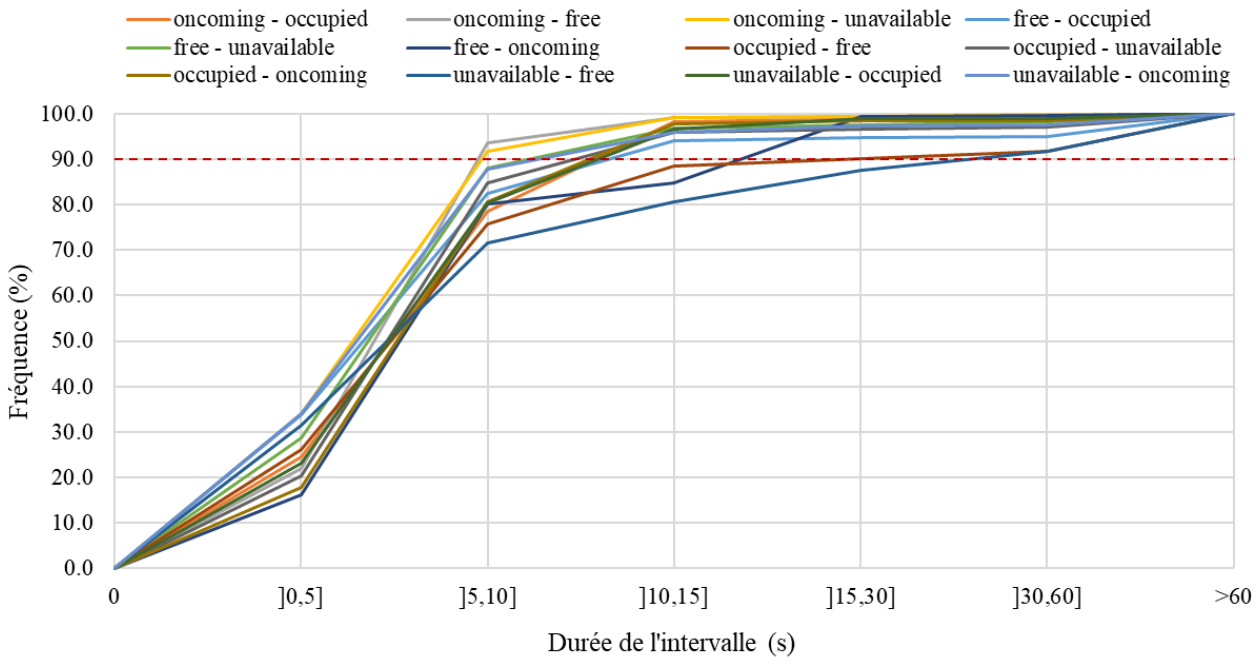


Figure 4-19 Distribution cumulée des intervalles de temps entre deux points consécutifs de statuts différents

On considère donc que si l'intervalle de temps séparant deux points consécutifs de statuts différents est inférieur à 60 secondes alors la moitié de cet intervalle est rattachée au groupe de points précédent et l'autre moitié au groupe suivant (Figure 4-18 (b)). Cette méthode est choisie afin de répartir l'éventuelle erreur (d'au plus 30 secondes) aux deux groupes et non uniquement à un seul groupe. Dans le cas où il est supérieur à 60 secondes, on considère alors que le taxi n'est pas en activité pendant cet intervalle.

La Figure 4-20 présente un schéma récapitulatif des différents critères de regroupement de deux points consécutifs présentés dans ce chapitre.

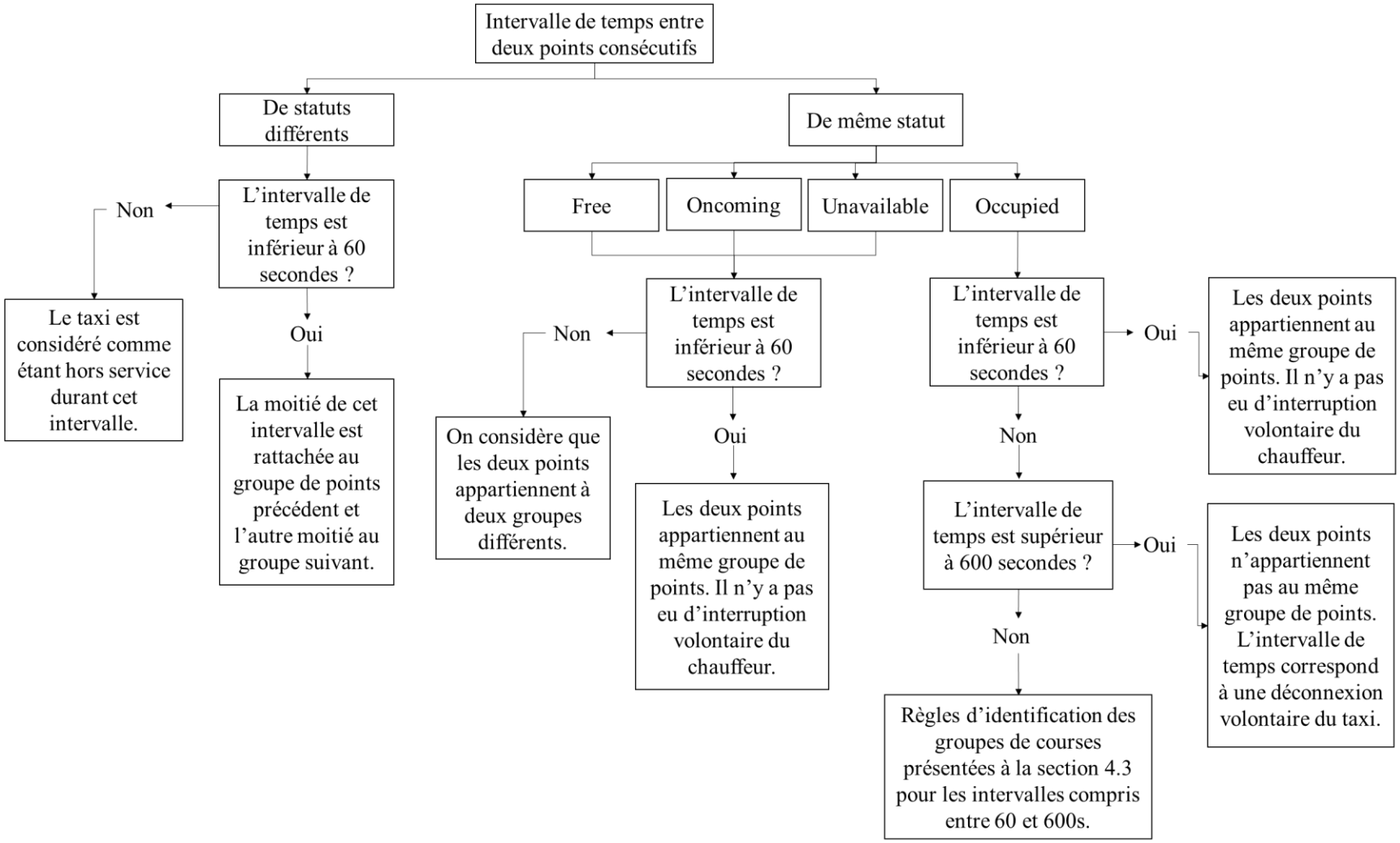


Figure 4-20 Schéma récapitulatif des critères de regroupement de deux points consécutifs

CHAPITRE 5 INDICATEURS

Rappelons-le, l'objectif de cette étude est le développement d'un tableau de bord permettant de visualiser et d'analyser des indicateurs concernant l'offre et la demande de déplacements en taxi.

Les Chapitre 3 et Chapitre 4 ont présenté les données disponibles et les méthodes de traitement de ces dernières dans le but de calculer les indicateurs. Le présent chapitre présente ces indicateurs.

Tel que mentionné à la section 2.1.3.6 de la revue de littérature, le projet de recherche fait suite aux travaux de Lacombe (2016) et Laviolette (2017). Ces derniers ont mis en évidence les principaux indicateurs qu'il serait pertinent de calculer.

Dans un premier temps les principaux indicateurs retenus ainsi qu'une classification de ces derniers sont détaillées. Les échelles spatiales et temporelles retenues pour l'analyse sont également présentées. Puis les défis méthodologiques liés au calcul de ces indicateurs sont mis en évidence.

5.1 Segmentation des indicateurs

5.1.1 Principaux indicateurs

Dans un premier temps il est essentiel d'identifier les principaux indicateurs qu'il peut être pertinent de calculer pour le tableau de bord. Cette identification est en partie présentée à la section 2.1.3.6 dans le Tableau 2-2 présentant les indicateurs d'offre et de demande en déplacements de taxi identifiés dans la littérature. Lacombe (2016) et Laviolette (2017) classifient ces indicateurs selon plusieurs objets d'analyse, à savoir la course, le véhicule de taxi, le chauffeur et le poste d'attente. Les indicateurs sont par la suite catégorisés selon qu'ils sont des indicateurs d'offre ou de demande.

Dans le présent projet, il est suggéré de catégoriser uniquement selon l'objet d'analyse. Le Tableau 5-1 présente les six objets d'analyse retenus et les principaux indicateurs qui y sont associés. Les six objets d'analyse sont la course, le véhicule de taxi, le chauffeur, l'intermédiaire en service, le poste d'attente et les origines et destinations des courses. L'objet course étant le plus étudié, les origines et destinations des courses sont étudiées à part afin de ne pas surcharger la visualisation des indicateurs de l'objet course.

1. **Course**

Les principaux indicateurs associés à l'objet course permettent de décrire la demande en déplacements de taxi en présentant le nombre de courses et les caractéristiques moyennes des courses (distance, durée, vitesse).

2. **Véhicule**

L'étude du véhicule de taxi permet de décrire l'offre en présentant notamment le nombre de véhicules actifs ainsi que leur durée de service. Aussi, l'utilisation et la productivité de ces véhicules peut être mise en évidence grâce à l'analyse des durées et distances parcourues à vide.

3. **Chauffeur**

L'étude des chauffeurs de taxi permet d'analyser leur comportement et leur productivité. Les chauffeurs ne sont pas analysés individuellement (mêmes si les données du Registre le permettent). Ce sont uniquement les caractéristiques d'un chauffeur moyen qui sont calculées. Des indicateurs sur le nombre de chauffeurs actifs, le nombre de courses réalisées par chauffeur ou encore la durée des quarts de travail sont ainsi calculés.

4. **Intermédiaires**

Les données du Registre permettent d'avoir accès aux données de l'ensemble des intermédiaires en service. Il est donc intéressant d'analyser les opérations des différents intermédiaires opérant sur l'île de Montréal. Le nombre d'intermédiaires actifs, le nombre de taxis affiliés à ces intermédiaires ou encore la répartition des courses peuvent être identifiés.

5. **Poste d'attente**

Les postes d'attente de taxis sont gérés par les différents arrondissements de Montréal et sont localisés de sorte à assurer un service rapide et efficient aux clients (Ville de Montréal, 2020c). Il est donc intéressant de pouvoir vérifier si la localisation des postes est pertinente en analysant leur taux d'utilisation.

6. **Origines et Destinations**

L'analyse de cet objet permet de mettre en évidence la répartition (spatiale et temporelle) de la demande et de l'offre sur le territoire de l'île de Montréal.

Tableau 5-1 Liste des principaux indicateurs

Course	Nombre de courses Distance de course Durée de course Vitesse de course
Véhicule	Nombre de véhicules actifs Durée de service Distance parcourue à vide / durée à vide Distance parcourue en course / durée en course Taux de disponibilité Nombre de véhicules-heures productifs Heures de disponibilité par zone
Chauffeur	Nombre de chauffeurs actifs Nombre de courses par chauffeur Durée des quarts de travail
Intermédiaire	Nombre d'intermédiaires actifs Nombre de taxis affiliés aux intermédiaires
Poste d'attente	Part des courses avec extrémité aux postes d'attente Taux d'utilisation des postes d'attente
Origines - Destinations	Densité des Origines et Destinations

5.1.2 Formes d'analyse

Si les principaux indicateurs qu'il serait pertinent de représenter ont pu être mis en évidence, il est également nécessaire d'identifier et de définir la forme que prendra l'indicateur. La Figure 5-1 illustre les quatre formes possibles pour l'analyse des indicateurs mises en évidence par Lacombe (2016).

La première est la statistique descriptive (Figure 5-1 (a)). C'est la forme de base d'un indicateur, soit une valeur totale ou une valeur moyenne. Par exemple le nombre de courses effectuées durant le mois d'avril 2019 est une statistique descriptive. Le coefficient de variation ou l'écart-type peuvent être renseignés afin d'apporter une précision sur la dispersion des données.

La deuxième forme (Figure 5-1 (b)) illustre la distribution d'une statistique dans la population à l'étude. Par exemple, un histogramme de distribution fréquentielle peut être choisi pour représenter la distribution du nombre de courses effectuées durant le mois d'avril 2019 dans la population des véhicules de taxi, avec en abscisse le nombre de courses et en ordonnée le nombre de véhicules (ou proportion des véhicules).

La troisième forme (Figure 5-1 (c)) permet d'illustrer la répartition dans le temps d'une statistique descriptive. Par exemple, le nombre de courses effectuées par véhicule de taxi pourrait être représenté pour plusieurs mois de l'année. Cela permet notamment de mettre en évidence d'éventuelles variations selon les mois ou saisons.

Enfin, la quatrième forme (Figure 5-1 (d)) concerne la répartition spatiale des données. La répartition des origines ou des destinations des courses peut par exemple être mise en évidence sur une carte du territoire à l'étude.

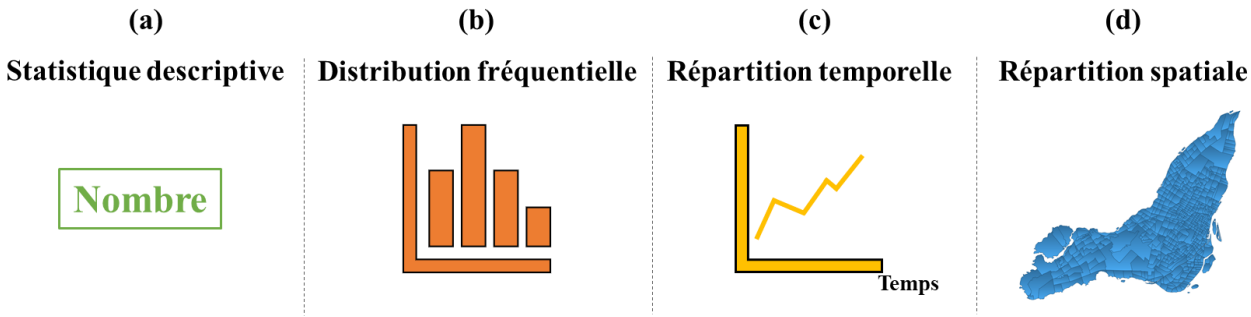


Figure 5-1 Quatre formes d'analyse des indicateurs identifiées par Lacombe (2016) : (a) statistique descriptive ; (b) distribution fréquentielle ; (c) répartition temporelle ; (d) répartition spatiale

5.1.3 Déclinaisons

Les indicateurs peuvent donc être analysés sous plusieurs formes. La Figure 5-2 présente une schématisation de la classification d'un indicateur particulier. Les différentes déclinaisons possibles pour le nombre total de courses du mois y sont présentées. Ainsi, en plus de pouvoir être analysés sous les quatre formes mentionnées précédemment, un indicateur principal peut être décliné en plusieurs indicateurs secondaires. Trois catégories principales ont été identifiées. Ainsi, il est possible de classifier :

- **Spatialement** (ex : une valeur pour chaque secteur de recensement) ;
- **Temporellement** (ex : par jour, par type de jour, jours ouvrables / non ouvrables, ouvrés, plage horaire) ;
- **Par "objet"** (ex : par véhicule, par chauffeur).

Il est également possible de combiner ces classifications. Ainsi, chaque indicateur principal peut être décliné en un grand nombre d'indicateurs secondaires, tel qu'illustré ici pour le nombre de courses du mois.

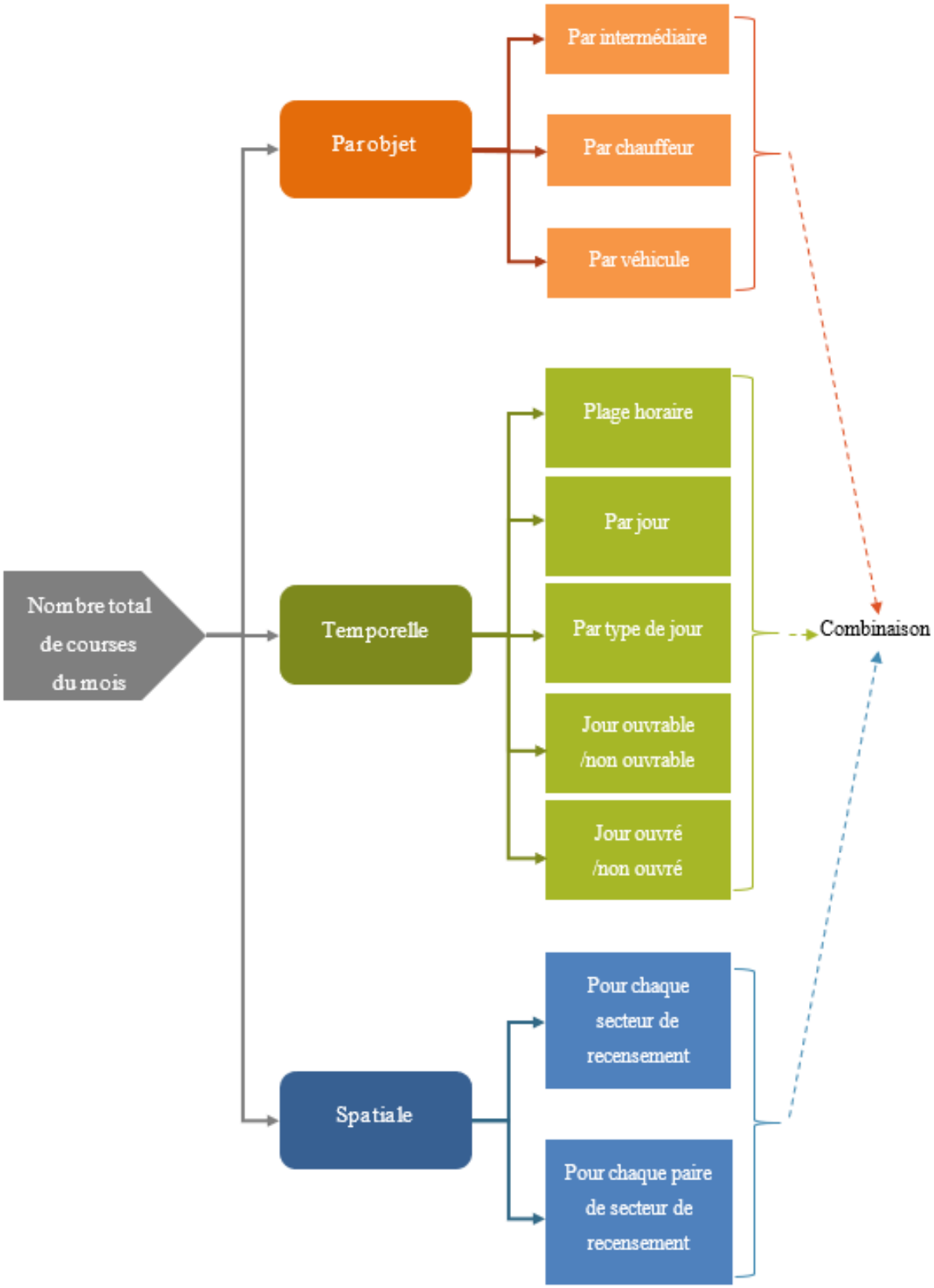


Figure 5-2 Illustration des déclinaisons possibles pour le nombre total de courses du mois

Les tableaux Tableau 5-2 à Tableau 5-4 présentent des exemples de déclinaisons possibles pour trois des indicateurs principaux. Le nombre total de courses par mois, le nombre total de véhicules actifs par mois et le total des heures en service sont ainsi déclinés par objet, spatialement et temporellement, donnant lieu à de nombreux indicateurs secondaires qui pourraient être calculés et présentés sur le tableau de bord.

Tableau 5-2 Exemples de déclinaisons possibles pour le nombre total de courses par mois

Indicateur principal	**Indicateurs secondaires**	**Déclinaison**
Nombre total de courses par mois	Nombre moyen de courses par chauffeur	Par objet
	Nombre moyen de courses par véhicule	Par objet
	Nombre moyen de courses par jour ouvrable	Temporelle
	Nombre moyen de courses en période de pointe du matin	Temporelle
	Nombre total de courses par heure pour chaque jour d'une semaine (choix de la semaine)	Temporelle
	Nombre moyen de courses par heure pour chaque jour d'une semaine (moyenne des 4 semaines du mois)	Temporelle
	Nombre de courses commençant dans chaque secteur de recensement	Spatiale
	Nombre de courses par paire de secteur de recensement	Spatiale
	Nombre moyen de courses par jour ouvrable par véhicule	Combinaison : temporelle et par objet
	Nombre moyen de courses par jour ouvrable par intermédiaire	Combinaison : temporelle et par objet
	Nombre moyen de courses commençant dans chaque secteur de recensement par jour ouvrable	Combinaison : Spatiale et temporelle

Tableau 5-3 Exemples de déclinaisons possibles pour le nombre total de véhicules actifs par mois

Indicateur principal	**Indicateurs secondaires**	**Déclinaison**
Nombre total de véhicules actifs par mois	Nombre moyen de véhicules actifs par intermédiaire	Par objet
	Nombre moyen de véhicules actifs en période de pointe du matin	Temporelle
	Nombre moyen de véhicules actifs par jour ouvrable	Temporelle
	Nombre moyen de véhicules actifs par heure par jour ouvrable	Temporelle
	Nombre moyen de véhicules actifs par jour ouvrable par secteur de recensement	Combinaison : temporelle et spatiale
	Nombre moyen de véhicules actifs en période de pointe par secteur de recensement	Combinaison : temporelle et spatiale
	Nombre moyen de véhicules actifs par heure par secteur de recensement	Combinaison : temporelle et spatiale
	Nombre moyen de véhicules actifs par intermédiaire par jour ouvrable	Combinaison : Par objet et temporelle

Tableau 5-4 Exemples de déclinaisons possibles pour le total des heures en service par mois

Indicateur principal	**Indicateurs secondaires**	**Déclinaison**
Total des heures en service par mois	Total des heures en service par jour	Temporelle
	Total des heures en course par jour ouvrable	Temporelle
	Durée moyenne en service par véhicule par jour ouvrable	Combinaison : Par objet et temporelle
	Durée moyenne en course par véhicule par jour ouvrable	Combinaison : Par objet et temporelle

Enfin pour chaque déclinaison, on peut calculer des **totaux**, des **moyennes**, des **ratios** ou encore des **pourcentages,** tel qu'illustré dans le Tableau 5-5.

Tableau 5-5 Exemple de déclinaisons d'indicateurs selon le type de calcul : total, moyenne, pourcentage ou ratio

Total	Moyenne	Pourcentage	Ratio
Total des km parcourus à vide par mois	Distance moyenne parcourue à vide	% de la distance parcourue à vide	Ratio de la distance parcourue à vide sur la distance totale
Total des heures passées sans client par mois	Durée moyenne passée sans client	% de la durée à vide	Ratio de la durée sans client sur la durée totale

Ainsi, au lieu de calculer le nombre total de kilomètres parcourus à vide, il peut être plus pertinent d'identifier le pourcentage de la distance parcourue à vide. Le Tableau 5-6 présente les déclinaisons possibles pour cet indicateur.

Tableau 5-6 Déclinaison du pourcentage de la distance parcourue à vide

Indicateur principal	Indicateurs secondaires	Déclinaison
% de la distance parcourue à vide	% de la distance parcourue à vide par chauffeur	Par objet
	% de la distance parcourue à vide par véhicule	Par objet
	% de la distance parcourue à vide par jour ouvrable	Temporelle
	% de la distance parcourue à vide en période de pointe du matin	Temporelle
	% de la distance parcourue à vide par heure pour chaque jour d'une semaine (choix de la semaine)	Temporelle
	% de la distance parcourue à vide par heure pour chaque jour d'une semaine (moyenne des 4 semaines du mois)	Temporelle
	% de la distance parcourue à vide dans chaque secteur de recensement	Spatiale
	% de la distance parcourue à vide par jour ouvrable par véhicule	Combinaison : temporelle et par objet
	% de la distance parcourue à vide par jour ouvrable par intermédiaire	Combinaison : temporelle et par objet

Les principaux indicateurs présentés précédemment pouvant être déclinés de différentes manières et sous plusieurs formes, la liste des indicateurs qui peuvent être calculés est donc très longue. Une

multitude d'indicateurs peuvent être calculés à partir des données dont on dispose, or l'utilisation d'un tableau de bord permet d'assurer les multiples représentations d'un indicateur. En effet, le tableau de bord conçu étant interactif, la liberté peut donc être laissée à l'utilisateur de la plateforme quant au choix de ce qu'il souhaite visualiser. Cependant s'il est théoriquement possible de tout afficher, un équilibre doit être trouvé afin de ne pas noyer l'utilisateur dans trop d'informations. En effet, présenter trop d'indicateurs nuirait à l'analyse. Les partenaires ont donc été sollicités afin de déterminer quels sont les indicateurs les plus pertinents dans un objectif d'analyse des performances du taxi et de prise de décision. Ainsi, parmi les indicateurs que les partenaires souhaitent pouvoir visualiser dans le tableau de bord, on peut citer par exemple:

- Nombre de chauffeurs actifs
- Nombre de véhicules actifs
- Nombre de chauffeurs par véhicule
- Nombre de chauffeurs affiliés à chaque intermédiaire en service
- Nombre moyen de courses effectuées en période de pointe, par jour, par mois, par année
- Distance moyenne parcourue par course en période de pointe, par jour, par mois, par année
- Nombre de courses effectuées ayant une destination en dehors de la Province
- Nombre de courses ayant comme origine un poste d'attente
- Distance parcourue sans client à bord
- Période de temps avec et sans client à bord

Il est en effet essentiel d'identifier les besoins des futurs utilisateurs de la plateforme afin qu'elle soit la plus pertinente possible dans un objectif de prise de décision. Ferreira, Poco, Vo, Freire et Silva (2013) ont ainsi conduit des entretiens avec des spécialistes du domaine (spécialistes des sciences sociales, ingénieurs en circulation et économistes) qui ont utilisé l'ensemble de données des taxis de New York dans leurs recherches afin de mieux comprendre leurs besoins et de concevoir un outil qui répond à leurs besoins (Ferreira et al., 2013).

5.1.4 Echelles temporelles et spatiales

Il a été montré à la section précédente que les indicateurs peuvent être déclinés spatialement et temporellement. Il est donc essentiel de déterminer les échelles spatiales et temporelles possibles

et pertinentes pour l'analyse des indicateurs. Lacombe (2016) et Laviolette (2017) présentent dans leur travaux les différents niveaux spatiaux et temporels identifiés dans la littérature et mis en application dans le cadre de leurs études. Pour le présent projet, différents niveaux ont été retenus. Le Tableau 5-7 présente ces niveaux. Huit niveaux d'analyse temporelle ont été retenus, allant de l'heure à l'année. Et quatre niveaux pour l'analyse spatiale sont retenus, allant du secteur de recensement au territoire complet de l'île de Montréal. L'analyse peut donc être très ciblée : à une zone restreinte pendant certaines heures de la journée, ou au contraire être menée sur l'ensemble du territoire et ce, pendant un laps de temps élevé.

Tableau 5-7 Echelles temporelles et spatiales retenues pour l'analyse des indicateurs, tiré de Laviolette (2017)

Période d'analyse	**Zone d'analyse**
1. Heure 2. Groupe d'heures (i.e. période de pointe du matin) 3. Jour 4. Groupe de jours (i.e. weekend, jours ouvrés) 5. Semaine 6. Mois 7. Saison 8. Année	1. Secteur de recensement (SR) 2. Secteur municipal (SM) 3. Arrondissement 4. Ile de Montréal (territoire complet)

Dans son prototype de tableau de bord, Laviolette (2017) choisit pour chaque indicateur les niveaux temporels d'analyse ainsi que la zone d'analyse qu'il considère comme optimale. Ainsi, selon les indicateurs, les options temporelles et spatiales varient. Cependant, dans le présent projet, tous les niveaux spatiaux et temporels sont disponibles pour chaque indicateur. Le choix du niveau est en effet laissé à l'utilisateur. Ces options de filtres sont détaillées à la section 6.3.2 du Chapitre 6.

5.1.5 Synthèse des déclinaisons

La Figure 5-3 présente une synthèse des différentes déclinaisons possibles mentionnées précédemment. Un indicateur principal peut ainsi être décliné spatialement, temporellement ou par objet. Ces trois déclinaisons peuvent également être combinées entre elles. Les niveaux d'analyse spatiale et temporelle ont été définis. Enfin pour chaque indicateur secondaire obtenu, quatre formes de visualisations sont possibles, ce qui donne lieu à de multiples possibilités d'analyse.

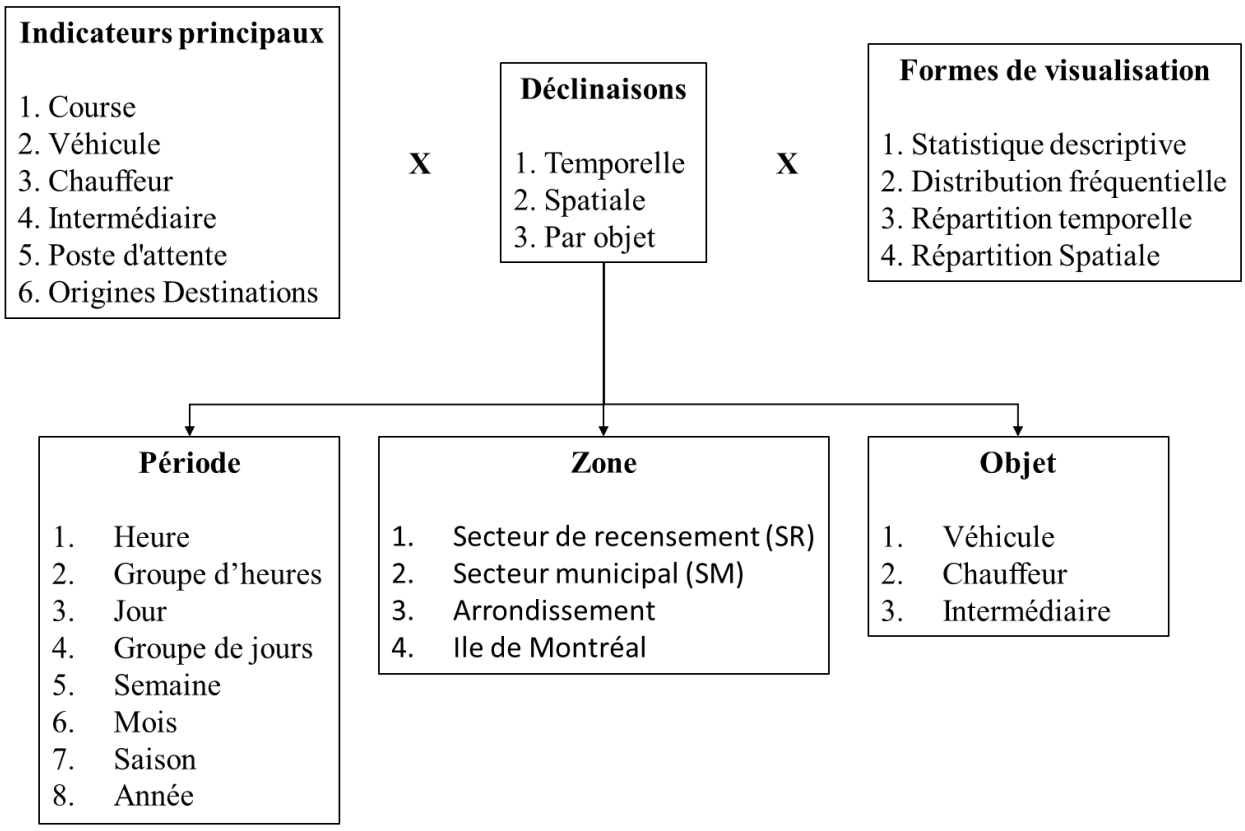


Figure 5-3 Synthèse des niveaux d'estimation et d'analyse pour les indicateurs

L'exemple du nombre de courses est repris pour illustrer différentes déclinaisons possibles.

Une première forme d'indicateur peut être la statistique descriptive de base : le nombre de courses du mois d'avril 2019 s'élève à 441 833 courses. Ici, le mois d'avril 2019 est choisi comme période temporelle, l'indicateur principal concerne la course et la forme de visualisation choisie est la statistique descriptive.

Notons que l'on considère que le nombre de courses d'une période donnée correspond aux courses qui ont débuté pendant cette période. Dans le cas du mois d'avril 2019, cela correspond aux courses qui ont débuté après minuit le 1er avril jusqu'à avant minuit le 30 avril 2019. Les courses débutées le 31 mars avant 23:59:59 (23 heures 59 minutes et 59 secondes) mais se terminant après minuit le 1er avril ne comptent pas comme des courses du mois d'avril. De même, les courses débutées le 30 avril avant 23:59:59 mais se terminant après minuit le 1er mai sont considérées comme des courses du mois d'avril.

La répartition temporelle du nombre de courses peut ensuite être visualisée à la Figure 5-4. L'histogramme présente le nombre de courses selon les jours du mois d'avril 2019. Les barres représentant les dimanches sont mises en évidence en orange. On peut ainsi observer que le nombre de courses est plus faible les dimanches que les autres jours de la semaine. La seule exception, mise en évidence en noir, concerne le lundi 22 avril, jour du mois d'avril où est enregistré le plus faible nombre de courses. C'est à cette date que le nombre minimum de courses réalisées au cours des jours du mois d'avril est atteint. Cependant, le lundi 22 avril 2019 correspondait au lundi de Pâques, jour férié à Montréal, et ceci explique donc le faible nombre de courses enregistré.

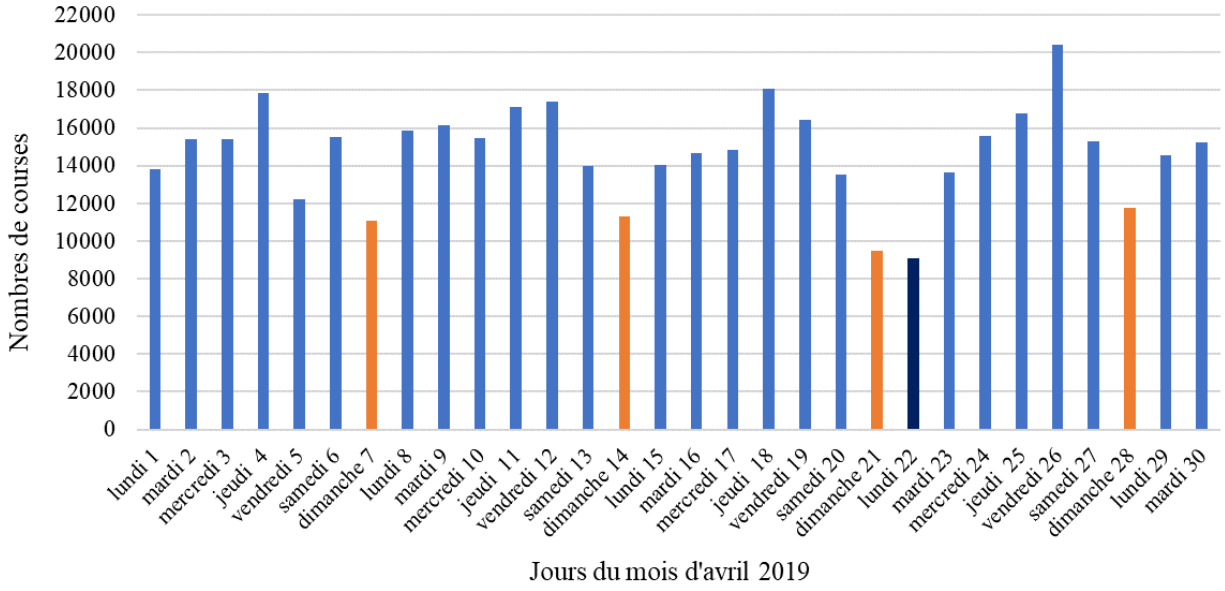


Figure 5-4 Histogramme de distribution du nombre de courses selon les jours du mois d'avril 2019

La période temporelle peut également être limitée à une seule semaine du mois d'avril, tel que présenté à la Figure 5-5. Le jeudi de cette première semaine est le jour qui comptabilise le plus de courses et le dimanche celui qui en comptabilise le moins.

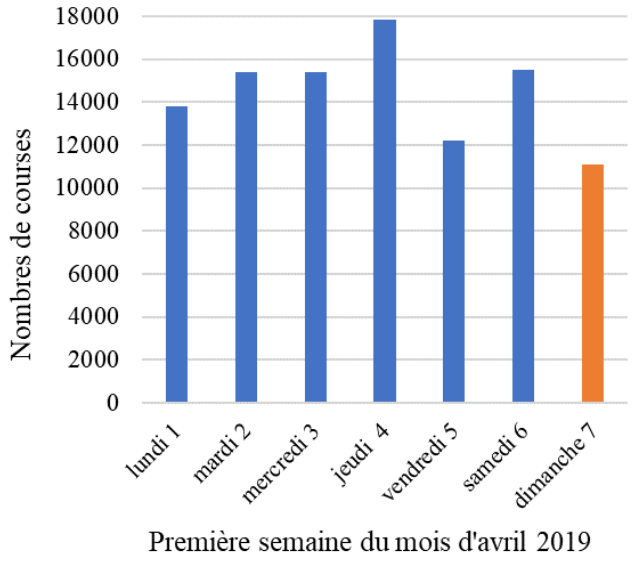


Figure 5-5 Histogramme de distribution du nombre de courses selon les jours de la première semaine du mois d'avril 2019

Il est également possible de présenter le nombre total de courses pour chaque jour du mois d'avril. Pour la distribution de la Figure 5-6, les courses des lundis du mois sont sommées, celles des mardis du mois et ainsi de suite. Cependant une limite à cette visualisation concerne le nombre inégal de chaque type de jour dans un mois. En effet, s'il semble que le nombre de courses des mardis est plus élevé, cela est notamment dû au fait qu'il y a 5 mardis dans le mois d'avril 2019. Le mois ne présente que 4 occurrences de mercredi, jeudi, vendredi, samedi et dimanche contre 5 occurrences de lundis et de mardis.

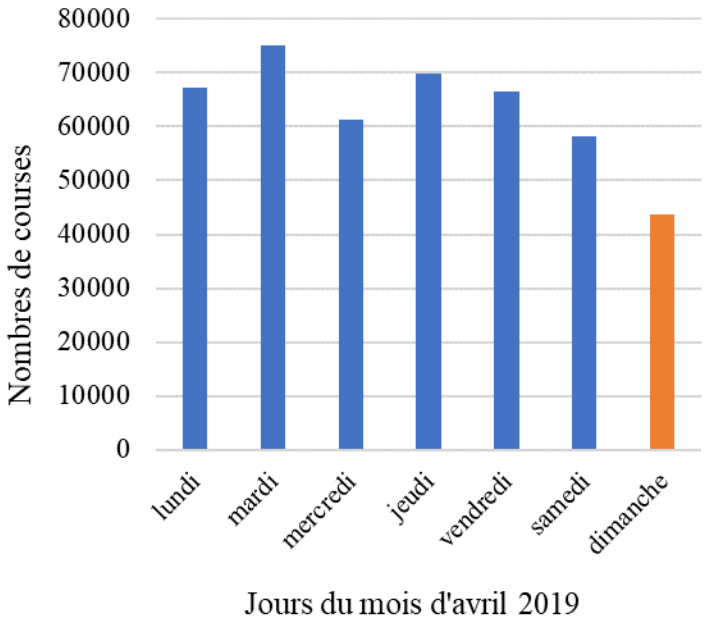


Figure 5-6 Histogramme de distribution du nombre de courses total pour chaque jour de la semaine

Face à la limite identifiée pour la répartition précédente, il peut donc être plus pertinent de représenter le nombre de course pour une semaine moyenne. La Figure 5-7 illustre cette répartition du nombre de courses pour une semaine moyenne du mois d'avril 2019. En moyenne, sur les quatre semaines du mois, les jeudi et vendredi sont les jours de la semaine recensant le plus de courses. Le dimanche est quant à lui celui qui en compte le moins.

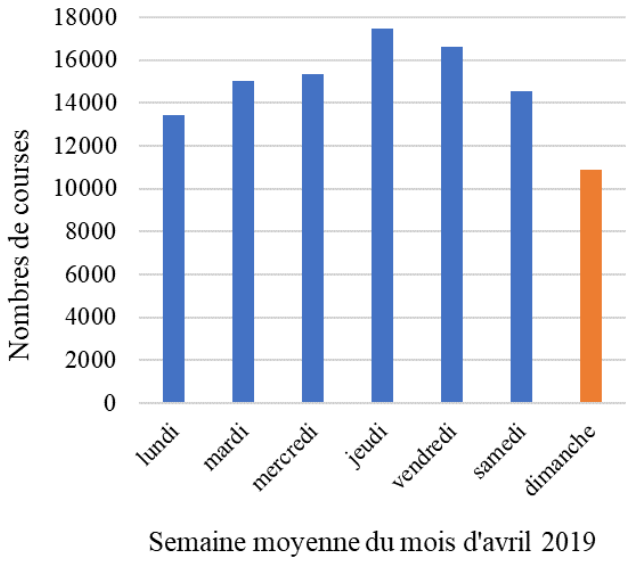


Figure 5-7 Histogramme de distribution du nombre de courses pour chaque jour d'une semaine moyenne d'avril 2019

Enfin, une répartition fréquentielle de cet indicateur peut également être réalisée. Ces répartitions sont présentées aux Figure 5-8 et Figure 5-9. Le nombre de véhicules de taxi effectuant un certain nombre de courses pour le mois d'avril 2019 est représenté dans la Figure 5-8. La Figure 5-9 présente quant à elle la proportion de véhicules (au lieu du nombre de véhicules). Ainsi, 497 véhicules de taxis ont réalisé entre une et cent courses lors du mois d'avril 2019, soit environ 30% des véhicules. Et un véhicule a quant à lui assuré plus de 1000 courses, soit environ 0.1% des véhicules. Il convient de rappeler qu'un véhicule de taxi peut être partagé par plusieurs chauffeurs. L'objet véhicule pourrait donc être remplacé par l'objet chauffeur afin d'obtenir une autre déclinaison de l'indicateur.

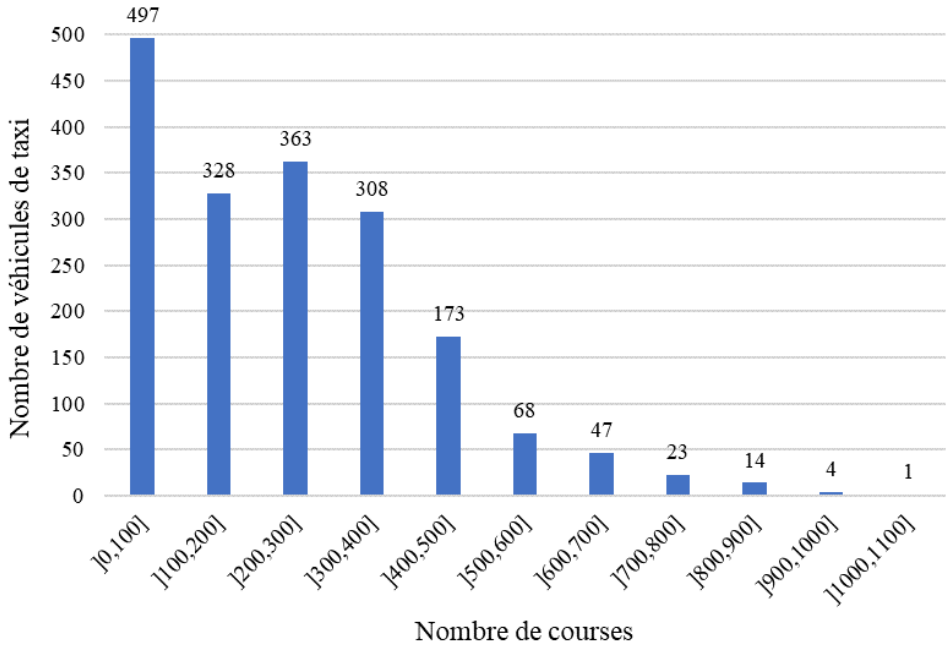


Figure 5-8 Nombre de véhicules de taxi effectuant un certain nombre de courses pour le mois d'avril 2019

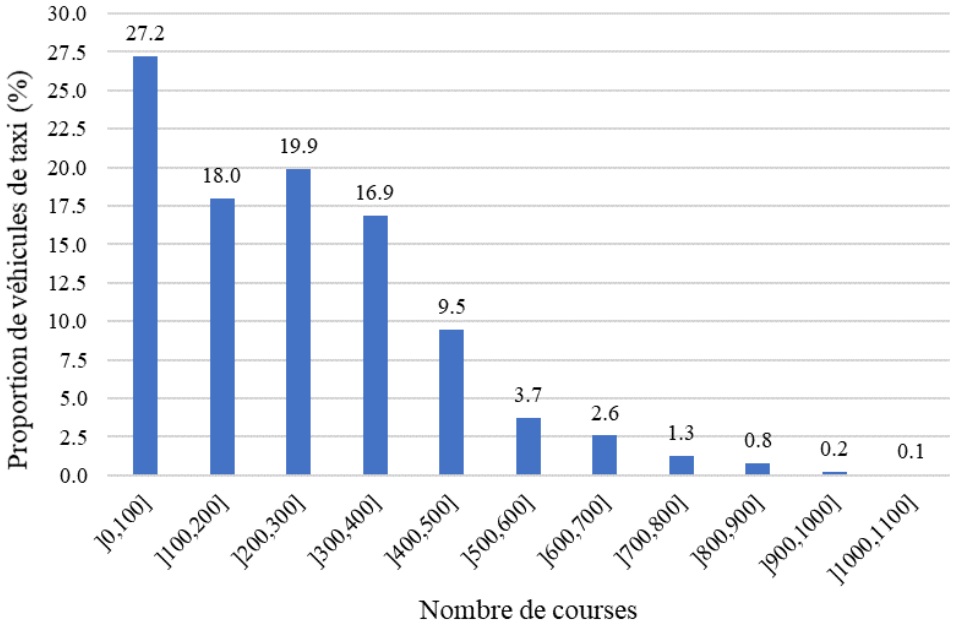


Figure 5-9 Proportion de véhicules de taxi effectuant un certain nombre de courses pour le mois d'avril 2019

5.2 Synthèse des méthodes de calcul des indicateurs et limitations

La section précédente a permis de mettre en avant les enjeux liés au choix des indicateurs, notamment en ce qui concerne les formes de visualisation et les périodes d'analyse spatiale et temporelle. Ce sont désormais les enjeux liés au calcul des indicateurs à partir des données de taxis qui seront présentés dans cette partie.

5.2.1 Indicateurs liés aux courses

5.2.1.1 Validité des courses

L'un des premiers défis concerne la validité des données. Par exemple, si l'on souhaite analyser la demande en déplacements de taxi, il est essentiel de s'assurer de la validité des courses identifiées dans les données. Une méthode de validation des courses basée sur des valeurs seuils de durée, distance et vitesse moyenne a pour cela été réalisée et est décrite à la section 4.4. Cette méthode permet d'éliminer des courses considérées comme non valides, soit environ 9% des courses brutes. Ces courses non valides présentaient des caractéristiques de durée et de distance très faibles ou au contraire très élevées qui auraient eu un impact significatif sur les statistiques descriptives des indicateurs liés aux courses.

5.2.1.2 Courses régulières et non régulières

Un autre enjeu des indicateurs de courses est lié au statut *unavailable*. En effet, lorsque le statut d'un taxi est *unavailable*, il peut par exemple s'agir d'une course de transport adapté. Comme mentionné dans la partie 3.1.3, plusieurs types de courses sont possibles lorsque le statut du taxi est *unavailable*. Et les informations actuellement disponibles dans le Registre ne permettent pas de discerner les différents types de trajets. Ainsi, seule une partie des courses effectuées par les taxis peut être identifiée. Ces courses, identifiées comme courses « régulières », sont celles pour lesquelles le statut indiqué est *occupied.* La méthode d'identification des courses régulières est détaillée dans le Chapitre 4.

Ainsi, seule une partie des activités des taxis peut réellement être identifiée. Cependant, certains intermédiaires en service sont spécialisés dans le transport adapté. Ce service constitue donc la majorité de leurs activités de transport par taxi. Il est donc essentiel de pouvoir identifier l'ensemble des courses si l'on souhaite pouvoir analyser toutes les activités des taxis.

Le statut *oncoming* indique normalement que le taxi est en route pour rejoindre un client. Ainsi, il pourrait servir à identifier les courses non régulières. En effet, si un groupe de statut *unavailable* est précédé d'un groupe de statut *oncoming*, cela peut donc être un indicateur que le groupe de statut *unavailable* correspond bien à une course particulière.

Le Tableau 5-8 présente le pourcentage des groupes de statut *unavailable* qui sont précédés d'un groupe de statut *oncoming* pour les différents intermédiaires en service. Si pour certains intermédiaires tels que l'Intermédiaire2 et l'Intermédiaire3, le pourcentage des groupes de statut unavailable qui sont précédés d'un groupe de statut *oncoming* est supérieur à 20%, il est très faible pour la grande majorité des intermédiaires en service. De plus, pour les intermédiaires en service mis en évidence en orange, ce pourcentage est particulièrement faible voire nul alors que ces opérateurs assurent des services de transport adapté. Les courses de transport adapté constituent donc la grande majorité de leurs activités comme le témoigne le Tableau 5-9. Parmi les groupes de statut, ce sont ceux de statut unavailable qui sont majoritaires. L'opérateur identifié par Intermédiaire20 effectue uniquement des courses de transport adapté puisqu'aucun groupe de statut *occupied* n'est relevé, indiquant qu'aucune course régulière n'est effectuée. Ainsi même si ces intermédiaires effectuent des courses de transport adapté il n'est pas possible de les identifier puisqu'elles ne sont pas précédées d'une période de statut *oncoming*. Cela peut être lié au fait que les courses sont commandées à l'avance et que le chauffeur n'indique donc pas qu'il est en route pour chercher le client (comme il pourrait le faire lorsqu'une course commandée vient de lui être attribuée).

Tableau 5-8 Pourcentage des groupes de statut unavailable qui sont précédés d'un groupe de statut oncoming

	Nombre de groupes de statut *unavailable*	**Nombre de groupes de statut *unavailable* précédés d'un groupe de statut *oncoming***	**Pourcentage des groupes de statut *unavailable* précédés d'un groupe de statut *oncoming***
Intermédiaire1	17819	1	0.01%
Intermédiaire2	449	101	22.49%
Intermédiaire3	13648	3525	25.83%
Intermédiaire4	9153	1613	17.62%
Intermédiaire5	26197	3859	14.73%
Intermédiaire6	6364	37	0.58%
Intermédiaire7	10425	108	1.04%
Intermédiaire8	13848	2053	14.83%
Intermédiaire9	3306	15	0.45%
Intermédiaire10	14911	2413	16.18%
Intermédiaire11	251	0	0.00%
Intermédiaire17	10	0	0.00%
Intermédiaire18	2404	0	0.00%
Intermédiaire19	3171	25	0.79%
Intermédiaire20	2866	0	0.00%
Intermédiaire22	215	0	0.00%
Intermédiaire21	82	0	0.00%

Tableau 5-9 Proportion des groupes de statut pour le mois d'avril 2019 pour les intermédiaires de transport adapté

	Intermédiaire7	**Intermédiaire9**	**Intermédiaire19**	**Intermédiaire20**
free	22%	37%	28%	42%
occupied	31%	1%	1%	0%
oncoming	1%	0%	1%	0%
unavailable	46%	61%	70%	58%

5.2.1.3 Calcul de quelques indicateurs liés aux courses

La Figure 5-10 présente le schéma méthodologique du calcul du nombre moyen de courses par véhicule. Tel que mentionné dans la section 5.2.1.2, seules les courses régulières sont considérées. A partir de la base de données optimisée un filtre sur les statuts est réalisé afin de conserver uniquement les statuts *occupied*. La durée totale est par la suite calculée en réalisant la différence de temps entre l'horodatage de fin de course et l'horodatage de début de course. La distance totale est quant à elle un attribut qui figure déjà dans la base de données optimisée (la structure de cette base de données optimisée a été détaillée à la section 3.4). La vitesse moyenne est alors calculée en réalisant le rapport de la distance totale et de la durée totale. Il est nécessaire de préalablement calculer ces attributs afin de pouvoir appliquer les règles de validations des courses. Ces dernières, détaillées à la section 4.4, sont ensuite appliquées aux groupes de statut *occupied*. Une fois la validation effectuée, le nombre de courses effectuées par un même véhicule est identifié. Enfin deux moyennes peuvent être calculées. La moyenne du nombre de courses peut soit être réalisée sur l'ensemble des véhicules ayant effectué au moins une course (N_o) ou alors être réalisée sur l'ensemble des véhicules en opération (N_T).

Il convient également de rappeler que les courses d'une période donnée correspondent aux courses commençant pendant cette période (soit les courses dont l'horodatage du point d'origine est compris dans la période d'analyse souhaitée).

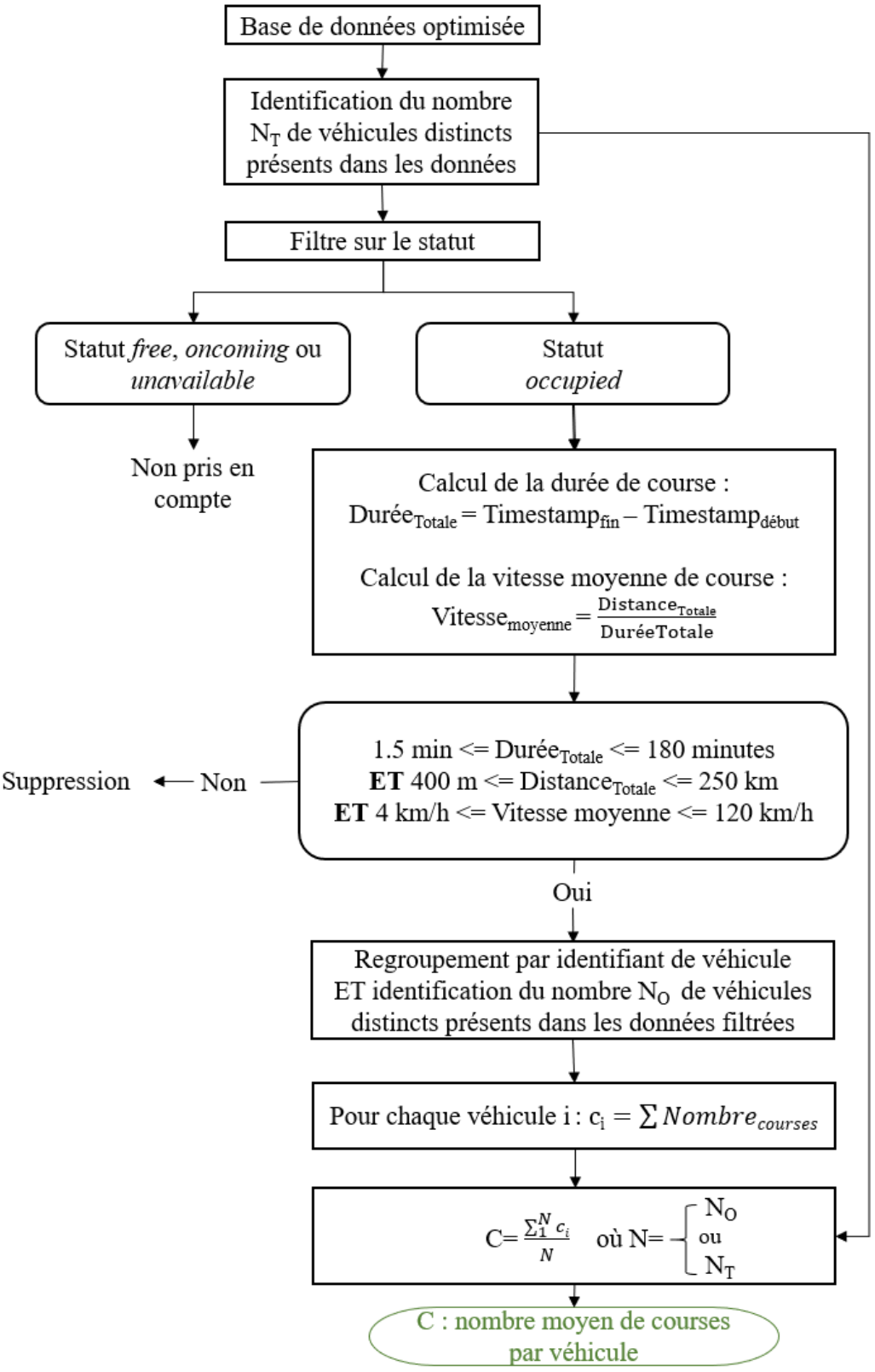


Figure 5-10 Schéma méthodologique du calcul du nombre moyen de courses par véhicule

La Figure 5-11 présente le schéma méthodologique du calcul de la durée moyenne de course et du temps d'attente moyen.

Pour le calcul du temps d'attente moyen, un filtre sur le statut est réalisé afin de ne conserver que les groupes de statut *oncoming* qui sont suivis d'un groupe de statut *occupied*. En effet, à la suite d'erreurs de manipulations ou de l'annulation d'une course par le client, des groupes de statut *oncoming* suivis de groupe de statut *free* sont observés dans les données. Ces groupes ne doivent donc pas être considérés dans le calcul des temps d'attente moyen pour les courses régulières. En calculant la différence de temps entre l'horodatage de début et l'horodatage de fin du groupe, la durée de chaque période d'attente peut être identifiée. La moyenne peut alors être réalisée sur l'ensemble des périodes d'attente.

Pour le calcul de la durée moyenne de course, un filtre est effectué sur les statuts afin de ne conserver que les statuts *occupied*. Les règles de validation des courses sont alors appliquées afin de ne conserver que les courses valides. Et la durée des courses est déterminée en calculant la différence de temps entre l'horodatage de début et l'horodatage de fin de groupe. La moyenne peut alors être réalisée sur l'ensemble des courses.

La Figure 5-12 présente le schéma méthodologique du calcul des distances et vitesses moyennes de course. Le processus est similaire au calcul de la durée moyenne détaillé précédemment.

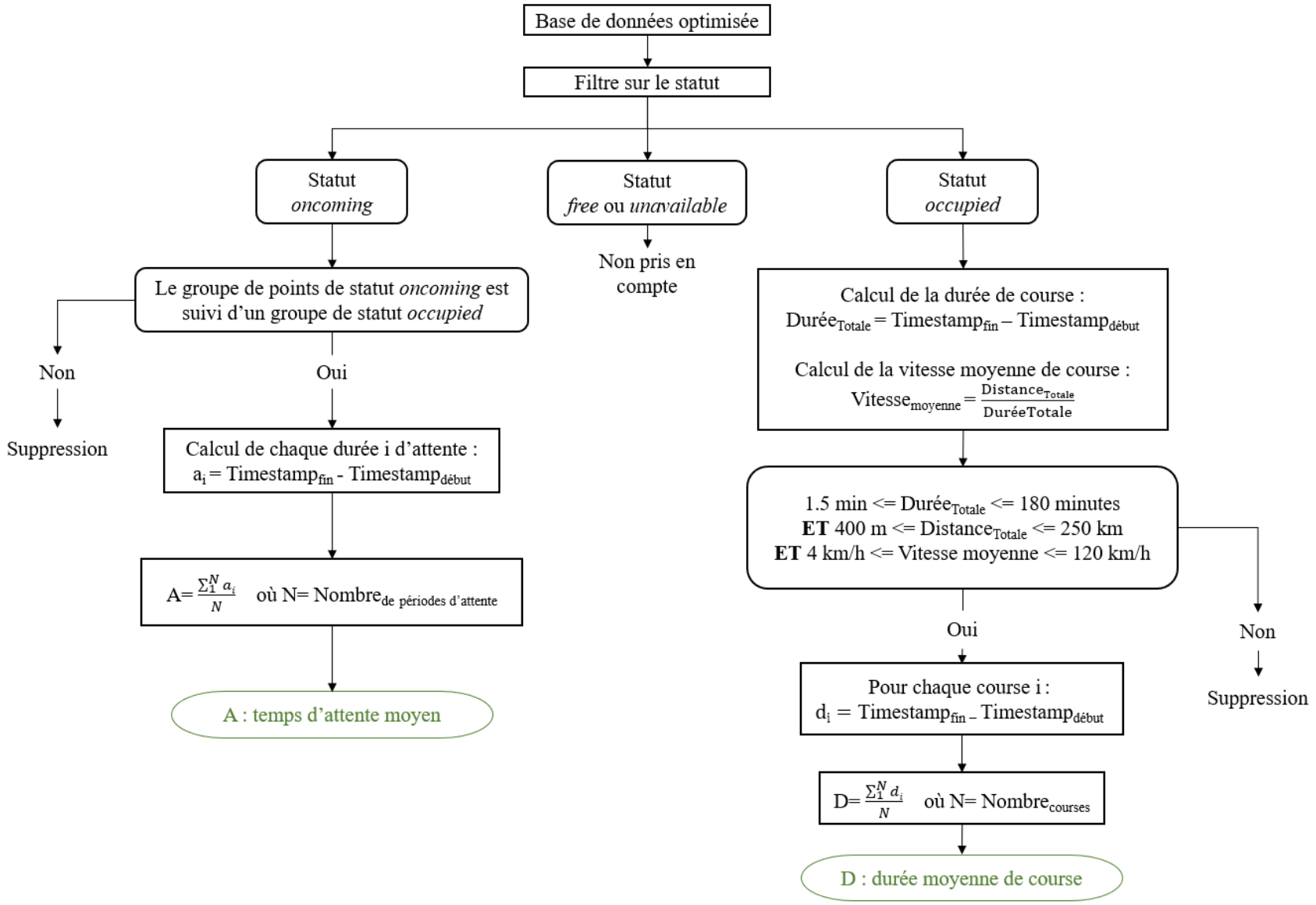


Figure 5-11 Schéma méthodologique du calcul de la durée moyenne de course et du temps d'attente moyen

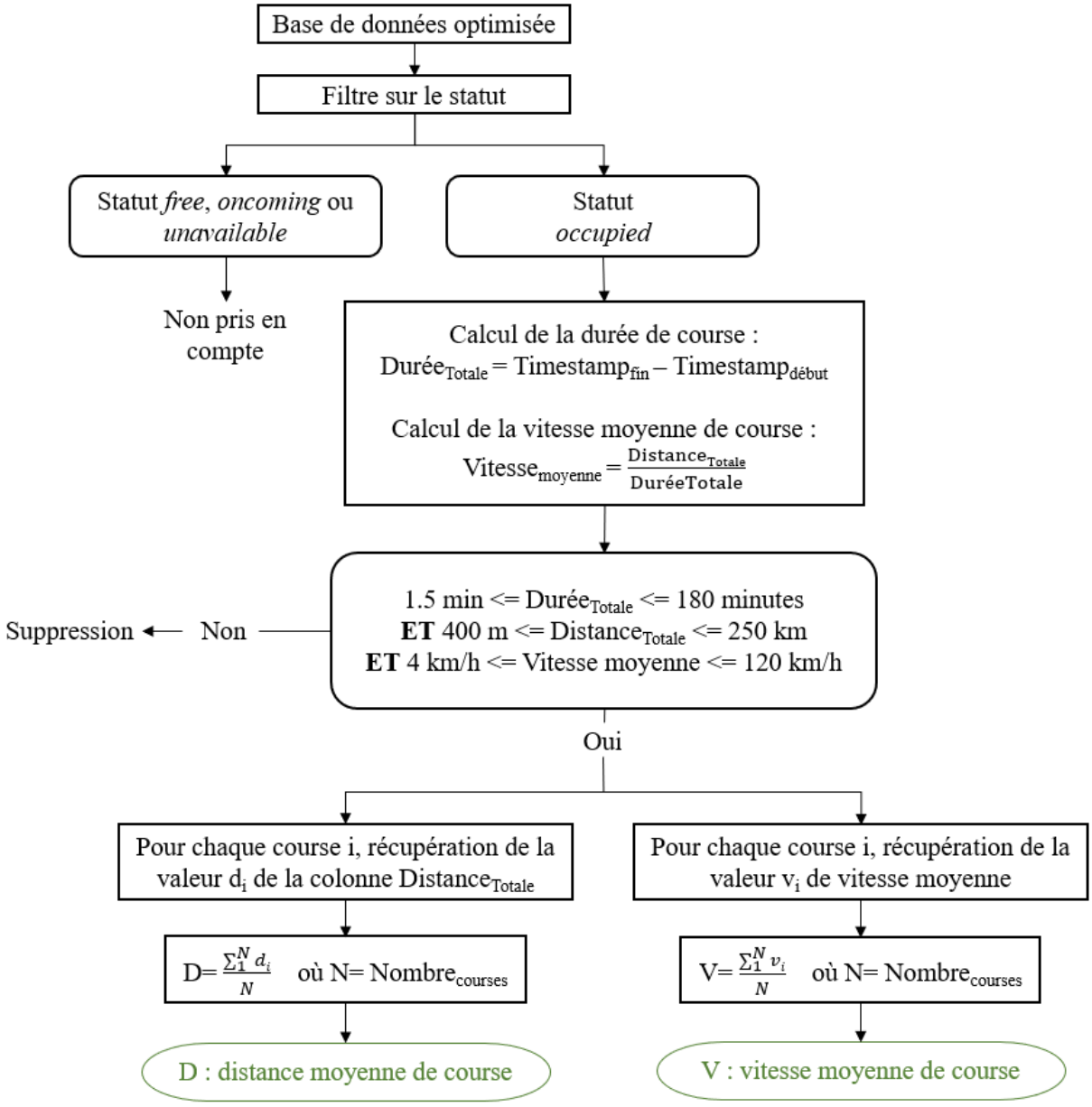


Figure 5-12 Schéma méthodologique du calcul des distances et vitesses moyennes de course

5.2.2 Indicateurs liés aux véhicules

5.2.2.1 Enjeu du statut unavailable

Tel que mentionné précédemment, le statut *unavailable* présente en enjeu dans la détermination des activités des taxis. En effet, une incertitude existe sur l'activité effectuée par le taxi lors de ce

statut. On ne peut être certain que c'est une course non régulière. En effet, il est également possible que certains chauffeurs indiquent ce statut lorsqu'ils prennent une pause ou qu'ils ne souhaitent pas se rendre disponible à répondre à une course commandée.

Ainsi, si l'on souhaite par exemple déterminer la durée passée à vide d'un véhicule ou la distance parcourue à vide, doit-on considérer que tout groupe de statut unavailable implique qu'un client est à bord ?

La Figure 5-13 présente le schéma méthodologique du calcul de la distance moyenne parcourue à vide. On considère ici que seuls les groupes de statut *free* et *oncoming* doivent être pris en compte dans le calcul de la distance parcoure à vide par véhicule. Il serait cependant pertinent d'effectuer ce calcul pour les deux cas de figure, soit en considérant également que le groupes de statut *unavailable* doivent être pris en compte dans le calcul de la distance parcourue à vide.

L'enjeu est le même pour la détermination du temps ou de la distance passée en course. Deux cas sont à considérer : les groupes de statut *unavailable* et *occupied* sont considérés ou alors seuls les groupes de statut *occupied* (les courses régulières) sont pris en considération.

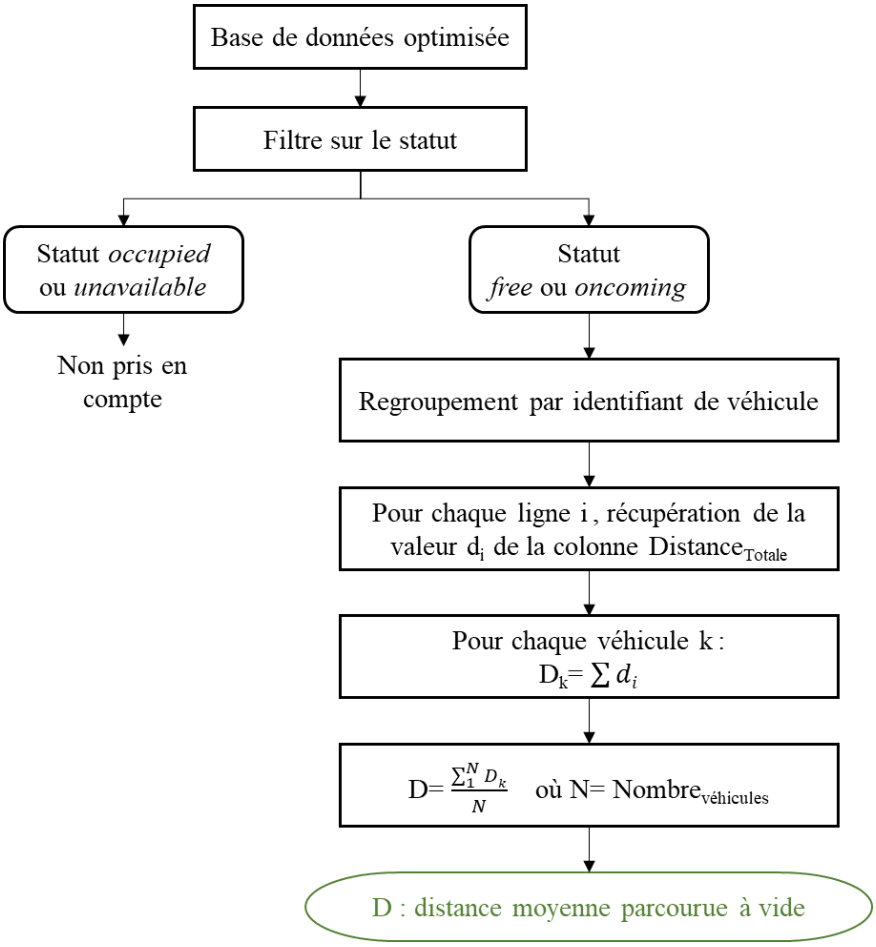


Figure 5-13 Schéma méthodologique du calcul de la distance moyenne parcourue à vide

De la même manière, lorsque l'on souhaite identifier la durée de service des véhicules, soit la durée pendant laquelle le véhicule est en opération, il est essentiel de pouvoir identifier s'il est en pause ou s'il effectue une course de transport adapté.

La Figure 5-14 présente le schéma méthodologique du calcul de la durée moyenne de service par véhicule. D'après l'ordonnance du BTM, dès qu'un taxi est en service il doit se connecter au Registre et envoyer ses données. Ainsi toute période de donnée reçue doit donc correspondre à une période où le taxi est en service. On considère donc que les groupes de statut correspondent à des périodes où le statut est en service. Il serait cependant pertinent une fois de plus, d'effectuer ce calcul pour le cas où les groupes de statut *unavailable* sont identifiés à des pauses (et donc des périodes hors service).

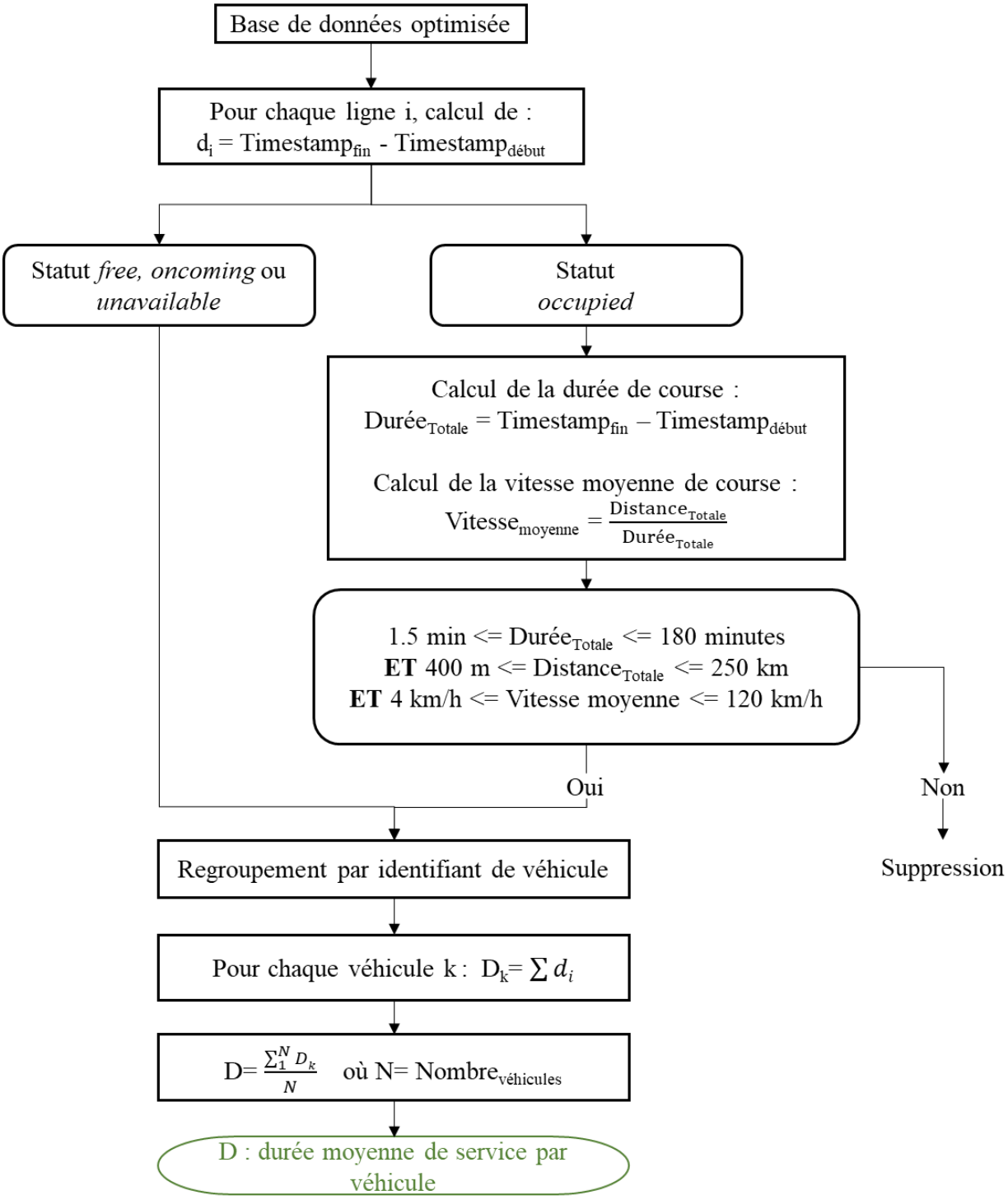


Figure 5-14 Schéma méthodologique du calcul de la durée moyenne de service par véhicule

5.2.2.2 Véhicules actifs

Une autre limite présente dans les données du Registre concerne l'identification de la proportion d'objets actifs. En effet, avec les données dont on dispose, il est possible d'obtenir le nombre de véhicules, de conducteurs ou encore de permis qui sont actifs, c'est-à-dire qui sont en opération. Cependant, si le pourcentage de permis actifs pendant une période d'analyse spécifique doit être identifié, il est nécessaire de connaître la population de référence, c'est-à-dire tous les permis qui étaient valables pendant la période d'étude mais qui n'ont pas forcément été actifs. Toutefois, la base de données des permis ne fournit pas le statut de validité ou d'invalidité du permis. Un contrôle de la validité des permis est effectué par le BTM mais l'information n'est pas renseignée dans le Registre. Il en est de même pour les chauffeurs et véhicules. On ne peut savoir si les chauffeurs et véhicules renseignés dans les données sont encore autorisés à offrir un service de taxi. Ainsi, seul le nombre d'objets en activité peut être identifié.

La Figure 5-15 présente le schéma méthodologique du calcul du nombre de véhicules actifs. On considère ici que dès lors qu'une donnée est reçue pendant la période d'analyse, alors le véhicule est considéré comme étant actif. Un véhicule s'étant connecté au moins une fois au Registre durant la période d'analyse considérée est donc considéré comme ayant été actif pendant cette période.

Cependant cette notion d'activité est variable. Considère-t-on qu'un taxi qui ne s'est connecté que pendant 5 secondes durant le mois a été actif pendant ce mois ? Ou doit-il avoir été en service pendant plusieurs jours ? Le choix peut donc être laissé à l'utilisateur de la plateforme quant à la durée minimum (ou au pourcentage minimum de la période d'analyse) pendant laquelle un véhicule doit être en service pour être considéré comme actif.

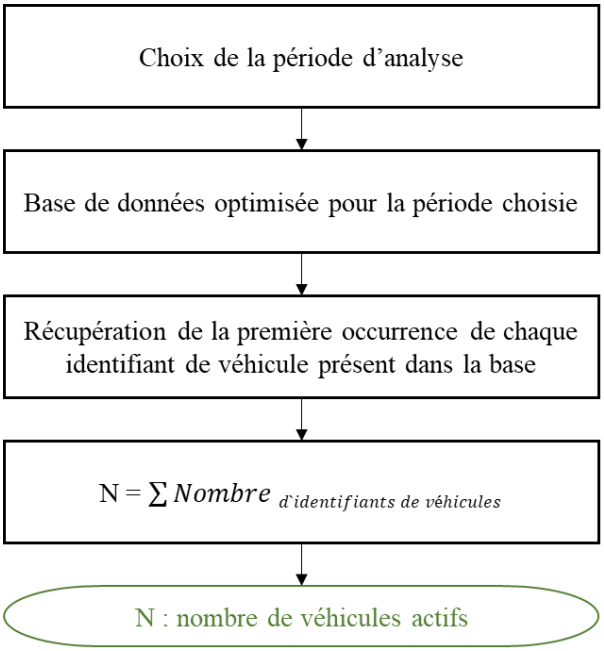


Figure 5-15 Schéma méthodologique du calcul du nombre de véhicules actifs

5.2.3 Indicateurs liés aux chauffeurs

5.2.3.1 Chauffeurs actifs

Tel que mentionné précédemment, la base de données des chauffeurs n'est pas mise à jour régulièrement. Un chauffeur qui a été ajouté à cette base demeure ainsi dans les données même s'il n'est plus en activité ou qu'il n'assure plus de service de taxi. Ainsi, le nombre d'objets « chauffeurs » renseigné dans la base de données des chauffeurs n'est pas un indicateur du nombre de chauffeurs qui peuvent réellement assurer un service de taxi. Seuls les chauffeurs qui ont été actifs pendant la période d'analyse considérée peuvent être identifiés. Tout comme pour les véhicules actifs mentionnés précédemment, le choix peut donc être laissé à l'utilisateur de la plateforme quant à la durée minimum requise (ou au pourcentage minimum de la période d'analyse) pour considérer qu'un chauffeur a été actif pendant la période d'étude.

5.2.3.2 Nombre de chauffeurs par véhicule

Un permis de propriétaire de taxi est associé à un unique véhicule. Cependant, ce dernier peut être partagé par plusieurs chauffeurs. La Figure 5-16 présente le schéma méthodologique du calcul du

nombre moyen de chauffeurs actifs par véhicule. Une fois de plus, on s'intéresse aux chauffeurs et véhicules actifs et non à l'ensemble des chauffeurs et véhicules renseignés dans les bases de données. La base de données optimisées contient des attributs correspondant à l'identifiant de véhicule et à l'identifiant du chauffeur du taxi. Les couples ($Identifiant_{Chauffeur}$, $Idenfiant_{Véhicule}$) sont identifiés pour chaque véhicule en activité. Ainsi, si deux couples sont identifiés pour un même véhicule, cela signifie que deux chauffeurs de taxi partagent le véhicule.

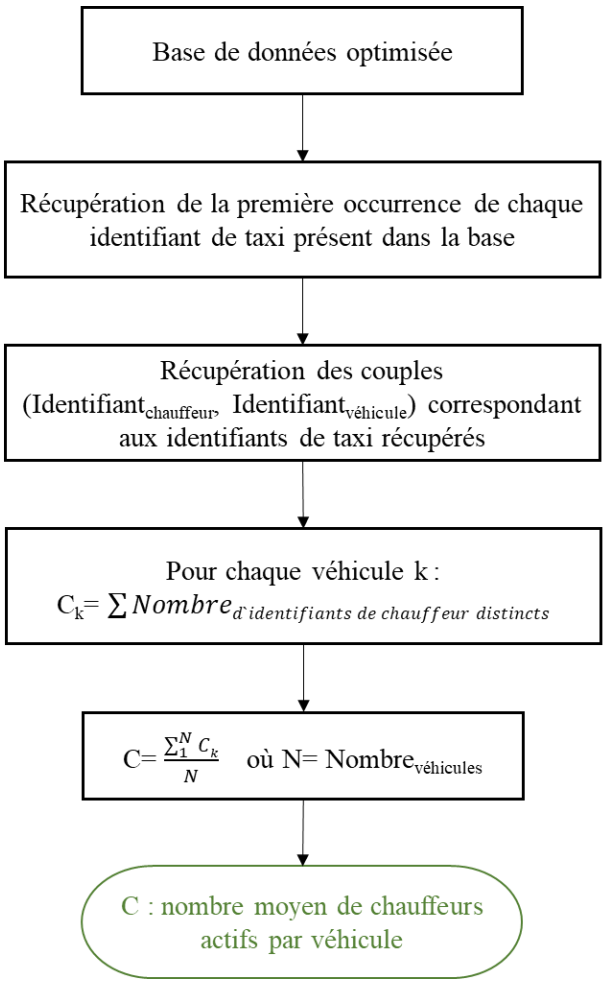


Figure 5-16 Schéma méthodologique du calcul du nombre moyen de chauffeurs actifs par véhicule

5.2.4 Indicateurs liés au poste d'attente

Les données disponibles pour la localisation des postes d'attente n'indiquent qu'un unique point pour chaque zone de postes d'attentes ainsi que le nombre de places. Il est supposé que le point indiqué corresponde au centroïde du poste d'attente. On ne dispose pas des coordonnées spatiales de chaque place de stationnement au sein du poste d'attente. Or, si l'on souhaite étudier le taux d'utilisation des postes d'attente il est essentiel de pouvoir déterminer combien de places sont occupées.

La Figure 5-17 présente le schéma méthodologique du calcul des taux d'utilisation des postes d'attente. Un filtre sur les statuts est réalisé afin de conserver les statuts *free* uniquement. En effet, lorsque le taxi est à la recherche d'un nouveau client, il peut se rendre à un poste d'attente afin d'attendre qu'un client souhaite y commencer une course. Parmi ces groupes de statut, les coordonnées des points GPS du groupe sont analysées afin d'identifier si des points correspondent à des emplacements de poste d'attente. On considère pour cela qu'une place de stationnement fait 5.5 mètres de long (Bourdeau, 2014). Un cercle de diamètre égal à la longueur d'une place de stationnement multipliée par le nombre de places disponibles dans le stationnement est tracé autour de chaque localisation de poste d'attente, tel qu'illustré à la Figure 5-18. Un taxi libre et à l'arrêt à l'intérieur d'une telle zone est considéré comme occupant une place de poste d'attente. On vérifie également que le nombre de taxis repérés dans la zone de stationnement pendant une certaine période n'est pas supérieur au nombre de places de stationnement dont dispose le poste d'attente. Le nombre de taxis identifiés et donc de places occupées pendant une certaine période est ensuite comparé au nombre de places disponibles dans le poste d'attente.

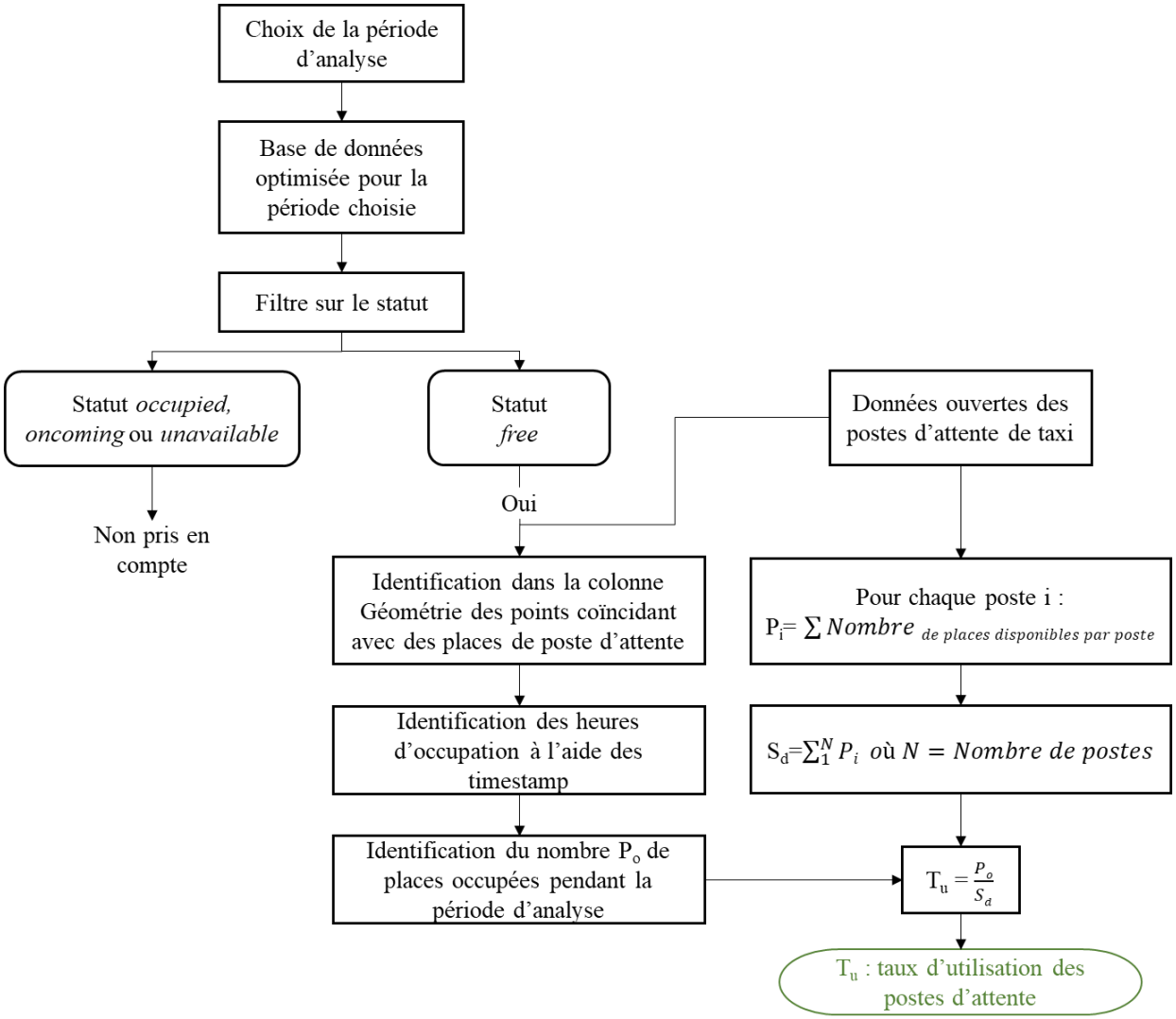


Figure 5-17 Schéma méthodologique du calcul des taux d'utilisation des postes d'attente

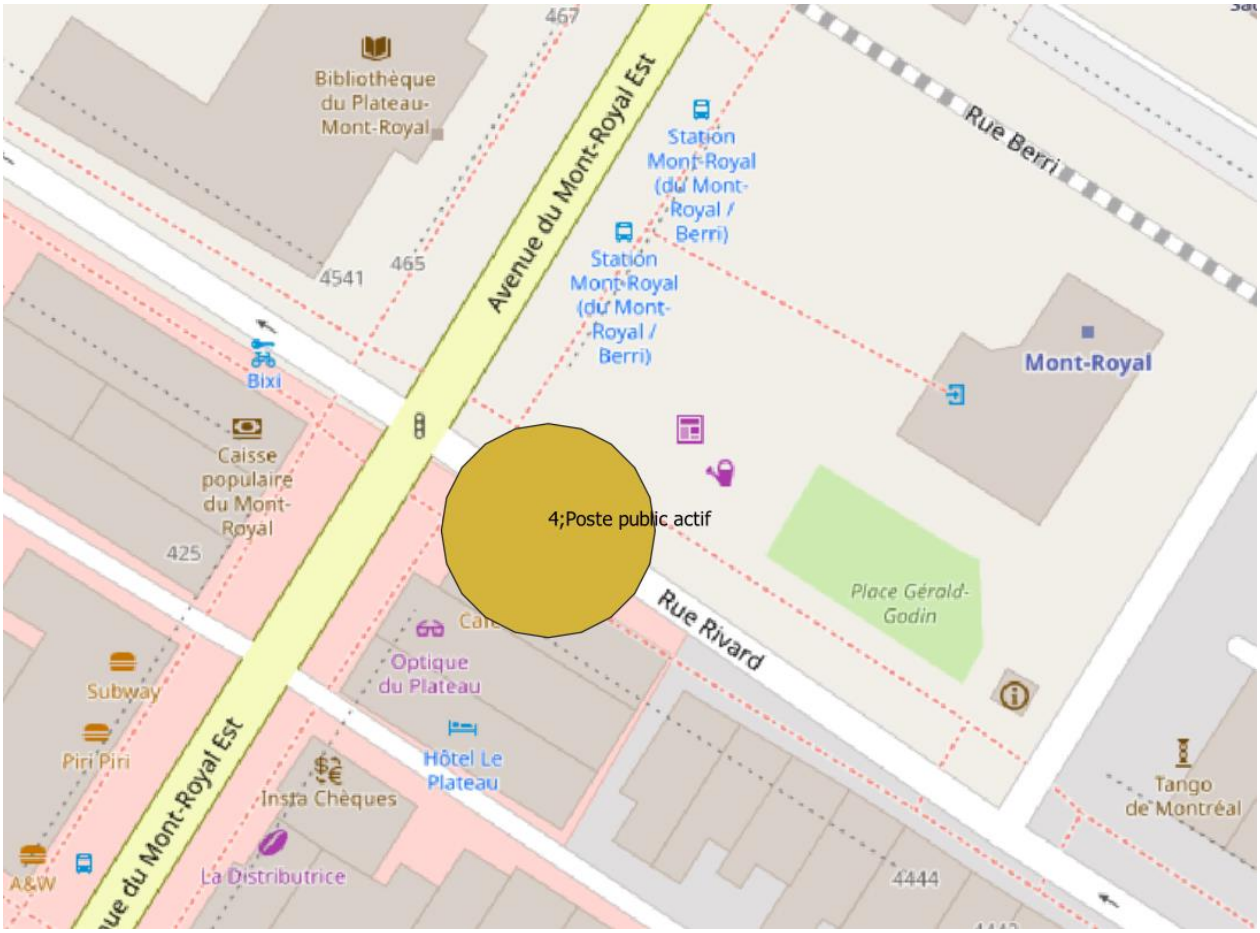


Figure 5-18 Exemple de zone tampon réalisée autour d'un poste d'attente de taxi indiquant le nombre de places disponibles (4) et le type de poste d'attente (Poste public actif)

5.2.5 Indicateurs liés aux origines et destinations

Les origines et destinations des courses peuvent être représentées directement sur la carte du territoire de l'île de Montréal. Cependant, le nombre de points peut être très conséquent si l'utilisateur souhaite par exemple visualiser les origines des courses pour un mois complet et le temps de chargement des données sur la carte serait considérable. De plus, en termes d'analyse il est plus pertinent d'agréger des points selon des zones plutôt que de visualiser chaque entité.

Ainsi il est proposé de réaliser des cartes choroplèthes des origines ou destinations en comptabilisant le nombre d'origines ou de destinations par zone. Ces zones peuvent être les secteurs de recensement, les secteurs municipaux ou encore les arrondissements.

La Figure 5-19 présente le schéma méthodologique de la création d'une carte choroplèthe des origines et destinations. Dans un premier temps les courses régulières sont identifiées. Une intersection est ensuite réalisée entre les points d’origine ou de destination des courses valides et les zones à l’étude. Le nombre de points par zone est alors identifié et une gradation des couleurs des zones peut alors être établie selon le nombre de points contenus dans la zone.

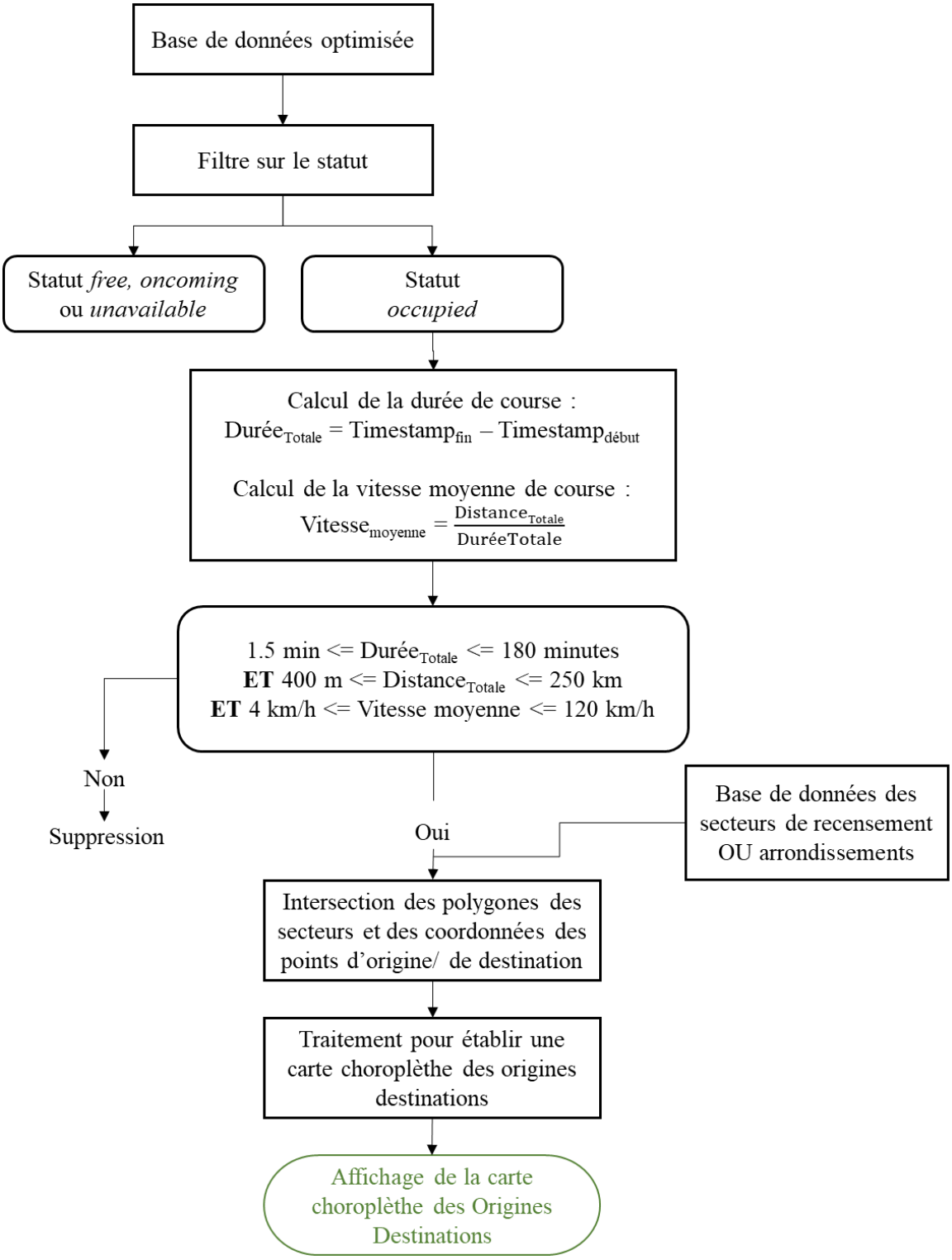


Figure 5-19 Schéma méthodologique de la création d'une carte choroplèthe des origines et destinations

CHAPITRE 6 PRESENTATION DE LA PLATEFORME DE VISUALISATION

Le principal objectif de ce projet de recherche est de concevoir un tableau de bord permettant d'assurer le suivi et l'analyse des indicateurs d'offre et de demande de déplacements par taxi.

La méthodologie développée précédemment dans les Chapitre 3 et Chapitre 4 établit les processus technologiques nécessaires au calcul des indicateurs de performance et de suivi qui sont présentés dans le Chapitre 5. S'il est essentiel que ces indicateurs soient correctement calculés, il est tout aussi nécessaire qu'ils soient correctement représentés dans le tableau de bord. En respectant les règles de visualisation mises en évidence dans la revue de littérature au Chapitre 2, le présent chapitre décrit comment ces indicateurs pourraient être présentés dans le tableau de bord afin d'assurer une consultation intelligible des données.

Dans un premier temps, la mise à jour de certains des processus de traitement lors de leur intégration au tableau de bord est détaillée. Notamment, la question de la fréquence d'importation des données est abordée. Puis, des recommandations de design permettant d'assurer une bonne visualisation de l'information sont suggérées. Enfin, la structure du tableau de bord est présentée : de la page d'accueil de faits saillants aux distributions plus détaillées.

6.1 Intégration des processus au tableau de bord

Dans un premier temps, afin de pouvoir analyser et traiter les données, seuls des échantillons de ces dernières ont été importés. Des méthodes de traitement des données ont été établies à partir de ces échantillons. Les règles de validation des courses ont par exemple été établies à l'aide des données du mois d'avril 2019, tel que détaillé à la section 4.4. Si certaines hypothèses sont émises, toutes les règles ne sont cependant pas fixes. En effet, rien de garantit que les courses observées au mois d'avril sont représentatives des courses de tous les autres mois de l'année. Il ne serait donc pas pertinent d'appliquer exactement les mêmes règles de validation. C'est pourquoi ces dernières varient selon les données observées, tel qu'expliqué à la section 4.3.1, lors de l'identification des bornes de distance et durée des courses qualifiées de complètes.

L'un des enjeux de la conception d'un tableau de bord pour un flux de données continu est ici mis en évidence. Les méthodes de traitement qui doivent être automatisées doivent surtout rester

valables dans le temps. Elles doivent s'adapter à d'éventuelles évolutions des données et des comportements observés.

6.1.1 Importation quotidienne

Le flux de données étant continu, les données doivent donc être importées au fur et à mesure dans la plateforme. Dans le cadre du tableau de bord, il a été établi que les données seraient importées quotidiennement et non en continu. En effet, puisque le tableau de bord conçu n'est pas une plateforme pour l'analyse en temps réel, il n'est pas nécessaire de récupérer les données à chaque 10 minutes.

Ainsi chaque journée de données sera importée une fois la journée complétée. Par exemple, les données de toute la journée du mardi seront importées dans la plateforme le mercredi matin, celles du mercredi le jeudi matin et ainsi de suite. Le script d'importation des données est donc lancé quotidiennement afin de récupérer les données du jour précédent.

6.1.2 Mise à jour hebdomadaire

Tel que mentionné précédemment, le tableau de bord n'étant pas une plateforme pour l'analyse en temps réel, les données en temps réel ne seront donc pas disponibles.

La plateforme sera mise à jour de manière hebdomadaire. Ainsi, si les données sont importées quotidiennement, les processus de traitement de ces données ne sont pas appliqués directement. En effet, tel que mentionné à la section 4.3.1, afin de déterminer les courses régulières il est nécessaire de déterminer des valeurs seuils à partir de la population de référence de la semaine de données dont on souhaite déterminer les courses. Les courses seront donc construites par semaine. Ainsi, pour l'étude d'une semaine, les données du dimanche précédent la semaine, celles de chaque jour de la semaine considérée (du lundi au dimanche) et celles du lundi de la semaine suivante, seront requises. En effet, des courses peuvent commencer le dimanche soir et se terminer le lundi matin. Neuf jours de données sont donc récupérés pour identifier les courses d'une semaine donnée et donc celles de chaque jour de cette semaine.

Les processus de regroupement des points en groupe de statut sont donc appliqués à chaque semaine, une fois l'importation des données de tous les jours de la semaine terminée. Ils sont donc appliqués au courant de la semaine suivant celle dont les données sont traitées.

Une fois cette semaine traitée, soit lorsque la base de données optimisées de la semaine entière est complétée, le choix de cette semaine en tant que période d'analyse est alors rendu disponible sur le tableau de bord. Par exemple, si un utilisateur consulte le tableau de bord au courant de la première semaine du mois de juillet, il ne pourra choisir cette même semaine comme période d'analyse pour les indicateurs. En effet, les données disponibles sur le tableau de bord s'arrêtent à l'avant dernière semaine de juin ou à la dernière semaine de juin selon si le processus de traitement de cette dernière est terminé.

6.1.3 Validations

Enfin, dans un objectif de vérification de la qualité des données et afin d'assurer un contrôle des données qui sont importées, des validations de ces dernières sont mises en place. Si une anomalie est détectée dans les données importées, un message d'erreur peut alors être envoyé afin d'informer les individus gérant le tableau de bord.

Par exemple, une validation des statuts de taxi est implémentée. En effet, ces derniers ne peuvent normalement prendre que quatre valeurs différentes : *free, oncoming, occupied* et *unavailable*. Si une valeur différente de ces quatre valeurs est enregistrée, le processus de validation mis en place peut indiquer qu'une valeur inconnue a été relevée. Ainsi, pour chaque attribut des données, des règles de validation sont établies.

De plus, un attribut intitulé « sourceID » est ajouté aux données importées dans la plateforme. Cet attribut indique d'où proviennent les données. En effet, si pour le moment, seules les données du Registre sont disponibles, il est possible que les flux des données de transport adapté, de taxi collectif ou des courses hospitalières soient rendues disponibles. Il est donc important de préciser la source des données. En effet si les différentes sources de données ne présentent pas le même format ou les mêmes règles, il est essentiel de pouvoir différencier les données afin de les traiter séparément. Pour l'instant cet identifiant indique : « BTM » puisque les données disponibles proviennent uniquement du Registre des taxis.

Il est important de prévoir lors de la conception de la plateforme les éventuelles modifications ou options qui seront par la suite rajoutées. Il est enfin plus simple de prévoir et d'anticiper les évolutions plutôt que de devoir effectuer des modifications ultérieurement.

Enfin, certaines données sont converties afin de limiter l'espace requis pour le stockage des données. Ainsi, le champ « operator » qui est une chaine de caractères correspondant au courriel de l'intermédiaire en service auquel est rattaché le taxi, est converti en en identifiant unique. En effet, cet identifiant unique sous forme de nombre occupe moins d'espace mémoire. Dès qu'un nouvel opérateur est détecté dans les données un nouvel identifiant est alors généré. Par exemple, si la valeur identifiée dans le champ « operator » est atlas@taxi.ca, au lieu de conserver cette valeur telle quelle dans la base de données optimisée, l'identifiant unique correspondant à cet intermédiaire (par exemple : 1) sera enregistré et seule la correspondance des valeurs distinctes des attributs à leur identifiant unique est stockée.

6.2 Charte de design

Tel que mentionné dans la revue de littérature dans la section 2.2 sur la visualisation des données, certaines pratiques sont recommandées pour assurer une visualisation pertinente de l'information.

Quelques recommandations qui ont été suivies pour le design de l'interface sont donc proposées :

- Il est suggéré d'utiliser l'orange et le bleu lorsqu'une comparaison à l'aide de couleurs est requise. En effet, la différence entre ces deux couleurs est perceptible même par ceux atteints de troubles de la vision des couleurs.

  Sinon, une palette de couleur adaptée à ceux souffrant de troubles de la vision des couleurs pourrait être implémentée et affectée à un bouton du tableau de bord permettant de mettre à jour les couleurs du tableau de bord si sélectionné.

- Il est suggéré de ne pas utiliser plus de cinq nuances d'une certaine couleur. Ainsi, pour les cartes choroplèthes pas plus de 5 classes doivent être établies. En effet au-delà de 5 nuances, il devient compliqué de distinguer les différentes nuances sans fournir un effort ou sans examiner avec minutie ce qui est contraire à l'objectif d'une visualisation pertinente et efficace.

  Aussi, il est préférable de choisir une couleur et d'en faire varier son intensité plutôt que d'utiliser des couleurs différentes, car même ceux atteints de troubles de la vision des couleurs peuvent détecter des intensités distinctes de la même couleur.

- Il est recommandé d'utiliser des diagrammes en barres pour les distributions. En effet, la longueur est l'un des attributs de pré-attention les plus efficaces à traiter. Ainsi, les différences de hauteur ou longueur entre les différentes barres d'un tel diagramme sont traitées et analysées rapidement par l'utilisateur.
- Il est fortement recommandé de ne pas utiliser de diagramme circulaire (« pie chart »). En effet, ce type de représentation déforme souvent les informations et nous force à comparer des zones ou des angles, ce qui est complexe et non évident.
- Pour les mêmes raison les visualisations et graphiques en 3D sont à éviter puisqu'ils surchargent les représentations sans ajouter d'information utile.
- Finalement, tout élément visuel des schémas et des graphiques qui ne sont pas essentiels pour comprendre les informations qui y sont présentées et qui détournent l'attention de l'observateur ("déchets graphiques" ou « chart Junk »), doivent être évités, tels que les bordures et arrières plans.

6.3 Structure de la plateforme

Dans le cadre de l'étude, le tableau de bord est intitulé **O**util de **C**onsultation du **T**axi pour l'**A**nalyse et la **V**isualisation d'**I**ndicateurs, soit **OCTAVI**.

Le tableau de bord (OCTAVI) est encapsulé dans la plateforme Transition de la Chaire Mobilité (Transition, 2020). La structure générale du tableau de bord est donc définie par celle de la plateforme globale. Par exemple, le même thème sombre (« dark ») doit être utilisé et la segmentation de la fenêtre de visualisation est également fixée. Cette dernière est composée de 4 parties illustrées dans la Figure 6-1 :

- 1 : « Top Menu » : le menu supérieur permet de se déconnecter de la plateforme ainsi que de choisir la langue de la plateforme. Il est pour l'instant uniquement possible de choisir entre l'anglais et le français.
- 2 : « Left Menu » : le menu de gauche permet de naviguer entre les différentes vues proposées et notamment de retourner à la page d'accueil.
- 3 : « Full Size Panel » : le panneau principal présente les différents indicateurs.

- 4 : « Right Panel » : le panneau de droite permet de choisir les filtres tels que la sélection de la période d'analyse, de la période temporelle ou des objets d'études. Ces filtres, introduits à la section 5.1.4, sont illustrés à la section 6.3.2.

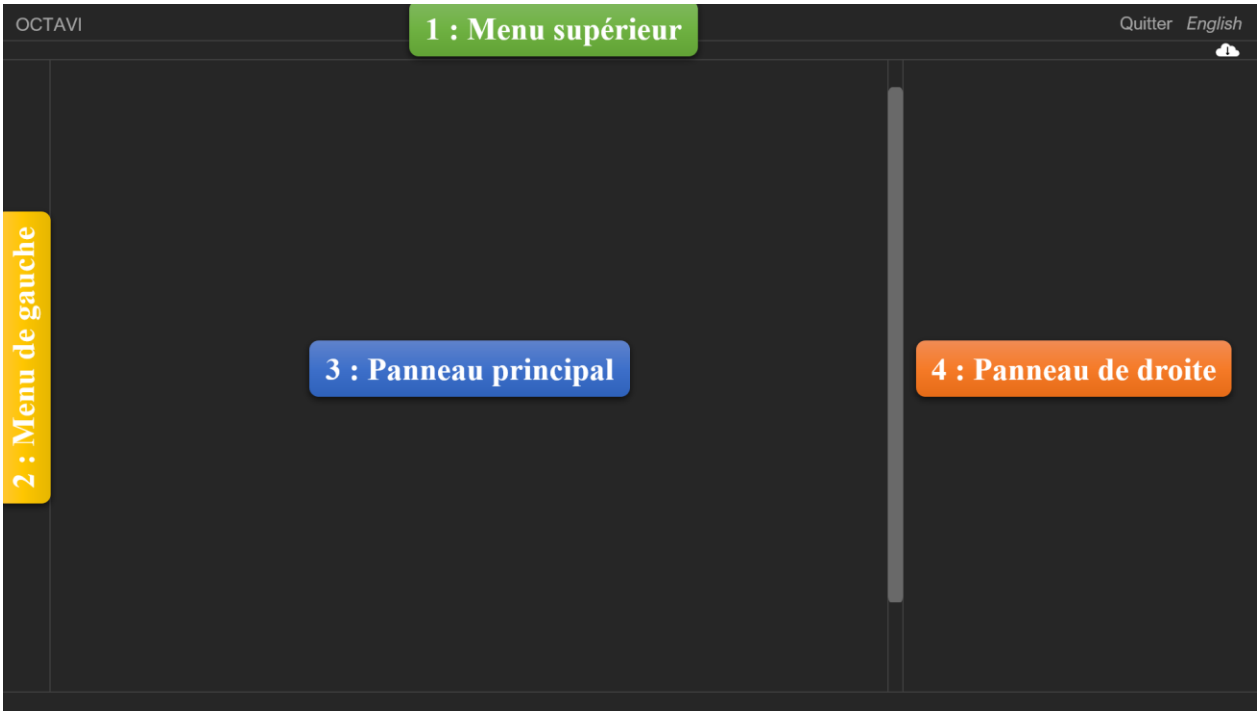


Figure 6-1 Maquette illustrant l'organisation de la fenêtre du tableau de bord

La structure du tableau de bord proposée est en forme d'entonnoir : du plus global au plus détaillé. La page d'accueil est une page de faits saillants. Ces faits saillants correspondent à des indicateurs qui sont tous sous la forme d'une statistique descriptive. Si l'utilisateur souhaite obtenir plus d'informations, il peut cliquer sur la statistique descriptive et des graphiques plus détaillés sont alors disponibles. Ces derniers sont des graphiques interactifs où l'utilisateur peut sélectionner la période d'analyse et/ou la zone d'analyse ainsi que d'autres paramètres. En permettant à l'utilisateur de naviguer vers des écrans distincts ou vers différentes instances d'un même écran pour accéder à des informations supplémentaires, il peut approfondir son analyse si nécessaire. Mais si seule une vue d'ensemble est nécessaire, alors la page des faits marquants répond au besoin. Cette hiérarchie de l'accès à l'information, du plus global au plus détaillé, constitue une bonne pratique de visualisation (Few, 2006; MacEachren, 1994).

Avant de pouvoir visualiser la page d'accueil, une authentification est requise pour accéder au tableau de bord. La Figure 6-2 présente une capture d'écran de cette page d'authentification. Un nom d'utilisateur ou courriel ainsi qu'un mot de passe sont requis. Lors de la première connexion il est d'abord nécessaire de créer un compte. Cela permet notamment d'enregistrer les préférences de chaque utilisateur et de les afficher lors de la prochaine connexion de l'utilisateur (par exemple les derniers filtres choisis ou la langue sélectionnée).

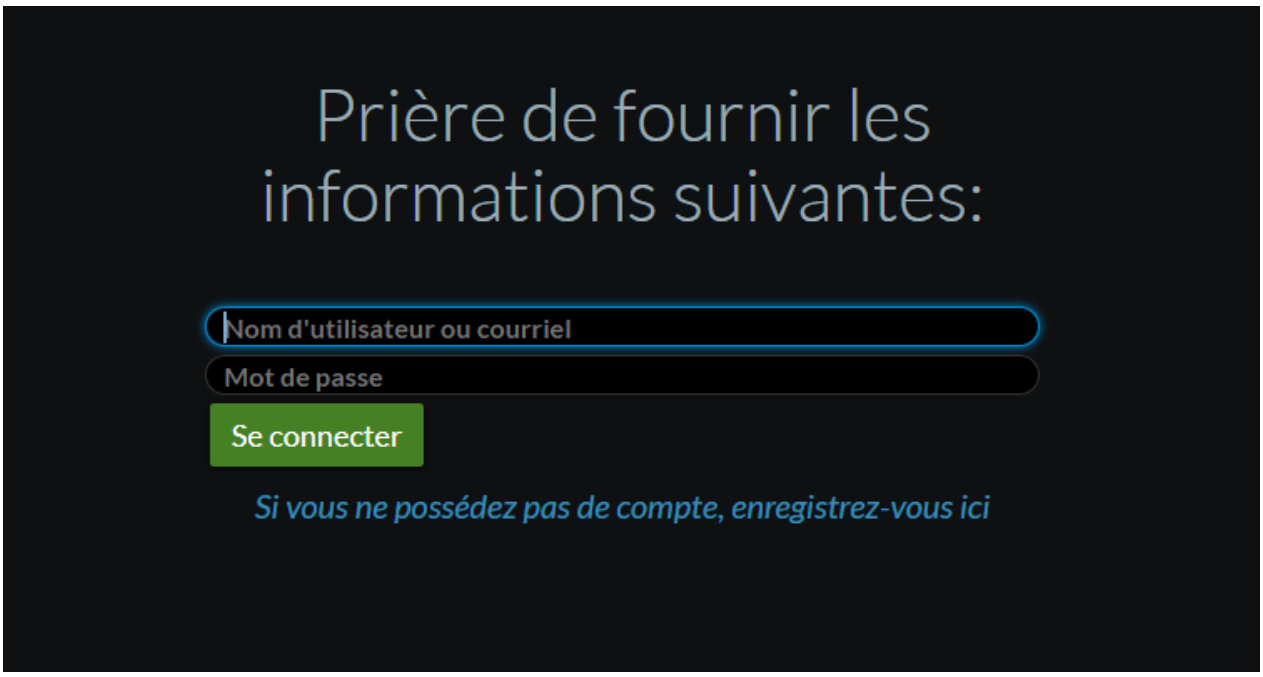


Figure 6-2 Capture d'écran de la page d'authentification

6.3.1 Page des faits saillants

La Figure 6-3 présente la page des faits saillants du tableau de bord. Les faits saillants présentés sont au nombre de 6, correspondant à chaque objet d'étude identifié à la section 5.1.1, à savoir la course, le véhicule, le chauffeur, l'intermédiaire en service, le poste d'attente et les origines-destinations des courses.

En utilisant le même type de représentation pour la page des faits saillants, soit uniquement des statistiques descriptives (sous la forme de gros chiffres), on assure une cohérence aux utilisateurs.

Ces derniers peuvent ainsi utiliser la même stratégie perceptive pour interpréter les données, permettant ainsi un gain de temps et d'énergie, selon Stephen Few dans *Information Dashboard Design* (Few, 2006).

Dans ce premier niveau, seule la période temporelle d'analyse peut être modifiée par l'utilisateur. La sélection de la période d'analyse se fait dans le panneau de droite de la fenêtre de visualisation. Le mois d'avril 2019 au complet est ici sélectionné. Le calcul des indicateurs présentés s'est donc fait sur ce mois. Ainsi, 441 833 courses régulières ont été effectuées au cours du mois d'avril 2019. Tel que mentionné dans la section 5.2.1.2, seule une partie des courses réalisées par les taxis peut être identifiée. Il est donc précisé que ce sont uniquement les courses régulières qui sont calculées. Parmi ces courses, 431 701 avaient une destination sur l'île de Montréal. Pendant cette période, 2673 véhicules et 3076 chauffeurs étaient actifs, c'est-à-dire qu'ils ont été en activité et ce sont donc connectés au Registre au moins une fois dans le mois. Le nombre de chauffeurs actifs étant plus élevé que celui des véhicules, cela illustre le fait qu'un véhicule est partagé par plusieurs chauffeurs. Au cours de ce même mois, 347 postes d'attente ont été visités au moins une fois, c'est-à-dire qu'au moins une des places de ces postes d'attente a été occupée par un véhicule en attente d'un prochain client. Ce premier niveau offre donc un aperçu général de l'offre et de la demande en déplacements de taxi.

Figure 6-3 Page des faits saillants du tableau de bord OCTAVI

En raison de l'ampleur des données, il est essentiel que le tableau de bord puisse permettre la synthèse des résultats et offrir un aperçu global des données. Mais il est également nécessaire de permettre aux utilisateurs une exploration flexible des données afin de passer des résumés agrégés à des analyses plus désagrégées. Il est donc possible d'affiner les requêtes grâce à une sélection et une analyse spatio-temporelle à différentes échelles. Ces échelles présentées à la section 5.1.4, sont intégrées au tableau de bord sous la forme de filtres qui peuvent être sélectionnés par l'utilisateur.

6.3.2 Filtres spatio-temporels

La Figure 6-4 présente les trois catégories de filtres disponibles. La première (a), concerne les objets : il est possible d'affiner l'analyse des indicateurs présentés par véhicule, par chauffeur ou par intermédiaire. Il est également possible de les combiner. La zone d'analyse (b) peut aussi être affinée par secteur de recensement, municipal ou par arrondissement. La plus fine résolution est les secteurs de recensement et la plus grande l'ensemble du territoire de l'île de Montréal. Enfin, la période temporelle d'analyse (c) peut être spécifiée au niveau de l'heure, du jour, de la semaine, du mois ou de l'année.

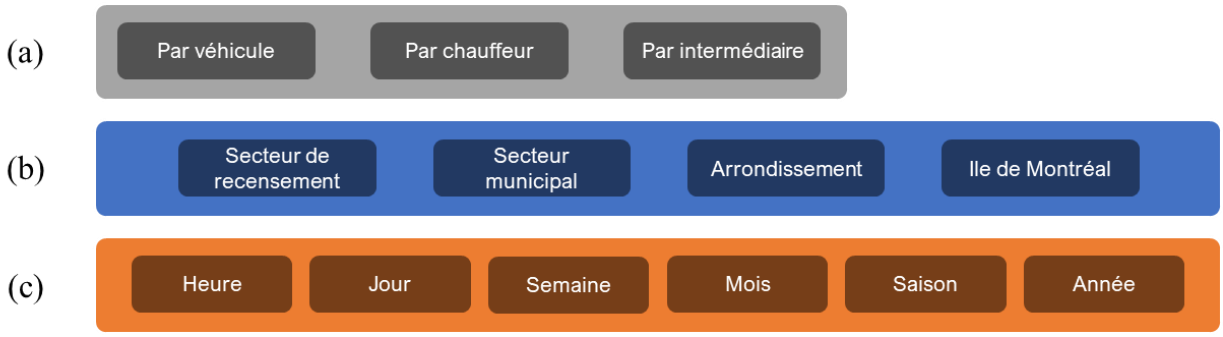


Figure 6-4 Les trois catégories de filtres : (a) par objet ; (b) spatiaux ; (c) temporels

La Figure 6-5 présente le détail des différents niveaux d'échelle temporelle. En effet, si l'utilisateur souhaite affiner l'analyse selon l'heure (b) plusieurs choix s'offrent à lui. Il peut soit analyser l'ensemble des 24 heures de la journée, soit par groupes d'heures : de 0 à 6 heures, de 6 à 9 heures, de 9 à 15 heures, de 15 à 18 heures et de 18 heures à minuit. Plusieurs périodes peuvent être sélectionnées simultanément. Par exemple, si l'utilisateur souhaite étudier les périodes de pointe du matin et de l'après-midi, il peut sélectionner respectivement les périodes 6 à 9h et 15 à 18h.

L'analyse du jour (c) peut être faite selon les jours de la semaine, du lundi au dimanche. L'ensemble des jours de la semaine peut être sélectionné ou uniquement certains jours. De plus, il est aussi possible de faire l'analyse selon les jours ouvrés de la période d'analyse sélectionnée ou selon les fins de semaine. Enfin, l'analyse peut être faite selon les groupes de mois correspondant aux quatre saisons (d) de l'année.

Les différents filtres qui peuvent être sélectionnés par l'utilisateur varient selon la période d'analyse sélectionnée. Par exemple, si l'utilisateur choisit comme période d'analyse le mois d'avril 2019, seuls les filtres sur les heures, les jours et la semaine seront disponibles. S'il choisit comme période d'analyse les mois d'avril et de mai 2019, il lui sera également possible de faire des analyses moyennes sur le mois. Enfin, s'il sélectionne l'année 2019, alors les filtres sur les saisons seront disponibles et il lui sera également possible d'effectuer des moyennes sur l'année au complet.

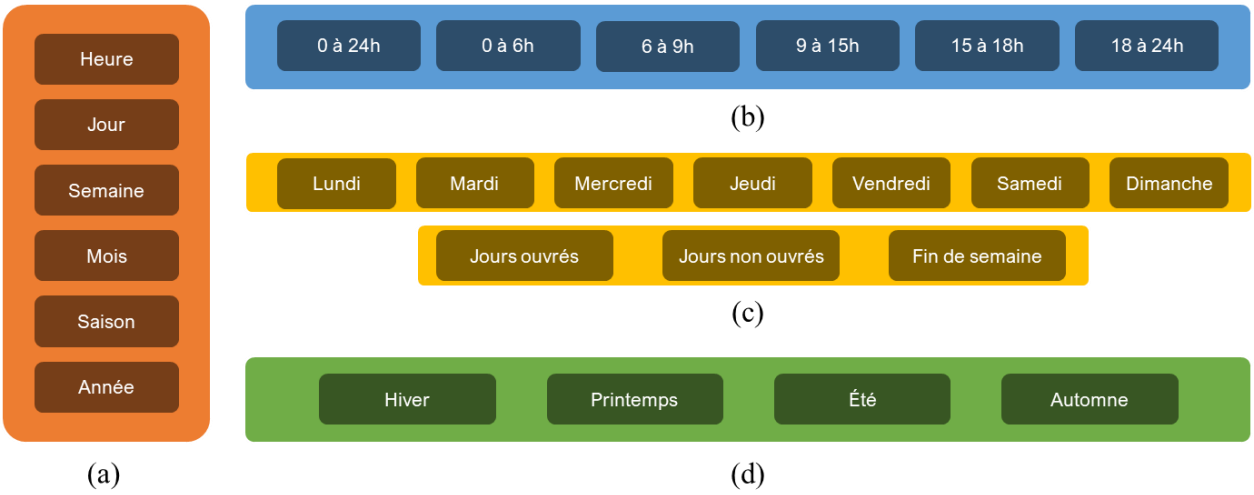


Figure 6-5 Détails des filtres temporels : (a) premier niveau ; (b) détail des groupes d'heures ; (c) détail des types de jours ; (d) détail des saisons

6.3.3 Les différents niveaux de détail

Afin d'accéder aux différents niveaux de visualisation, l'utilisateur doit cliquer sur une des six statistiques descriptives présentées sur la page d'accueil, soit la page des faits saillants. En effet, pour chaque objet présenté, d'autres vues plus détaillées sont disponibles. Ainsi, si l'utilisateur souhaite avoir plus de précision sur le nombre de courses régulières, et donc sur l'objet course en général, il lui suffit de cliquer sur la statistique correspondante et une nouvelle vue lui est alors présentée. Cette vue correspond au second niveau de visualisation.

Cette hiérarchie des indicateurs est détaillée dans la suite en prenant pour exemple l'objet course.

6.3.3.1 Exemple pour l'objet Course

Le second niveau de visualisation et d'analyse est atteint lorsque l'utilisateur souhaite obtenir davantage de détails sur un des faits saillants du premier niveau et qu'il clique sur ce dernier. D'autres statistiques descriptives sont présentées, tel qu'illustré dans la Figure 6-6. Le nombre de courses en moyenne par jour pour la période d'analyse sélectionnée ainsi que la durée, distance et vitesse moyenne de course peuvent être visualisées. Pour le mois d'avril 2019, ce sont 14 728 courses qui sont réalisées en moyenne par jour. Et ces courses durent en moyenne 13,9 minutes pour une distance moyenne parcourue de 5,6 km et une vitesse moyenne de 23 km/h. L'écart-type peut également être visualisé afin d'évaluer la dispersion autour des moyennes calculées.

(a)

(b)

Figure 6-6 2ème niveau de visualisation : (a) pour l'objet course ; (b) visualisation des écart-types

Cette analyse peut alors être affinée selon les filtres disponibles. La Figure 6-7 présente les indicateurs filtrés selon le groupe d'heures 6 à 9h, soit selon la période de pointe du matin. Ainsi, pour le mois d'avril 2019 et durant la période de pointe du matin, environ 1974 courses sont réalisées en moyenne par jour. Ces courses durent en moyenne 15.8 minutes pour une distance parcourue moyenne de 6,3 km environ et une vitesse moyenne de 23.4 km/h.

Figure 6-7 2ème niveau de visualisation pour l'objet course : analyse selon le groupe d'heures 6 à 9h

La Figure 6-8 présente les indicateurs filtrés selon le groupe d'heures 15 à 18h, soit selon la période de pointe de l'après-midi. Ainsi, pour le mois d'avril 2019 et durant la période de pointe de l'après-midi, environ 2405 courses sont réalisées en moyenne par jour. Ces courses durent en moyenne 15.1 minutes pour une distance parcourue moyenne de 5,3 km environ et une vitesse moyenne de 19.9 km/h.

Figure 6-8 2ème niveau de visualisation pour l'objet course : analyse selon le groupe d'heures 15 à 18h

Si l'utilisateur souhaite obtenir encore plus de précision et affiner son analyse, un troisième niveau de visualisation est disponible. Une fois de plus, il lui suffit de cliquer sur la statistique dont il souhaite obtenir plus d'informations afin d'atteindre le troisième niveau de visualisation. Par exemple, si l'utilisateur souhaite poursuivre l'analyse sur le nombre de courses, il peut cliquer sur la statistique descriptive présentant le nombre moyen de courses réalisées par jour. Des distributions temporelles et fréquentielles liées au nombre de courses lui sont alors présentées, telles qu'illustrées sur la Figure 6-9. Un premier histogramme illustre le nombre de courses réalisées chaque jour pour la période d'analyse sélectionnée, soit le mois d'avril 2019. Un second histogramme présente le nombre de courses qui ont été réalisées par les véhicules de taxi durant cette période. Et un troisième histogramme présente le nombre de courses réalisées par les chauffeurs de taxi.

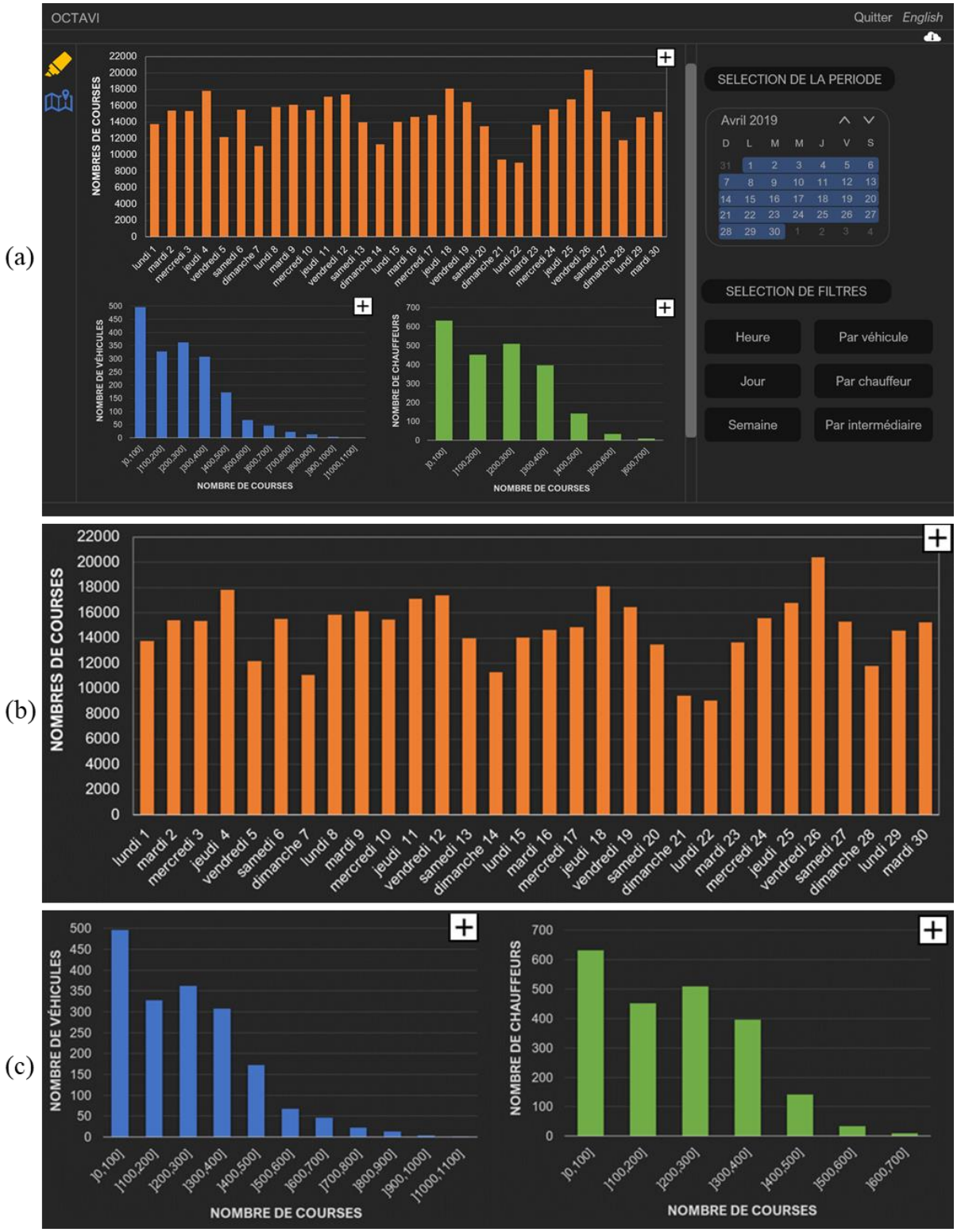
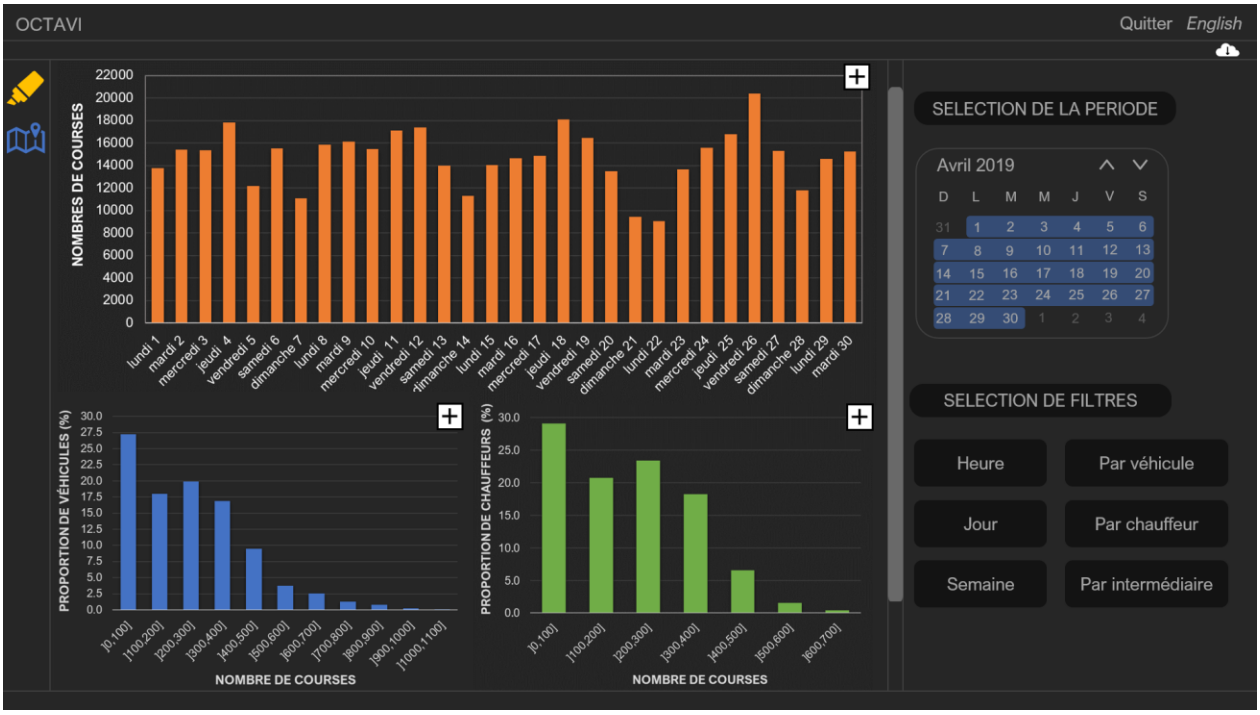


Figure 6-9 3ème niveau de visualisation pour l'objet course : (a) distributions liées au nombre de courses régulières ; (b) zoom sur la distribution temporelle ; (c) zoom sur les distributions fréquentielles

L’icône « plus » située dans le coin supérieur droit des distributions permet de sélectionner le type de calcul à afficher : soit le nombre total ou le pourcentage. La Figure 6-10 présente ces mêmes distributions lorsque le pourcentage est sélectionné par l’utilisateur pour la deuxième et troisième distribution. Ainsi, près de 27,5% des véhicules de taxi et environ 30% des chauffeurs ont effectué entre une et cent courses durant le mois d’avril 2019.



(a)

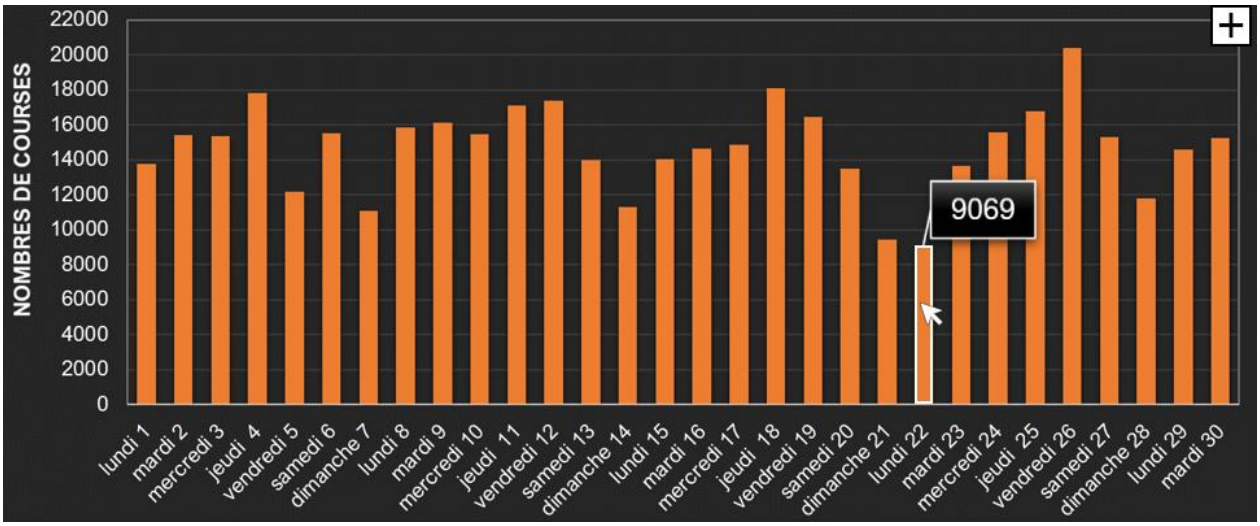
(b)

Figure 6-10 3ème niveau de visualisation pour l'objet course : (a) pourcentage sélectionné pour les distributions fréquentielles; (b) zoom sur les distributions fréquentielles

La Figure 6-11 présente un exemple d'interactivité du tableau de bord. Lorsque l'utilisateur déplace le curseur de sa souris sur l'une des barres de valeurs des histogrammes de distribution, la valeur correspondante s'affiche sous forme d'un « pop-up ». Dans l'exemple présenté, l'utilisateur

souhaite connaitre le nombre de courses réalisés lors de la journée du lundi 22 avril. La barre de la journée est mise en évidence et la valeur de 9069 courses peut être visualisée dans le « pop-up » correspondant.

(a)



(b)

Figure 6-11 Exemple d'interactivité du tableau de bord : (a) affichage des valeurs des barres ; (b) zoom sur la distribution temporelle

Ainsi, trois niveaux de visualisation sont disponibles pour l'objet d'étude. Le premier niveau (page des faits saillants) présente une statistique descriptive générale de l'objet. Puis, dans le deuxième niveau, cette statistique est déclinée en plusieurs autres statistiques descriptives. Enfin, le troisième niveau présente des distributions fréquentielles et temporelles pour chacune des statistiques du deuxième niveau. Cette hiérarchie de visualisation présentée pour l'objet course s'applique également aux objets véhicule, chauffeur, intermédiaire et poste d'attente.

Enfin, deux icônes sont présentes sur le menu de gauche du tableau de bord. La première icône, en forme de surligneur, permet de regagner la page des faits saillants (« highlights »). Ainsi, si l'utilisateur se trouve au 3ème niveau de visualisation et qu'il souhaite retourner au premier niveau, il lui suffit de cliquer sur cette icône. La seconde icône, représentant une carte permet quant à elle d'atteindre la carte des origines et destinations.

6.3.3.2 Origines et Destinations

L'analyse de l'objet des origines et destinations diffère puisqu'une analyse spatiale est proposée. La Figure 6-12 présente la carte choroplèthe des destinations des courses régulières pour le mois d'avril 2019. L'objet destination a été sélectionné dans le panneau de droite ainsi que les secteurs de recensement. Le nombre de destinations est donc calculé par secteur de recensement. Les cinq classes sont déterminées selon le mode de classification de ruptures naturelles (« Jenks »). Ce mode permet de regrouper en classes en minimisant l'écart à l'intérieur des classes et en maximisant l'écart entre les classes (Laviolette, 2017). Enfin, l'icône « plus » située dans le coin supérieur droit de la carte permet de sélectionner le type de calcul à afficher : soit le nombre total ou la densité (nombre d'origines ou de destinations par km^2).

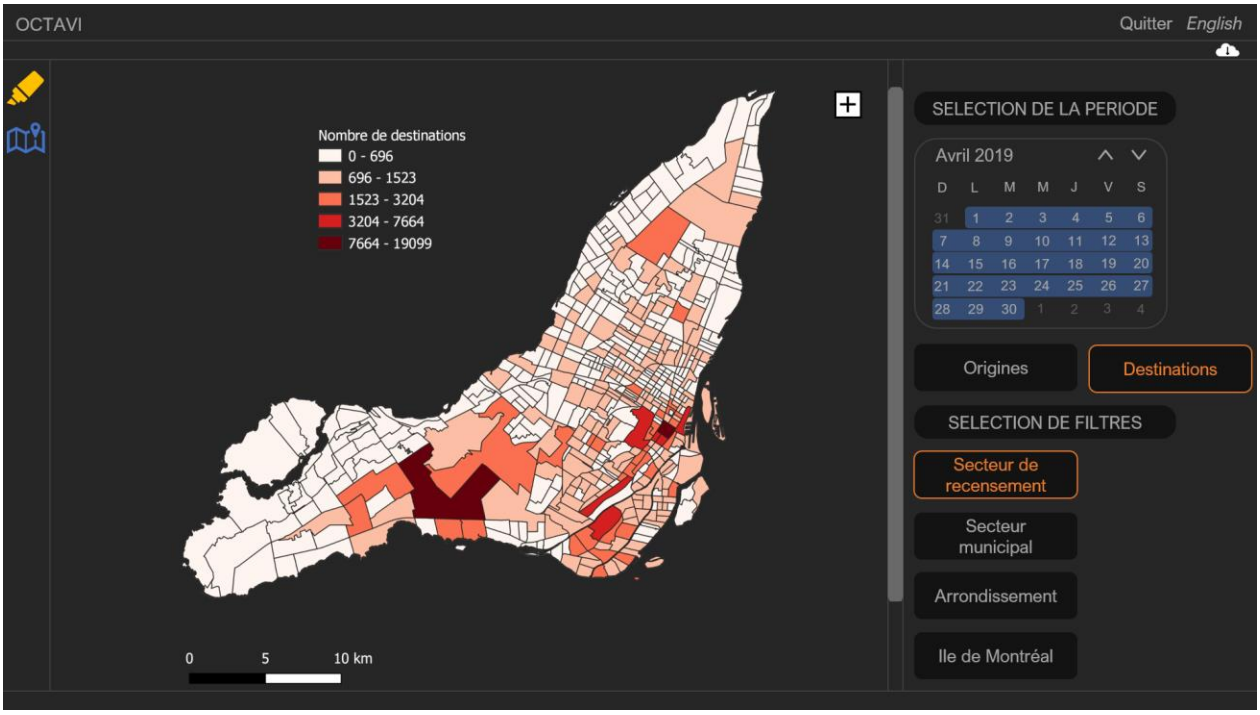


Figure 6-12 Carte du nombre de destinations

Il est possible pour l'utilisateur de zoomer sur la carte afin d'assurer une meilleure visualisation des différents secteurs. Il peut également obtenir la valeur du nombre de destinations (ou d'origines selon le filtre sélectionné) pour chaque secteur en déplaçant le curseur de la souris sur le secteur en question, tel qu'illustré à la Figure 6-13. La carte présente une zone zoomée des secteurs du centre-ville de l'île de Montréal et le nombre de destinations du secteur présentant le plus de destinations y est indiqué dans un pop-up.

Figure 6-13 Carte du nombre de destinations zoomée sur le centre-ville de Montréal

6.3.4 Volet comparatif

Un volet comparatif permet également de comparer sur la même fenêtre de visualisation un ou plusieurs indicateurs selon deux périodes d'analyses distinctes. Ce volet d'analyse est disponible pour tous les indicateurs et à chacun des trois niveaux de visualisation.

La Figure 6-14 présente la comparaison des faits saillants (1er niveau) des mois d'avril 2019 et 2020. L'impact de la pandémie de Covid et du confinement sont ainsi mis en évidence puisque le nombre de courses régulières du mois d'avril 2020 est réduit de près de 61% par rapport à celui d'avril 2019. De même, une baisse d'environ 46% des véhicules actifs et de 51% des chauffeurs actifs est observée, malgré la présence de trois nouveaux intermédiaires en service.

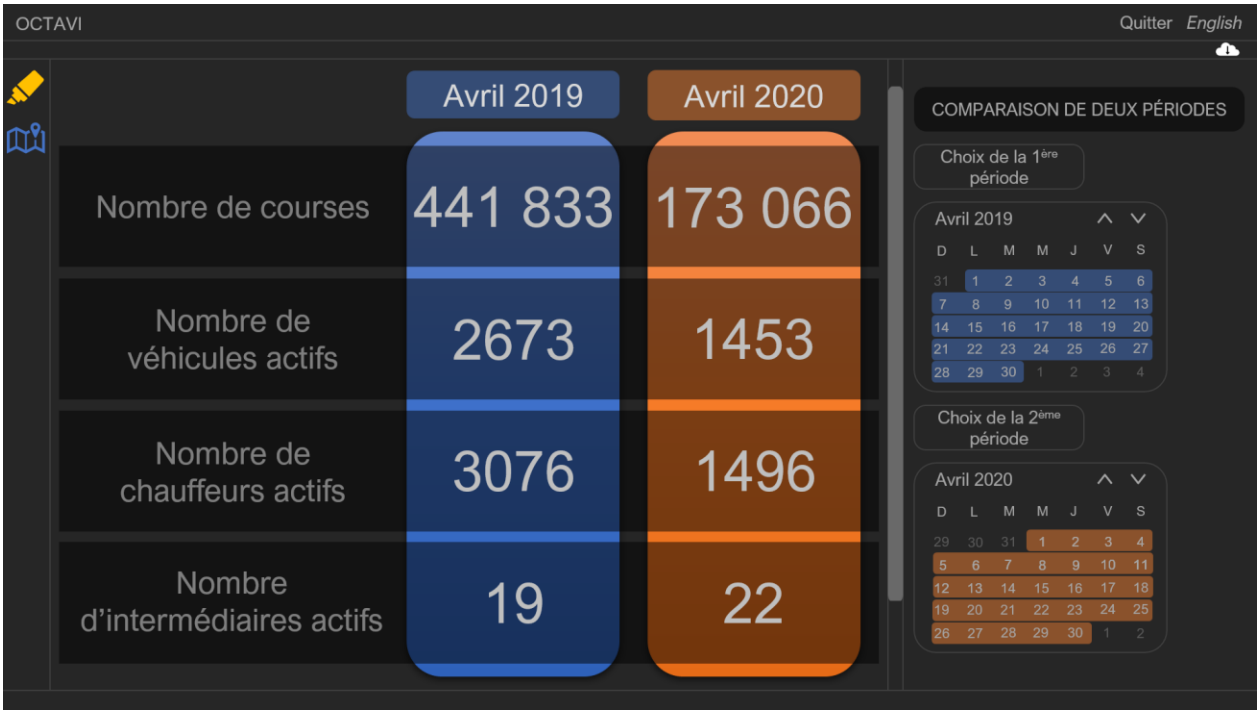


Figure 6-14 Comparaison des faits saillants pour les mois d'avril 2019 et 2020

La Figure 6-15 présente une comparaison pour le 2ème niveau de visualisation. Les statistiques descriptives des courses régulières du mois d'avril 2019 réalisées en période de pointe de matin et de l'après-midi peuvent être comparées. Plus de courses sont réalisées en période de pointe de l'après-midi. De plus, si durant cette période la durée de la course est relativement semblable à celle de la pointe du matin, la distance moyenne parcourue est quant à elle plus faible (et donc la vitesse moyenne observée également), témoignant de la congestion plus élevée des axes routiers en après-midi liée au retour à la maison.

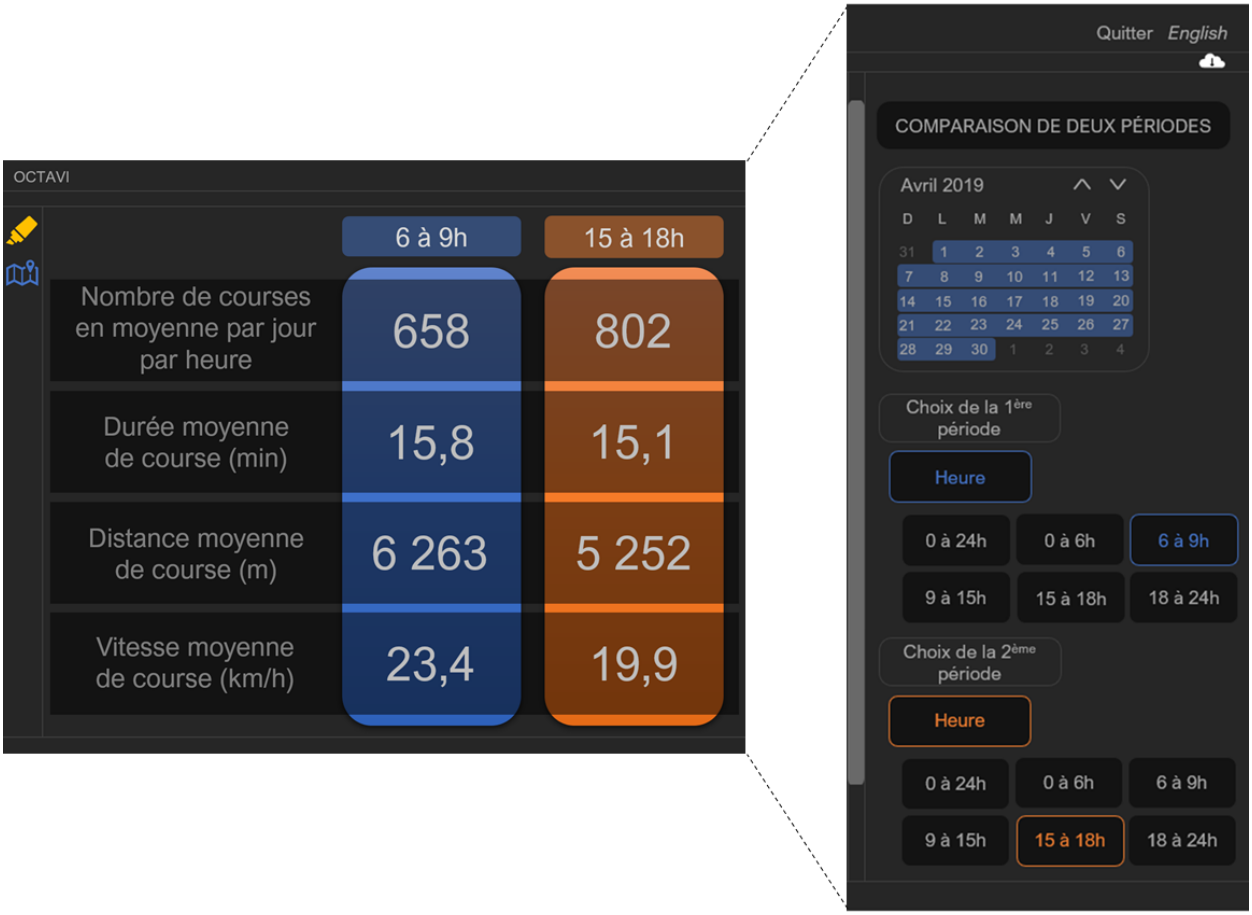


Figure 6-15 2ème niveau de visualisation : Comparaison des statistiques descriptives des courses régulières des périodes de pointe du matin et de l'après-midi

Enfin, la Figure 6-16 illustre ce volet comparatif dans le cas des cartes des origines et destinations. La période d'analyse choisie est la période de pointe du matin (6 à 9h) de la journée du jeudi 4 avril 2019. Les cartes présentant la densité des origines (à gauche) et celle des destinations (à droite) sont représentées côte à côte pour permettre la comparaison. Le centre-ville de Montréal est à la fois un lieu de production et d'attraction des déplacements en taxi durant la période de pointe du matin.

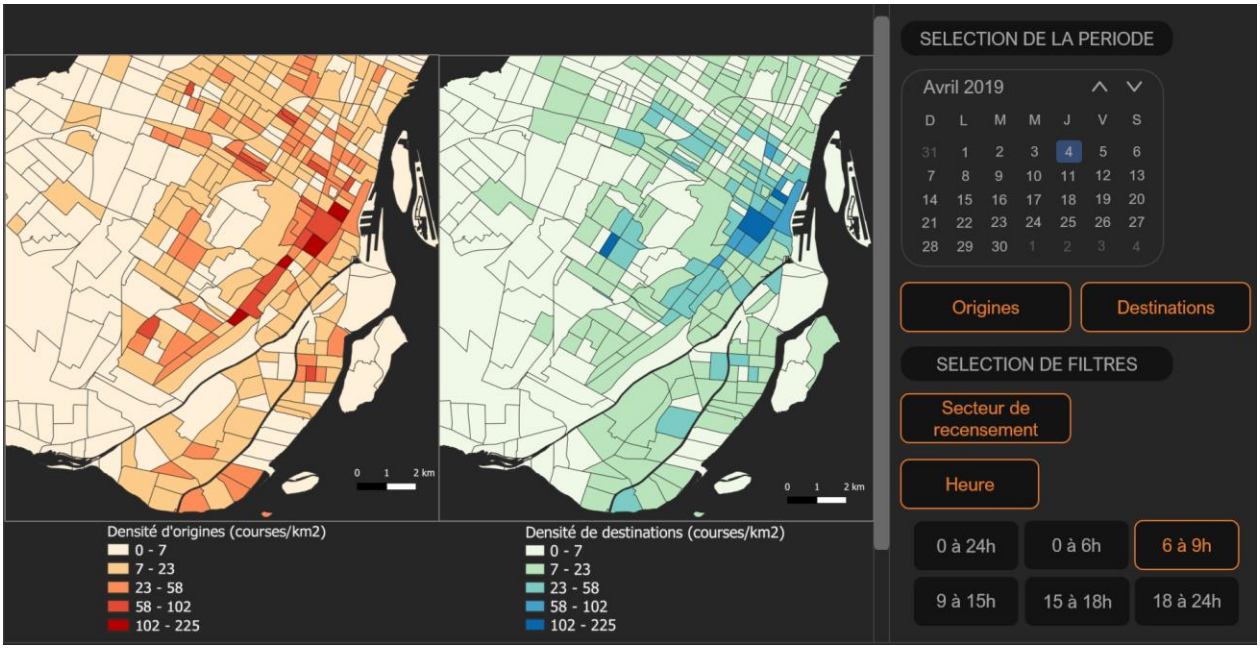


Figure 6-16 Comparaison des densités d'origines et de destinations pour la journée du 4 avril 2019 lors de la pointe du matin

Finalement, pour tous les indicateurs et visualisations, il est possible d’exporter les données des requêtes sous forme de fichiers de différents formats. Les distributions peuvent ainsi être exportées sous forme d’image ou d’un document PDF. Et les données utilisées dans ces distributions peuvent être exportées sous forme de fichier CSV ou d’un fichier EXCEL afin d’être analysées à l’aide d’autres outils.

CHAPITRE 7 CONCLUSION ET RECOMMANDATIONS

Ce dernier chapitre conclut ce projet de recherche dont l'objectif est le développement d'une plateforme de consultation et d'analyse de l'industrie du taxi. Cette synthèse se décline en quatre parties. Dans un premier temps, les méthodes établies pour atteindre l'objectif ainsi que les principaux résultats de ce mémoire sont présentés. Puis, les limitations de la méthodologie et des analyses effectuées ainsi que les contributions apportées au sujet sont détaillées. Enfin, les perspectives de recherche sur l'amélioration et l'utilisation de l'outil de visualisation sont mises en évidence.

7.1 Synthèse de la recherche

Tout d'abord, la revue de littérature a permis de mettre en évidence les enjeux liés à la conception d'une plateforme de visualisation et d'analyse de l'industrie du taxi. Pour cela, la première partie de la revue s'est intéressée au secteur des taxis. Un rappel des évènements marquants de l'histoire du taxi ainsi que la définition du service ont d'abord été énoncés afin de mieux appréhender le rôle du taxi dans la mobilité quotidienne et notamment de son interaction avec le transport en commun. Une synthèse des thématiques d'analyses récentes liées à ce secteur a par la suite été menée dans le but de dresser un inventaire complet des connaissances sur le secteur des taxis. Les récentes améliorations des différents modèles, tels que ceux de répartition des courses, de prévision de la demande ou de recherche du prochain client, permises par la disponibilité des données GPS de taxis ont été mises en exergue. Enfin une synthèse des indicateurs pertinents permettant de caractériser l'offre et la demande de déplacements en taxi est réalisée. Dans la deuxième partie de la revue, l'accent a été mis sur la visualisation des données. Le tableau de bord en tant qu'outil de prise de décision a été défini. Enfin les différentes caractéristiques de visualisation essentielles à la conception d'un tableau de bord communiquant les informations de manière efficace et juste ont été mises en avant.

Le troisième chapitre présente la méthodologie de traitement des données. Dans un premier temps, les principaux points de la structure organisationnelle du taxi à Montréal y sont détaillés afin de fournir un contexte aux données utilisées dans le cadre de ce projet. Le Registre des taxis y est ensuite présenté. Grâce à cette plateforme de données ouvertes, il est désormais possible d'avoir un flux continu de données de tous les taxis en service à Montréal. Ce sont ces données GPS, collectées

passivement par les taxis indépendants ou associés à des intermédiaires en service à Montréal, qui sont utilisées dans le présent mémoire. Une description de ces données et notamment des quatre statuts disponibles pour caractériser l'activité d'un taxi en service est alors fournie. Puis, dans l'objectif de pouvoir estimer une variété d'indicateurs, des procédures systématiques et automatisées de traitement des données ont été établies. Les principales étapes du processus général de traitement des données sont résumées : de l'extraction des données brutes au traitement de ces dernières jusqu'à la visualisation des indicateurs. Plus particulièrement, la méthodologie de construction d'une base de données optimisée permettant de minimiser l'espace requis pour le stockage des données et d'accélérer le calcul des indicateurs y est explicitée.

Le quatrième chapitre s'intéresse à la résolution d'un enjeu majeur des données du Registre, à savoir l'irrégularité des intervalles de temps entre deux points consécutifs. En effet, si selon l'ordonnance du BTM les données GPS doivent être envoyées au Registre aux 5 secondes, en réalité des intervalles supérieurs à 5 secondes sont observés. Le chapitre traite donc des méthodes d'identification des groupes de statut correspondant aux activités des taxis. Plus particulièrement, une méthode d'identification des courses régulières se basant sur l'étude d'une population de référence est développée.

Le cinquième chapitre concerne les indicateurs. Dans un premier temps, les principaux indicateurs qu'il serait pertinent d'estimer ont été identifiés à partir de la littérature et des travaux de Lacombe (2016) et Laviolette (2017). Ces indicateurs sont classifiés selon l'objet d'analyse. Six objets d'analyse ont été retenus, à savoir la course, le véhicule de taxi, le chauffeur, l'intermédiaire en service, le poste d'attente et les origines et destinations des courses. Puis les déclinaisons possibles de ces principaux indicateurs ainsi que les diverses formes de visualisation réalisables sont détaillées. Les échelles spatiales et temporelles retenues pour l'analyse sont également présentées. Enfin, les défis méthodologiques liés au calcul de ces indicateurs sont mis en évidence.

Finalement, le sixième chapitre présente le tableau de bord assurant la visualisation et l'analyse des indicateurs d'offre et de demande en déplacements de taxi. En respectant les règles de visualisation mises en évidence dans la revue de littérature au Chapitre 2, ce chapitre décrit comment les indicateurs sont présentés dans le tableau de bord afin d'assurer une consultation intelligible des données. Dans un premier temps, la mise à jour de certains des processus de traitement lors de leur intégration au tableau de bord y est détaillée. Puis la structure du tableau de bord est présentée : de

la page d'accueil de faits saillants à des distributions fréquentielles et temporelles plus détaillées. La structure en entonnoir du tableau de bord assure à l'utilisateur une exploration flexible des données afin de passer des résumés agrégés à des analyses plus désagrégées. Il lui en effet possible d'affiner ses requêtes grâce à la sélection de filtres spatio-temporels. Afin d'illustrer les différents niveaux de visualisation des données du tableau de bord, l'objet course est analysé ainsi que les origines et destinations. Enfin, le volet comparatif permettant de comparer sur la même fenêtre de visualisation un ou plusieurs indicateurs selon deux périodes d'analyses distinctes est explicité. Les faits saillants des mois d'avril 2019 et 2020 y sont comparés, révélant notamment l'impact de la pandémie et du confinement sur l'industrie du taxi.

7.2 Limitations

Plusieurs hypothèses ont été émises afin de réaliser le traitement des données du Registre et l'estimation des indicateurs à partir de ces données. En effet, des limitations liées à la qualité des données et à leur structure ont été identifiées. Or les hypothèses posées influencent la qualité des analyses et des résultats. Il convient donc de rappeler les principales limitations rencontrées au cours de ce projet. Ces dernières sont listées ci-dessous :

1. **Absence de règles de validation dans le Registre**

   Les tables des véhicules, des chauffeurs et des permis présentent de nombreuses entrées de données erronée liées à une erreur humaine et qui auraient pu être évitées si des validations des entrées étaient appliquées. Seule la table des taxis n'en présente pas puisqu'elle est générée automatiquement par le Registre à partir de la combinaison des tables des chauffeurs, des véhicules et des permis.

   Ainsi les numéros de permis sont par exemple souvent mal renseignés dans la plateforme car aucune validation n'est faite sur les caractères à remplir. Pourtant puisque les numéros de permis respectent un même format, l'utilisation d'une expression régulière correspondant au format de numéro de permis délivré par les autorités permettrait de contrôler le remplissage de ce champ. L'entrée d'un caractère ou d'une expression ne respectant pas le format autorisé serait donc refusée et l'utilisateur serait avisé de son erreur. Puisque cette étape de validation n'est pas présente dans le Registre, de nombreux numéros de permis renseignés sont erronés. Si l'on souhaite utiliser cette information, il est donc

nécessaire de croiser les données du Registre avec une autre source de données, à savoir la liste des permis valides dont dispose le BTM. Cela rajoute des étapes et complexifie l'analyse de données puisqu'en plus de devoir croiser les sources de données, il faut également s'assurer de la mise à jour de la source externe de données et de sa validité. L'absence de validation complique donc l'automatisation des processus et requiert des opérations supplémentaires de traitement et nettoyage des données.

2. **Mauvais fonctionnement des dispositifs à bord et erreurs de manipulation humaine**

   De nombreuses erreurs liées à une mauvaise manipulation du dispositif à bord ou à un mauvais fonctionnement de ce dernier peuvent être relevées. Par exemple, des occurrences seules de statut *occupied* peuvent être observées. Or il n'est pas possible qu'une course de taxi ne corresponde qu'à un point de donnée. Si les occurrences seules de ce type de statut peuvent être aisément supprimées, il arrive que le chauffeur se trompe de statut sur une plus longue période. En effet, de nombreuses courses de 30 secondes de durée ou d'une distance inférieure à 100 mètres sont observées ou au contraire, des courses durant plus de 10 heures, témoignant d'une absence complète de changement de statut de la part du chauffeur. Ces erreurs imposent donc la nécessité d'établir des règles de validation des courses. Or, si ces dernières peuvent permettre d'éliminer la majorité des courses non valides, il est possible que des courses valides mais particulièrement courtes ou longues soient éliminées car considérées comme non valides par les règles définies.

3. **Irrégularité d'envoi des données**

   L'une des limitations majeures des données du Registre concerne l'irrégularité d'envoi des données GPS. En effet, selon l'ordonnance du BTM les données d'un taxi en service doivent être envoyées à chaque 5 secondes. Or, dans la réalité, les données sont envoyées à des intervalles très irréguliers. L'enjeu est donc de pouvoir identifier à quoi correspondent les périodes d'absence de données. En effet, un intervalle de temps élevé entre deux points consécutifs peut soit être dû à une interruption volontaire du taxi qui prend par exemple une pause, ou alors à une perte de données GPS. Cela impacte fortement l'identification des périodes d'activité des taxis et particulièrement l'identification des courses régulières.

4. Absence de précision sur le statut unavailable

Le statut *unavailable* présente un enjeu considérable. En effet, lorsque le taxi est *unavailable*, il peut soit être en opération pour un service de taxi collectif pour une société de transport en commun, soit fournir des services de taxi adapté, ou encore effectuer des contrats corporatifs ou hospitalier. Toutefois, les informations actuellement disponibles dans le Registre ne permettent pas de différencier et d'identifier quel type de course été réalisé. Ainsi, seule une partie des trajets effectués par les taxis peut être identifiée avec les données disponibles dans le Registre. Seuls les trajets réguliers peuvent être comptabilisés pour l'instant, c'est-à-dire lorsque le statut indique *occupied*. Une des limites du Registre des taxis est ainsi mise en évidence puisqu'on ne peut identifier l'ensemble des courses réalisées par les taxis en exploitant uniquement les données du Registre. Les flux de données du transport adapté et du taxi collectif seraient nécessaires pour croiser les données et ainsi identifier les différents types de courses.

Ainsi seule une partie des activités des taxis peut réellement être identifiée. Cependant certains intermédiaires en service sont spécialisés dans le transport adapté. Ce service constitue donc la majorité de leurs activités de transport par taxi. Il est donc essentiel de pouvoir identifier l'ensemble des courses si l'on souhaite pouvoir analyser toutes les activités des taxis.

Cela impacte également le calcul de certains indicateurs tels que la distance et durée à vide et la distance et durée en course. En effet, considère-t-on qu'un statut unavailable implique forcément la présence d'un client ? Une incertitude existe sur l'activité effectuée par le taxi lors de ce statut. On ne peut être certain que c'est une course non régulière. En effet, il est également possible que certains chauffeurs indiquent ce statut lorsqu'ils prennent une pause ou qu'ils ne souhaitent pas se rendre disponible pour répondre à une course commandée.

5. Localisation exacte des places des postes d'attente

Les données disponibles pour la localisation des postes d'attente n'indiquent qu'un unique point pour chaque zone de postes d'attentes ainsi que le nombre de places. Il est supposé que le point indiqué corresponde au centroïde du poste d'attente. On ne dispose pas des coordonnées spatiales de chaque place de stationnement au sein du poste d'attente. Or, si l'on souhaite étudier le taux d'utilisation des postes d'attente il est essentiel de pouvoir

déterminer combien de places sont occupées. Ainsi, une hypothèse doit être émise quant à la zone de poste d'attente. Une zone d'un rayon proportionnel au nombre de places disponibles dans le poste est considérée autour de ce dernier. Cependant, cette zone établie peut en réalité chevaucher d'autres espaces ne correspondant pas au poste d'attente. L'utilisation des postes d'attente peut ainsi être surestimée.

6. **Objets actifs**

Une autre limite présente dans les données du Registre concerne l'identification de la proportion d'objets actifs. En effet, avec les données dont on dispose, il est possible d'obtenir le nombre de véhicules, de conducteurs ou encore de permis qui sont actifs, c'est-à-dire qui sont en opération. Cependant, si le pourcentage de permis actifs pendant une période d'analyse spécifique doit être identifié, il est nécessaire de connaître la population de référence, c'est-à-dire tous les permis qui étaient valables pendant la période d'étude mais qui n'ont pas forcément été actifs. Toutefois, la base de données des permis ne fournit pas le statut de validité ou d'invalidité du permis. Un contrôle de la validité des permis est effectué par le BTM mais l'information n'est pas renseignée dans le Registre. Il en est de même pour les chauffeurs et véhicules. On ne peut savoir si les chauffeurs et véhicules renseignés dans les données sont encore autorisés à offrir un service de taxi. Ainsi, seul le nombre d'objets en activité peut être identifié et non le pourcentage d'objets valides en activité.

7.3 Contributions

Ce projet de recherche s'inscrit dans la continuité des travaux de recherche de Lacombe (2016) et Laviolette (2017). Cependant, si seul un échantillon des taxis avait pu être analysé dans les travaux précédents, c'est désormais l'ensemble des données des taxis opérant sur l'île de Montréal qui peut être analysé. En effet, le Registre des taxis, mis en place par le BTM, assure la collecte des données GPS de tous les taxis opérant sur l'île. Le projet de recherche permet de mettre en valeurs ces données grâce au développement d'un outil de visualisation pour l'analyse des indicateurs d'offre et de demande de déplacements par taxis.

Tout d'abord, une synthèse des travaux de recherche sur l'industrie du taxi et particulièrement sur les travaux utilisant les données GPS permet de mettre en évidence l'impact positif de la

disponibilité nouvelle de ces données GPS sur les modélisations des activités de taxis. De plus, la liste des indicateurs pertinents sur les activités des taxis identifiée par Lacombe (2016) a été complétée en intégrant les travaux les plus récents. Surtout, une revue des bonnes pratiques de visualisation a été réalisée. Elle a permis de mettre en évidence les enjeux liés à la conception d'une plateforme de visualisation et plus particulièrement du tableau de bord en tant qu'outil de prise de décision. Jusqu'à présent l'attention était uniquement portée sur l'estimation des indicateurs et non sur la qualité de représentation de ces derniers dans le but de faciliter l'analyse.

Pour ce qui est des contributions méthodologiques, elles se concentrent principalement sur l'automatisation des processus de traitement systématique des données et d'estimation des indicateurs. En effet, la disponibilité d'un flux continu de données implique que les méthodes de traitement de ces données restent valables dans le temps et s'adaptent aux éventuelles évolutions des données. Les nombreuses étapes de nettoyage et prétraitement des données ont permis de mettre en évidence les défis liés à l'utilisation des données GPS et, notamment, l'identification des périodes de perte de données GPS liées à un dysfonctionnement des dispositifs de relevé. De plus, le processus de construction d'une base de données optimisée est bénéfique pour tout projet impliquant l'utilisation d'importants volumes de données. Aussi, les données GPS de taxis présentant une structure semblable à celle des données du Registre, cela assure une reproductibilité des méthodes développées. Particulièrement, une méthode d'identification des courses à partir de critères seuils tirés des distributions de distance et durée d'une population de courses de référence, est proposée.

Enfin, le tableau de bord conçu permet d'assurer une analyse à plusieurs niveaux des activités des taxis de Montréal. La structure en entonnoir du tableau de bord et l'interactivité assure à l'utilisateur une exploration flexible des données afin de passer des résumés agrégés à des analyses plus désagrégées. Il lui en effet possible d'affiner ses requêtes grâce à la sélection de filtres spatio-temporels et de comparer plusieurs périodes grâce au volet comparatif. De plus, le design de ce tableau de bord suit les principes de visualisation identifiés dans la littérature afin de représenter l'information de la manière la plus pertinente et intelligible possible. L'utilisation de ce tableau de bord permettra donc de brosser un portrait plus représentatif que ceux établis précédemment des taxis opérant sur l'île Montréal et favorisera donc la prise de décisions quant à l'opération et à la planification de ce service.

7.4 Perspectives

Dans le contexte actuel où l'on cherche à réduire la dépendance à l'automobile privée et à mettre en place une mobilité plus durable, tel que dans le concept de mobilité en tant que service, il est essentiel de pouvoir analyser la performance actuelle des services de taxi. En plus de son rôle dans la mobilité, l'analyse de son interaction avec les autres modes de transport urbains serait donc pertinente. Ainsi, inclure des indicateurs de complémentarité de ce mode avec les autres modes de transport disponibles en ville permettrait de mettre en évidence d'éventuelles stratégies de coopération des modes dans le but d'améliorer la planification intermodale.

De plus, à nouveau dans un objectif de mobilité durable, il pourrait être pertinent de proposer des scénarios de répartition des taxis aux clients afin de mettre en évidence le nombre de kilomètres à vides qui pourraient être sauvés si la répartition des taxis était optimale sur l'ensemble du territoire.

Également, il peut être intéressant de caractériser les activités des chauffeurs de taxi et d'identifier si des motifs ou des comportements similaires peuvent être mis en évidence. Surtout, une éventuelle différence de comportement du chauffeur s'il est affilié à un intermédiaire ou s'il est indépendant pourrait être mise en évidence. De même, il serait possible d'étudier si des comportements similaires sont observés pour tous les chauffeurs d'un même intermédiaire.

Les indicateurs identifiés et calculés pour ce projet peuvent être améliorés en intégrant notamment d'autres jeux de données qui peuvent expliquer la demande en déplacement de taxi, tels que les dates de spectacles, festivals, compétitions sportives, congrès au autres évènements spéciaux. Des données sur les clients permettraient également de mieux identifier les besoins de la clientèle et donc d'améliorer l'offre. De plus, de telles données permettraient également d'améliorer les processus d'identification des courses valides.

En ce qui concerne les méthodes développées, les éléments méthodologiques liés à l'automatisation des processus doivent être approfondis. Par exemple, plutôt que d'utiliser des seuils de distance et de durée d'une population de référence qui varient de manière hebdomadaire pour l'identification des courses, des caractéristiques annuelles ou saisonnières peuvent être utilisées si l'on observe peu de variabilité entre les mois. Des travaux futurs pourraient évaluer cette variabilité. De plus, les règles de validation des courses pourraient être améliorées en prenant en compte la dimension spatiale.

Il est avant tout indispensable d'obtenir les flux de données du transport adapté afin de pouvoir identifier l'ensemble des courses réalisées par les taxis et non uniquement les courses régulières.

Il serait également essentiel de poursuivre l'optimisation des processus développés afin d'améliorer la rapidité de calcul des indicateurs.

Enfin, il pourrait être pertinent d'évaluer la performance du tableau de bord conçu quant à la prise de décisions, par exemple par le biais d'entretiens auprès des acteurs du secteurs. Les indicateurs présentés dans la plateforme pourraient être alors améliorés, tant au niveau de leur estimation que de leur représentation dans le tableau de bord, à la suite des retours des utilisateurs projetés.

RÉFÉRENCES

Alexander, M., & Walkenbach, J. (2010). *Excel dashboards & reports* Mr. Spreadsheet's bookshelf. Tiré de https://polymtl.on.worldcat.org/oclc/664571368

Alisoltani, N., Zargayouna, M., & Leclercq, L. (2020). A Sequential Clustering Method for the Taxi-Dispatching Problem Considering Traffic Dynamics. *IEEE Intelligent Transportation Systems Magazine*, 0-0. doi:10.1109/MITS.2020.3014444

Allio Michael, K. (2012). Strategic dashboards: designing and deploying them to improve implementation. *Strategy & Leadership, 40*(5), 24-31. doi:10.1108/10878571211257159

Andreeva, J., Dzhunov, I., Karavakis, E., Kokoszkiewicz, L., Nowotka, M., Saiz, P., & Tuckett, D. (2012). Designing and developing portable large-scale JavaScript web applications within the Experiment Dashboard framework. *Journal of Physics: Conference Series, 396*(5), 052069. doi:10.1088/1742-6596/396/5/052069

Andrienko, G., Andrienko, N., & Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *SIGKDD Explor. Newsl., 9*(2), 38–46. doi:10.1145/1345448.1345455

Austin Jr, A. B. (2011). *The taxicab as public transportation in Boston.* (Massachusetts Institute of Technology).

The Taxi Exchange Point Data Extraction Guide (2017).

The Taxi Exchange Point Operator's Guide (2017).

Berman, J. J. (2018). *Principles and practice of big data : preparing, sharing, and analyzing complex information* (Second edition.[e] éd.). Tiré de https://polymtl.on.worldcat.org/oclc/1046065689

Billhardt, H., Fernández, A., Ossowski, S., Palanca, J., & Bajo, J. (2019). Taxi dispatching strategies with compensations. *Expert Systems with Applications, 122*, 173-182.

Bischoff, J., Maciejewski, M., & Sohr, A. (3-5 June 2015 2015). *Analysis of Berlin's taxi services by exploring GPS traces.* Communication présentée à 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (p. 209-215).

Bourdeau, J.-S. (2014). *Méthodologie d'analyse automatisée des stationnements.* (École Polytechnique de Montréal).

Brath, R., & Peters, M. (2004). Dashboard Design: Why Design is Important. *DM Direct, 85*.

Bureau du Taxi de Montréal. (2020a). Documentation technique. Tiré de http://www.registretaximontreal.ca/documentation-technique/

Bureau du Taxi de Montréal. (2020b). Propriétaire et chauffeur. Tiré de http://www.registretaximontreal.ca/proprietaire-chauffeur/

Bureau du Taxi de Montréal. (Automne 2018). Le registre des taxis est officiellement lancé. *Taxi le journal, Volume 30 l n°3*. Tiré de http://ville.montreal.qc.ca/pls/portal/docs/page/bur_taxi_fr/media/documents/jlt_automne_2018.pdf

Card, M. (1999). *Readings in information visualization: using vision to think*: Morgan Kaufmann.

Contributeurs de Wikipédia. (2020). Shapefile. Tiré de http://fr.wikipedia.org/w/index.php?title=Shapefile&oldid=169394394

Cooper, J., Mundy, R., & Nelson, J. (2010). *Taxi! : urban economies and the social and transport impacts of the taxicab* Transport and society. Tiré de http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=476262

Correia, P. (2006). *Guide pratique du GPS*: Editions Eyrolles.

Dandl, F., Bracher, B., & Bogenberger, K. (2017). *Microsimulation of an autonomous taxi-system in Munich.* Communication présentée à 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (p. 833-838).

Darbéra, R. (2010). Transports publics et taxis: concurrence ou complémentarité?

Darbéra, R. (2017). Business--models for the taxi of the future.

Djenouri, Y., Belhadi, A., Lin, J. C., Djenouri, D., & Cano, A. (2019). A Survey on Urban Traffic Anomalies Detection Algorithms. *IEEE Access, 7,* 12192-12205. doi:10.1109/ACCESS.2019.2893124

Eckerson, W. W. (2011). *Performance dashboards : measuring, monitoring, and managing your business* Finance professional collection, (2nd ed.e éd.). doi:https://doi.org/10.1002/9781119199984

Egbelu, P. J., & Tanchoco, J. M. A. (1984). Characterization of automatic guided vehicle dispatching rules. *International Journal of Production Research, 22*(3), 359-374. doi:10.1080/00207548408942459

Facebook Inc. (2020). React. Tiré de https://fr.reactjs.org/

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2149-2158. doi:10.1109/TVCG.2013.226

Few, S. (2006). *Information dashboard design: The effective visual communication of data*: O'Reilly Media, Inc.

Ge, Y., Liu, C., Xiong, H., & Chen, J. (2011). *A taxi business intelligence system.* Communication présentée à Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (p. 735-738).

Ge, Y., Xiong, H., Liu, C., & Zhou, Z.-H. (2011). *A taxi driving fraud detection system.* Communication présentée à 2011 IEEE 11th International Conference on Data Mining (p. 181-190).

Geisberger, R., Sanders, P., Schultes, D., & Delling, D. (2008). *Contraction hierarchies: Faster and simpler hierarchical routing in road networks.* Communication présentée à International Workshop on Experimental and Efficient Algorithms (p. 319-333).

Geneste, A. (2017). *Caractérisation de l'offre et de la demande de transport dans un système de taxis électriques.* (École Polytechnique de Montréal).

Loi concernant le transport rémunéré de personnes par automobile, T-11.2 C.F.R. (2020).

Harris, R. L. (2000). *Information graphics: A comprehensive illustrated reference*: Oxford University Press.

Hartikainen, A., Pitkänen, J., Riihelä, A., Räsänen, J., Sacs, I., Sirkiä, A., & Uteng, A. (2019). WHIMPACT: Insights from the world's first Mobility-as-a-Service (MaaS) system: Ramboll.

Healey, C., & Enns, J. (2011). Attention and visual memory in visualization and computer graphics. *IEEE transactions on visualization and computer graphics, 18*(7), 1170-1188.

Healey, C. G., Booth, K. S., & Enns, J. T. (1996). High-speed visual estimation using preattentive processing. *ACM Trans. Comput.-Hum. Interact., 3*(2), 107–135. doi:10.1145/230562.230563

Hodges, G. R. G. (2020). *Taxi!: a social history of the New York City cabdriver*: JHU Press.

Hu, Y., Yang, Y., & Huang, B. (2015). *A comprehensive survey of recommendation system based on taxi GPS trajectory.* Communication présentée à 2015 International Conference on Service Science (ICSS) (p. 99-105).

Kamargianni, M., & Matyas, M. (2017). *The Business Ecosystem of Mobility-as-a-Service*. Communication présentée à 96th Transportation Research Board (TRB) Annual Meeting Washington DC, 8-12.

Kamimura, J., Ogawa, M., Wakayama, H., Iga, N., Shiota, N., & Yano, M. (2013). *D-Taxi: Adaptive area recommendation system for taxis by using DiRAC.* Communication présentée à 2013 International Conference on Connected Vehicles and Expo (ICCVE) (p. 507-508).

King, D. A., Peters, J. R., & Daus, M. W. (2012). *Taxicabs for improved urban mobility: are we missing an opportunity?* :

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*: Sage.

Kosara, R., Miksch, S., & Hauser, H. (2002). Focus+ context taken literally. *IEEE Computer Graphics and Applications, 22*(1), 22-29.

Kourtit, K., & Nijkamp, P. (2018). Big data dashboards as smart decision support tools for i-cities – An experiment on stockholm. *Land Use Policy, 71*, 24-35. doi:https://doi.org/10.1016/j.landusepol.2017.10.019

Kuang, W., An, S., & Jiang, H. (2015). Detecting Traffic Anomalies in Urban Areas Using Taxi GPS Data. *Mathematical Problems in Engineering, 2015*, 809582. doi:10.1155/2015/809582

Lacombe, A. (2016). *Méthodologie d'analyse et de suivi d'un système de transport par taxi.* (École Polytechnique de Montréal, Montréal). Accessible par WorldCat.org. Tiré de http://publications.polymtl.ca/2261/

Laviolette, J. (2017). *Planification stratégique d'un système de transport par taxi.* (École Polytechnique de Montréal, Montréal). Accessible par WorldCat.org. Tiré de https://publications.polymtl.ca/2738/

Lee, W. K., & Sohn, S. Y. (2017). Taxi vacancy duration: a regression analysis. *Transportation Planning and Technology, 40*(7), 771-795. doi:10.1080/03081060.2017.1340025

Li, Y., & Voege, T. (2017). Mobility as a service (MaaS): Challenges of implementation and policy required. *Journal of transportation technologies, 7*(2), 95-106.

Liao, Z., Yu, Y., & Chen, B. (25-26 Oct. 2010 2010). *Anomaly detection in GPS data based on visual analytics.* Communication présentée à 2010 IEEE Symposium on Visual Analytics Science and Technology (p. 51-58). doi:10.1109/VAST.2010.5652467

Little, A. (2018). The Future of Mobility 3.0: Reinventing Mobility in the era of disruption and creativity: Academic Press.

Liu, W., Zheng, Y., Chawla, S., Yuan, J., & Xing, X. (2011). *Discovering spatio-temporal causal interactions in traffic data streams*. Communication présentée à Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA. Tiré de https://doi.org/10.1145/2020408.2020571

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography, 43*, 78-90. doi:https://doi.org/10.1016/j.jtrangeo.2015.01.016

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of geographical systems, 14*(4), 463-483.

Loshin, D. (2001). *Enterprise knowledge management: The data quality approach*: Morgan Kaufmann.

Lyons, G., Hammond, P., & Mackay, K. (2019). The importance of user perspective in the evolution of MaaS. *Transportation Research Part A: Policy and Practice, 121*, 22-36. doi:https://doi.org/10.1016/j.tra.2018.12.010

Maas Alliance. (2020). MaaS Alliance. Tiré de https://maas-alliance.eu/

MacEachren, A. M. (1994). *Some truth with maps: A primer on symbolization and design*: Assn of Amer Geographers.

Matheus, R., Janssen, M., & Maheshwari, D. (2020). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly, 37*(3), 101284. doi:https://doi.org/10.1016/j.giq.2018.01.006

Mathieu, Y. D. (2020). *Déplacements par taxi : caractérisation et études de compétitivité.* (Ecole Polytechnique de Montréal, Montréal).

Meirelles, I. (2013). *Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations*: Rockport publishers.

Moreira-Matias, L., Fernandes, R., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (14-16 Nov. 2012 2012). *An online recommendation system for the taxi stand choice problem (Poster).* Communication présentée à 2012 IEEE Vehicular Networking Conference (VNC) (p. 173-180). doi:10.1109/VNC.2012.6407427

New York City Taxi & Limousine Commission. (2018a). Simple-to-Use Visualizations for Trip Trends. Tiré de https://medium.com/@NYCTLC/simple-to-use-visualizations-for-trip-trends-6dd35ae1f247

New York City Taxi & Limousine Commission. (2018b). TLC FastDash. Tiré de https://tlcanalytics.shinyapps.io/tlc_fast_dash/

New York City Taxi & Limousine Commission, Bloomberg, M., & Yassky, D. (2014). *2014 Taxicab Fact Book*.

New York City Taxi & Limousine Commission, De Blasio, B., & Joshi, M. (2016). *2016 TLC Factbook*.

New York City Taxi & Limousine Commission, De Blasio, B., & Joshi, M. (2018). *2018 TLC Factbook*.

Node.js. (2020). Node.js. Tiré de nodejs.org

OSRM. (2020). OSRM Open Source Routing Machine. Tiré de http://project-osrm.org/

Pan, G., Qi, G., Wu, Z., Zhang, D., & Li, S. (2012). Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, *14*(1), 113-123.

Pedro, K. (2020). Initiez-vous au fonctionnement des API. Tiré de https://openclassrooms.com/fr/courses/6573181-adoptez-les-api-rest-pour-vos-projets-web/6816951-initiez-vous-au-fonctionnement-des-api

Pele, N., & Morency, C. (2014). *When, where and how taxis are used in Montreal*.

Phiboonbanakit, T., & Horanont, T. (2016). *How does taxi driver behavior impact their profit? discerning the real driving from large scale GPS traces*. Communication présentée à Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, Heidelberg, Germany. Tiré de https://doi.org/10.1145/2968219.2968417

PostgreSQL. (2020a). About PostgreSQL. Tiré de https://www.postgresql.org/about/

PostgreSQL. (2020b). Window Functions. Tiré de https://www.postgresql.org/docs/current/tutorial-window.html

Qu, M., Zhu, H., Liu, J., Liu, G., & Xiong, H. (2014). *A cost-effective recommender system for taxi drivers*. Communication présentée à Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, New York, USA. Tiré de https://doi.org/10.1145/2623330.2623668

Riascos, A. P., & Mateos, J. L. (2020). Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. *Scientific Reports*, *10*(1), 4022. doi:10.1038/s41598-020-60875-w

Salanova, J. M., Estrada, M., Aifadopoulou, G., & Mitsakis, E. (2011). A review of the modeling of taxi services. *Procedia - Social and Behavioral Sciences*, *20*, 150-161. doi:https://doi.org/10.1016/j.sbspro.2011.08.020

Savage, T. H., & Vo, H. T. (2013). *Yellow cabs as red corpuscles*. Communication présentée à 2013 IEEE International Conference on Big Data (p. 22-28).

Skyscraper Source Media. (2020). Montreal Skyscraper Map. Tiré de https://skyscraperpage.com/cities/maps/?cityID=22

Smith, G. (2020). Making Mobility-as-a-Service—Towards Governance Principles and Pathways. *Service. Towards Governance Principles and Pathways*.

Smith, V. S. (2013). Data Dashboard as Evaluation and Research Communication Tool. *New Directions for Evaluation, 2013*(140), 21-45. doi:10.1002/ev.20072

Sochor, J., Arby, H., Karlsson, I. M., & Sarasini, S. (2018). A topological approach to Mobility as a Service: A proposed tool for understanding requirements and effects, and for aiding the integration of societal goals. *Research in Transportation Business & Management, 27*, 3-14.

Statistique Canada. (2018). Secteur de recensement : définition détaillée. Tiré de https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/ct-sr/def-fra.htm

Tessier, M.-A. (2015). *Développement d'indicateurs d'analyse et de suivi de la congestion routière.* (École polytechnique de Montréal, Montréal). Accessible par WorldCat.org. Tiré de http://publications.polymtl.ca/1957/

The R Foundation. (2020). The R Project for Statistical Computing. Tiré de https://www.r-project.org/

Transition. (2020). Transition. Tiré de transition.city

Treisman, A. (1985). Preattentive processing in vision. *Computer vision, graphics, and image processing, 31*(2), 156-177.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*: Graphics Press.

Tufte, E. R. (2006). *Beautiful Evidence*: Graphics Press.

Veloso, M., Phithakkitnukoon, S., & Bento, C. (2011). *Urban mobility study using taxi traces.* Communication présentée à Proceedings of the 2011 international workshop on Trajectory data mining and analysis (p. 23-30).

Vij, A., Ryan, S., Sampson, S., & Harris, S. (2020). Consumer preferences for Mobility-as-a-Service (MaaS) in Australia. *Transportation Research Part C: Emerging Technologies, 117*, 102699. doi:https://doi.org/10.1016/j.trc.2020.102699

Vila, R. A., Estevez, E., & Fillottrani, P. R. (2018). *The design and use of dashboards for driving decision-making in the public sector*. Communication présentée à Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland. Tiré de https://doi.org/10.1145/3209415.3209467

Règlement sur le transport par taxi RCG 10-009 C.F.R. (2017).

Ville de Montréal. (2020a). Bureau du taxi de Montréal. Tiré de https://montreal.ca/unites/bureau-du-taxi-de-montreal

Ville de Montréal. (2020b). Maximum 40 km/h. Tiré de https://ville.montreal.qc.ca/portal/page?_pageid=8957,101461781&_dad=portal&_schema=PORTAL

Ville de Montréal. (2020c). Postes d'attente de taxi. Tiré de http://donnees.ville.montreal.qc.ca/dataset/postes-attente-taxi

W3Schools. (2020). JSON - Introduction. Tiré de https://www.w3schools.com/js/js_json_intro.asp

Wade, A. (2018, 10th August 2018). August 1897 – The London Electrical Cab. Tiré de https://www.theengineer.co.uk/august-1897-london-electrical-cab/

Wagner, D., & Willhalm, T. (2007). *Speed-Up Techniques for Shortest-Path Computations*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, F., & Ross, C. L. (2019). New potential for multimodal connection: exploring the relationship between taxi and transit in New York City (NYC). *Transportation, 46*(3), 1051-1072.

Wang, H., Zou, H., Yue, Y., & Li, Q. (2009). *Visualizing hot spot analysis result based on mashup*. Communication présentée à Proceedings of the 2009 International Workshop on Location Based Social Networks, Seattle, Washington. Tiré de https://doi.org/10.1145/1629890.1629900

Wang, Y., Zhu, Y., He, Z., Yue, Y., & Li, Q. (2011). Challenges and opportunities in exploiting large-scale GPS probe data. *HP Laboratories, Technical Report HPL-2011-109, 21*.

Ware, C. (2004). *Information visualization : perception for design* (2nd ed.[e] éd.). San Francisco: Morgan Kaufman.

Wexler, S., Shaffer, J., & Cotgreave, A. (2017). *The big book of dashboards : visualizing your data using real-world business scenarios*. doi:10.1002/9781119283089

Winkelmann, R. (2008). *Econometric analysis of count data*: Springer Science & Business Media.

Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., . . . Ye, J. (2018). *Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach.* Communication présentée à Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (p. 905-913).

Yang, C., & Gonzales, E. J. (2014). Modeling taxi trip demand by time of day in New York City. *Transportation Research Record, 2429*(1), 110-120.

Yang, C., & Gonzales, E. J. (2017). Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data. Dans P. Thakuriah, N. Tilahun, & M. Zellner (édit.), *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics* (p. 405-425). Cham: Springer International Publishing.

Yang, H., Lau, Y. W., Wong, S. C., & Lo, H. K. (2000). A macroscopic taxi model for passenger demand, taxi utilization and level of services. *Transportation, 27*(3), 317-340. doi:10.1023/A:1005289504549

Yuan, N. J., Zheng, Y., Zhang, L., & Xie, X. (2013). T-Finder: A Recommender System for Finding Passengers and Vacant Taxis. *IEEE Transactions on Knowledge and Data Engineering, 25*(10), 2390-2403. doi:10.1109/TKDE.2012.153

Zhang, D.-Z., Peng, Z.-R., & Sun, D. J. (2014). A comprehensive taxi assessment index using floating car data. *Journal of Harbin Institute of Technology, 21*(1), 7-16.

Zhang, D., Li, N., Zhou, Z.-H., Chen, C., Sun, L., & Li, S. (2011). *iBAT: detecting anomalous taxi trajectories from GPS traces.* Communication présentée à Proceedings of the 13th international conference on Ubiquitous computing (p. 99-108).

Zhang, J. (2012). *Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC*. Communication présentée à Proceedings of the ACM SIGKDD International Workshop on Urban Computing, Beijing, China. Tiré de https://doi.org/10.1145/2346496.2346521

Zhang, L., Ahmadi, M., Pan, J., & Chang, L. (2012). *Metropolitan-scale taxicab mobility modeling.* Communication présentée à 2012 IEEE Global Communications Conference (GLOBECOM) (p. 5404-5409).

Zhang, M., Liu, J., Liu, Y., Hu, Z., & Yi, L. (1-3 Nov. 2012 2012). *Recommending Pick-up Points for Taxi-drivers Based on Spatio-temporal Clustering.* Communication présentée à 2012 Second International Conference on Cloud and Green Computing (p. 67-72). doi:10.1109/CGC.2012.34

Zhang, Y. (2014). *How Do Taxis Work in Beijing? An Exploratory Study of Spatio-Temporal Taxi Travel Pattern Using GPS Data.* (UCLA).

Zheng, Z., Rasouli, S., & Timmermans, H. (2014). Evaluating the accuracy of GPS-based taxi trajectory records. *Procedia Environmental Sciences, 22*, 186-198.

Zhou, X., Rong, H., Yang, C., Zhang, Q., Khezerlou, A. V., Zheng, H., . . . Liu, A. X. (2020). Optimizing Taxi Driver Profit Efficiency: A Spatial Network-Based Markov Decision Process Approach. *IEEE Transactions on Big Data, 6*(1), 145-158. doi:10.1109/TBDATA.2018.2875524

ANNEXE A DEFINITION DES ATTRIBUTS DE LA TABLE DES PERMIS (ADS)

Tableau A-1 Définition des attributs de la table des permis (ADS).

Attribut	**Type de variable**	**Définition**
Id	Entier	Identifiant du permis
Numero	Chaine de caractères	Numéro de permis (ADS) tel que délivré par la CTQ
Added_at	Chaine de caractère	Date à laquelle le permis a été ajouté au Registre
Last_update_at	Chaine de caractère	Date de la dernière modification
Insee	Chaine de caractères	Zone associée au permis (3 zones possibles pour Montréal : 102005 – 102011 – 102012)
Vehicle_id	Entier	Identifiant du véhicule associé au permis
Owner_name	Chaine de caractères	Nom du détenteur du permis (Peut être un individuel ou une entreprise)
Owner_type	Chaine de caractères	Type de détenteur. 2 valeurs sont possibles : « Company » OU « Individual »
Vdm_vignette	Chaine de caractères	Numéro de vignette assigné par le BTM
Nom_zupc	Chaine de caractères	Nom de la zone associée au permis (3 zones possibles pour Montréal : A5 – A11 – A12)
Added_by_name	Chaine de caractères	Nom de la personne ayant renseigné le permis dans le Registre

*Certains champs requis pour des raisons administratives uniquement n’ont pas été présentés dans les tableaux .

ANNEXE B DEFINITION DES ATTRIBUTS DE LA TABLE DES VEHICULES

Tableau B-1 Définition des attributs de la table des véhicules

Attribut	**Type de variable**	**Définition**
licence_plate	Chaine de caractères	Plaque d'immatriculation du véhicule. Pour le Québec, la plaque d'immatriculation est une combinaison alphanumérique de 6 caractères.
constructor	Chaine de caractères	Constructeur automobile du véhicule.
model	Chaine de caractères	Modèle du véhicule.
color	Chaine de caractères	Couleur du véhicule.
type_	Chaine de caractères	Type de véhicule. Les valeurs possibles sont : « sedan », « station_wagon », « normal » ou « mpv ».
nb_seats	Entier	Nombre de sièges dans le véhicule.
air_con	Booléen	Le véhicule est équipé de la climatisation.
amex_accepted	Booléen	Le véhicule accepte la carte American Express (aucun montant minimum n’est requis).
baby_seat	Booléen	Le véhicule est équipé d’un siège pour bébé.
bank_check_accepted	Booléen	Le véhicule accepte les chèques bancaires nationaux (les chèques bancaires étrangers peuvent néanmoins être refusés).
bike_accepted	Booléen	Le véhicule peut transporter un vélo.
credit_card_accepted	Booléen	Le véhicule accepte les paiements par carte de crédit (aucun montant minimum n’est requis).

Tableau B-1 (Suite) Définition des attributs de la table des véhicules

Attribut	Type de variable	Définition
dvd_player	Booléen	Le véhicule dispose d'un lecteur dvd à la disposition des clients pendant le trajet.
electronic_toll	Booléen	Le véhicule est équipé d'un dispositif électronique permettant d'utiliser des cabines de péage express sur les routes à péage.
every_destination	Booléen	Conformément à la réglementation française, les taxis peuvent refuser de servir les clients dont la destination ne se trouve pas dans leur zone. Certains taxis acceptent toute destination en dehors de leur zone. Le Booléen « every_destination » doit être faux par défaut et vrai pour les taxis qui renoncent à leur droit de refuser le service aux clients en fonction de leur destination.
fresh_drink	Booléen	Le taxi offre des rafraîchissements.
gps	Booléen	Le véhicule est équipé d'un système de navigation GPS.
luxury	Booléen	Il s'agit d'un véhicule de luxe.
nfc_cc_accepted	Booléen	Le véhicule accepte les paiements par carte de crédit NFC.
pet_accepted	Booléen	Le véhicule peut accueillir des animaux de compagnie (des chats ou des petits chiens ; les autres animaux de grande taille ou insolites peuvent être refusés).
special_need_vehicle	Booléen	Véhicule accessible aux fauteuils roulants, tel que défini dans la directive EU/678/2011. Véhicule construit ou aménagé spécifiquement pour accueillir une ou plusieurs personnes assises dans leur fauteuil roulant au cours du déplacement.

Tableau B-1 (Suite) Définition des attributs de la table des véhicules

Attribut	Type de variable	Définition
tablet	Booléen	Le véhicule dispose d'une tablette numérique à la disposition des clients pendant le trajet.
wifi	Booléen	Le véhicule dispose du Wi-Fi gratuit à bord.
cpam_conventionne	Booléen	Le véhicule a une convention avec la sécurité sociale pour le transport des patients.
date_dernier_ct	Chaine de caractères	Date du dernier contrôle technique obligatoire au format "AAAA-MM-JJ".
date_validite_ct	Chaine de caractères	Date d'expiration du dernier contrôle technique obligatoire au format "AAAA-MM-JJ".
engine	Chaine de caractères	Type de moteur du véhicule.
horse_power	Entier	Puissance du moteur.
model_year	Entier	Année du modèle de véhicule.
relais	Booléen	Vrai si le véhicule est un véhicule de remplacement provisoire d'un véhicule homologué.
taximetre	Chaine de caractères	Marque et modèle du taximètre.
horodateur	Chaine de caractères	Marque et modèle de l’horodateur.
added_by_name	Chaine de caractères	Nom de la personne qui a ajouté la licence.

ANNEXE C DEFINITION DES ATTRIBUTS DE LA TABLE DES CHAUFFEURS

Tableau C-1 Définition des attributs de la table des chauffeurs

Attribut	Type de variable	Définition
departement	Chaine de caractères	L'objet « departement » est constitué du numéro d'identification et du nom de la collectivité locale. Pour le Québec, le département doit toujours être défini à 660-Montréal.
professional_licence	Chaine de caractères	Numéro de licence professionnelle du conducteur. Attention : cet identifiant n'est pas unique au niveau national : deux collectivités locales peuvent attribuer chacune le même numéro à des conducteurs différents.
last_name	Chaine de caractères	Nom de famille du conducteur.
first_name	Chaine de caractères	Prénom du conducteur.
birth_date	Chaine de caractères	Date de naissance du conducteur au format "AAAA-MM-JJ".
added_by_name	Chaine de caractères	Nom de la personne qui a ajouté la licence.

ANNEXE D DEFINITION DES ATTRIBUTS DE LA TABLE DES TAXIS

Tableau D-1 Définition des attributs de la table des taxis

Attribut	Type de variable	Définition
vehicle_id	integer	Identifiant du véhicule.
ads_id	integer	Identifiant du permis (ADS).
driver_id	integer	Identifiant du chauffeur.
id	string	Identifiant généré par le serveur du Registre pour le triplet véhicule/permis/conducteur.
ban_begin	date	La date à laquelle l'interdiction commence.
ban_end	date	La date à laquelle l'interdiction prend fin.
rating	**float**	La moyenne des évaluations des dernières courses du taxi. Elle est calculée par le serveur du Registre et se situe entre 0 et 5.

ANNEXE E PREMIER SCENARIO

Dans le scénario n°1 de la méthodologie d'identification des regroupements des points GPS consécutifs de même statut occupied, on estime d'abord le temps nécessaire pour effectuer la distance entre les deux points. La vitesse considérée pour le calcul de cette durée correspond à la moyenne des vitesses instantanées des deux points. Le détail de ce scénario est présenté dans les Figure E-1 et Figure E-2.

Dans un premier temps, on détermine le temps requis, t_R, pour parcourir la distance séparant les deux points et effectuer le débarquement du client et l'embarquement d'un autre client. Ce temps t_R est ensuite comparé à l'intervalle de temps Δ_{Temps} séparant les deux points, tel qu'illustré dans la Figure E-1. Si Δ_{Temps} est inférieur à $t_{R,}$ alors les deux points appartiennent à la même course. En effet, cela signifie que l'intervalle de temps séparant les deux points est inférieur au temps requis pour parcourir la distance les séparant et effectuer l'embarquement et le débarquement d'un client. Il n'y a donc pas pu avoir de changement de client entre les deux. Les deux points appartiennent donc à la même course.

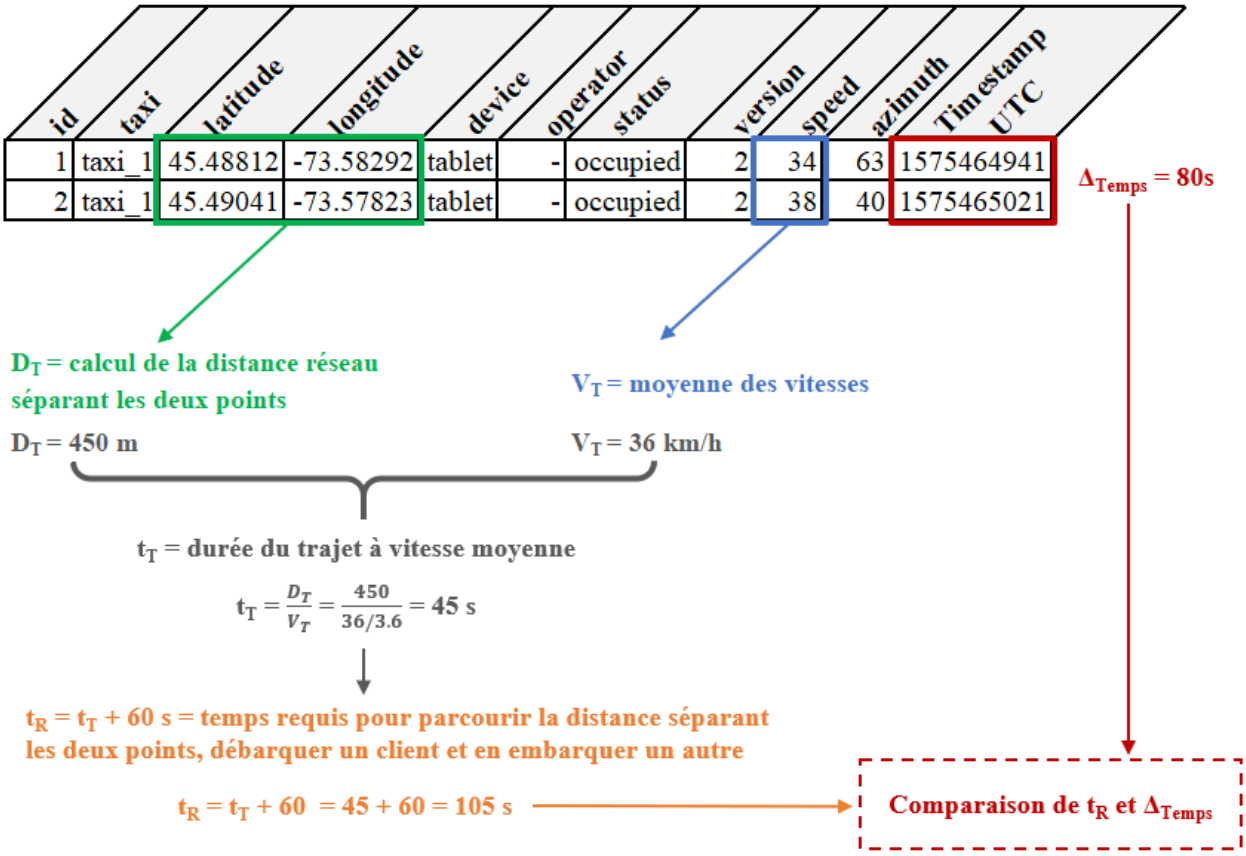


Figure E-1 Détail du scénario 1 de regroupement des points GPS consécutifs séparés d'un intervalle de 60 à 600 secondes : calcul du temps requis pour parcourir la distance séparant les deux points, débarquer un client et en embarquer un autre.

Quand est-il du cas où Δ_{Temps} est supérieur à t_R? Il est possible que l'intervalle de temps entre les deux points soit suffisamment grand pour qu'un éventuel changement de client ait lieu mais sans que ça n'en soit réellement le cas. L'interruption peut être liée à une perte de données au sein de la même course. Il est donc nécessaire de réaliser une autre validation.

Dans le cas où Δ_{Temps} est supérieur à t_R, les vitesses avant et après interruption sont comparées à la vitesse moyenne durant l'interruption. La vitesse moyenne d'interruption correspond à la moyenne des vitesses instantanées des deux points. Les vitesses moyennes avant et après l'interruption correspondent respectivement aux moyennes des vitesses instantanées des points précédents l'interruption et des points suivants l'interruption, telles qu'illustrées dans la Figure E-2. Un des enjeux de cette méthode concerne le nombre de points considérés pour la vitesse moyenne avant et

après interruption. Réalise-t-on un calcul à partir des cinq points précédents et suivants l'interruption ? D'uniquement deux ou trois points ? Il est donc nécessaire de déterminer le nombre de points suffisants pour effectuer une analyse précise. De plus comment compare-t-on les vitesses ? Dans quelle plage de vitesse considère-t-on que la vitesse moyenne d'interruption est semblable à celles avant et après interruption ? A plus ou moins 5 kilomètres-heures près, plus ou moins 3 kilomètres-heures ? Il est donc également essentiel de définir les intervalles de vitesses à employer pour comparer les vitesses moyennes.

Toutes ces interrogations sont révélatrices des nombreux enjeux présents dans ce scénario.

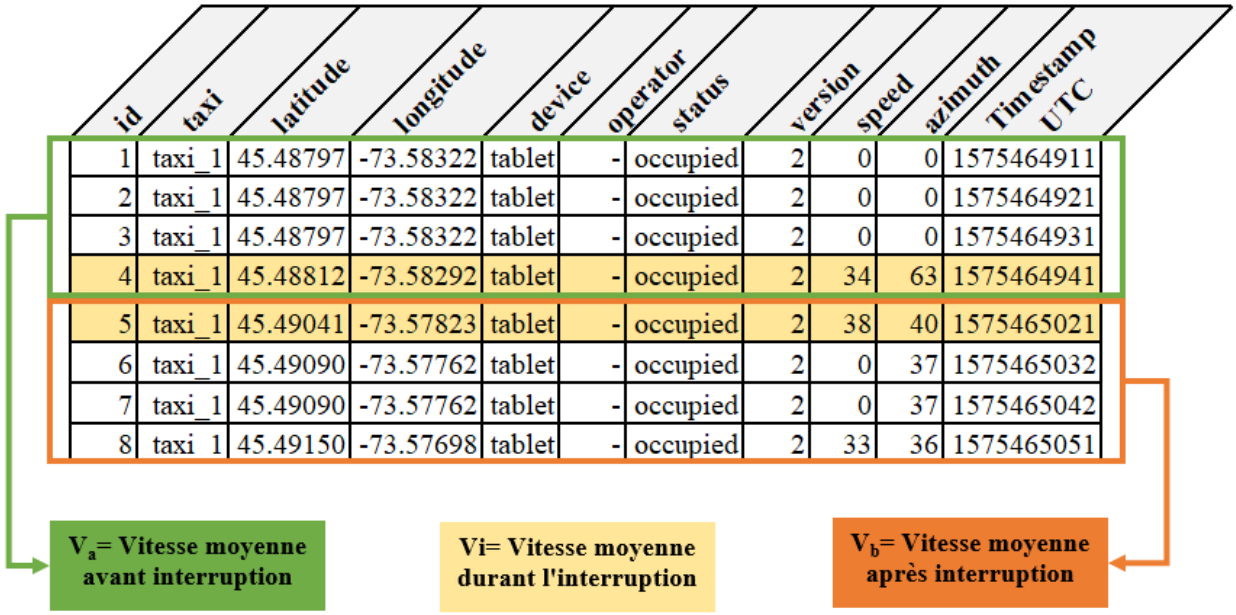


Figure E-2 Détail du scénario 1 de regroupement des points GPS consécutifs séparés d'un intervalle de 60 à 600 secondes : comparaison des vitesses moyennes avant et après interruption à la vitesse moyenne d'interruption.

En raison de la trop grande incertitude des hypothèses et méthodes de comparaison des vitesses, le scénario n°1 n'a donc pas été retenu pour déterminer les regroupements de points.