

Titre: Validation des temps de parcours des bus à partir de données
Title: massives

Auteur: Imed Essid
Author:

Date: 2023

Type: Mémoire ou thèse / Dissertation or Thesis

Référence: Essid, I. (2023). Validation des temps de parcours des bus à partir de données
Citation: massives [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie.
<https://publications.polymtl.ca/54406/>

 **Document en libre accès dans PolyPublie**
Open Access document in PolyPublie

URL de PolyPublie: <https://publications.polymtl.ca/54406/>
PolyPublie URL:

**Directeurs de
recherche:** Martin Trépanier, & Lijun Sun
Advisors:

Programme: Maîtrise recherche en génie industriel
Program:

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Validation des temps de parcours des bus à partir de données massives

IMED ESSID

Département de mathématiques et de génie industriel

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

Août 2023

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Validation des temps de parcours des bus à partir de données massives

présenté par **Imed ESSID**

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

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

Philippe DOYON-POULIN, président

Martin TRÉPANIÉ, membre et directeur de recherche

Lijun SUN, membre et codirecteur de recherche

Marc-Olivier THIBAUT, membre externe

REMERCIEMENTS

Mes remerciements les plus sincères vont en premier lieu à mes deux directeurs de recherche, professeurs Martin Trépanier et Lijun Sun. Leur dévouement, leur sens de partage, leurs conseils précieux et leur soutien inconditionné m'ont conduit à achever mon parcours dans les plus bonnes conditions.

J'aimerais aussi adresser à mes encadrants ma profonde gratitude pour l'aide financière qu'ils m'ont offerte afin que je puisse me concentrer sur mon travail et mes études.

Je suis également reconnaissant envers les membres de l'équipe exo, leur expertise était un atout et une valeur ajoutée pour ma personne et pour la qualité de mes travaux de recherche.

Je ne peux pas oublier de dire un mot de remerciement à Amaury Philippe pour toute l'aide apportée.

Enfin, ce parcours qui touche à sa fin était marqué par des collègues qui ont fait preuve de soutien et d'amabilité tout au long de ces deux années d'études. Merci pour vos remarques constructives ainsi qu'à votre bienveillance.

RÉSUMÉ

Persuader davantage de gens d’emprunter le transport en autobus demeure tout un défi. Tous les exploitants de transport en commun ne cessent de chercher à améliorer la planification de leur offre de transport, surtout au cours des dernières années avec l’apparition de sources de données nouvelles et variées.

C’est dans ce contexte que s’inscrit la présente étude en collaboration avec EXO en ayant à notre disposition trois sources de données : les horaires planifiés (GTFS), les traces GPS des bus et les données provenant du fournisseur HERE. L’objectif est donc de déterminer les principales caractéristiques des forces et des faiblesses de ces données, et voir dans quel cadre elles peuvent être utilisées pour une amélioration de la planification.

La méthodologie développée dans ce mémoire se compose de trois étapes : l’extraction des données, la validation des données extraites à partir de certains critères et une comparaison des temps de parcours issus de ces deux types de données pour l’interprétation et l’analyse.

En s’appuyant sur les exemples des lignes de bus étudiées, voici les conclusions qui s’en dégagent :

- les horaires planifiés sont faits pour ne pas être dépassés. Un ajustement de ces horaires demeure possible, dans la mesure où des données correspondant aux besoins sont fournies.
- Les données GPS complètes et non erronées permettent d’ajuster les horaires au besoin ; dans le cas contraire, ces données peuvent fausser considérablement l’étude.
- Les niveaux de confiance associés aux vitesses sur les segments HERE varient selon la période de la journée. Les vitesses fournies par HERE pour les heures de pointe sont globalement des vitesses observées alors que, pour les heures hors pointe, les vitesses sont modélisées.
- Le ratio entre les temps parcours HERE et les temps parcours GPS a été proposé pour évaluer la pertinence des données HERE. Quatre lignes de bus ont fait l’objet d’une analyse comparée. Le meilleur résultat observé est celui de la ligne 27 (Lorraine - Sainte-Thérèse dans la direction Bois-des-Filion via Lorraine) si l’on considère la valeur du ratio concentré à plus que 80% entre 0.75 à 0.90. Cette analyse nous a permis de conclure que, premièrement, les données HERE représentent des temps routiers en voiture et qu’ils doivent donc être ajustés pour représenter des temps en bus et que, deuxièmement, la pertinence des données HERE dépend de la période de la journée et la zone géographique visée.

Ce mémoire se termine par une conclusion générale et une explication claire des limites et des recommandations pour les futures recherches.

Mots clés : GTFS, GPS, HERE, Planification

ABSTRACT

Persuading more people to take bus transportation remains a challenge. All public transport operators are constantly seeking to improve the planning of their transport offer, especially in recent years with the emergence of new and varied data sources.

It is in this context that the present study in collaboration with EXO takes place by having at our disposal three sources of data: the planned timetables (GTFS), the GPS tracks of the buses and the data from the HERE supplier. The objective is therefore to determine the main characteristics of the strengths and weaknesses of these data, and to see in what context they can be used to improve planning.

The methodology developed in this thesis consists of three steps: data extraction, validation of extracted data based on certain criteria and a comparison of travel times from these two types of data for interpretation and analysis.

Based on the examples of the bus lines studied, here are the conclusions that emerge:

- The scheduled times are made not to be exceeded. An adjustment of these schedules remains possible, insofar as data corresponding to the needs are provided.
- Complete and non-erroneous GPS data allow schedules to be adjusted as needed; otherwise, these data can significantly distort the study.
- Confidence levels associated with speeds on HERE segments vary depending on the time of day. The speeds provided by HERE for peak hours are globally observed speeds while, for off-peak hours, the speeds are modeled.
- The ratio between HERE travel times and GPS travel times was proposed to assess the pertinence of HERE data. Four bus lines were the subject of a comparative analysis. The best result observed is that of line 27 (Lorraine - Sainte-Thérèse in the direction of Bois-des-Filion via Lorraine) if we consider the value of the ratio concentrated at more than 80% between 0.75 and 0.9. This analysis allowed us to conclude that, firstly, the HERE data represents road times by car and that they must therefore be adjusted to represent times by bus and that, secondly, the pertinence of the HERE data depends on the period of the day and the geographical area targeted.

This dissertation ends with a general conclusion and a clear explanation of the limitations and recommendations for future research.

Keywords: GTFS, GPS, HERE, Planning

TABLE DES MATIÈRES

REMERCIEMENTS	iii
RÉSUMÉ	iv
ABSTRACT	vi
TABLE DES MATIÈRES	vii
LISTE DES TABLEAUX	ix
LISTE DES FIGURES	x
LISTE DES SIGLES ET ABRÉVIATIONS	xii
CHAPITRE 1 INTRODUCTION	1
1.1 Mise en contexte	1
1.2 Objectifs de ce mémoire	2
1.3 Contenu	2
CHAPITRE 2 REVUE DE LITTÉRATURE	3
2.1 Temps de parcours de bus : définition	3
2.1.1 Types de temps de parcours	4
2.1.2 Fiabilité des temps de parcours	4
2.1.3 Mesure de temps de parcours	6
2.2 Temps de parcours routiers	7
2.3 Temps de parcours de transport en commun	10
CHAPITRE 3 MÉTHODOLOGIE	14
3.1 Méthodologie générale	14
3.2 GTFS	16
3.2.1 Structure	16
3.2.2 Extraction de données GTFS	17
3.3 GPS	18
3.3.1 Structure	18
3.3.2 Synchronisation du temps GPS	19
3.3.3 Extraction des données GPS	20

3.3.4	Nettoyage des données GPS	20
3.3.5	Scénarios d'appariement	24
3.4	Données HERE	29
3.4.1	Structure	29
3.4.2	Précision	30
3.4.3	Outil d'extraction Larochele	31
3.5	Extraction de données HERE	32
3.6	Développement d'un outil de visualisation permettant la comparaison des temps de parcours mesurés	34
CHAPITRE 4 EXPÉRIMENTATION ET RÉSULTATS		37
4.1	GTFS Ligne 22 - Bois-des-Filion - Sainte-Thérèse	37
4.2	GPS Ligne 22 - Bois-des-Filion - Sainte-Thérèse	41
4.3	HERE Ligne 22 - Bois-des-Filion - Sainte-Thérèse	45
4.4	Analyse comparée Ligne 22 - Bois-des-Filion - Sainte-Thérèse	51
4.5	Expérimentation Autres lignes	55
CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS		58
5.1	Synthèse du travail	58
5.2	Limitations	59
RÉFÉRENCES		60

LISTE DES TABLEAUX

Tableau 2.1	Synthèse des articles sur les temps de parcours routiers	9
Tableau 2.2	Synthèse des articles sur les temps de parcours des transports en commun	13
Tableau 3.1	Description des informations recomposées d'un voyage	17
Tableau 3.2	Structure d'un enregistrement GPS	19
Tableau 3.3	Les champs de la table «VITESSE» de base de données, HERE . . .	30
Tableau 3.4	Les niveaux de confiance HERE	31

LISTE DES FIGURES

Figure 2.1	Trajet d'une ligne de bus (n arrêts)	3
Figure 2.2	Embarquement et débarquement des passagers aux(n) arrêts desservis	5
Figure 3.1	Méthodologie générale	15
Figure 3.2	Voyages planifiés sur la ligne exo-22 pour la date 2020-10-12	18
Figure 3.3	Exactitude limitée du positionnement des données GPS	21
Figure 3.4	Méthode du polygone	21
Figure 3.5	Méthode de la maille de polygones	22
Figure 3.6	Identification de tracé planifié de voyage en question	22
Figure 3.7	Découpage d'un tracé sur (n) tronçons	23
Figure 3.8	Traçage d'un (n) polygones	23
Figure 3.9	Maille de polygones	24
Figure 3.10	Séquence d'enregistrements GPS valide selon la maille de polygones .	25
Figure 3.11	Séquence d'enregistrements GPS non valide (1)	26
Figure 3.12	Séquence d'enregistrements GPS non valide (2)	26
Figure 3.13	Séquence d'enregistrements GPS non valide (3)	27
Figure 3.14	Séquence d'enregistrements GPS non valide (4)	28
Figure 3.15	Enregistrements effectués à la station de départ ou d'arrivée malgré l'état d'arrêt de bus	28
Figure 3.16	Études de cas : exemples	33
Figure 4.1	Projection de tracé de la ligne-direction à l'étude depuis le site http://geo- json.io/	38
Figure 4.2	Horaires de départ de la ligne-direction à l'étude	38
Figure 4.3	Distribution des temps de parcours planifiés de la ligne-direction à l'étude	39
Figure 4.4	Distribution des temps de parcours au cours d'une journée de service selon les périodes de la journée	40
Figure 4.5	Couverture de données GPS	41
Figure 4.6	Distribution des temps de parcours réalisés de la ligne-direction à l'étude	42
Figure 4.7	Répartition des voyages selon la comparaison entre l'horaire de départ prévu et l'horaire réel de départ.	43
Figure 4.8	Visualisation des valeurs des retards au moment de départ	44
Figure 4.9	Heures pour lesquelles les données HERE ont été extraites	45
Figure 4.10	Distribution des temps de parcours HERE de la ligne-direction à l'étude	46

Figure 4.11	Pourcentage de kilomètres avec des vitesses HERE auxquels des niveaux de confiance égaux ou supérieures à 30 sont attribués par rapport à la longueur de la ligne	49
Figure 4.12	Pourcentage de kilomètres avec des vitesses HERE auxquels des niveaux de confiance égaux à 5 sont attribués par rapport à la longueur de la ligne	50
Figure 4.13	Comparaison des temps de parcours réalisés aux temps de parcours planifiés	51
Figure 4.14	Diagrammes "boîte à moustaches" de temps de parcours GPS et temps de parcours GTFS	52
Figure 4.15	Diagrammes "boîte à moustaches" de temps des parcours GPS, GTFS et HERE	53
Figure 4.16	Distribution du ratio entre les temps de parcours fournis par HERE et GPS	55
Figure 4.17	Distribution du ratio entre les temps de parcours fournis par HERE et GPS -Ligne A	56
Figure 4.18	Distribution du ratio entre les temps de parcours fournis par HERE et GPS -Ligne B	57
Figure 4.19	Distribution du ratio entre les temps de parcours fournis par HERE et GPS -Ligne C	57

LISTE DES SIGLES ET ABRÉVIATIONS

APC	Automatic Passenger Counting
AVL	Automatic Vehicle Location
BlueTOAD	Bluetooth Travel Time Origin and Destination
CSV	Comma-separated values
GPS	Global Positioning System
GTFS	General Transit Feed Specification
STEWARD	Statewide Traffic Engineering Ware-house for Regionally Archived Data

CHAPITRE 1 INTRODUCTION

1.1 Mise en contexte

En ces temps où les changements climatiques nous montrent l'urgence de réduire nos empreintes environnementales et réduire l'émission de gaz à effet de serre (GES), un des moyens privilégiés pour y parvenir est le transport en commun. Cependant, attirer les citoyens vers les transports en commun n'est souvent pas une tâche facile. Il s'agit d'un objectif complexe qui pose de grands défis. Les attentes des usagers des services de transport en commun s'articulent autour de 4 axes fondamentaux : la sécurité, le confort, l'accessibilité, sans oublier l'élément le plus important, la fiabilité des temps de parcours. En utilisant les réseaux d'autobus urbains, les utilisateurs cherchent en effet à arriver à leur destination à l'heure et, au pire cas, avec un retard acceptable. Le temps de parcours réel doit égaler le temps de parcours annoncé avec une faible marge d'erreur qui les réconfortera dans leur choix d'utiliser les transports publics. Dans le cas contraire, les usagers se verront obligés de délaisser le transport collectif au profit de la voiture (transport privé), même si ce dernier coûte plus cher. Selon le Centre d'études sur les réseaux, les transports, l'urbanisme et les constructions publiques (2018), 45% des utilisateurs de transport confirment pouvoir accepter un retard ne dépassant pas dix minutes pour un temps indiqué de trente minutes. Ceci nous amène à souligner l'importance de fournir à l'utilisateur une information pertinente sur le temps que prendra un bus pour faire un parcours. Mission difficile à réaliser pour les sociétés de transport en commun, surtout dans le cas où le bus ne pourra pas disposer d'un site propre (voie réservée) et qu'il devra circuler dans des zones de circulation mixte. Sans compter l'achalandage, l'expérience des chauffeurs et la météo, la prédiction de temps de parcours devrait aussi prendre en considération l'état de la circulation routière. Pour ce faire, il est indispensable d'obtenir des jeux de données suffisamment représentatifs des conditions de circulation afin que l'horaire puisse mieux refléter les temps de parcours entre les arrêts. Les planificateurs utilisent plusieurs types de données pour déterminer les temps de parcours en vue de fabriquer les horaires. Le choix du type de données dépend généralement de la précision souhaitée et des ressources disponibles. Parmi ces données, on trouve : les données issues des enquêtes, les données de transaction des cartes à puce et les données de localisation GPS des bus qui en sont équipés. Il y a également des fournisseurs privés qui fournissent des données qui représentent l'état de la circulation routière. Chacune de ces données présente des avantages et des inconvénients qui lui sont propres des points de vue de la précision et du coût.

1.2 Objectifs de ce mémoire

Cette recherche consiste à proposer une méthodologie et un outil permettant de comparer les temps de parcours de bus de transport en commun obtenus à partir de trois sources : les horaires planifiés (GTFS), les traces GPS des bus et les données provenant du fournisseur HERE. L'objectif est de déterminer les caractéristiques de ces données et dans quel contexte elles peuvent être utilisées pour évaluer la fiabilité des temps de parcours.

De cet objectif découlent ces sous-objectifs :

- améliorer les méthodes et outils informatiques d'extraction et de traitement des données utilisées ;
- développer une interface de visualisation permettant la comparaison des temps de parcours mesurés ;
- analyser la pertinence d'utiliser des données de vitesse sur des segments routiers provenant d'un fournisseur commercial dans le cadre de mesure du temps de parcours de bus.

1.3 Contenu

Ce mémoire fait état d'un projet qui a été réalisé en partenariat avec l'entreprise exo, un organisme public de transport en commun québécois qui exploite des services de transport collectif par trains et autobus, ainsi que le transport adapté pour desservir principalement la couronne nord et la couronne sud de Montréal, au Québec, Canada. Nous nous intéressons ici au transport collectif de voyageurs par autobus sur le secteur des Laurentides, qui comprend les villes de Blainville, Bois-des-Filion, Boisbriand, Deux-Montagnes, Lorraine, Mirabel, Pointe-Calumet, Rosemère, Sainte-Anne-des-Plaines, Saint-Eustache, Sainte-Marthe-sur-le-Lac, Sainte-Thérèse, Saint-Jérôme et Saint-Joseph-du-Lac.

Ce mémoire comporte cinq (5) chapitres. Ce premier chapitre est consacré à l'introduction, où l'on a défini le contexte de l'étude, la problématique et les objectifs. Le chapitre 2 est une revue de littérature permettant de rappeler les travaux menés jusqu'à ce jour dans la modélisation du trafic routier et le calcul des temps de parcours sur un réseau de transport collectif par autobus. Le chapitre 3 décrit premièrement la méthodologie générale, deuxièmement les données et troisièmement les étapes de traitement permettant d'atteindre les sous-objectifs. Le chapitre 4 expose les résultats d'application de la méthodologie proposée sur quelques lignes d'autobus dans le secteur des Laurentides. Le chapitre 5 conclut ce mémoire. Il montre les contributions de cette recherche, mais aussi ses limitations et ses perspectives d'amélioration pour de futures recherches.

CHAPITRE 2 REVUE DE LITTÉRATURE

Cette revue de littérature a pour objectif d’explorer les travaux concernant les différentes sources de données et méthodes d’estimation de temps de parcours pour le transport routier, qu’il s’agisse de la circulation automobile ou des services de transport en commun par bus, utilisées dans les études précédentes. Nous allons également mettre en évidence les avantages et les limites de chaque méthode et discuter des résultats obtenus. Cette revue de littérature permettra de mieux comprendre les techniques actuellement disponibles pour l’estimation des temps de parcours et de fournir des recommandations pour les futures recherches dans ce domaine.

2.1 Temps de parcours de bus : définition

Comme la figure 2.1 le montre, un trajet d’une ligne de bus est constitué de n arrêts entre deux terminus, dans une direction (Ferrettini et al., 2023).



FIGURE 2.1 Trajet d’une ligne de bus (n arrêts)

Voici quelques définitions des termes de temps de parcours tirés de Bourbonnais (2022).

1. Le temps qu’un bus doit prendre pour faire le parcours complet, appelé le temps de parcours de bus, est le temps écoulé depuis le terminus de départ jusqu’au terminus d’arrivée.
2. Le temps de parcours de bus moyen, sur un itinéraire pendant une période donnée, correspond à la moyenne arithmétique des temps de parcours sur cet itinéraire durant cette période.
3. Le temps de parcours de segment ou inter-arrêts est le temps de parcours entre deux arrêts consécutifs qui n’inclut pas les temps d’arrêt.
4. Le temps de parcours de bus sur une ligne est défini comme égal à la somme des temps

de parcours de tous les segments de cette ligne et les temps durant lesquels le bus est à l'arrêt complet, permettant l'embarquement et le débarquement des passagers.

2.1.1 Types de temps de parcours

Les temps de parcours diffèrent par leur mode de mesure ou d'estimation (Pesteil et al., 2008). Pour une ligne, dans une direction, on peut mentionner trois variantes :

- Le temps de parcours planifié est le temps de parcours diffusé aux usagers dans les horaires officiels. Ce temps est calculé en fonction de la distance du déplacement, le nombre d'arrêts et d'usagers, et la vitesse commerciale.
- Le temps de parcours de bus estimé est le temps que l'on prévoit que le bus va prendre pour parcourir un trajet déterminé, en tenant compte des facteurs tels que les embouteillages, les perturbations de la circulation, les accidents, etc. Ce temps peut être estimé en utilisant des données de circulation en temps réel et les données GPS de localisation des bus en temps réel.
- Le temps de parcours de bus réellement effectué est le temps réalisé pour parcourir un trajet déterminé. Il est généralement obtenu à partir des données GPS, données issues des boucles électromagnétiques ou des capteurs vidéo.

2.1.2 Fiabilité des temps de parcours

Fournir une information fiable et exacte sur le temps de parcours aux usagers de la route est crucial pour garantir leur pleine satisfaction envers le service offert. Une information fiable, fournie au bon moment, permet aux usagers de planifier leurs déplacements en toute confiance d'atteindre leur destination à temps. Pour les usagers habituels ou fréquents d'une ligne de bus, le risque qui leur semble se poser pour l'arrivée à l'heure provient d'un départ très en retard du bus ou d'une augmentation abusive du temps à bord de véhicule. Si le bus part à l'heure exacte indiquée, le seul risque d'arriver en retard à destination, à prendre en compte, sera une augmentation abusive du temps à bord de véhicule. Pour une ligne de bus offre un service avec N arrêts entre la station de départ A et la station d'arrivée B, le temps à bord du véhicule est égal à la somme des temps de parcours inter-arrêt et les temps d'embarquement et de débarquement des passagers aux arrêts desservis Bourbonnais (2022).

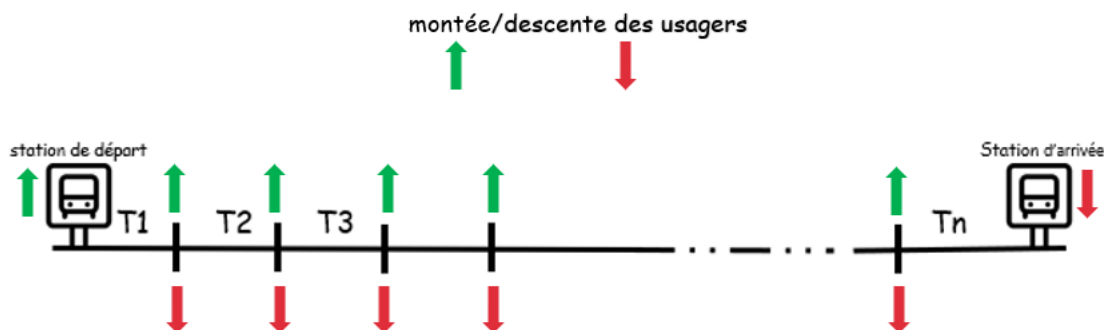


FIGURE 2.2 Embarquement et débarquement des passagers aux(n) arrêts desservis

Le temps de parcours inter-arrêt (entre deux arrêts consécutifs) inclut le temps à vitesse programmée, les temps d'accélération et de décélération, ainsi que les temps d'arrêt aux feux de circulation ou en congestion.

Si le bus suit un itinéraire établi entièrement en site propre, la somme des temps de parcours inter-arrêt prend une valeur prédéterminée de manière fixe ou se situe dans une zone prédéterminée de manière fixe. Le temps d'attente aux arrêts correspond à la somme des temps nécessaires à faire monter et descendre les usagers à chaque arrêt. Il varie selon le nombre d'usagers montant et descendant du bus à chaque arrêt. Le temps à bord de véhicule pour un voyage sur une ligne dans une direction varie selon plusieurs facteurs. Certains sont intrinsèquement liés à la ligne de bus comme :

- la variabilité du nombre de passagers (Shalaby et Farhan, 2004) qui peut avoir comme conséquence une variation de temps nécessaire à la montée et à la descente des usagers aux arrêts ;
- la disparité des comportements des chauffeurs, surtout face à des situations inattendues
- et les caractéristiques des autobus(Pesteil et al., 2008).

D'autres facteurs extérieurs, tels que les conditions de trafic, sont une des causes majeures de la variabilité du temps de parcours. Peuvent aussi s'ajouter les conditions météorologiques(Zhou et al., 2017) ou toute autre situation anormale (cas d'accident de la route). Tous les exploitants de transport en commun visent à prédire de manière efficace le temps de trajet de leurs véhicules afin de mieux en planifier les trajets dans un premier temps et dans un deuxième temps assurer la satisfaction des usagers. Dans la poursuite de cet objectif, les exploitants de transport en commun ont mis en exergue comme priorité la nécessité de renforcer la collecte des données servant d'analyser la réalité du service délivré et à résoudre les problèmes liés à l'estimation du temps de parcours d'un bus entre un point

A et un point B.

2.1.3 Mesure de temps de parcours

Cette étude s'intéresse au traitement et à la visualisation de données historiques de circulation. Ainsi, en combinant ces dernières avec d'autres données, on pourra obtenir davantage d'informations. L'usage classique de ces données a pour but une planification d'un réseau de transport en commun plus efficace. Nous présentons ici brièvement trois types de données décrites plus en détail au chapitre 3, et qui sont utilisées dans les travaux décrits à la suite, dans cette revue.

Données de localisation des véhicules

Les données provenant des systèmes de localisation automatiques des véhicules (Automated Vehicle Location, AVL) sont le plus souvent sous la forme de données GPS (Global Positioning System). Elles sont connues sous plusieurs termes qui sont couramment utilisés comme les données de localisation GPS des bus, les relevés GPS des bus, ou les traces GPS des bus. Elles sont considérées comme des empreintes de déplacements, enregistrées grâce à un système de localisation de la position à la surface de la Terre installé dans le bus. Une trace GPS est donc une succession d'enregistrements de positions reflétant un déplacement. Chaque enregistrement d'une trace GPS est associé à un instant : la trace GPS est dite horodatée. Chaque enregistrement renferme habituellement les éléments suivants : la date, l'heure de lecture, la longitude, la latitude, un identifiant du bus et des identifiants de ligne. L'acquisition de ces enregistrements s'effectue habituellement à intervalles réguliers, toutes les 5 secondes par exemple. Il est aussi parfois possible que l'acquisition de ces enregistrements se limite à chaque arrêt de bus ou à chaque ouverture de portes, selon la précision choisie pour la mesure et le budget attribué.

Données présentant le service planifié

Normalisé par Google en 2005, le GTFS (General Transit Feed Specification) est un format commun pour les informations d'horaires des transports publics et également de données géographiques associées (Nazem, 2014). C'est un flux de données qui permet aux agences de partager leurs données et aux développeurs d'intégrer ces données et de développer des applications au service des usagers du transport en commun (par exemple l'application Transit de Google Maps, TransitApp et Mapnificent). La norme GTFS concerne les données théoriques, c'est-à-dire les horaires planifiés, autrement dit, la norme représente le service qui

a été prévu par la société exploitante du transport en commun. Le format gtfs est un jeu de données qui est composé d'une série de fichiers textes au format CSV qui sont généralement rassemblés dans des zips fournis par les agences de transport. Malgré le soin considérable apporté par les exploitants de transport en commun à la planification, il est possible que certains ajustements soient faits et que le service réellement fourni puisse différer légèrement de ce qui a été planifié face à des situations imprévues. Pour une ligne de bus offrant un service avec N arrêts entre la station de départ A et la station d'arrivée B, les ajustements puissent être apportés au niveau des heures d'arrivée et de départ de bus depuis les N arrêts et de tracé entre les deux terminus. Ces ajustements sont décrits dans les données General Transit Feed Specification in RealTime (GTFS - RT). Une extension de la norme GTFS, a été développée pour permettre aux agences de transport de fournir des mises à jour en temps réel sur l'état de leur service (horaires/positions véhicules/alertes). Pour fournir aux usagers l'option la mieux adaptée pour arriver à destination, certains calculateurs de trajets prennent en considération les données GTFS – RT (Jariyasunant, Mai et Sengupta, 2011).

Données de trafic routier

Les données de trafic routier constituent d'immenses bases de données contenant les données historiques de circulation sur le réseau routier. Cela sert de base à aider les planificateurs et les décideurs du transport collectif à surmonter les problèmes de congestion pour la réalisation d'une planification opérationnelle plus efficace. Ses données de circulation routière examinent les valeurs de vitesse moyenne pour une journée typique. Elles fournissent aussi des projections complètes des temps de conduite basées sur la collecte de milliards de points de données réunis à partir de l'observation des vitesses des véhicules s'étalant sur plusieurs années, y compris les enregistrements de données de sondes GPS. Dans la section suivante, nous décrivons plus en détail les travaux relatifs à l'utilisation de ces données.

2.2 Temps de parcours routiers

Dans cette section, nous recensons certains travaux de recherche touchant la prévision des temps de parcours routiers. Plusieurs types de données ont été largement étudiés ces dernières années lesquelles ont débouché sur des résultats prometteurs. Cette section présente des travaux de recherche qui ont utilisé ces données. Le travail de Loustau (2009) porte sur la modélisation des temps de parcours en s'appuyant sur les données historiques sur les conditions de circulation, issues principalement de véhicules flottants, recueillies sur une période de 7 ans par le ministère des Transports du Québec. Les résultats obtenus lors de cette étude ont été évalués par comparaison avec des résultats obtenus à l'aide d'un autre

type de données (traces GPS des véhicules de Communauto). Pour prévoir le temps de parcours sur des tronçons routiers, Saunier et Morency (2011) ont utilisé les données de capteurs Bluetooth et de bandes vidéo. Au cours de ce travail, ils ont également abordé la comparaison de ces données avec les traces GPS et les données de véhicules flottants selon le coût, les couvertures spatiale et temporelle. En utilisant les données de temps de parcours sur les artères de la Floride ainsi que sur des tronçons d'autoroutes recueillies à l'aide de véhicules instrumentés, (Kondly et al., 2018) ont comparé ces différentes sources obtenues par des méthodes différentes : (STEWARD (Statewide Traffic Engineering Warehouse for Regionally Archived Data), INRIX, BlueTOAD (Bluetooth Travel-time Origination and Destination), et HERE afin de trouver celles qui reflètent au mieux les temps de parcours. Les résultats de cette comparaison suggèrent que les données provenant de fournisseur HERE offrent une meilleure estimation des temps de parcours sur l'autoroute par rapport aux données des autres fournisseurs. Pour le cas des autoroutes congestionnées, HERE présente une mesure précise et ayant un certain sens contrairement aux autres données qui donne habituellement une valeur sous-estimée de temps de parcours effectivement réalisé. Des surestimations de la valeur du temps de parcours sont parfois constatées, mais en restant dans la marge de tolérance admissible. Par contre, dans le cas inverse, le cas pour les autoroutes non congestionnées, STEWARD et INRIX sont favorisés par rapport aux autres dans le but d'effectuer une estimation plus précise. Pour les artères, l'analyse se limite à une comparaison entre les données HERE et celles de BlueTOAD (les seules données disponibles). Les résultats de cette analyse ont suggéré qu'aucune de ces données ne peut assurer une bonne prédiction du temps de parcours, probablement en raison de la petite taille des données étudiées, surtout en situation de congestion.

Dans le même contexte, pour se faire une idée de la réalité des performances de données de vitesses fournies par INRIX, Kim et Coifman (2014) ont utilisé les données de circulation des véhicules collectées à l'aide des boucles électromagnétiques comme données de référence. L'étude couvre 14 millions de kilomètres de l'autoroute inter-États qui traverse les états du Kentucky et de l'Ohio. Il est ressorti de cette étude que les vitesses fournies par INRIX accusent toujours un certain retard (près de 6 minutes) par rapport aux vitesses obtenues par l'utilisation des boucles électromagnétique. Dans l'étude de Zhang et Rice (2003), les données employées pour construire un modèle linéaire de prévision des temps de parcours sur les autoroutes sont issues à la fois de capteurs installés sur les véhicules et d'un ensemble de capteurs encastrés dans la chaussée (détecteurs à boucle). Dans Kwon, Coifman et Bickel (2000), les données utilisées proviennent de deux sources : les données sur le débit et le taux d'occupation fournies par les boucles électromagnétiques implantées sous la chaussée et les données historiques sur les temps de parcours. À partir d'une analyse de régression linéaire

basée sur ces données, les auteurs ont montré que les temps de parcours issus des boucles magnétiques assurent une prédiction plus précise à court terme (jusqu'à 20 minutes), alors que les données historiques sont plus pertinentes pour les prédictions à long terme. Le tableau suivant résume les articles qui ont été étudiés.

TABLEAU 2.1 Synthèse des articles sur les temps de parcours routiers

Auteurs	types de données	Résultats
Loustau (2009)	les données historiques sur les conditions de circulation, issues principalement de sources flottantes	
Saunier et Morency(2011)	les Données de capteurs Bluetooth et de bandes vidéo	
Kondly et al(2018)	STEWARD (Statewide Traffic Engineering Warehouse for Regionally Archived Data), INRIX, BlueTOAD (Bluetooth Travel-time Origination and Destination), HERE	les données HERE offrent une meilleure estimation des temps de parcours sur l'autoroute par rapport aux autres données
Kim and Coifman (2014)	données de circulation des véhicules collectées à l'aide des boucles électromagnétiques, données INRIX	les vitesses fournies par INRIX accusent toujours un certain retard (près de 6 minutes) par rapport aux vitesses obtenues par l'utilisation des boucles électromagnétique
Zhang and Rice(2003)	données fournies par les capteurs installés sur les véhicules et d'un ensemble de capteurs encastrés dans la chaussée (détecteurs à boucle)	
Kwon,Coifman and BickelLes(2000)	les données sur le dé-bit et le taux d'occupation fournies par les boucles électromagnétiques implantées sous la chaussée et les données historiques sur les temps de parcours	les données historiques sont plus pertinentes pour les prédictions à long terme

2.3 Temps de parcours de transport en commun

Cette section décrit les travaux de recherche visant à estimer les temps de parcours en transport en commun. Plusieurs types de données ont été étudiés ces dernières années, lesquelles ont débouché sur des résultats prometteurs. Dans le cadre du projet à Blacksburg, en Virginie dont l'objectif était de maximiser l'utilisation des transports publics grâce à une meilleure information délivrée à l'utilisateur, quatre algorithmes de prédiction de l'heure d'arrivée des bus ont été proposés dans l'étude de Lin et Zeng (1999). Les paramètres d'entrée sur lesquels l'estimation du temps de parcours se base diffèrent d'un algorithme à l'autre. Le premier algorithme détermine l'heure d'arrivée d'autobus au prochain arrêt en utilisant sa position en temps réel et une matrice prédéfinie des temps de parcours. Le deuxième algorithme utilise, en plus de données utilisés par le premier, les horaires planifiés. Différemment de deux premiers algorithmes qui supposent que le bus roule à une vitesse constante même avec un retard déjà existant, les algorithmes 3 et 4 prennent en considération la valeur de ce retard en basant sur la réalité que les conducteurs adaptent leur vitesse en fonction de ce dernier. Ce qui diffère entre les deux est que l'algorithme 4 prend aussi en considération le temps d'attente avant l'embarquement à un arrêt. Les résultats d'évaluation de tous les algorithmes ont indiqué que l'algorithme 4 présente la meilleure performance de l'estimation des temps de parcours, l'erreur moyenne étant de 2,0 minutes environ. L'erreur maximale est de 3,6 minutes. En accédant à des données historiques, aussi bien qu'en temps réel de positionnement de bus, Weigang et al. (2002) ont développé un algorithme capable d'estimer les heures d'arrivée de bus aux prochains arrêts. La position exacte d'un véhicule et sa vitesse sur la route sont les deux données utilisées comme paramètres d'entrée de l'algorithme. Cette vitesse est obtenue en interpolant deux valeurs : la vitesse instantanée du véhicule et la moyenne arithmétique des vitesses enregistrées pour des périodes de temps similaires. Plus la distance entre la position de véhicule et l'arrêt d'autobus est élevée, plus la formule mathématique utilisée donne plus de poids à la vitesse issue des données historiques par rapport à la vitesse instantanée du véhicule et inversement. Il est à noter que les déviations dues à un accident de la circulation ou à des travaux imprévus, les travaux d'entretien du réseau routier et les temps d'attente dans les encombrements ne sont pas pris en considération dans cette étude. Les auteurs Yu et al. (2017) ont eu recours à trois types de données : les données de localisation des véhicules (AVL), les données de comptage automatique des passagers à bord (APC) en temps réel et les données météorologiques. Les données météorologiques comprenaient la température moyenne enregistrée durant l'heure considérée, le taux moyen de précipitation enregistré durant l'heure considérée et l'épaisseur de la neige au sol pour la journée en question. Deux types de modèles ont été comparés

dans cette étude : des modèles de survie et des modèles de régression linéaire. La profondeur de neige et la température n'ont pas été des facteurs mathématiquement significatifs pour expliquer la variable de temps de parcours. Cependant, ces variables ont été incluses dans le modèle car elles ont permis d'améliorer l'ajustement global du modèle.

Pour la prédiction de temps de parcours sur une ligne de bus, (Yang et al., 2016) ont basé leurs calculs sur les informations de circulation les plus récentes disponibles. Ils ont utilisé des données sur l'état de la circulation provenant des vitesses de déplacement des bus précédents sur la même ligne de bus. À part les données du trafic routier, le modèle développé au cours de cette étude prend également en considération la période de la journée, les conditions météorologiques. Morency et al. (2011) ont mis l'accent sur d'autres données pour fournir de l'information de planification sur les temps de parcours : des données de circulation de fournisseurs de données tels INRIX et RealTraffic. En comparant les temps de parcours réels avec les temps de parcours estimées à partir des données INRIX et RealTraffic, Morency et ses collègues ont trouvé que les données fournies par INRIX sont plus fidèles à la réalité. Pour obtenir une estimation précise du temps de parcours d'un bus pour un segment donné, Sun et al. (2016) ont utilisé une méthode de clustering et des filtres de Kalman qui prennent en paramètre les retards pour tous les bus de toutes les lignes sur le même segment à la même période. Le modèle a été mis à l'essai à Nashville, Tennessee, et a permis de diminuer les erreurs de la prédiction de l'heure d'arrivée de 25 % en moyenne et les erreurs à 15 minutes de 47 %. Trépanier et Phillippe (2019), lors d'un projet de la ville de Sherbrooke qui a dépassé le simple cadre théorique pour arriver jusqu'à l'implémentation, ont établi un modèle de prédiction de retards d'autobus grâce aux données opérationnelles à disposition fournies par la Société de transport de Sherbrooke. Une table de données opérationnelles pour le trajet donné d'un jour précis est une séquence d'enregistrements à chaque arrivée/départ d'arrêt. Chaque enregistrement reprend les champs suivants : l'identification du bus, sa géolocalisation, l'heure, l'arrêt où il se trouve et son retard réel au moment de l'enregistrement. Selon eux, pour aller plus loin dans l'exactitude des prédictions, la prise en considération des données de circulation, les données d'achalandage ou encore les données d'événements spéciaux pourraient être la bonne solution. Le service de transport par bus de la Ville indienne de Chennai a fait l'objet de plusieurs travaux de recherche. Nous avons choisi de citer deux recherches qui analysent la problématique de la prédiction de temps de parcours de bus. La première étude a été effectuée par Ramakrishna et al. (2008). Dans le but de prédire le temps de parcours sur une route déterminée, cette étude a utilisé les données GPS et les données sur le nombre de passagers des autobus issues d'une collecte manuelle. Par la suite, les chercheurs ont confirmé l'amélioration apportée par les données sur le nombre de passagers au niveau de l'exactitude de la prédiction du temps de

trajet par rapport à l'utilisation des seules données GPS. La deuxième étude compte parmi les premières tentatives pour prédire en temps réel l'arrivée des autobus aux arrêts. Dans leur étude, Vanajakshi et al. (2009) ont présenté une technique de filtrage Kalman appliquée à des données GPS avec des résultats assez satisfaisants. Cristobal et al. (2019) ont proposé un modèle basé sur la similarité des profils. Ayant les données historiques sur les temps de parcours, les auteurs ont utilisé un algorithme de clustering k-medoid afin d'avoir des profils types. Ensuite, les auteurs terminent avec une comparaison de profil du bus dont ils veulent prédire le temps de parcours avec l'ensemble des profils construits pour prédire le temps de parcours. Les résultats ont indiqué que le modèle présente une marge d'erreur absolue moyenne de 13 % (calculée en additionnant toutes les erreurs absolues et en les divisant par le nombre d'erreurs). Cette marge d'erreur peut varier en fonction de la durée et de la longueur du trajet. Wall et Dailey (1999) ont été les premiers à introduire le modèle du filtre de Kalman pour prédire avec précision l'heure d'arrivée des bus. Les chercheurs ont combiné les données AVL avec des données historiques dans leur algorithme, ils ont utilisé un modèle de filtre de Kalman pour déterminer la position exacte du bus à partir des données AVL et ont utilisé une technique d'estimation statistique pour estimer le temps de parcours. Les résultats de cette étude ont été très satisfaisants. L'heure d'arrivée de bus est déterminée avec une marge d'erreur inférieure à 12 %. Xinghao et al. (2013) ont présenté un modèle qui permet d'estimer en temps réel le temps d'arrivée de bus à l'arrêt basée sur les données AVL et les données RFID. Les données RFID réelles n'étaient pas disponibles au moment de l'étude, les chercheurs ont utilisé des caméras pour collecter des données similaires à celles des systèmes RFID. Seuls les taxis et les bus ont été ciblés pour une surveillance par caméra, aucune autre catégorie de véhicules n'a été incluse. Avant de développer leur modèle, les auteurs ont d'abord confirmé la validité de l'estimation des vitesses des bus à partir de celles des taxis avec une étude de régression. Ils ont observé qu'en situation de congestion, les bus et les taxis avaient des vitesses similaires, et que le coefficient de régression était proche de 1,00. En revanche, dans des conditions de trafic fluide, les taxis avaient une vitesse plus élevée que les bus, ce qui se traduisait par un coefficient supérieur à 1,00. Le modèle qu'ils ont créé utilise comme paramètres le temps de roulage (déduit à partir de la vitesse de roulage obtenue à partir des caméras de surveillance), les temps d'attente aux feux, ainsi que les délais d'accélération et de décélération. Les auteurs ont fini par tester le modèle proposé sur deux lignes de bus afin d'évaluer sa performance. Pour cela, ils l'ont comparé à un modèle basé uniquement sur des données historiques AVL. Les résultats ont montré que La performance de modèle utilise des données de sources multiples a surpassé la performance observée avec le modèle utilisant uniquement les données AVL.

TABLEAU 2.2 Synthèse des articles sur les temps de parcours des transports en commun

Auteurs	Données utilisées	Résultats
Lin et Zeng (2017)	Position en temps réel, matrice prédéfinie des temps de parcours, données sur les retards, temps d'attente avant l'embarquement à un arrêt	Estimation des temps de parcours avec une erreur moyenne de 2,0 minutes environ. L'erreur maximale est de 3,6 minutes.
Weigang et al (2002)	position de véhicule, La vitesse instantanée du véhicule, données historiques	
Yu et al. (2017)	les données de localisation des véhicules (AVL), les données de comptage automatique des passagers à bord (APC) en temps réel, les données météorologiques	les données météorologiques ont permis d'améliorer l'ajustement global du modèle.
Morency et al (2011)	données de circulation de fournisseurs de données INRIX et RealTraffic	
Sun et al (2016)	retards pour tous les bus de toutes les lignes sur le même segment à la même période	Le modèle a permis de diminuer les erreurs de la prédiction de l'heure d'arrivée de 25% en moyenne et les erreurs à 15 minutes de 47% .
Ramakrishna et al (2008)	données GPS et données sur le nombre de passagers des autobus issues d'une collecte manuelle	l'amélioration apportée par les données sur le nombre de passagers au niveau de l'exactitude de la prédiction du temps de trajet par rapport à l'utilisation des seules données GPS
Cristobal et al (2019)	les données historiques sur les temps de parcours	le modèle présente une marge d'erreur absolue moyenne de 13%
Wall et Dailey (1999)	données AVL, données historiques	L'heure d'arrivée de bus est déterminée avec une marge d'erreur inférieure à 12%
Xinghao et al (2013)	données AVL et les données RFID	La performance de modèle utilise des données de sources multiples a surpassé la performance observée avec le modèle utilisant uniquement les données AVL

CHAPITRE 3 MÉTHODOLOGIE

3.1 Méthodologie générale

Afin de pouvoir atteindre les objectifs énoncés, il est crucial de suivre une démarche structurée en plusieurs étapes (voir figure 3.1).

1. Extraction des voyages GTFS planifiés sur la ligne-direction à l'étude pour la période d'étude choisie. Les horaires de départs, les temps de parcours et tracés des bus sont les données sur lesquelles se basent les étapes qui suivent.
2. Extraction à partir de données représentant le service effectivement fourni, la séquence GPS correspondant à chaque voyage planifié sur la ligne direction à l'étude pour la période d'étude choisie.
3. Comme des erreurs peuvent se glisser dans les données GPS, une vérification des traces GPS extraites qui précède normalement toute utilisation de ces derniers doit être effectuée.
4. Cette étape est divisée en deux parties :
 - identification suivie d'une extraction des tronçons HERE constituant le trajet planifié ;
 - extraction des données de vitesses HERE appropriées pour les différents tronçons en se basant sur les horaires planifiés.
5. Visualisation sur graphiques des caractéristiques des différentes données.
6. Analyse et comparaison des temps de parcours obtenus des trois sources de données

En collaboration avec exo, le cadre applicatif de ce projet est le réseau de transport collectif d'autobus dans la région des Laurentides, une région du Québec au nord de Montréal. Le projet bénéficie d'un accès fourni par exo à trois bases de données, pour la région des Laurentides :

1. les données GTFS présentant le service planifié offert par la CITLA pour l'année 2020 ;
2. les données GPS de bus du CITLA (exo-Laurentides) pour la période de février 2020 à février 2021 représentant le service effectivement fourni ;
3. les données de trafic routier HERE pour les années 2018 à 2020.

Les analyses sont faites sur le mois d'octobre 2020.

3.2 GTFS

Nous avons reçu de exo deux bases de données en format CSV : l'une réunit les voyages sur le secteur des Laurentides pendant 2020 nommé «gtfs2020_voyages.zip», tandis que l'autre contient chaque arrivée et chaque départ d'arrêt sur le même secteur et pendant la même période (année 2020) nommé «gtfs2020_passagesarrets.zip». Les fichiers sont très gros, atteignant plusieurs dizaines de gigaoctets.

3.2.1 Structure

Commençons par une brève description de la base de données de voyages. Le tableau qui suit présente les champs de cette base de données. En utilisant les deux champs Date et trip_id, il est possible d'identifier un voyage de façon unique. Autrement dit, deux voyages dans la même journée de service ne peuvent pas avoir la même valeur de trip_id.

TABLEAU 3.1 Description des informations recomposées d'un voyage

Champ	Description
Date	Journée de service
trip_id	Identifiant unique du voyage pour une journée de service
route_id	Identifiant unique du parcours pour une journée de service
direction_id	Identifiant unique de la direction
agency_id	Identifiant unique de l'exploitant
departure_time	Heure de départ depuis le Terminus de départ
arrival_time	Heure d'arrivée au Terminus d'arrivée
dist_traveled	Distance parcourue entre les deux terminus
Duration	Temps total entre le départ du terminus de départ et l'arrivée au Terminus d'arrivée.
nb_stops	Nombre d'arrêts desservis lors de voyage
start_stop_name	Nom de la station de départ
end_stop_name	Nom de la station d'arrivée
geom	Le parcours emprunté par le bus sur la totalité de l'itinéraire

Passant à la deuxième base de données, celle qui présente les passages-arrêts, chaque ligne de la table présente une arrivée ou un départ d'arrêt, chaque arrivée ou départ d'arrêt est caractérisé par l'identifiant de voyage auquel il appartient et un numéro qui précise l'ordre d'arrêt lors de voyage. Ces deux éléments constituent l'identification unique d'un passage arrêt. On trouve aussi d'autres attributs comme le nom de l'emplacement qui doit être compréhensible pour les usagers, la longitude et la latitude de l'emplacement et l'heure d'arrivée et celle de départ qui sont toujours identiques.

3.2.2 Extraction de données GTFS

Notre but premier est de rendre les données GTFS facilement utilisables pour nos besoins techniques. Pour ce faire, nous avons développé un algorithme dont les paramètres sont la ligne, la direction et la période permettant de sortir la séquence de voyages planifiés. Ensuite, en utilisant la colonne géométrie, on ajoute une colonne « identifiant du tracé » pour donner une identification au tracé utilisé, ce qui nous permet de savoir si la ligne utilise un seul tracé ou plusieurs tracés, ce qui est possible pour certaines lignes dans le but de répondre à des besoins particuliers ou ponctuels variant au courant d'une journée. Prenons un exemple pour la ligne 22, direction Bois-des-Filion - Sainte-Thérèse et pour la date de 12 octobre 2020 : on trouve 16 voyages planifiés de 08 :15 le matin jusqu'à 23 :15 le soir. Pour tous les voyages, la

ligne utilise un même tracé.

date	jour	route_id	direction_id	trip_short_name	departure_time	arrival_time	duration(min)	start_stop_name	end_stop_name	nb_stops	geom_id	dist_traveled
2020-10-12	Lundi	"22"	0	"2"	08:15:00	08:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"4"	09:15:00	09:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"6"	10:15:00	10:40:00	25	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"8"	11:15:00	11:40:00	25	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"10"	12:15:00	12:40:00	25	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"12"	13:15:00	13:40:00	25	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"14"	14:15:00	14:41:00	26	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"16"	15:14:00	15:40:00	26	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"18"	16:14:00	16:40:00	26	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"20"	17:15:00	17:41:00	26	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"22"	18:15:00	18:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"24"	19:15:00	19:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"26"	20:15:00	20:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"28"	21:15:00	21:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"32"	0	"30"	22:15:00	22:38:00	23	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-12	Lundi	"22"	0	"32"	23:15:00	23:37:00	22	Gare Sainte-Thérèse	Câ'te de Terrebonne / George-VI	42	geom_0	10.61

FIGURE 3.2 Voyages planifiés sur la ligne exo-22 pour la date 2020-10-12

En continuant sur l'exemple pris, on trouve cette table décrivant les passages arrêts du premier départ à 08 :00 le matin. Il s'agit d'une première étape importante, qui nous permet d'avoir une idée détaillée sur le service planifié pour la ligne-direction en question. En d'autres mots, cette étape nous permet de présenter la couverture spatiale et temporelle de la ligne-direction en question. Les données générées à la suite de cette étape vont nous aider lors de la deuxième étape qui consiste à extraire et nettoyer les données GPS. Une sous-étape de vérification et nettoyage de données GPS extraites est extrêmement nécessaire vu que contrairement aux données présentant le service planifié, les données GPS peuvent contenir différents types d'erreurs.

3.3 GPS

Les données GPS collectées par les autobus qui circulent sur le réseau routier des Laurentides ont été fournies par exo pour cette étude. Les données pour la période allant du premier octobre 2020 au 31 octobre 2020 ont été extraites, comprenant plus de trente millions d'enregistrements GPS.

3.3.1 Structure

Chaque enregistrement reprend les champs décrits ci-dessous. Ces champs sont séparés par une virgule (,).

TABLEAU 3.2 Structure d'un enregistrement GPS

Champ	Description	Exemple
Latitude longitude de la position du bus		45.734719166666 -73.823122500000
Date de l'enregistrement	DD/MM/AAAA	2020-01-08 00 :00 :00.000
date et l'heure GPS de la lecture	à la seconde près DD/MM/AAAA HH :MM :SS	2020-01-08 22 :40 :57.000
Azimut	Angle horizontal mesuré dans le sens horaire à partir d'une direction donnée (le nord par exemple)	346.620000000000
Agency_id	Information sur l'agence de transport	CITLA
route_short_name	Version courte du nom d'itinéraire(ne donne aucune indication sur les lieux desservis)	23
Trip_short_name	l'identification unique d'un trajet au cours d'une journée de service pour une ligne dans une direction	36
Direction_id	la direction du trajet	Sens

3.3.2 Synchronisation du temps GPS

L'heure obtenue à l'aide d'un système GPS est le temps GPS. En date de 6 janvier 2010, le temps GPS a été réglé sur l'UTC (Temps universel coordonné) qui est une échelle de temps standard. Contrairement au GPS, l'UTC est occasionnellement mis à jour par une incrémentation ou décrémentation d'une seconde intercalaire (*leap second*) par rapport au Temps atomique international (TAI) qui est un temps stable basé sur l'observation de transitions atomiques, élaboré par le Bureau International des Poids et Mesures (BIPM) pour rester cohérent avec la rotation de la Terre qui définit le temps universel. Depuis le 31 décembre 2016, le temps GPS est de 18 secondes de plus que le temps UTC. Dans cette étude, la conversion de l'heure GPS en heure locale de Montréal s'est effectuée en deux étapes : d'abord, la conversion de l'heure GPS en UTC, puis une conversion à partir de l'UTC en heure locale. La première étape se fait en soustrayant 18 secondes de l'heure GPS pour obtenir l'heure UTC. Vient ensuite la deuxième étape, la conversion est effectuée en utilisant le langage de programmation Python et la fonction `timezone` de la librairie `Pytz`.

3.3.3 Extraction des données GPS

Cette méthode s'appuie sur les données GTFS déjà traitées lors de la première partie. Ce qui est attendu, pour chaque ligne de la table de voyages GTFS, est d'identifier son équivalent (séquence d'enregistrements) dans la table de données GPS (traces GPS). Comme les traces GPS sont des données enregistrées à l'aide d'un dispositif d'enregistrement électronique et que des erreurs peuvent se produire, il est possible de ne pas trouver l'équivalent d'un voyage planifié dans la base de données GPS. Dans le cas inverse, on peut croire qu'un voyage a été ajouté par suite d'une cause (étrangère) imprévisible et inévitable. Pour les deux bases de données, celle qui présente les voyages planifiés et celle qui présente les voyages effectivement fournis, chaque voyage ou séquence d'enregistrements est caractérisé par une ligne (Route_id), une direction (Direction_id) et son identifiant unique au cours d'une journée de service (trip_short_name). Ces trois éléments constituent l'identification unique d'un voyage. À l'aide de ces trois éléments, on peut identifier, pour chaque voyage de la table GTFS, sa séquence d'enregistrements dans la table de données GPS par une série de filtrage. Une sous-étape de vérification et nettoyage de données GPS extraites est nécessaire vu que contrairement aux données présentant le service planifié, les données GPS peuvent contenir différents types d'erreurs.

3.3.4 Nettoyage des données GPS

Prenons un exemple facile pour vraiment comprendre la méthode utilisée dans cette partie de la méthodologie dans sa forme la plus simple. Supposons que nous voulions comparer un voyage GTFS planifié avec son équivalent GPS réellement effectué. En mettant à notre disposition, comme données les horaires de passage planifiés aux arrêts et les traces GPS de bus, la première idée qui vient à l'esprit est d'abord de sortir à partir de l'horaire planifié l'heure de départ depuis le terminus de départ et l'heure d'arrivée au terminus d'arrivée et à partir des traces GPS, le premier et le dernier enregistrement, ensuite, de déduire le temps de parcours planifié et celui réellement réalisé et enfin, de les comparer pour estimer l'adéquation entre le service proposé et celui planifié. Comme on l'a évoqué dans le chapitre précédent, il est possible qu'un conducteur ait pris un autre trajet que celui planifié. Dans ce cas, la comparaison n'a guère de sens. Il est indiscutable qu'une vérification de similarité entre les deux trajets est nécessaire avant de sortir les deux temps de parcours. La vérification s'effectue en contrôlant si les points GPS sont situés sur le tracé de référence (planifié). Cette vérification (point sur une ligne droite) n'est pas parfaite sur le plan pratique ; par exemple, comme l'illustre la figure ci-après, un point GPS appartient à une route spécifique, ce n'est pas forcément qu'il appartient à son tracé. Deux raisons peuvent expliquer cette situation ;

premièrement, comme tout type de capteur, le capteur GPS présente une exactitude limitée et son degré d'exactitude du positionnement GPS horizontal se trouvait dans un rayon de 1 m et, deuxièmement, le tracé de référence ne représente que l'axe central de la route, il ne prend pas la largeur exacte de la route en compte.

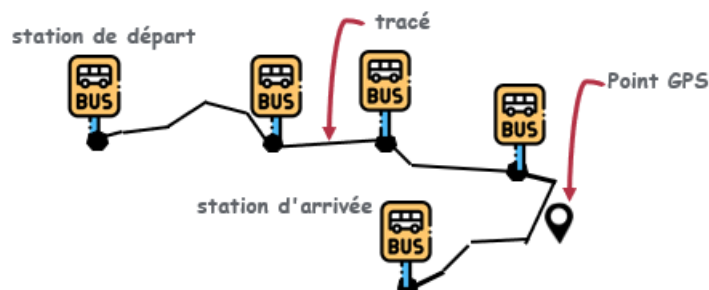


FIGURE 3.3 Exactitude limitée du positionnement des données GPS

L'idée est alors de créer un polygone qui englobe le tracé planifié et, par suite, la vérification sera effectuée par ce polygone à la place du tracé.

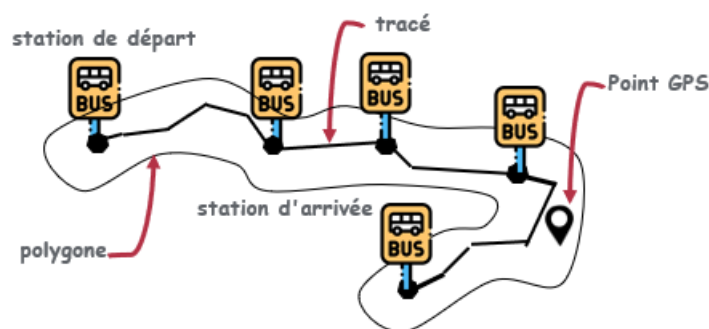


FIGURE 3.4 Méthode du polygone

À ce stade, l'information recueillie par cette méthode est limitée, n'est pas encore suffisante pour dire que le trajet pris par le conducteur soit équivalent au trajet planifié. Donc, même si on trouve que tous les points GPS sont inclus dans le polygone, cela ne garantit pas l'équivalence de deux trajets. Possible que les points GPS ne présentent qu'une partie du trajet. Pour faire face à ce problème, une modification a été apportée à la solution initiale. Cette modification consiste à remplacer le polygone par une maille de polygones.

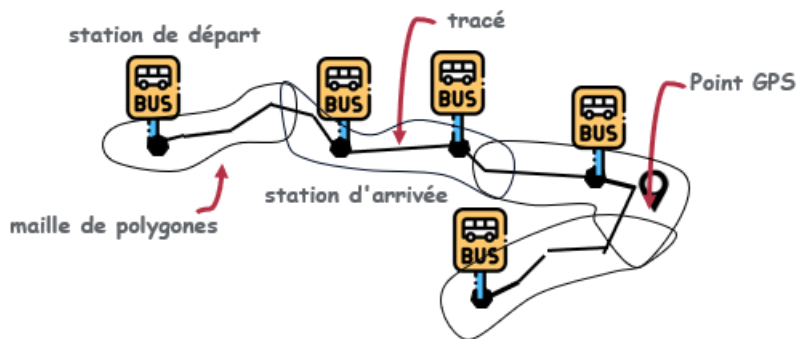


FIGURE 3.5 Méthode de la maille de polygones

La distribution des points GPS sur tous les polygones permet de confirmer que le trajet pris par le conducteur correspond au trajet planifié. Dans ce cas, la comparaison des temps de parcours mesurés fait sens. Dans la partie qui suit, l'algorithme est expliqué plus en détail : Pour un voyage choisi, la première étape consiste à sortir son tracé planifié de la base de données GTFS et le transformer en geojson.io.

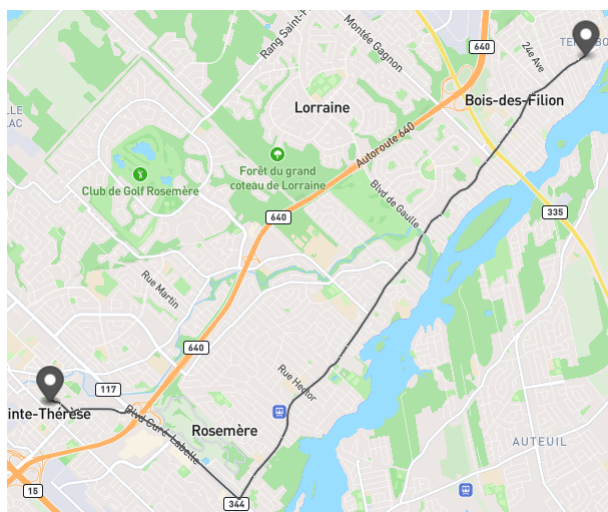


FIGURE 3.6 Identification de tracé planifié de voyage en question

La deuxième est le découpage de tracé sur plusieurs tronçons. Il n'y a pas d'exigence particulière concernant le nombre de tronçons, mais de préférence la division sera en tronçons de même taille.

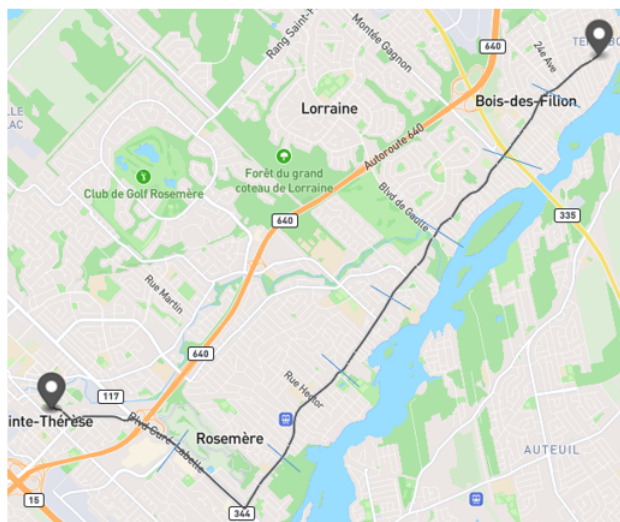


FIGURE 3.7 Découpage d'un tracé sur (n) tronçons

La troisième étape est d'entourer chaque tronçon par un polygone en commençant par le premier tronçon qui inclut le point de départ et ainsi de suite, jusqu'au dernier qui inclut le point d'arrivée. On obtient N polygones numérotés de 0 à N-1. Concernant la taille du polygone, la seule contrainte est de s'assurer que ce dernier englobe la totalité du tronçon par rapport à sa largeur et à sa longueur.

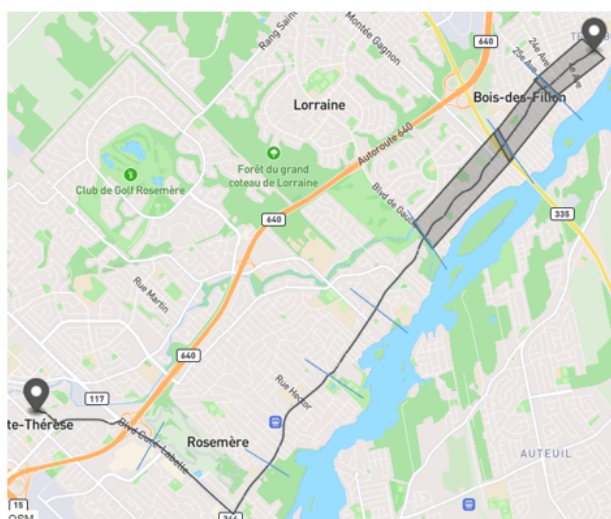


FIGURE 3.8 Traçage d'un (n) polygones

Finalement, en supprimant le tracé, le point de départ et le point d'arrivée, on obtient la maille de polygones qui représente le tracé original. L'exportation de cette maille de polygones dans le format du fichier *.geojson est nécessaire pour le travail qui reste à faire.

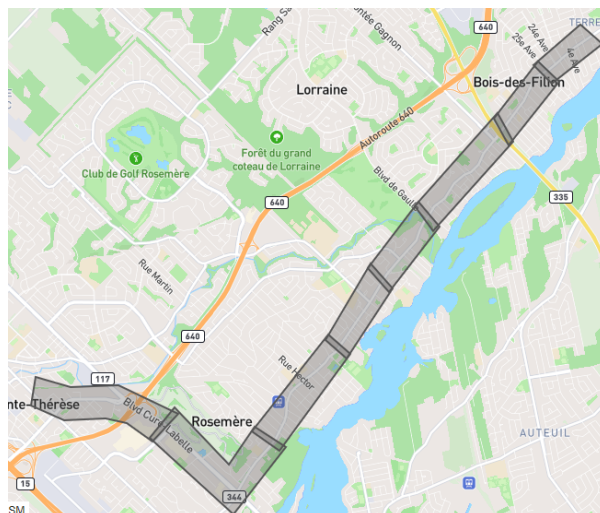


FIGURE 3.9 Maille de polygones

La question est de savoir si le tracé réalisé est contenu dans cette maille de polygones. La réponse à ce problème est fournie par l'application d'un algorithme Python en trois étapes comme suit :

1. Balayage de la séquence d'enregistrements GPS, point par point et à chaque point, déterminer à quel polygone ce point appartient en utilisant les champs latitude et longitude. Si un point n'appartient à aucun polygone, il est hors du tracé planifié et on lui donne la valeur NaN.
2. Ajout d'une colonne « Num_Polygone » à la séquence d'enregistrements GPS (Data frame).
3. Tri de la séquence d'enregistrements GPS (Data frame) par ordre croissant de la date et de l'heure de l'enregistrement.
4. Conversion de la colonne « Num_Polygone » en une liste « Num_Polygone ».

Dans cette étude, le nombre de tronçons ou le nombre de polygones n'a pas été considéré comme une contrainte mais il est intéressant de noter que plus on agrandit le nombre de polygones, plus on recherche la précision de la comparaison entre le tracé réalisé et celui planifié.

3.3.5 Scénarios d'appariement

Cette sous-section décrit les différents scénarios auxquels nous avons été confrontés, et la façon dont nous avons réagi face à chacun d'entre eux. Pour une séquence d'enregistrements

GPS, une première sélection peut s'effectuer sur la base du nombre d'éléments uniques dans la liste « Num_Polygone » ($nb_elements_uniq$). Deux cas peuvent se présenter :

- Soit ($nb_elements_uniq$) = N : Deux exigences remplies permettent de classer la séquence d'enregistrements GPS en tant que valide , premièrement il n'existe pas des points hors de la maille de polygones puis, deuxièmement, les points GPS sont distribués sur tous les polygones. On peut affirmer que le trajet pris par le chauffeur de bus est le même que celui planifié.

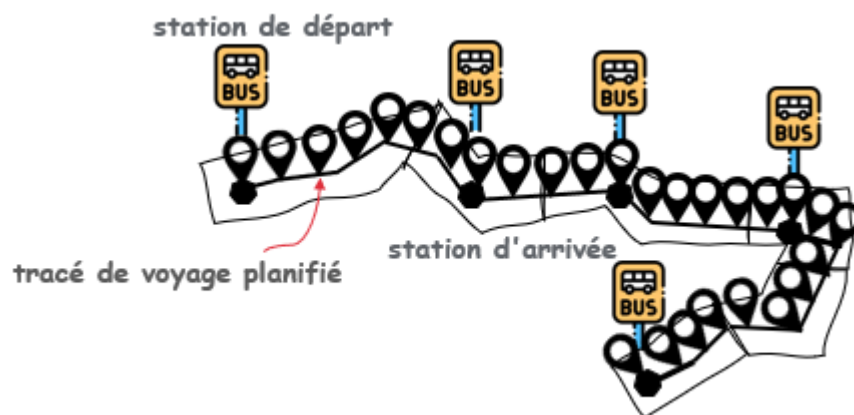


FIGURE 3.10 Séquence d'enregistrements GPS valide selon la maille de polygones

- Soit ($nb_elements_uniq$) = $N + 1$: Une des exigences n'est pas respectée, il existe des points GPS hors de la maille de polygones. L'interprétation est comme suit : la séquence d'enregistrements GPS contient de l'information utile pour notre analyse, considérée comme valide, mais il existe une partie de l'information qui est fausse ou trompeuse, cette partie est susceptible d'affecter la valeur de temps de parcours réellement réalisé. Voilà deux scénarios qui risquent de se produire et qui sont facilement détectables :
1. Le capteur GPS est en mode de fonctionnement ON avant que le bus arrive à la station de départ pour commencer son trajet. Donc le trajet qui peut fausser le calcul de temps du parcours effectivement effectué est probablement un trajet entre le garage d'autobus et la station de départ. Ce scénario est détecté lorsqu'on trouve que la liste «Num_Polygone» commence par une série de valeurs nulles.

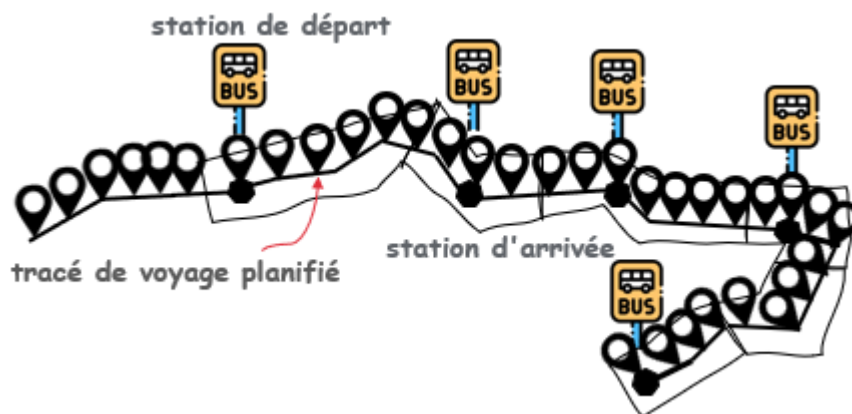


FIGURE 3.11 Séquence d'enregistrements GPS non valide (1)

2. Le capteur GPS est en mode de fonctionnement "On" après que le bus a fait une ligne dans une direction. Donc le trajet qui peut fausser le calcul de temps de parcours effectivement effectué est probablement un trajet vers une autre destination s'il s'agit d'un bus « en transit » ou un trajet vers le garage d'autobus. Ce scénario est détecté lorsqu'on trouve que la liste « Num_Polygone » finit par une série de valeurs nulles.

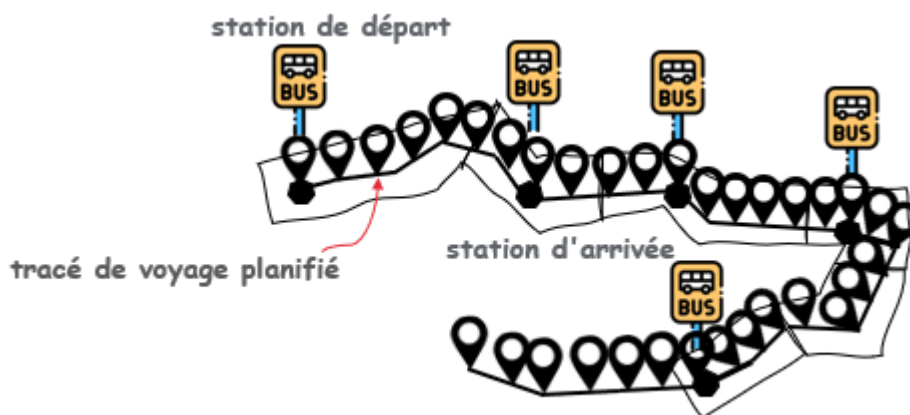


FIGURE 3.12 Séquence d'enregistrements GPS non valide (2)

La solution proposée pour faire face à ces problèmes consiste à supprimer les enregistrements effectués hors du tracé planifié, qui sont les enregistrements effectués avant que le bus arrive à la station de départ pour commencer son trajet pour le premier scénario et les enregistrements effectués après que le bus a fait une ligne dans une direction pour le deuxième scénario.

- Soit $(nb_polygones_uniq) < N$: c'est le pire cas que l'on puisse rencontrer, la distribution des points GPS se limite sur quelques polygones et non pas à toutes. On ne peut donc juger si le trajet pris par le chauffeur de bus est le même que celui planifié. La séquence d'enregistrements GPS ne nous permet pas d'avoir une information utile pour notre analyse. La séquence est considérée comme étant non valide dans le contexte de notre travail. Toutefois, deux scénarios possibles pourraient expliquer ce résultat :

1. La géométrie du trajet enregistré est différente de celle du trajet planifié : il est possible qu'un conducteur ait pris un autre trajet que celui planifié s'il se trouve face à une route bloquée, à une route qui ne lui permet plus de poursuivre dans la direction jusque-là parcourue. Le blocage de la route peut être la cause d'accident, d'incendie, d'une tempête de neige.

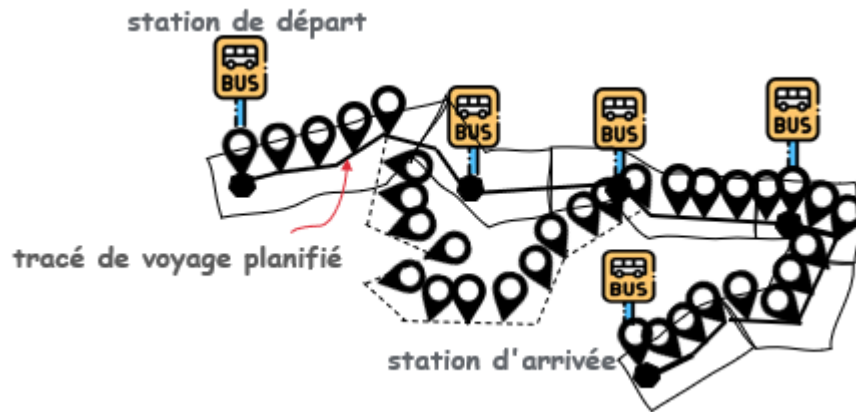


FIGURE 3.13 Séquence d'enregistrements GPS non valide (3)

2. La géométrie du trajet enregistré ne présente qu'une partie de celle du trajet planifié : Ce type de problème peut être lié à une anomalie au niveau de capteur GPS.

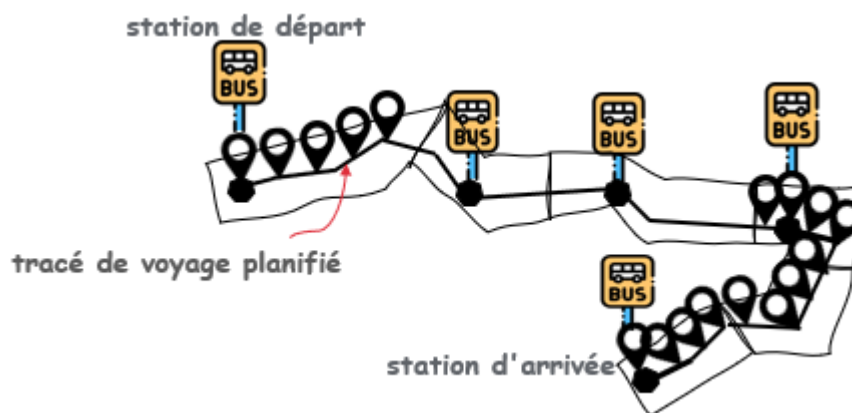


FIGURE 3.14 Séquence d'enregistrements GPS non valide (4)

Avant d'entamer la partie suivante (visualisation), les séquences GPS qui ont été jugées valides doivent passer par une autre étape de nettoyage. Une séquence GPS valide peut être touchée par d'autres types d'erreurs qui ne sont pas enregistrées et ne peuvent pas être réduites par la maille de polygones. Ces erreurs concernent l'existence des enregistrements effectués à la station de départ avant que le bus commence son trajet et des enregistrements effectués à la station d'arrivée après avoir fait une ligne dans une direction. Ces enregistrements sont détectés toujours au début ou fin de la séquence d'enregistrements GPS où on trouve des enregistrements GPS ayant la même latitude et longitude, ce qui veut dire que le capteur GPS est en mode de fonctionnement ON malgré que le bus est dans un état d'arrêt.

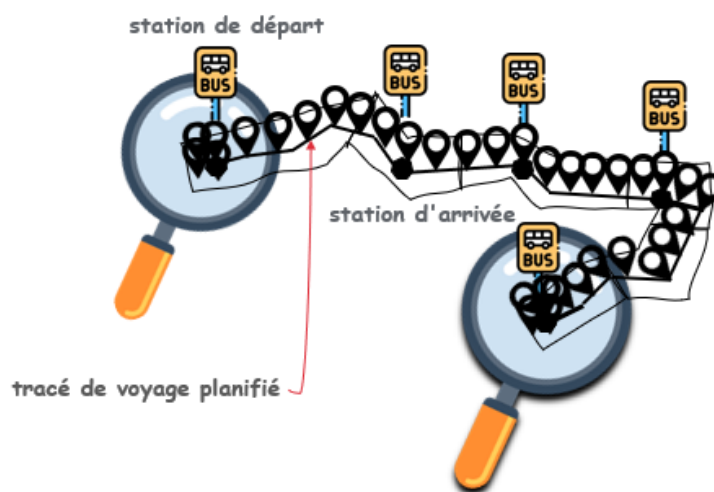


FIGURE 3.15 Enregistrements effectués à la station de départ ou d'arrivée malgré l'état d'arrêt de bus

Face à ce type de problème, dans le premier cas, l'algorithme cherche l'enregistrement à partir duquel le bus était en état de marche, ensuite il supprime tous les enregistrements qui le précèdent. Dans le deuxième cas, l'algorithme cherche l'enregistrement à partir duquel le bus était en état d'arrêt, ensuite il supprime tous les enregistrements qui le suivent. Une fois la séquence d'enregistrements GPS a passé tous les tests, l'algorithme renvoie le temps du parcours réellement réalisé pour ce voyage en calculant la différence entre la date et l'heure des premiers et derniers enregistrements. L'algorithme peut traiter plusieurs séquences à la fois, il suffit de lui passer en paramètres la table de voyages planifiés et la table de voyages réellement effectués correspondante à l'état brut, sans oublier de lui fournir la maille de polygones pour chaque tracé existant. Généralement, une ligne-direction possède entre 1 à 4 tracés différents. Suite à ce procédé, on obtient une base de données qui décrit réellement la différence en temps de parcours entre le service planifié et le service réellement effectué pour la ligne direction en question.

3.4 Données HERE

Les données utilisées pour cette étude sont des données historiques du trafic automobile achetées à HERE, un fournisseur de données de circulation à l'échelle mondiale. Les données HERE sont constituées des vitesses des véhicules enregistrées sur chacun des tronçons routiers au Québec, comptabilisées aux heures et aux 5 minutes. Or, comme le réseau routier québécois est d'une longueur totale d'à peu près 325 000 km, se découpant en quelque 1,5 millions de tronçons directionnels de longueur moyenne approximativement de 250 m, HERE peut fournir grosso modo à exo entre 1 et plus de 12 milliards de nouvelles lignes de données par mois, selon la fréquence horaire ou aux 5 minutes ; ce qui résulte en une croissance mensuelle des volumes de plus de 1 To.

3.4.1 Structure

Pour le portrait géographique, chacun des tronçons routiers au Québec est défini par un identifiant unique, un toponyme officiel et en vigueur, la longueur du tronçon (m) et un indice aux contraintes administratives locales, telles les arrondissements et les limites municipales. Pour le portrait de la circulation par heure, chaque tronçon est décrit heure par heure pile de la journée (12h00 ou 15h00, par exemple) par plusieurs attributs. Le tableau 3.3 recense ces différents champs ainsi que leur description.

TABLEAU 3.3 Les champs de la table «VITESSE» de base de données, HERE

Champ	Description
Datetime	L'horodatage (date et heure)
Road Segment_ID	Numéro identifiant le tronçon routier
Free Flow Speed	Vitesse de véhicule en circulation fluide, calculée à partir de 80e percentile de mesures de vitesses prises la nuit en fin de semaine sur quelques années. Si les mesures prises ne sont pas suffisantes pour atteindre cet objectif, cette vitesse est calculée en tenant compte de la limite de vitesse permise celle qui est autorisée par la loi pour garantir la sécurité.
Mean Speed	Vitesse qui correspond à la moyenne arithmétique des vitesses observées
Sample Count	Nombre des vitesses observées
Confidence	Référence à la précision de la valeur de vitesse moyenne attribuée
Speeds MIN and MAX	Valeurs minimum et maximum des vitesses observées
Percentiles	Percentiles de distribution des vitesses
Gap Filling	Prend la valeur Y ou N, permet de savoir si la valeur de la vitesse moyenne attribuée est basée ou non sur des observations
Speed Limit	Vitesse maximale permise par le Code de la sécurité routière, affichée sur les panneaux de signalisation en bordure des routes.
Standard Deviation	Écart-type de l'ensemble des vitesses observées en kilomètres par heure avec une décimale (0000,0). S'il s'agit d'une seule observation alors l'écart-type est nul.

La combinaison de deux colonnes Datetime et Road Segment_ID permet d'identifier de manière unique chaque ligne de la table.

3.4.2 Précision

Le niveau de confiance attribué à la valeur de vitesse moyenne de tel tronçon du réseau routier à telle heure est calculé à la base de nombre d'observations effectuées sur ce tronçon à cette heure-là et la variance de l'ensemble des vitesses observées. Les valeurs possibles que le niveau de confiance peut prendre et leurs significations sont résumées dans le tableau suivant :

TABLEAU 3.4 Les niveaux de confiance HERE

Valeur de niveau de confiance	Signification
5	«Gap-filled only»
10,15	«Suggestive»
20,25	«Highly suggestive»
30,35	«Confident»
40	«Highly confident»

On peut distinguer deux groupes de valeurs : le premier comprend les valeurs finissant par 0 (10,20,30,40) ; le deuxième inclut les valeurs finissant par 5 (5,15,25,35). Les valeurs du premier groupe sont des valeurs attribuées à des vitesses observées tandis que celles du deuxième groupe sont des valeurs attribuées à des vitesses modélisées. Dans le cas où HERE ne fournit pas d'observations suffisantes pour pouvoir estimer correctement la vitesse moyenne de tel tronçon du réseau routier à telle heure, HERE propose à sa clientèle une vitesse modélisée ; cette dernière peut être déterminée suivant l'une des méthodes suivantes :

1. La première consiste à estimer la vitesse moyenne de ce tronçon à cette heure en utilisant une interpolation entre les deux vitesses enregistrées 15 minutes précédant et suivant cette heure à condition que ses vitesses soient disponibles. Le niveau de confiance attribué à cette vitesse est le minimum entre les deux niveaux de confiance attribués aux deux vitesses utilisées, moins 5.
2. La deuxième est appliquée dans les cas où ses vitesses enregistrées 15 minutes précédant et suivant cette heure ne sont pas disponibles, s'appuie sur les données historiques de vitesses. Le niveau de confiance attribué à cette vitesse est le niveau le plus faible, égale à 5.

3.4.3 Outil d'extraction Larochelle

En se trouvant face à des problèmes de stockage et d'exploitation de données, exo a fait appel à Larochelle Groupe Conseil qui offre des services conseils spécialisés en technologie de l'information et en solutions d'affaires pour proposer, mettre en place et gérer une plateforme de données moderne d'infonuagique apte à fournir les capacités de traitement recherchées tout en satisfaisant des contraintes liées au budget. Le choix a été porté sur la plateforme snowflake pour des raisons techniques et économiques. Cette plateforme inclut également l'API Open Source Routing Machine (OSRM) requis pour le map-matching.

3.5 Extraction de données HERE

Comme nous l'avons déjà mentionné, la base de données HERE Traffic est stockée sur la plateforme snowflake. Sur cette plateforme chaque requête SQL envoyée coûte de l'argent en fonction de sa durée d'exécution, donc l'idée qui consiste d'abord à extraire les données dans des fichiers Excel puis les filtrer selon le besoin est une idée que l'on ne doit pas prendre. Exactement comme nous avons fait avec les données GPS, l'extraction des données HERE se fait à la base des données GTFS. Le principal objectif de notre nouvel algorithme est d'extraire les données des vitesses qui correspondent exactement à nos besoins afin de garantir la réussite dans la durée et d'éviter de payer des frais pour des données qui ne servent à rien. L'algorithme commence donc, comme première étape essentielle par définir la couverture spatiale et temporelle des données à extraire. À partir d'une table de voyages pour une ligne-direction qui ne possède qu'un seul tracé pour un jour donné, l'algorithme renvoie les heures d'arrivée et de départ et la séquence d'arrêts desservis(lat, long). Une fois cette étape réalisée il est possible de passer à la deuxième étape, l'algorithme fait le traçage de trajet d'autobus à l'aide de service match d'une des instances de l'API OSRM, en utilisant comme paramètres la séquence d'arrêts desservis (lat, long). La route tracée est une route qui épouse le mieux la série de points employés comme ressources (les arrêts d'autobus). L'API découpe la route en legs, lesquels contiennent les annotations de type nodes : pour chaque leg, il s'agit en réalité des identifiants des points qui sont disponibles dans la base de données HERE (la table de réseau routier). Il est important de vérifier que le trajet tracé est le même que le trajet planifié avant de passer à l'étape suivante. L'étape suivante est de faire le pont entre l'API OSRM et la base de données snowflake, il s'agit de reconstituer les tronçons HERE pour chaque leg donné à l'aide d'un algorithme qui reçoit deux paramètres :

- une liste d'identifiants de points (par exemple, ceux retournés par l'API OSRM sous les annotations)
 - une liste correspondante qui distingue les nœuds des vertex et retourne une liste de segments, i.e. des listes de points qui correspondent de manière unique à des tronçons HERE.
- Une fois cette étape achevée, ayant identifié les tronçons HERE constituant le trajet planifié, il reste qu'extraire les données des vitesses correspondantes. Les données des vitesses utilisées pour ce projet sont des données heure par heure (0 à 24 h.) et comme les heures de départ planifiées ne sont pas forcément des heures justes, l'algorithme suit une approche simple pour l'identification des données de vitesses à extraire.

Supposons que **HdHd** :**MdMd** :SS et **HaHa** :**MaMa** :SS sont respectivement l'heure de départ et l'heure d'arrivée de bus.

— 1ère Cas **HdHd** :00 :00==**HaHa** :00 :00

On calcule les deux pourcentages suivants : $P1 = \%$ temps de parcours entre **HdHd** :00 :00 et **HdHd** :30 :00 et $P2 = \%$ temps de parcours entre **HdHd** :30 :00 et **HaHa** :00 :00

Si $P1 > P2$: On extrait les données de vitesse HERE pour l'heure **HdHd** :00 :00

Si $P2 > P1$: On extrait les données de vitesse HERE pour l'heure qui suit **HaHda** :00 :00

— 2ème Cas **HdHd** :00 :00 != **HaHa** :00 :00

On calcule les deux pourcentages suivants : $P1 = \%$ temps de parcours entre **HdHd** :00 :00 et **HdHd** :30 :00, $P2 = \%$ temps de parcours entre **HdHd** :30 :00 et **HaHa** :00 :00, $P3 = \%$ temps de parcours entre **HaHa** :00 :00 et **HaHa** :30 :00 et $P4 = \%$ temps de parcours entre **HaHa** :30 :00 et l'heure pile qui suit **HaHda** :30 :00

Si $\max(P1, P2, P3, P4) = P1$: On extrait les données de vitesse HERE pour l'heure **HdHd** :00 :00

Si $\max(P1, P2, P3, P4) = P2$: On extrait les données de vitesse HERE pour l'heure **HaHa** :00 :00

Si $\max(P1, P2, P3, P4) = P3$: On extrait les données de vitesse HERE pour l'heure **HaHa** :00 :00

Si $\max(P1, P2, P3, P4) = P4$: On extrait les données de vitesse HERE pour l'heure pile qui suit **HaHda** :30 :00

les deux exemples ci-dessus permettent de mieux comprendre l'approche.

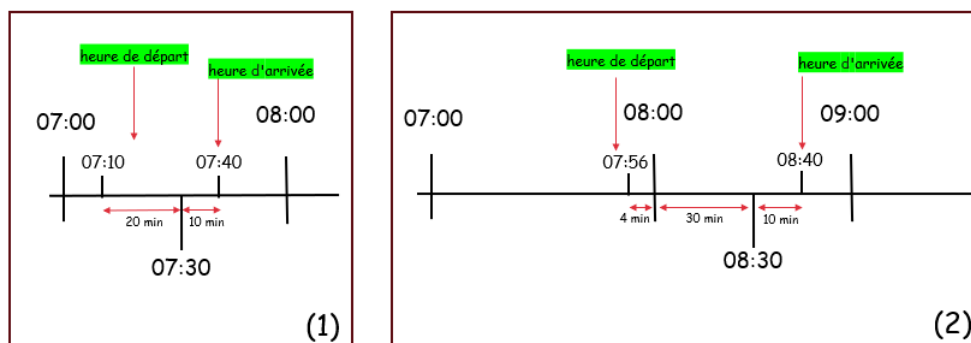


FIGURE 3.16 Études de cas : exemples

Le premier exemple, c'est le cas où $HdHd :00 :00 == HaHa :00 :00$

$P1=66.66\% \parallel P2=33.33\% \rightarrow$ On extrait les données de vitesse HERE pour l'heure 07 :00 :00.

Le deuxième exemple, c'est le cas où $HdHd :00 :00 != HaHa :00 :00$

$P1=9\% \parallel P2=68.1\% \parallel P3=22.72\% \parallel P4=0\%$ OR $\max(P1,P2,P3,P4)=P2$

$\rightarrow$ On extrait les données de vitesse HERE pour l'heure 08 :00 :00 .

3.6 Développement d'un outil de visualisation permettant la comparaison des temps de parcours mesurés

Ce que nous avons voulu faire, ce n'est pas seulement la visualisation des données à l'aide des représentations graphiques, mais aussi fournir une interface utilisateur interactive pour gérer ces derniers. En d'autres mots, la création d'éléments interactifs tels que des curseurs, des menus déroulants, des boutons, etc., qui modifient les graphiques ou les visualisations en fonction des données d'entrée dans le cadre de l'analyse. Cet outil a été développé en Python. La polyvalence, la simplicité d'utilisation et les améliorations constantes obtenues grâce à une communauté très active ont motivé le choix de Python. Pour répondre à l'objectif de créer cette interface, deux bibliothèques open source ont été utilisées : Dash et Plotly Express.

Plotly Express est une librairie optimisée pour l'exploration des données, permet la visualisation d'un jeu de données puis, très rapidement, changer la vue dans des différentes représentations graphiques, par exemple des diagrammes de dispersion, des histogrammes, pour explorer facilement les données.

Dash est un cadre de travail qui permet de créer des applications interactives de visualisation à base web. Les principales étapes pour la réalisation de l'outil sont :

1. La génération de l'application Web à l'aide d'un navigateur.
2. Faire l'appel à Plotly pour générer les graphiques
3. L'ajout des composants interactifs (tels que des curseurs, des menus déroulants, etc.) aux graphiques pour permettre des mises à jour faciles.

L'application web établie contient cinq sections. Commençons par la première section, la section de données utilisateur, celle où figurent :

1. Deux listes déroulantes : la première permet d'identifier la ligne, la deuxième permet d'indiquer le sens de déplacement de la ligne choisie pour l'étude.
 - (0) - Voyage dans une direction (par exemple, voyage sortant).
 - (1) - Voyage dans la direction opposée (par exemple, voyage entrant)

2. Un sélecteur de dates qui permet d'indiquer le début et la fin de la période sous étude avec l'option d'exclure les samedis, dimanches et jours fériés reconnus
3. Un curseur de plage horaire permet de se limiter à l'étude des voyages effectués au cours d'une période spécifique de la journée de service.
4. Un bouton d'exécution : une pression sur ce bouton permet de déclencher l'exécution de l'algorithme en fonction des données saisies.

La deuxième section, la section de données présentant le service planifié – GTFS. Les données sont visualisées en utilisant un histogramme à barres qui indique la valeur du temps de parcours pour chaque voyage GTFS planifié sur la ligne de bus en question, dans la période en question. Pour obtenir de plus amples renseignements concernant les caractéristiques spatiales et temporelles pour chaque voyage, les données de voyages sont affichées dans une table. Un sélecteur de date donne la possibilité de choisir les voyages d'une journée de service précise.

La troisième section concerne les données présentant le service effectif – GPS, les mêmes types de graphiques utilisés dans la section précédente ont été utilisés pour visualiser et analyser ces données avec l'ajout d'autres éléments graphiques qui permettent de comparer ces données à celles du service planifié :

1. Deux graphiques à secteurs : le premier permet d'évaluer la couverture assurée par les données GPS. Plus précisément, ce dernier permet d'identifier le pourcentage des voyages planifiés possédant leurs traces GPS, le deuxième, fut le résultat de la comparaison de l'heure de départ de chaque voyage planifié et celle de son équivalent réellement, fournit des informations sur les pourcentages suivants : Le pourcentage des voyages enregistrés qui ont commencé à l'heure prévue, le pourcentage des voyages qui ont débuté avant l'heure de départ et le pourcentage de voyages qui ont subi des retards et qui ont commencé après l'heure prévue initialement. Ces pourcentages permettent d'analyser la performance de l'algorithme d'extraction et de nettoyer des données GPS.
2. Une représentation matricielle de type heatmap, permet de visualiser clairement les différences entre les horaires de départs planifiés et ceux enregistrés, d'identifier les voyages qui ont subi des retards importants, aux moments de départ. La palette de couleurs utilisée est variée du jaune au rouge pour représenter les valeurs basses à élevées.

Dans la quatrième section, nous présentons les données extraites de la base de données HERE. Les mêmes éléments graphiques de base ont été utilisés pour visualiser et analyser ces données. En plus de ça, nous avons ajouté un outil GEOJSON qui permet de générer une vue des niveaux de confiance associés aux segments HERE constituant le tracé de bus pour une période donnée ou un voyage donné. Un diagramme à barres empilées et une représentation

matricielle pour but de comparer dûment et en détail les niveaux de confiance associés aux segments HERE constituant le tracé des différents voyages.

Pour cette section permettant d'extraire les données de vitesses HERE, deux librairies Python ont été utilisées : la première bibliothèque assure, en premier lieu, la communication avec le moteur de recherche d'itinéraire OSRM et, en deuxième lieu, l'envoi et la réponse aux requêtes de routage à l'API du serveur à l'aide d'appels HTTP GET ou POST. La deuxième bibliothèque assure la connexion avec snowflake et permet d'effectuer toutes les requêtes SQL qui permettent d'extraire les données HERE comme si nous étions sur l'interface web.

Enfin, la dernière section s'appuie sur les précédentes. Elle présente une comparaison des temps de parcours issue de différentes bases de données. Pour ce faire, deux éléments graphiques sont utilisés :

1. Un graphique en boîte à moustaches où chaque boîte représente la distribution des temps de parcours pour un voyage particulier et une source de données spécifique.
2. Une représentation matricielle de type heatmap permet d'étudier le ratio temps de parcours GPS divisé par le temps de parcours GTFS dans le but de comparer le service réellement effectué au service planifié.
3. Un graphique en nuage de points étudiant le ratio temps de parcours GPS divisé par le temps de parcours HERE vise à voir si les temps de parcours HERE sont relativement représentatifs des temps de parcours effectifs.

Deux autres librairies sont indispensables pour développer cet outil. La première librairie, nommée CSV, est utilisée pour lire des données à partir de fichiers CSV. La deuxième librairie, nommée pandas, est utilisée pour extraire, préparer et éventuellement analyser des données.

CHAPITRE 4 EXPÉRIMENTATION ET RÉSULTATS

4.1 GTFS | Ligne 22 - Bois-des-Filion - Sainte-Thérèse

Le choix a été porté sur la ligne de bus 22 dans la direction Bois-des-Filion via Georges IV, pour la période d'octobre 2020 à l'exception des weekends et des jours fériés. L'algorithme d'extraction de données GTFS a permis, en premier lieu, de sortir les voyages GTFS planifiés sur la ligne choisie pour la période d'étude en question. Les données de 504 voyages ont été extraites à partir de la base de données fournie, constituant une nouvelle base de données nommée «gtfs_22_0_2020_10_voyages.xlsx» sur laquelle se base le reste de cette étude. Pour plus de clarté, le fichier est nommé comme suit :

gtfs_ Identifiant de l'itinéraire _direction_ année_mois _voyages
--

En deuxième lieu, cet algorithme a permis d'illustrer le nombre de tracés utilisés par cette ligne de bus en ajoutant à la base de données créée dans l'étape qui précède, une colonne « shape_id » contient l'identifiant de chaque tracé associé aux voyages. La ligne-direction à l'étude possède un seul tracé pour tous les voyages, peu importe la journée ou l'horaire de départ. La distance réelle parcourue le long du tracé (entre le première et le dernier point du tracé) est égale à 10.61 km. La figure 4.1 présente une projection de tracé de la ligne 22 dans la direction Bois-des-Filion via Georges IV depuis le site <http://geojson.io/>.

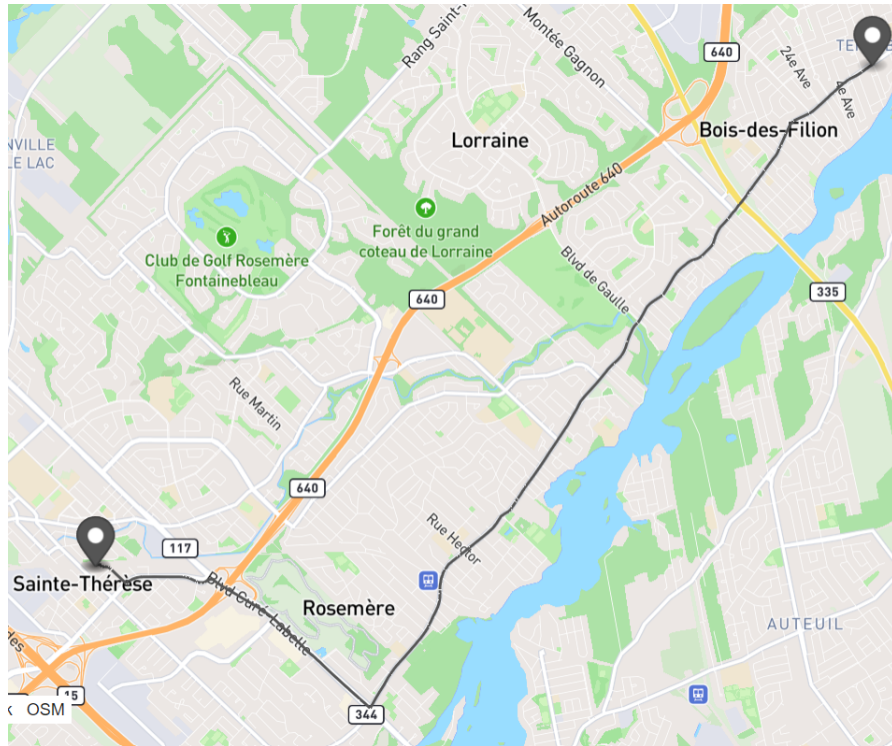


FIGURE 4.1 Projection de tracé de la ligne-direction à l'étude depuis le site <http://geojson.io/>

42 arrêts sont desservis par cette ligne de bus. Le bus fait la navette de la station Gare Sainte-Thérèse Quai 3 à la station George-VI / Côte de Terrebonne. Pour les jours de la semaine, la navette est en service du lundi au vendredi, de 6h36 à 23h15, les horaires de départ sont regroupés selon les périodes de la journée comme suit :

Pointe de matin (AM)	5H-8H59	06H36 06H56 07H33 07H54 08H35 08H59
Hors pointe (HP)	9H-14H59	10H15 11H15 12H15 13H15 14H13
Pointe de l'après-midi (PM)	15H-18H29	15H12 15H45 16H17 17H11 17H41 18H16
Hors pointe (HP)	19H-23H59	18H51 19H15 19H50 20H20 21H15 22H15 23H15

FIGURE 4.2 Horaires de départ de la ligne-direction à l'étude

Dans ce qui suit, nous présentons une description détaillée du service planifié, de point du vue des temps de parcours grâce à des graphiques facilement interprétables. La figure 4.3 est un graphique à barres qui illustre la distribution des temps de parcours au cours d'une journée de service, d'une journée à l'autre de la période d'étude.

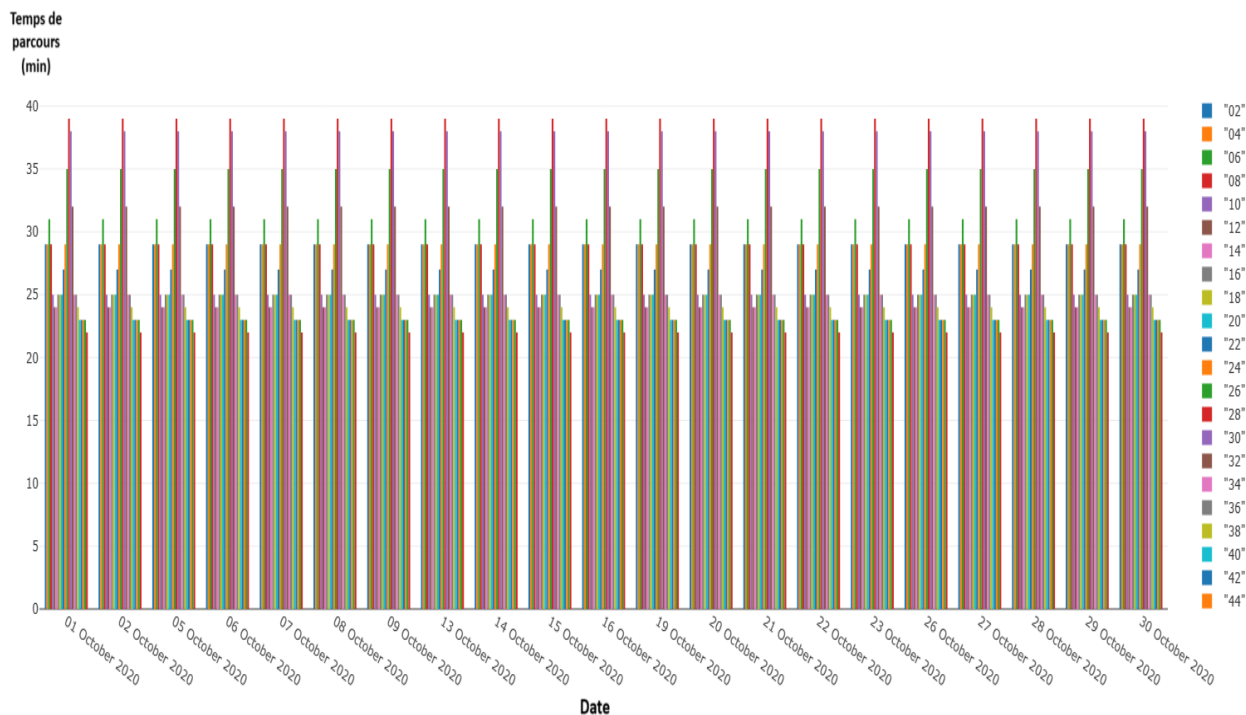


FIGURE 4.3 Distribution des temps de parcours planifiés de la ligne-direction à l'étude

Parmi les principales conclusions, on peut souligner que :

- Les voyages planifiés sur cette ligne de bus pour une journée de service entière ont des horaires plus fréquents et plus denses en période de pointe du matin et celle de l'après-midi allant respectivement de 9h à midi et de 13h à 16h.
- Le temps de parcours est le même pour un voyage à telle heure de départ, quel que soit le jour ouvrable de la semaine. Par exemple, pour un voyage qui part à 6h :36 du matin le temps du parcours est de 29 minutes peu importe le jour de la semaine.
- Pour une journée de service, les voyages présentent un temps de parcours qui varie en fonction de l'horaire de départ. Les voyages pendant les heures de pointe prennent plus de temps, ce qui peut être expliqué par la congestion de la circulation. La figure suivante devrait permettre de visualiser de manière plus claire comment varie le temps de parcours au cours d'une journée de service selon ses différentes périodes (AM, HP, PM)

date	route_id	direction_id	trip_short_name	departure_time	P	duration(min)	start_stop_name	end_stop_name	nb_stops	geom_id	dist_traveled
2020-10-06	22	0	2	06:36:00	Pointe AM	29	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	4	06:56:00		29	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	6	07:33:00		31	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	8	07:54:00		29	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	10	08:35:00		25	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	12	08:59:00		24	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	14	10:15:00	HP	24	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	16	11:15:00		25	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	18	12:15:00		25	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	20	13:15:00		25	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	22	14:13:00	Pointe PM	27	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	24	15:12:00		29	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	26	15:45:00		35	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	28	16:17:00		39	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	30	17:11:00		38	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	32	17:41:00		32	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	34	18:16:00	HP	25	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	36	18:51:00		25	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	38	19:15:00		24	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	40	19:50:00		23	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	42	20:20:00		23	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	44	21:15:00		23	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	46	22:15:00	HP	23	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61
2020-10-06	22	0	48	23:15:00		22	Gare Sainte-Th��r��se	C��te de Terrebonne / George-VI	42	geom_0	10.61

FIGURE 4.4 Distribution des temps de parcours au cours d'une journ  e de service selon les p  riodes de la journ  e

4.2 GPS | Ligne 22 - Bois-des-Filion - Sainte-Thérèse

Comme expliqué dans la description de la méthodologie que nous avons suivie, l'extraction de données GPS a été faite sur la base de données GTFS. Comme résultat, pour chaque voyage GTFS planifié, une séquence des enregistrements GPS a été attribuée. En deuxième étape, chacune de ses séquences a été appliquée au processus de nettoyage et de validation. Sur l'ensemble de 504 séquences GPS, 404 séquences ont été considérées comme valides. Toutes les autres séquences GPS n'ont pas répondu aux critères de validation. Elles ont été considérées comme erronées et non pertinentes pour les fins de la présente discussion, et par conséquent ces séquences ont été exclues de l'étude. Comme l'illustre la figure 4.5, les données GPS ne sont pas disponibles pour 12% de voyages, mais elles sont présentes pour tous les autres.

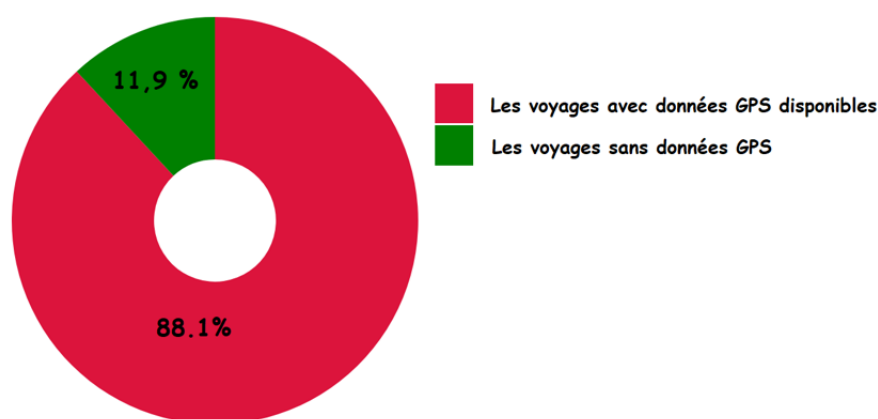


FIGURE 4.5 Couverture de données GPS

Dans la figure 4.6, nous regardons avec les données GPS les vrais temps de parcours. Ce qu'on peut tirer comme informations à partir de cette figure :

1. la plupart des temps de parcours enregistrés sont dans l'intervalle 22-39 minutes. Les limites de cet intervalle sont respectivement le minimum et le maximum des temps de parcours planifiés.
2. Des temps de parcours hors de cet intervalle sont enregistrés ; des temps de parcours inférieurs à la limite inférieure de cet intervalle et d'une fréquence moins élevée des temps de parcours supérieurs à la limite supérieure de cet intervalle ; ce qui s'explique notamment par le fait que les temps de parcours planifiés sont conçus pour ne pas être dépassés.

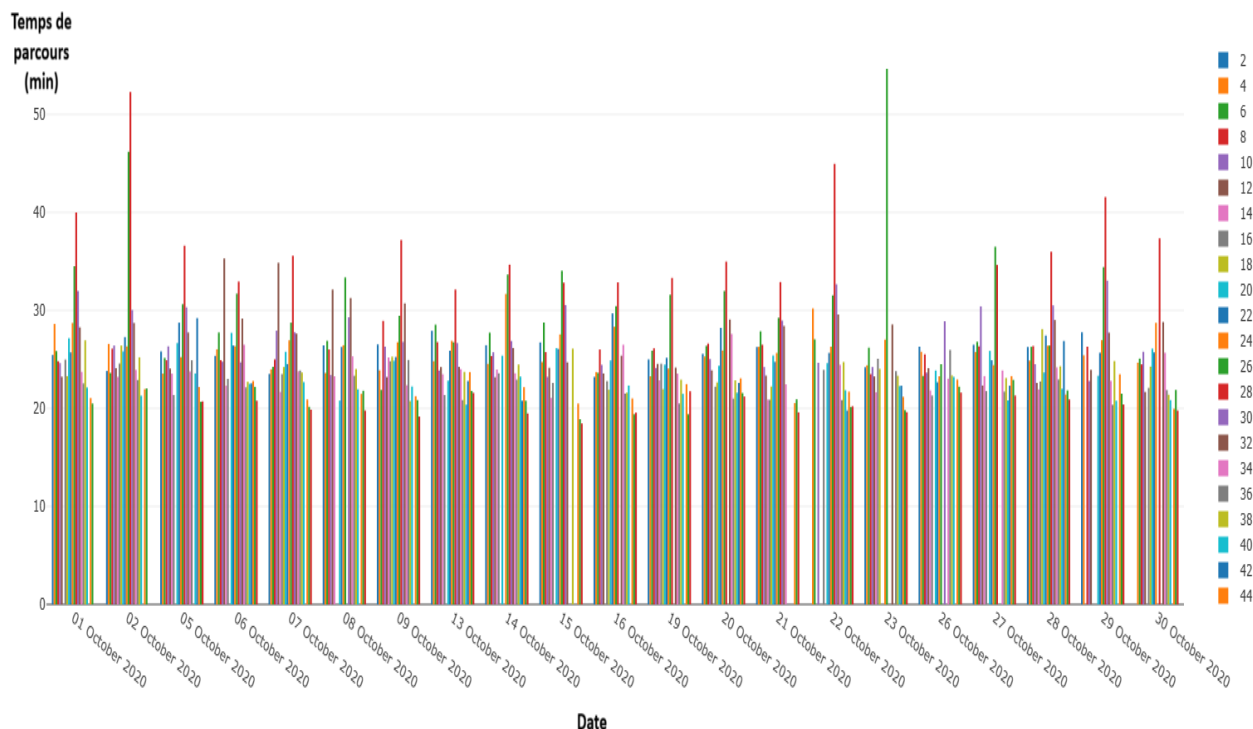


FIGURE 4.6 Distribution des temps de parcours réalisés de la ligne-direction à l'étude

Avant de comparer le service réellement effectué au service planifié, il a été nécessaire de vérifier l'exactitude de nos données afin de ne pas fausser les résultats de notre analyse. Autrement dit, il a été nécessaire de vérifier que la bonne et la plus précise séquence GPS a été attribuée à chaque voyage GTFS planifié. Cette vérification est effectuée en comparant l'heure de départ de chaque voyage planifié et celle de son équivalent réalisé. Les résultats de cette vérification sont présentés dans les figures 4.7 et 4.8. Le graphique 4.7 présente la répartition des voyages selon la comparaison entre l'horaire de départ prévu et l'horaire réel de départ. Un voyage est considéré comme réalisé avec un départ à l'heure prévue si la différence est d'une minute ou moins. On observe que 73.2% des voyages ont commencé à l'heure prévue, un pourcentage ne dépassant pas 1% ont débuté avant l'heure de départ prévue et 25.9% ont subi des retards et ont commencé après l'heure prévue. Il s'agit des résultats prometteurs montrant le bon fonctionnement du processus de nettoyage et de validation de traces GPS, mais aussi on ne peut pas passer sous silence le pourcentage de 25.9% de voyages qui ont commencé en retard par rapport à l'heure prévue. Il est indispensable de visualiser les valeurs de ces retards pour nous permettre de valider une de deux hypothèses suivantes :

- des retards légers, sont défendables sur la base que ces retards peuvent se produire en réalité ;

- des retards importants, qui sont en désaccord avec la réalité. Cela s'explique probablement par le fait que ces retards sont les résultantes des traces GPS imprécises et trompeuses.

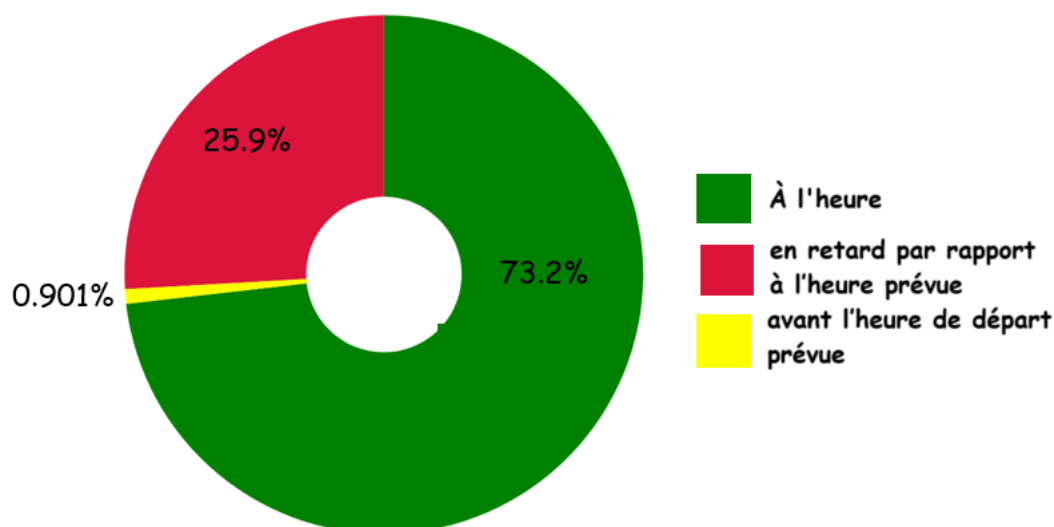


FIGURE 4.7 Répartition des voyages selon la comparaison entre l'horaire de départ prévu et l'horaire réel de départ.

Le graphique 4.8 est une représentation matricielle de type heatmap permettant de comprendre les valeurs issues de la comparaison entre l'horaire de départ prévu et l'horaire réel de départ. On observe que pour la plupart des voyages qui ont commencé en retard par rapport à l'heure prévue, les retards sont légers, défendables et ne peuvent pas nous faire douter du bon fonctionnement du processus de nettoyage et de validation des traces GPS. Les cellules de la matrice avec des points au centre représentent les voyages qui ne disposent pas de leurs séquences GPS.

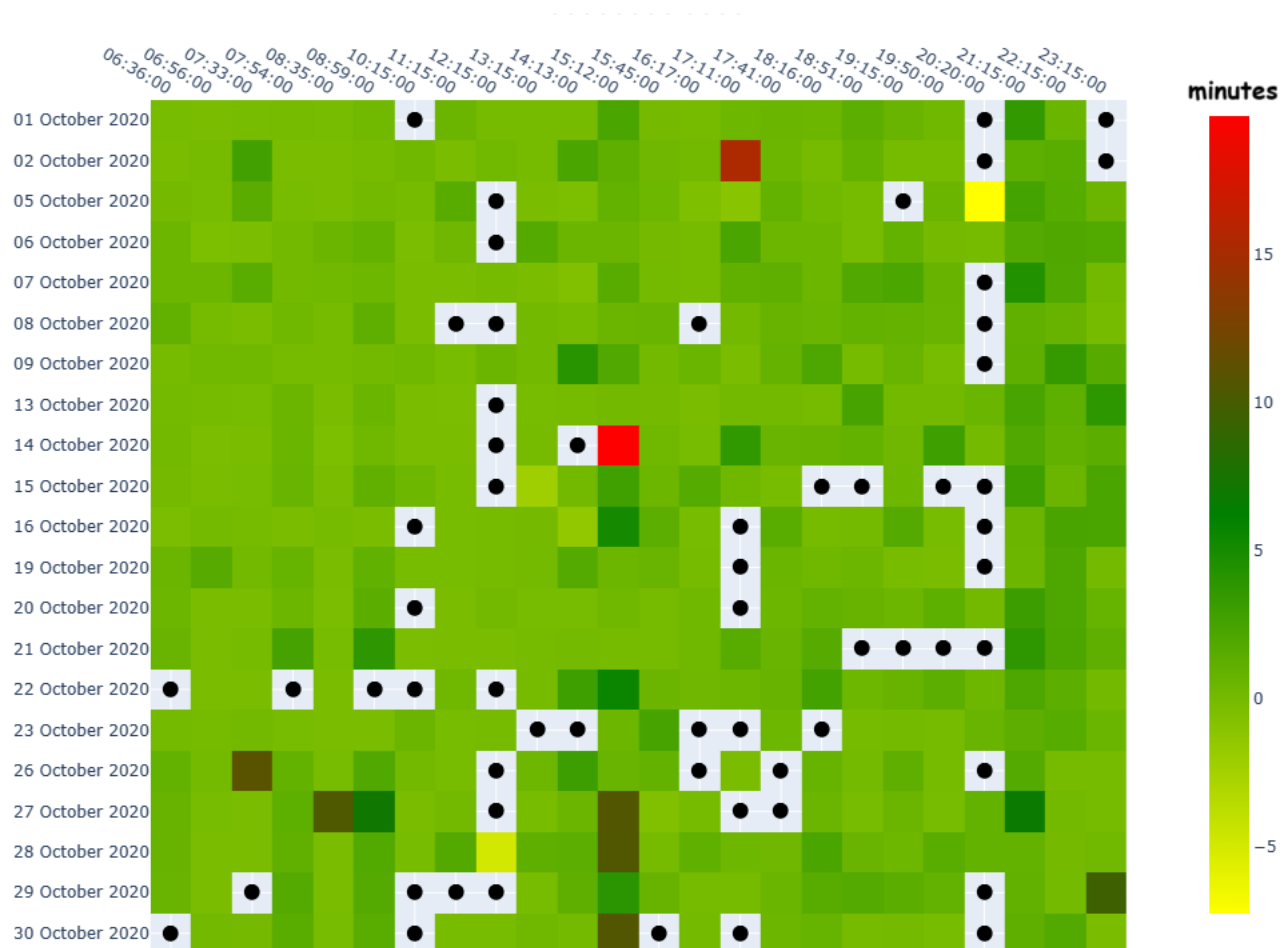


FIGURE 4.8 Visualisation des valeurs des retards au moment de départ

4.3 HERE | Ligne 22 - Bois-des-Filion - Sainte-Thérèse

Tout comme les données GPS, les données HERE ont été extraites sur la base de données GTFS. Premièrement, en utilisant le tracé planifié de la ligne-direction à l'étude, nous avons pu extraire les tronçons HERE constituant le même tracé. Deuxièmement, en utilisant les horaires planifiés et suivant l'approche décrite au chapitre méthodologie, section 3.5, nous avons pu extraire les données de vitesses HERE appropriées. La table qui suit indique, pour chaque voyage d'une journée de service, l'heure pour laquelle les données HERE ont été extraites.

Heure de départ HH:MM:SS	Heure d'arrive HH:MM:SS	Temps de parcours (min)	Heure pour laquelle les données HERE ont été extraites
06:36:00	07:05:00	29	07:00:00
06:56:00	07:25:00	29	07:00:00
07:33:00	08:04:00	31	08:00:00
07:54:00	08:23:00	29	08:00:00
08:35:00	09:00:00	25	09:00:00
08:59:00	09:23:00	24	09:00:00
10:15:00	10:39:00	24	10:00:00
11:15:00	11:40:00	25	11:00:00
12:15:00	12:40:00	25	12:00:00
13:15:00	13:40:00	25	13:00:00
14:13:00	14:40:00	27	14:00:00
15:12:00	15:41:00	29	15:00:00
15:45:00	16:20:00	35	16:00:00
16:17:00	16:56:00	39	17:00:00
17:11:00	17:49:00	38	18:00:00
17:41:00	18:13:00	32	18:00:00
18:16:00	18:41:00	25	18:00:00
18:51:00	19:16:00	25	19:00:00
19:15:00	19:39:00	24	19:00:00
19:50:00	20:13:00	23	20:00:00
20:20:00	20:43:00	23	21:00:00
21:15:00	21:38:00	23	21:00:00
22:15:00	22:38:00	23	22:00:00
23:15:00	23:37:00	22	23:00:00

FIGURE 4.9 Heures pour lesquelles les données HERE ont été extraites

La figure 4.10 présente les temps de parcours construits à partir des données de vitesses HERE extraites pour l'ensemble des voyages planifiés sur la ligne-direction à l'étude pour la période d'étude choisie qui disposent de leurs traces GPS.

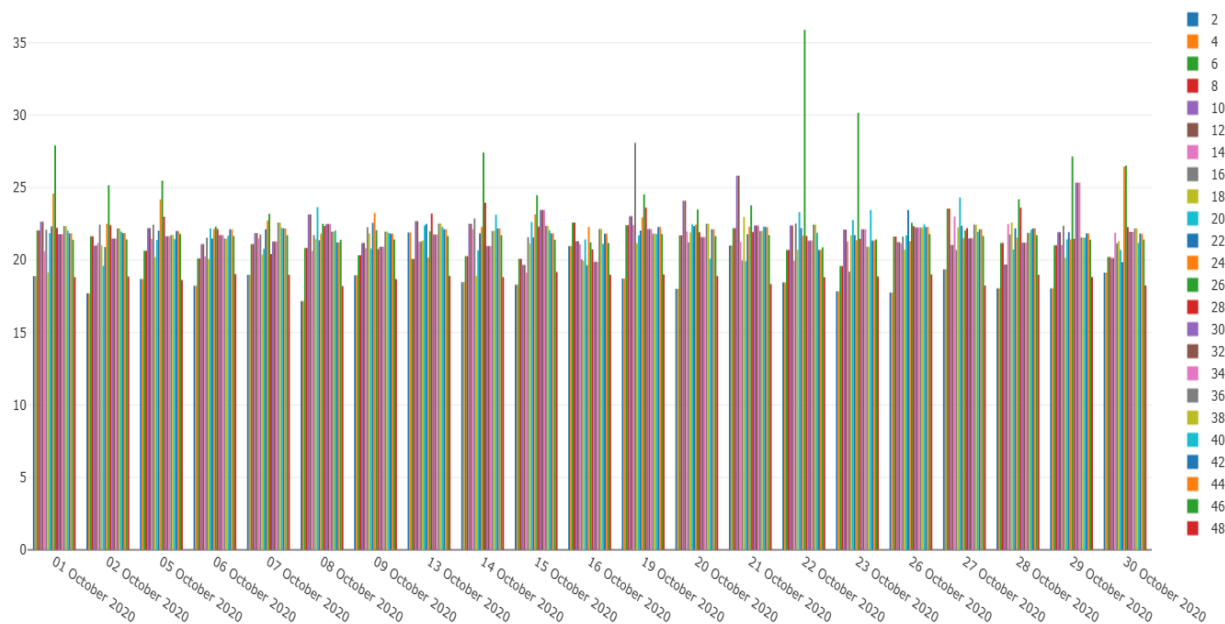


FIGURE 4.10 Distribution des temps de parcours HERE de la ligne-direction à l'étude

Ce graphe se lit de la façon suivante :

1. avec un pourcentage de près de 70%, les temps de parcours HERE sont des temps de parcours inférieurs à 22 minutes.
2. les temps de parcours HERE allant de 27 à 35 minutes sont des temps de parcours pour les heures de pointe.
3. Pour la même période de la journée ou pour le même voyage, le temps de parcours HERE est fixe pour différentes journées de service. Cette observation nous suggère de mener une analyse concernant l'origine de ces données et la confiance de ces dernières.

Analyse de confiance

Avant de discuter les résultats, il est utile de se rappeler la définition d'un niveau de confiance. Un niveau de confiance attribué à une valeur de vitesse moyenne d'un tronçon du réseau est calculé à la base de nombre d'observations effectuées sur ce tronçon à cette heure-là et la variance de l'ensemble des vitesses observées. Une vitesse avec un niveau de confiance égale à 5 est une vitesse modelisée, qui n'est pas basée sur des observations, mais sur les données historiques. Une vitesse avec un niveau de confiance égale à 30, est une vitesse observée et les observations faites pour le déduire ont donné des vitesses très similaires ou proches les unes des autres. Dans les pages qui suivent, les figures 4.11 et 4.12 donnent un aperçu des niveaux de confiance attribués aux données HERE extraites. Pour chaque voyage sur la ligne-direction à l'étude pour la période d'étude choisie :

- la figure 4.11 nous permet de voir le pourcentage de kilomètres avec des vitesses HERE auxquelles des niveaux de confiance égaux ou supérieurs à 30 sont attribués par rapport à la longueur de la ligne.
- La figure 4.12 nous permet de voir le pourcentage de kilomètres avec des vitesses HERE auxquelles des niveaux de confiance égaux à 5 sont attribués par rapport à la longueur de la ligne.

L'analyse de ces figures révèle que pour la ligne-direction à l'étude, HERE fournit des données de vitesses avec des niveaux de confiance qui varient considérablement selon la période de la journée :

- Pour la période hors pointe du soir (PM), peu importe le jour de service, on peut voir à la figure 4.11 que des tronçons HERE constituant plus que 90% de la longueur de la ligne possèdent des vitesses auxquelles des niveaux de confiance inférieurs à 30 sont attribués. Plus loin que ça, la figure 4.12 montre que la majorité de ces niveaux de confiance sont égaux à 5. Donc les vitesses fournies par HERE pour la période hors pointe du soir (PM) sont des vitesses qui ne sont pas basées ni sur des observations ni sur des vitesses enregistrées 15 minutes précédant et suivant l'heure donnée. Ce sont des vitesses basées sur des données historiques.
- Pendant les heures de pointe du matin et du soir, peu importe le jour de service, les vitesses fournies par HERE sont des vitesses observées, ce n'est pas pour toute la longueur de la ligne, mais quand même pour des tronçons constituant même jusqu'à 60% de la longueur de la ligne, bien qu'il existe des exceptions.
- Pour la période hors pointe du matin (AM), les données de vitesses fournies par HERE ne présentent pas un profil homogène. Comme on le voit sur la figure 4.12 (HP-AM) il n'existe pas de couleur dominante par rapport aux autres. Les vitesses des tronçons HERE constituant le trajet planifié peuvent être majoritairement des vitesses basées

sur des observations, ou au contraire, des vitesses basées sur des données historiques.

	Pointe AM						HP					Pointe PM						HP							
	06:36:00	06:56:00	07:33:00	07:54:00	08:35:00	08:59:00	10:15:00	11:15:00	12:15:00	13:15:00	14:13:00	15:12:00	15:45:00	16:17:00	17:11:00	17:41:00	18:16:00	18:51:00	19:15:00	19:50:00	20:20:00	21:15:00	22:15:00	23:15:00	
2020-10-01	36.05	36.05	28.45	28.45	8.30	8.30	50.87	15.68	67.09	22.88	3.98	37.46	57.52	52.64	18.33	18.33	18.33	0.98	0.98	0.46	0	0	0	0	
2020-10-02	20.75	20.75	19.35	19.35	7.19	7.19	20.88	26.41	8.83	66.36	30.52	47.34	24.84	13.31	0.97	0.97	0.97	0	0	0	0	0	0	0	
2020-10-05	35.95	35.95	25.57	25.57	12.92	12.92	16.44	3.74	27.29	26.03	7.18	6.79	23.74	36.10	0.23	0.23	0.23	15.14	15.14	10.56	5.38	5.38	0	10.56	
2020-10-06	33.00	33.00	40.61	40.61	15.91	15.91	16.94	16.39	29.63	0	22.46	58.88	21.19	41.00	19.97	19.97	19.97	4.16	4.16	9.79	0.40	0.40	0	9.31	
2020-10-07	43.02	43.02	14.01	14.01	0.33	0.33	10.69	15.09	53.50	15.98	4.40	13.21	2.03	41.48	29.70	29.70	29.70	0	0	0.88	0	0	0	0	
2020-10-08	28.63	28.63	23.49	23.49	1.66	1.66	32.96	15.59	16.60	21.56	27.05	22.30	19.48	49.28	14.18	14.18	14.18	15.07	15.07	0	1.90	1.90	0	11.77	
2020-10-09	5.44	5.44	28.06	28.06	16.90	16.90	9.31	10.47	0	29.34	30.60	26.00	46.87	59.39	28.56	28.56	28.56	9.17	9.17	3.22	0	0	0	5.57	
2020-10-13	36.15	36.15	35.00	35.00	5.80	5.80	15.06	18.96	27.41	14.02	3.18	58.69	31.94	26.85	12.01	12.01	12.01	0	0	0	0	0	0	0	
2020-10-14	23.02	23.02	31.24	31.24	3.98	3.98	18.23	4.60	70.15	33.06	59.12	53.22	51.79	19.15	61.97	61.97	61.97	17.33	17.33	0	0	0	0	7.19	
2020-10-15	32.03	32.03	16.81	16.81	37.08	37.08	35.33	22.72	25.97	18.25	16.31	28.80	32.68	16.03	9.85	9.85	9.85	2.22	2.22	0	0	0	0	6.12	
2020-10-16	34.05	34.05	18.53	18.53	16.11	16.11	17.63	22.94	43.72	6.60	36.11	27.58	56.72	32.59	37.84	37.84	37.84	1.17	1.17	16.74	0	0	1.89	14.89	
2020-10-19	29.76	29.76	24.46	24.46	21.80	21.80	14.54	15.39	36.15	16.26	0.18	10.85	14.40	1.59	10.07	10.07	10.07	10.07	10.07	16.99	0	0	0	0	
2020-10-20	20.67	20.67	23.98	23.98	39.22	39.22	17.76	16.84	17.09	3.88	14.08	47.04	21.39	16.22	5.47	5.47	5.47	1.89	1.89	16.83	0	0	0	0	
2020-10-21	22.72	22.72	24.97	24.97	7.61	7.61	16.38	53.57	22.63	53.28	5.75	58.05	1.74	17.88	2.62	2.62	2.62	13.20	13.20	0	1.16	1.16	0	14.13	
2020-10-22	22.63	22.63	24.55	24.55	21.25	21.25	24.69	17.35	13.54	16.93	13.36	29.38	77.08	66.26	18.07	18.07	18.07	0	0	0	12.45	12.45	11.05	0	
2020-10-23	47.60	47.60	40.19	40.19	2.71	2.71	21.40	83.87	8.22	17.51	4.32	22.81	37.51	28.77	0	0	0	13.01	13.01	2.27	2.78	2.78	0	0	
2020-10-26	25.78	25.78	35.05	35.05	33.95	33.95	29.86	19.21	25.71	24.03	1.49	13.38	19.86	32.29	24.62	24.62	24.62	2.20	2.20	0	0	0	0	0	
2020-10-27	15.96	15.96	27.06	27.06	32.24	32.24	17.89	51.44	15.43	9.44	0	37.07	20.92	54.96	8.54	8.54	8.54	0.57	0.57	1.14	0	0	0	14.19	
2020-10-28	24.63	24.63	24.27	24.27	47.30	47.30	15.91	11.84	13.06	54.47	16.74	64.08	63.15	48.86	21.13	21.13	21.13	16.79	16.79	6.04	0	0	0	0	
2020-10-29	20.88	20.88	57.40	57.40	10.17	10.17	20.64	16.44	40.51	41.80	28.12	33.07	46.77	73.89	30.54	30.54	30.54	24.41	24.41	10.57	0	0	0	0	
2020-10-30	18.71	18.71	33.14	33.14	33.78	33.78	0.57	24.61	37.00	62.82	51.83	59.98	57.12	48.90	7.25	7.25	7.25	0	0	15.37	0	0	0	9.81	

FIGURE 4.11 Pourcentage de kilomètres avec des vitesses HERE auxquels des niveaux de confiance égaux ou supérieures à 30 sont attribués par rapport à la longueur de la ligne

	Pointe AM						HP				Pointe PM						HP							
	06:36:00	06:56:00	07:33:00	07:54:00	08:35:00	08:59:00	10:15:00	11:15:00	12:15:00	13:15:00	14:13:00	15:12:00	15:45:00	16:17:00	17:11:00	17:41:00	18:16:00	18:51:00	19:15:00	19:50:00	20:20:00	21:15:00	22:15:00	23:15:00
2020-10-01	62.30	62.30	64.06	64.06	91.54	91.54	45.80	83.21	25.79	73.55	96.02	53.66	39.21	44.35	80.45	80.45	80.45	98.84	98.84	99.44	100.00	100.00	100.00	100.00
2020-10-02	76.69	76.69	69.49	69.49	92.59	92.59	77.15	65.98	87.64	28.41	63.65	36.11	71.04	86.00	82.37	82.37	82.37	100.00	100.00	100.00	99.01	99.01	100.00	100.00
2020-10-05	58.82	58.82	66.98	66.98	79.84	79.84	82.96	94.71	68.46	71.26	91.57	86.65	73.52	59.76	91.92	91.92	91.92	84.27	84.27	89.44	94.43	94.43	100.00	89.44
2020-10-06	62.87	62.87	54.53	54.53	83.21	83.21	65.58	83.21	65.64	100.00	70.68	37.69	73.87	55.31	79.28	79.28	79.28	89.44	89.44	87.42	99.60	99.60	100.00	90.37
2020-10-07	55.51	55.51	74.55	74.55	80.14	80.14	86.21	75.06	37.85	70.55	89.35	82.53	97.29	57.56	69.90	69.90	69.90	100.00	100.00	98.87	100.00	100.00	100.00	100.00
2020-10-08	68.49	68.49	74.68	74.68	97.64	97.64	58.10	83.65	82.54	73.76	69.82	74.87	79.42	50.06	85.60	85.60	85.60	84.13	84.13	100.00	89.93	89.93	100.00	84.16
2020-10-09	93.15	93.15	47.34	47.34	81.69	81.69	87.97	79.07	100.00	64.79	64.01	72.79	51.03	39.46	65.76	65.76	65.76	87.79	87.79	96.02	100.00	100.00	100.00	93.82
2020-10-13	44.62	44.62	51.89	51.89	92.94	92.94	83.21	64.20	71.87	81.23	94.63	29.97	67.57	71.59	87.26	87.26	87.26	100.00	100.00	100.00	100.00	100.00	100.00	100.00
2020-10-14	72.20	72.20	64.63	64.63	95.34	95.34	69.90	91.88	21.54	63.52	33.24	43.78	33.02	80.40	32.76	32.76	32.76	81.58	81.58	99.19	100.00	100.00	100.00	92.11
2020-10-15	55.67	55.67	71.09	71.09	48.35	48.35	31.58	73.46	71.75	80.47	76.74	70.21	47.13	56.90	87.61	87.61	87.61	96.90	96.90	100.00	100.00	100.00	100.00	93.79
2020-10-16	62.36	62.36	56.01	56.01	79.39	79.39	64.90	75.18	46.35	93.22	49.36	71.33	41.72	65.31	58.90	58.90	58.90	96.18	96.18	82.95	100.00	100.00	98.11	83.05
2020-10-19	52.11	52.11	52.38	52.38	77.62	77.62	83.82	82.06	61.13	83.03	97.93	81.02	81.72	90.01	88.29	88.29	88.29	89.44	89.44	82.58	100.00	100.00	100.00	100.00
2020-10-20	74.52	74.52	68.64	68.64	52.33	52.33	72.99	83.03	79.80	90.83	83.11	41.53	61.93	81.89	67.76	67.76	67.76	98.11	98.11	39.18	100.00	100.00	100.00	100.00
2020-10-21	66.67	66.67	25.25	25.25	90.01	90.01	83.21	23.80	71.92	35.57	89.37	31.30	97.79	81.95	93.40	93.40	93.40	85.43	85.43	100.00	98.52	98.52	100.00	84.28
2020-10-22	75.77	75.77	64.10	64.10	70.73	70.73	62.73	67.82	49.15	82.50	85.84	62.52	16.38	31.51	80.87	80.87	80.87	100.00	100.00	99.55	87.55	87.55	88.69	100.00
2020-10-23	49.25	49.25	36.13	36.13	97.29	97.29	77.09	10.23	89.93	74.61	95.23	75.30	54.65	70.68	100.00	100.00	100.00	74.56	74.56	97.25	96.44	96.44	100.00	100.00
2020-10-26	70.42	70.42	37.35	37.35	65.22	65.22	67.58	78.21	65.42	75.62	90.76	75.36	72.04	65.02	72.97	72.97	72.97	96.92	96.92	100.00	100.00	100.00	100.00	100.00
2020-10-27	80.69	80.69	51.87	51.87	67.41	67.41	81.76	47.43	60.50	87.25	99.60	62.12	77.05	42.08	69.34	69.34	69.34	99.02	99.02	96.76	100.00	100.00	100.00	85.06
2020-10-28	73.00	73.00	72.23	72.23	40.23	40.23	80.48	87.99	85.49	41.21	79.45	29.16	33.71	47.50	63.00	63.00	63.00	83.21	83.21	91.78	100.00	100.00	100.00	100.00
2020-10-29	76.74	76.74	25.41	25.41	81.05	81.05	70.83	82.49	58.24	51.39	64.48	66.15	47.72	22.22	65.35	65.35	65.35	75.21	75.21	88.27	100.00	100.00	100.00	100.00
2020-10-30	70.08	70.08	57.28	57.28	62.94	62.94	98.54	72.50	59.73	32.77	41.71	36.69	39.85	49.22	92.25	92.25	92.25	100.00	100.00	84.49	100.00	100.00	100.00	87.38

FIGURE 4.12 Pourcentage de kilomètres avec des vitesses HERE auxquels des niveaux de confiance égaux à 5 sont attribués par rapport à la longueur de la ligne

4.4 Analyse comparée | Ligne 22 - Bois-des-Filion - Sainte-Thérèse

Une fois que toutes les données nécessaires ont été extraites avec un niveau de détail approprié, une comparaison a été faite entre les temps de parcours issus de différents types de données. Les résultats de l'analyse comparée sont résumés dans les tableaux et graphiques qui suivent. Observons le graphique 4.13 pour voir à quel point les horaires de service planifié ont été respectés.

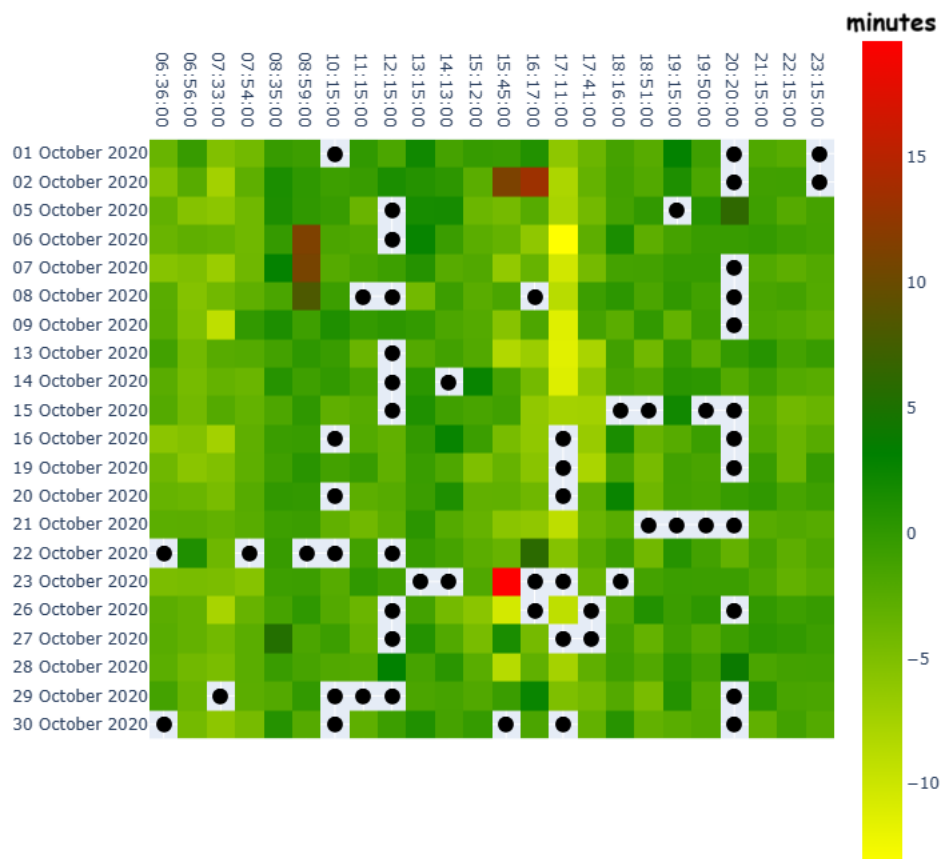


FIGURE 4.13 Comparaison des temps de parcours réalisés aux temps de parcours planifiés

Cette figure illustre visuellement les différences entre les valeurs de temps de parcours planifié et celui effectivement réalisé pour chaque journée de service et chaque voyage sur la période à l'étude. Cette figure permet de confirmer, premièrement, que les horaires planifiés ont été globalement respectés et, deuxièmement, pour les voyages dont le temps de parcours effectivement réalisé excède celui planifié, sont des voyages aux heures de pointe.

En continuant la comparaison entre le service réellement effectué et le service planifié sur la ligne-direction à l'étude, La figure 4.14 présente une autre méthode pour comparer les temps de parcours ; un graphique en boîte à moustaches. Sur cette figure, pour chaque voyage sur la ligne-direction à l'étude pour la période d'étude, une boîte à moustaches colorée en rouge résume la distribution de ses temps de parcours planifiés et une boîte à moustaches colorée en vert résume la distribution de ses temps de parcours réalisés.

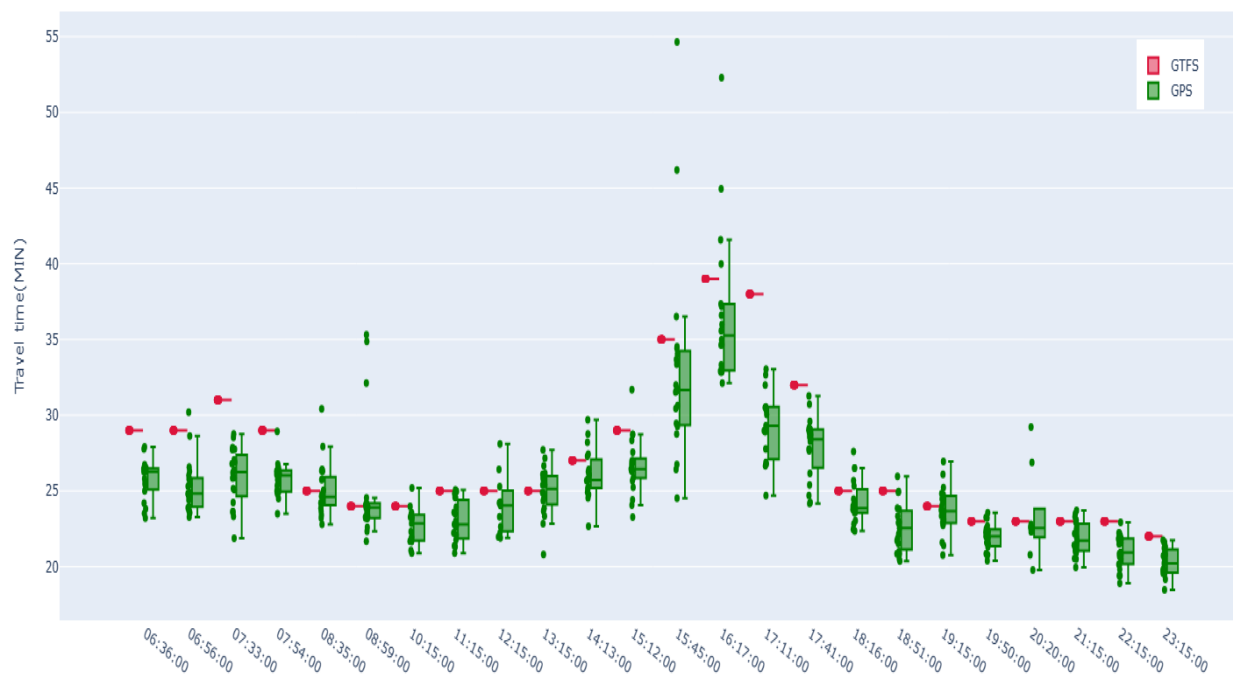


FIGURE 4.14 Diagrammes "boîte à moustaches" de temps de parcours GPS et temps de parcours GTFS

En examinant les boîtes à moustaches : autre information peut être ajoutée aux informations tirées de la figure 4.14 ; les boxplots avec des moustaches longues, sont des boxplots présentant les voyages aux heures de pointe, c'est ce qui nous permet de dire que les voyages aux heures de pointe possèdent des temps de parcours dispersés alors que les temps de parcours pour les voyages hors pointe présentent peu de variation et sont regroupées autour d'une valeur fixe. Dans l'analyse qui suit, nous avons placé sur le même graphique des boxplots présentant les données HERE. Comme le montre la figure 4.15, les boxplots HERE sont colorés en bleu.

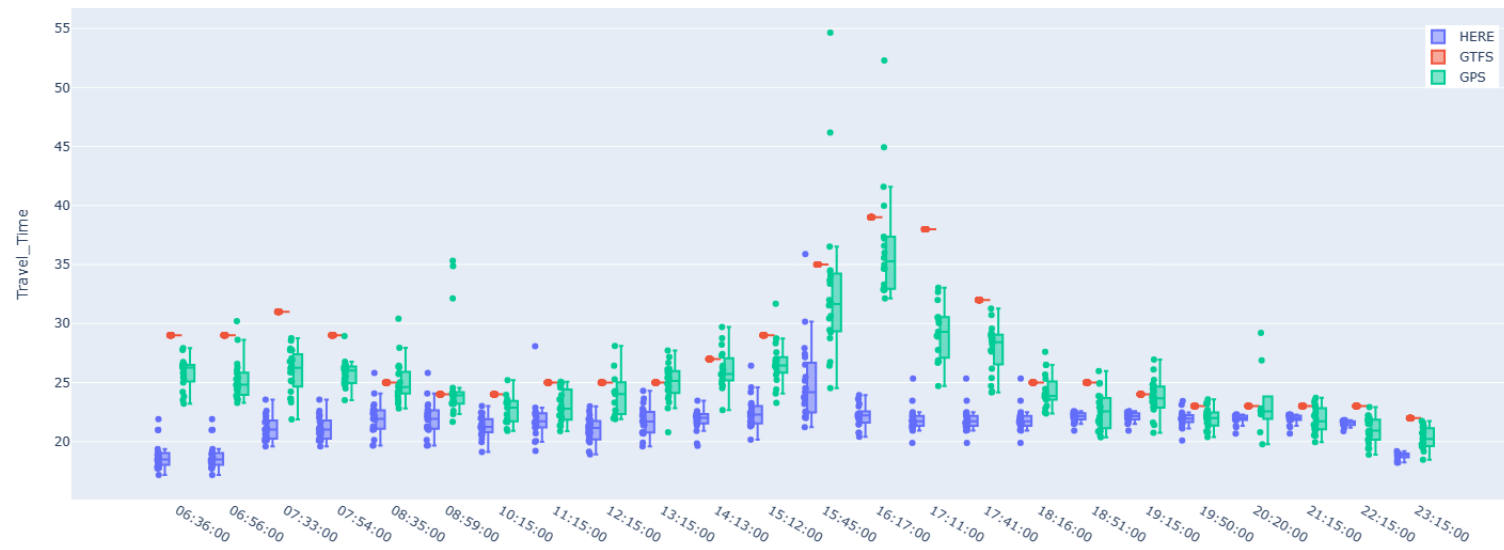


FIGURE 4.15 Diagrammes "boîte à moustaches" de temps des parcours GPS, GTFS et HERE

Nous avons focalisé notre attention sur deux points :

- En comparant les boxplots HERE, aux boxplots GPS, ces derniers présentent des moustaches moins longues. Ceci s'explique par le fait que les temps de parcours HERE pour un voyage sur la période à l'étude sont regroupés autour de la valeur médiane.
- Pour chaque voyage, la médiane des temps de parcours HERE est inférieure à celle des temps de parcours GPS, qui est à son tour inférieure à celle des temps de parcours GTFS. On peut expliquer cette constatation par le fait que le temps de parcours HERE ne prend pas en considération le temps nécessaire à faire monter et descendre les usagers à chaque arrêt vu que les véhicules utilisés par HERE sont soit des voitures, soit des camions.

À ce stade, on ne peut pas juger la pertinence d'utiliser les données de vitesse HERE dans le cadre de mesure du temps de parcours de bus :

- Ratio de comparaison entre les temps de parcours GPS par rapport aux temps de parcours HERE est faible, mais stable, qui peut juger les données HERE comme pertinentes.
- Ratio de comparaison entre les temps de parcours GPS par rapport aux temps de parcours HERE est faible et aussi non stable, qui peut juger les données HERE comme faiblement pertinentes ou non pertinentes.

Le graphe 4.16 a permis d'avoir une réponse à la question de pertinence précédemment posée : le ratio entre les temps de parcours fournis par HERE et GPS varie de 0.4 à 1.1, concentré à 46% entre 0.85 à 1. Cela signifie que le ratio de comparaison est non satisfaisant et montre les données HERE comme faiblement pertinentes pour la mesure du temps de parcours de bus sur la ligne-direction à l'étude.

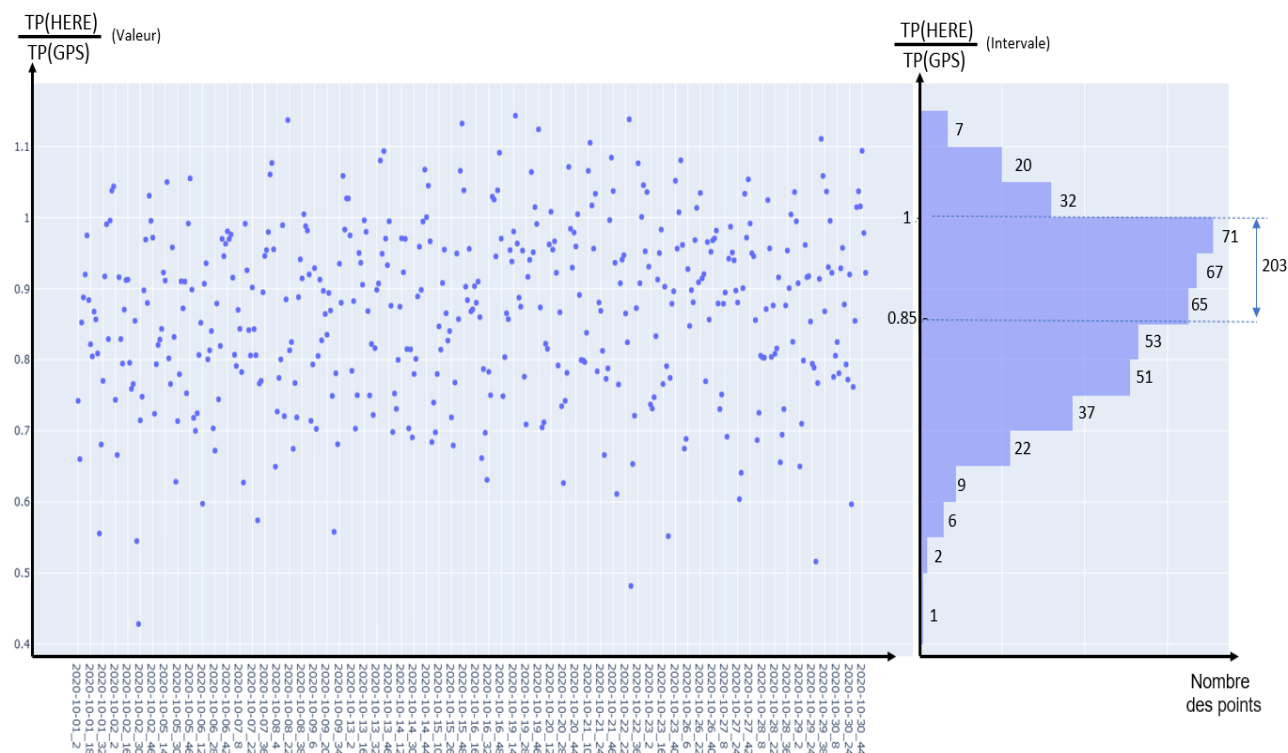


FIGURE 4.16 Distribution du ratio entre les temps de parcours fournis par HERE et GPS

4.5 Expérimentation | Autres lignes

Pour aller plus loin, pour voir si le ratio de comparaison entre les temps de parcours GPS par rapport aux temps de parcours HERE varie d'une ligne de bus à une autre, le même travail a été effectué ensuite auprès de 3 autres lignes de bus. Les trois lignes ajoutées dans cette partie sont :

- Ligne 27 - Lorraine - Sainte-Thérèse dans la direction Bois-des-Filion via Lorraine (A)
- Ligne 28 - Terrebonne - Laval (métro Cartier) dans la direction Métro Cartier (B)
- Ligne 90 - Saint-Eustache - Arthur-Sauvé et Industriel dans la direction Terminus/St-Eustache (C)

Dans les figures 4.17, 4.18 et 4.19, on observe :

- Pour la ligne (B), le ratio se range parmi ce qu'on trouve de pire, varie de 0.3 à 0.7, avec une grande instabilité.
- Pour la ligne(C), le ratio est concentré à plus que 80% entre 0.6 à 0.8.
- Pour la ligne (A), le ratio est concentré à plus que 80% entre 0.75 à 0.9. Si l'on compare avec les ratios trouvés avec les autres lignes , ce ratio est considéré comme le plus satisfaisant.

Les ratios pour les différentes lignes montrent que la pertinence d'utiliser les données de vitesse HERE pour la mesure du temps de parcours de bus varie d'une ligne à une autre. Juger cette pertinence est difficile si on ne fait pas une référence à une ligne de bus exacte.

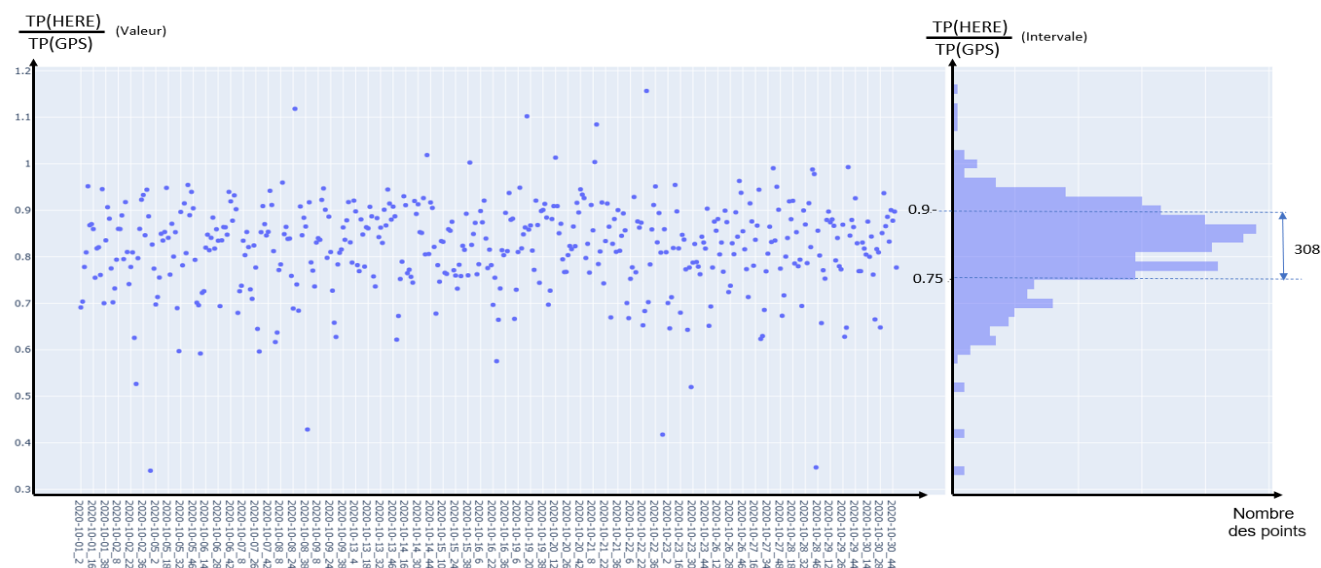


FIGURE 4.17 Distribution du ratio entre les temps de parcours fournis par HERE et GPS -Ligne A

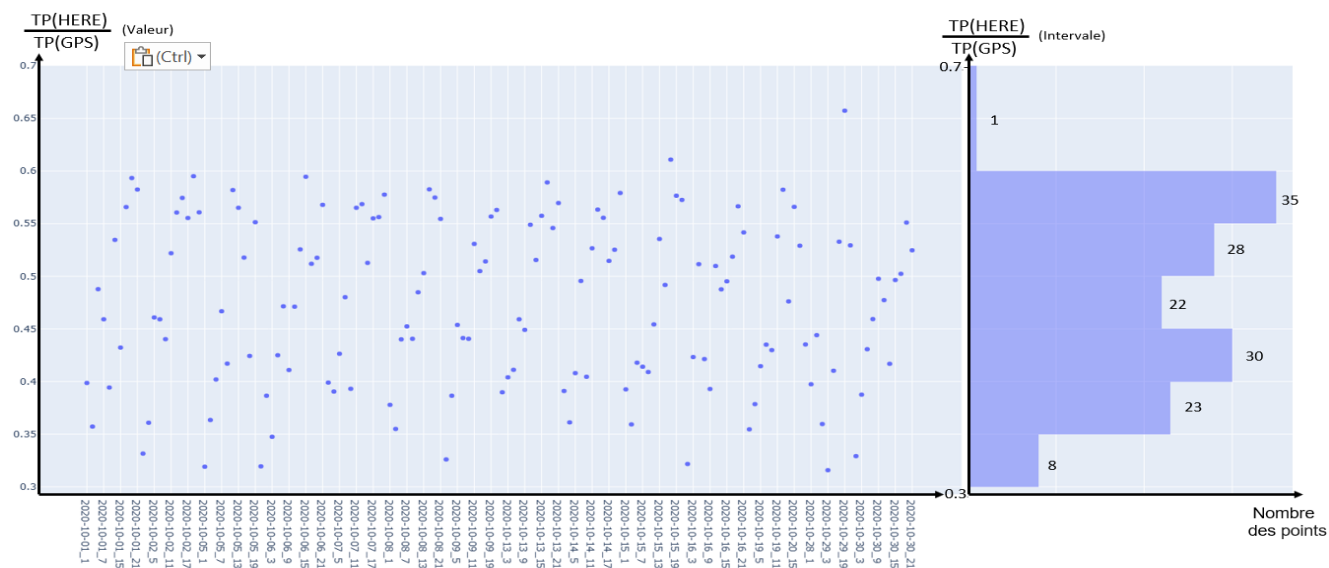


FIGURE 4.18 Distribution du ratio entre les temps de parcours fournis par HERE et GPS -Ligne B

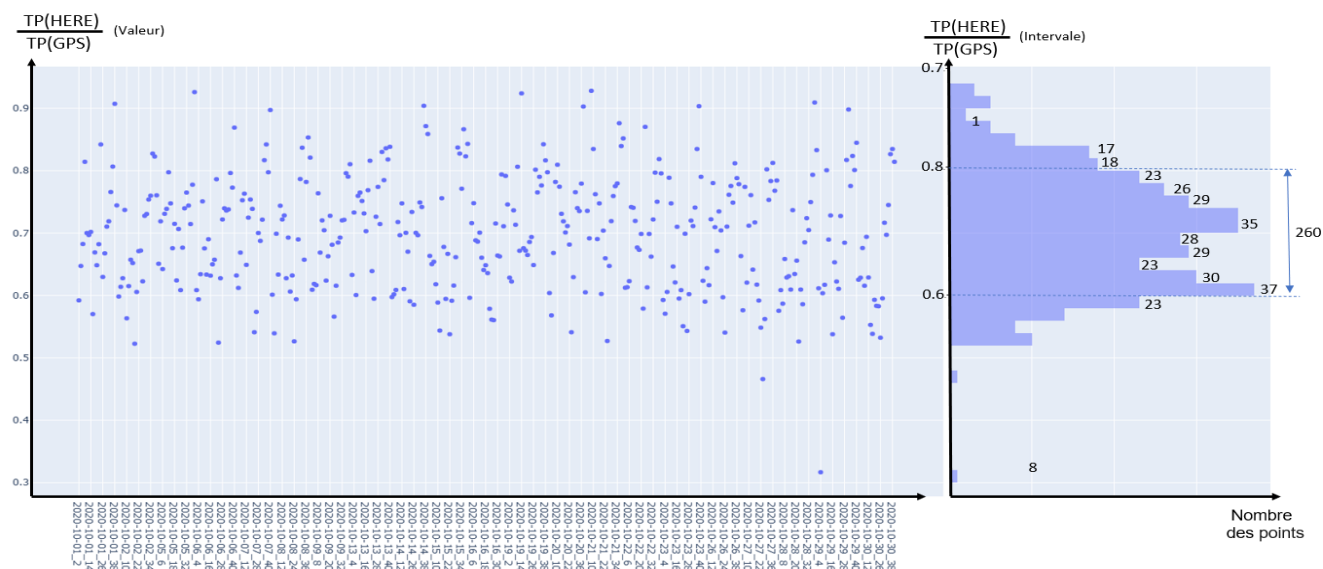


FIGURE 4.19 Distribution du ratio entre les temps de parcours fournis par HERE et GPS -Ligne C

CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS

Ce dernier chapitre présente la synthèse du travail, les limites et des recommandations pour les projets futurs.

5.1 Synthèse du travail

Disposant de trois types de données, les données GTFS, les données GPS et les données HERE , qui représentent respectivement le service de bus planifié, le service de bus réalisé et les données historiques de circulation. L’objectif de cette étude consistait en premier lieu, à déterminer les caractéristiques de ces données et en deuxième lieu, à comparer les temps parcours issus de ces données ; une première comparaison entre les temps parcours GTFS et les temps parcours GPS afin d’évaluer la planification et une deuxième comparaison les temps parcours GPS et les temps parcours HERE pour analyser la pertinence d’utiliser ces derniers dans le cadre de mesure du temps de parcours de bus.

Une méthodologie permettant de comparer les temps de parcours obtenus des trois sources de données a été développée, puis intégrée dans une interface utilisateur graphique qui ne demande aucune compétence spécifique pour la gérer. Suite à l’application de la méthodologie sur 4 lignes de bus, les conclusions qui se sont dégagées se résument comme suit :

- Les données GPS dans leur état brut peuvent contenir des erreurs ou des informations incomplètes, une étape de validation qui précède normalement toute utilisation de ces données est nécessaire et incontournable. Les données GPS soigneusement vérifiées et validées permettent d’ajuster les horaires planifiés au besoin.
- Les horaires planifiés sont faits pour ne pas être dépassés, un ajustement de ces horaires peut être bénéfique à la fois pour l’exploitant d’un réseau de transport en commun et pour les usagers.
- Les données HERE représentent des temps routiers en voiture, elles doivent donc être ajustées pour représenter des temps en bus.
- Les données HERE peuvent être considérées comme pertinentes pour des lignes de bus, mais pas pour d’autres.
- Les données HERE assurent une meilleure estimation des temps parcours pour les voyages aux périodes de pointe du matin et de l’après-midi. Ceci est expliqué par le fait que HERE fournit des données de vitesses avec un niveau de confiance élevé pour ces périodes.

Après avoir été testé, cette approche sera généralisée aux autres lignes du réseau de exo. Les limites liées aux données à disposition, et à la méthodologie suivie seront ensuite présentées, tout en proposant des améliorations possibles.

5.2 Limitations

Dans cette partie, nous expliquons les limitations concernant les données disponibles et la méthodologie suivie dans le cadre de cette étude et nous proposons des recommandations solidement argumentées pour des améliorations.

Commençons d’abord par les limitations concernant les données disponibles :

Une première limite est de ne pas disposer d’une base de données historiques sur les causes des retards ; cette base de données peut nous permettre d’exclure de cette étude les voyages qui ont subi des retards dus à des circonstances exceptionnelles ou inhabituelles qui touchent le bus et pas à la circulation comme : une agressivité physique détectée à bord, panne mécanique du véhicule, etc. L’utilisation de données temps parcours GPS de ces voyages peut avoir biaisé les résultats obtenus concernant la pertinence de données HERE.

La deuxième limite est liée aux données présentant le service réellement effectué à disposition : aucune des autres données fournies ne peut nous permettre de savoir si les traces GPS ont été soigneusement extraites. Avoir d’autres données présentant le service réellement effectué, comme par exemple les données de cartes à puce peut faire l’affaire. Pour un voyage GTFS planifié pour lequel on a extrait la trace GPS correspondante, la comparaison entre l’heure de départ et l’heure d’arrivée issues de la trace GPS respectivement avec l’heure de départ et l’heure d’arrivée issues de données de cartes à puce pour le même voyage permet de valider si la trace GPS a été soigneusement extraite.

Passons aux limitations concernant la méthodologie suivie :

Comme l’évoque la partie 3.3.4, la maille de polygones utilisée pour l’extraction des séquences GPS est créée manuellement sur le site <http://geojson.io>. Ce processus peut être long et fastidieux, surtout quand on travaille sur une ligne direction qui possède plus qu’un seul tracé ; généralement, une ligne direction possède entre 1 à 4 tracés différents. Si la création d’une seule maille de polygones prend environ 10 à 15 minutes, la création de 4 mailles dure environ 40 à 60 minutes. L’automatisation de cette tâche est préférable. Cette automatisation sera efficace pour éviter tout risque d’erreur humaine au niveau de la création de la maille de polygones qui pourrait biaiser les résultats recherchés, autrement dit cette automatisation permet aux utilisateurs travaillant sur la même ligne direction d’obtenir les mêmes résultats d’où l’uniformité et l’exactitude des résultats.

RÉFÉRENCES

- Cristóbal, T., Padrón, G., Quesada-Arencibia, A., Alayón, F., de Blasio, G., García, C. R. (2019). Bus travel time prediction model based on profile similarity. *Sensors*, 19(13), 2869.
- Ferrettini, G., Queyroi, F., Harzallah, M. (2023, January). Prédiction de temps de parcours de bus par chaînage des données d'entraînement. In *EGC 2023*.
- Kim, S., Coifman, B. (2014). Comparing INRIX speed data against concurrent loop detector stations over several months. *Transportation Research Part C : Emerging Technologies*, 49, 59-72.
- Kondyli, A., St. George, B., Elefteriadou, L. (2018). Comparison of travel time measurement methods along freeway and arterial facilities. *Transportation Letters*, 10(4), 215-228.
- Kwon, J., Coifman, B., Bickel, P. (2000). Day-to-day travel-time trends and travel-time prediction from loop-detector data. *Transportation Research Record*, 1717(1), 120-129.
- Lin, W. H., Zeng, J. (1999). Experimental study of real-time bus arrival time prediction with GPS data. *Transportation Research Record*, 1666(1), 101-109.
- Loustau, P. (2009). Modélisation des temps de parcours sur un réseau routier à l'aide de données de véhicules flottants. École Polytechnique de Montréal.
- Morency, C., Trépanier, M., Saunier, N. (2011). Évaluation des enjeux analytiques et des potentialités des données de temps de parcours pour mesurer la fiabilité du réseau routier montréalais.
- Nazem, M. (2014). Contributions à la modélisation des déplacements en transport collectif (Doctoral dissertation, École Polytechnique de Montréal).
- Pesteil, F. (2008). Les temps de parcours : Estimation, diffusion et approche multimodale.

CERTU, Éditions du Certu, France.

Philippe, A., Prudent, X., Trépanier, M. (2019). Prédiction des retards d'autobus urbains à Sherbrooke. CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport= Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation.

Ramakrishna, Y., Ramakrishna, P., Lakshmanan, V., Sivanandan, R. (2008). Use of GPS probe data and passenger data for prediction of bus transit travel time. In *Transportation Land Use, Planning, and Air Quality* (pp. 124-133).

Saunier, N., Morency, C. (2011). Comparing data from mobile and static traffic sensors for travel time assessment. In *Transportation and Development Institute Congress 2011 : Integrated Transportation and Development for a Better Tomorrow* (pp. 1178-1187).

Sun, F., Pan, Y., White, J., Dubey, A. (2016, May). Real-time and predictive analytics for smart public transportation decision support system. In *2016 IEEE International Conference on Smart Computing (SMARTCOMP)* (pp. 1-8). IEEE.

Vanajakshi, L., Subramanian, S. C., Sivanandan, R. (2009). Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. *IET intelligent transport systems*, 3(1), 1-9.

Wall, Z., Dailey, D. J. (1999, January). An Algorithm for Predicting the Arrival Time of Mass Transit Vehicle Using Automatic Vehicle Location Data. TRB No. 990870. In 78th annual meeting.

Weigang, L., Koendjibiharie, W., de M Juca, R. C., Yamashita, Y., MacIver, A. (2002, September). Algorithms for estimating bus arrival times using GPS data. In *Proceedings. The IEEE 5th International Conference on Intelligent Transportation Systems* (pp. 868-873). IEEE.

Xinghao, S., Jing, T., Guojun, C., Qichong, S. (2013). Predicting bus real-time travel time basing on both GPS and RFID data. *Procedia-Social and Behavioral Sciences*, 96, 2287-2299.

Yang, M., Chen, C., Wang, L., Yan, X., Zhou, L. (2016). Bus arrival time prediction using support vector machine with genetic algorithm. *Neural Network World*, 26(3), 205.

Yu, Z., Wood, J. S., Gayah, V. V. (2017). Using survival models to estimate bus travel times and associated uncertainties. *Transportation Research Part C : Emerging Technologies*, 74, 366-382.

Zhang, X., Rice, J. A. (2003). Short-term travel time prediction. *Transportation Research Part C : Emerging Technologies*, 11(3-4), 187-210.

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., Cao, R. (2017). Impacts of weather on public transport ridership : Results from mining data from different sources. *Transportation research part C : emerging technologies*, 75, 17-29.