CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Prédiction des retards d'autobus urbains à Sherbrooke

Amaury Philippe
Xavier Prudent
Martin Trépanier

Juillet 2019

CIRRELT-2019-25

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

Prédiction des retards d'autobus urbains à Sherbrooke

Amaury Philippe[1,*], Xavier Prudent[2], Martin Trépanier[1]

1. Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport et Département de mathématique et de génie industriel, Polytechnique de Montréal, C.P. 6079, Succursale Centre-ville, Montréal, Canada H3C 3A7
2. Civilia Groupe Conseil (www.civilia.ca)

Résumé. Le but de l'étude est d'établir un modèle de prédiction des retards d'autobus urbains, appliqué au cas de la ville de Sherbrooke. Même si les techniques de localisation de bus en temps réel se développent, la précision de la prédiction d'heure d'arrivée d'un bus à une station demeure souvent incertaine. Elle est pourtant capitale. Cela vaut aussi bien pour l'efficience opérationnelle du réseau, que pour la satisfaction de l'usager. Les répercussions à cette satisfaction sont nombreuses : un plus grand rayonnement des sociétés de transport, une part modale recentrée vers le transport en commun, et un impact positif sur l'environnement avec la réduction des gaz à effet de serre des automobiles, encore prédominantes à Sherbrooke. Grâce aux données opérationnelles à disposition, du matériel est disponible pour travailler sur la création de modèles de prédiction des retards d'autobus. À terme, ces modèles doivent être facilement utilisables pour le service opérationnel de la Société de Transport de Sherbrooke. Après avoir analysé les données fournies, une première méthode intuitive de prédiction est mise en place : le modèle de prédiction simple. À partir de celui-ci, des améliorations sont effectuées avec des méthodes d'apprentissage machine plus poussées : respectivement la régression linéaire, les arbres de décision, les forêts aléatoires et les réseaux de neurones. En outre, des variables adaptées au problème considéré sont déterminées pour alimenter ces différents modèles. Finalement, la méthode des forêts aléatoires est retenue car elle apporte la meilleure performance dans la grande majorité des cas, tout en étant stable, ce qui est capital pour l'utilisation opérationnelle. Enfin, des axes de réflexions complémentaires sont creusés pour donner des pistes de continuation envisageables à l'étude.

---

* Auteur correspondant : Amaury.Philippe@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2019

ABSTRACT

The goal of the study is to establish a prediction model for urban autobuses delays, applied on the case of Sherbrooke city, Quebec. Even though real-time localization techniques for buses are developing, the predictions for arrival times often remain uncertain. Nevertheless, this prediction is capital. This idea stands for the operational efficiency of the network, as well as for users' satisfaction. The consequences of that satisfaction are numerous : a higher influence of transportation companies, a modal share recentered on public transportation, and a positive impact on the environment with the reduction of cars' greenhouse gas emissions, which are still predominant in Sherbrooke. Thanks to the operational data given to us, material is available to work on in order to create prediction models for autobuses delays. In the end, these models have to be easily usable by the operational service of the Société de Transport de Sherbrooke.

After having analyzed the data, a first intuitive method of prediction is set : the model of simple prediction. From there, improvements are done with more advanced machine learning methods : respectively linear regression, decision trees, random forests and neural networks. Furthermore, relevant variables are determined as inputs for these different models. Eventually, the random forest method is chosen as it brings the best performance in the vast majority of cases, while being stable, which is capital for operational use. To end with, complementary reflections are studied in order to give possible ways of continuation for the project.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1 INTRODUCTION

1.1 Introduction générale

Devant les défis environnementaux et le réchauffement climatique, les villes cherchent à réduire leurs émissions de gaz à effet de serre. Ainsi, le développement d'un transport collectif durable est un enjeu majeur d'avenir. Cela est particulièrement vrai dans les villes québécoises. L'autobus semble être un bon compromis, notamment à Sherbrooke où il est le principal mode de transport alternatif à la voiture. Cependant, l'autobus se heurte souvent à la frustration de l'usager face à son adhérence à l'horaire variable. La solution à cela est de prédire les heures d'arrivée des autobus de manière précise et fiable. Des modèles corrects existent actuellement, mais ils peuvent être améliorés en tenant compte des contraintes opérationnelles. L'objet de la présente étude est d'utiliser les données opérationnelles, mises à disposition par la ville de Sherbrooke, pour créer un modèle amélioré de prédiction des retards des autobus de la ville. Pour cela, le logiciel utilisé est le logiciel statistique R. Il a l'avantage d'être intuitif et très complet grâce aux nombreuses bibliothèques et fonctions prédéfinies dont il dispose.

1.2 Présentation du cadre de travail

Le travail est effectué avec l'entreprise Civilia. Civilia groupe-conseil est une société spécialisée en mobilité urbaine et en infrastructures. Elle souhaite offrir un service adapté et personnalisé afin de permettre aux organisations de réaliser des projets novateurs et durables dans un monde en pleine transformation. Ses domaines d'intervention principaux sont la mobilité intégrée, le transport collectif, la planification des transports, la viabilité hivernale et les systèmes de transport intelligents. Le projet a été réalisé en alternance entre les bureaux de Civilia, situés à Longueuil (Québec), et le siège du CIRRELT (Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport) basé à Montréal. Une rencontre d'une journée dans les bureaux de la Société de Transport de Sherbrooke (STS) a également été organisée pour confronter les idées et exposer les résultats.

1.3 Présentation du réseau d'étude

En collaboration avec la STS, le cadre de ce projet est le réseau de transport de la ville de Sherbrooke, au Québec. Puisqu'il n'y a ni métro ni tramway, le transport en commun (TC) y est

exclusivement routier, par bus et taxibus à la demande. Le développement du transport en commun est un réel enjeu pour cette ville, puisqu'il ne représente actuellement qu'environ 5% de la part modale, très peu relativement aux 80% occupés par la voiture. À titre de comparaison, dans son Plan stratégique organisationnel 2025, la Société de Transport de Montréal vise une part modale de 28,1% pour le transport en commun en 2025. L'étude sera focalisée sur les 18 lignes de bus « standards » (hors minibus et taxibus), numérotées de 1 à 19, la 10 n'existant pas.

Voici le plan de la STS représentant le réseau.

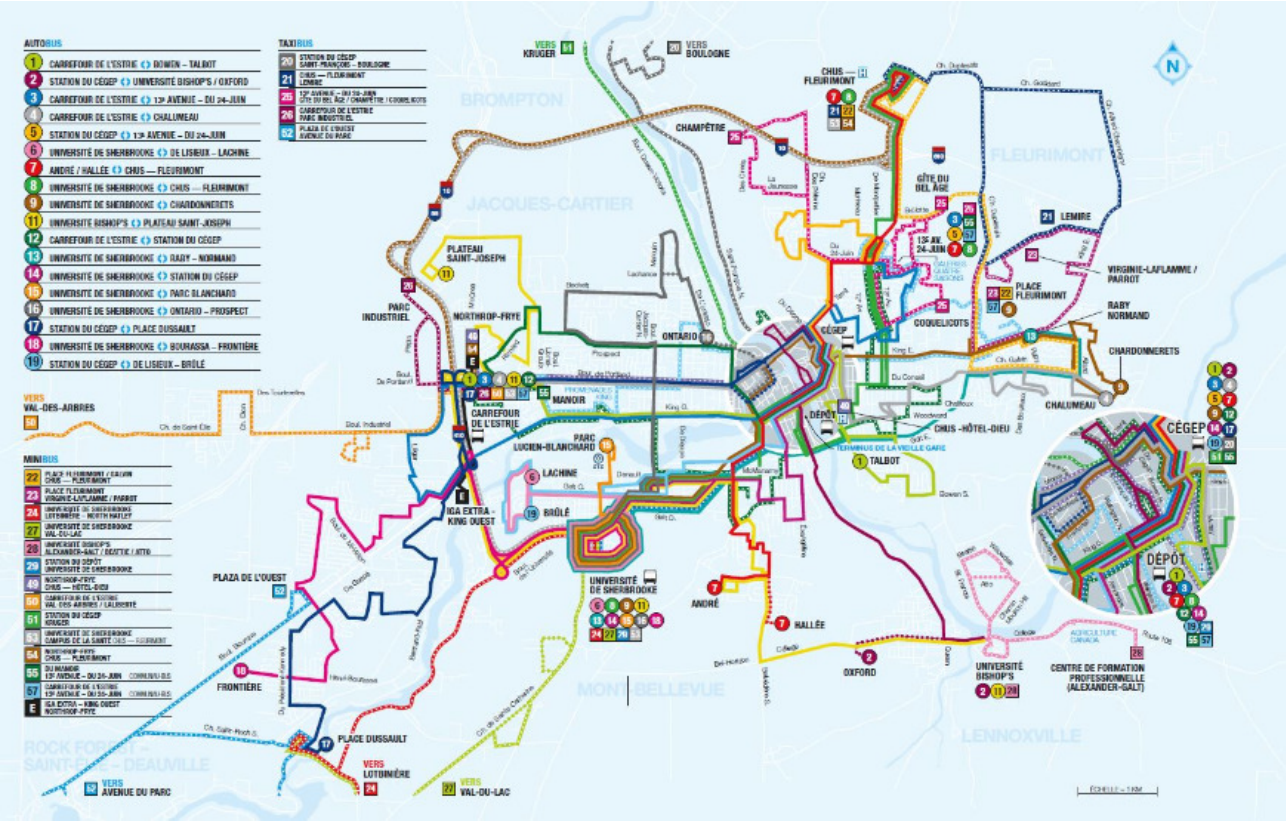

Figure 1.1 : Structure du réseau de transport en commun de la ville de Sherbrooke, Québec

Toutes ces lignes comportent un certain nombre d'arrêts, dans les deux directions. La structure de ces arrêts sera précisée dans le chapitre suivant.

1.4 Enjeux de l'étude et objectifs

Comme précisé dans l'introduction générale ci-dessus, les enjeux sont multiples.

L'objectif principal de l'étude est de mettre au point un outil opérationnel, utilisable en temps réel par la STS. Cet outil prendra la forme d'un modèle de prédiction à partir de différentes variables

d'entrée à déterminer. Le but du modèle est de contenir au mieux l'adhérence horaire dans un intervalle proche de la valeur nulle. Par adhérence horaire, il s'entend l'écart entre la prédiction du retard et le retard réel. L'objectif poursuivi est donc de chercher à estimer du mieux possible le retard de chaque bus sur chacun de ses arrêts futurs, à chaque instant.

L'idée est bien ici d'estimer le retard, et non pas particulièrement de chercher à réduire ce retard. Par la suite, l'étude pourrait néanmoins être un point de départ pour réviser les horaires de service et les réadapter aux contraintes opérationnelles. La finalité qui motive cette étude est d'atteindre un service client optimum ainsi que la satisfaction de l'usager, notions au cœur de toute entreprise de service.

1.5 Revue de littérature

1.5.1 Présentation

À l'aide des articles lus et présentés dans la bibliographie en annexe, il est constaté que plusieurs études ont déjà été réalisées sur les prédictions de retards pour les autobus. Tous les types de transports sont d'ailleurs concernés par ces recherches, mais l'accent est ici porté sur les autobus. Cependant, ces études restent la plupart du temps totalement théoriques et ne sont pas mises en opération à grande échelle dans des entreprises de transport. Cela est dû à plusieurs facteurs, notamment la difficulté d'obtenir de vastes échantillons de données de circulation en temps réel. Pour avoir de telles mesures, il faut posséder des outils précis de localisation, gérer la transmission de l'information, et avoir une capacité de stockage suffisante car le nombre de données devient vite immense. Tout cela coûte cher et demande de l'investissement. Enfin, la confidentialité de ces données est souvent un frein supplémentaire puisque les entreprises ne sont pas nécessairement enclines à diffuser leurs informations.

Le projet a la chance de bénéficier d'un accès privilégié à toutes ces données, pour la ville de Sherbrooke, grâce à l'accord avec la Société de Transport de Sherbrooke. Cela va permettre d'aborder la méthode de prédiction des retards en transport avec un angle inédit et résolument tourné vers l'opérationnel.

1.5.2 Synthèse des articles de la liste de références

Auteurs	Méthodes utilisées	Données d'entrée	Résultats
Amita et al. (2015)	Régression multilinéaire et réseau de neurones	Heures d'arrivée/de départ, vitesse moyenne, retards antérieurs	Réseau de neurones meilleur que la régression
Fan Jiang (2017)	Réseau de neurones à trois couches	Latitude, longitude, numéro de véhicule, heures et données historiques	Plus grande précision avec plus de données historiques, et en combinant les prédictions de petits segments
Jithendra et Naga Ravi (2015)	Non précisé	Temps d'attente, nombre de passagers montants, vitesse moyenne	Pas de résultats explicités, mais implantation dans une application mobile
Ma et al. (2017)	Algorithme à base de filtres de Kalman	Données historiques et données en ligne sur les retards et la congestion	Le modèle statistique est capable de prévoir les congestions
Sun et al. (2016)	Clustering et filtres de Kalman	Horizons étudiés entre 15 minutes et 2 heures, retards pour tous les bus de toutes les lignes sur le même segment à la même période	Plus l'horizon est élevé, moins la précision de la prédiction est bonne : elle devient faible à partir de 90 minutes et inutile pour plus de 2 heures

Yin et al. (2016)	*Machine à vecteur de support (MVS) de quatre types : fonction linéaire de Kernel, fonction polynômiale de Kernel, fonction sigmoïde de Kernel et fonction à base radiale *Réseau de neurones de structure 3-5-1	Temps de parcours des trois bus précédents de la même ligne sur le même segment, temps de parcours du même bus sur le segment précédent, temps de parcours des bus précédents des autres lignes sur le même segment, dichotomies pointe/hors-pointe et centre-ville/banlieue	Réseau de neurones meilleur que le MVS, sauf pour les périodes de pointe de banlieue. Plus on met de données en entrée, plus les prédictions sont précises.

Tableau 1.1 : Synthèse des six articles de la liste de références

1.5.3 Commentaires sur ces travaux

Comme vu dans le tableau précédent, six articles ont été étudiés au vu des thèmes abordés qui sont similaires à ceux de notre étude.

Ce qu'on remarque en premier, c'est qu'aucun auteur n'a de base de données opérationnelles conséquente sur les temps de parcours et les retards des autobus. De plus, les réseaux considérés sont soit théoriques, soit très limités. Par exemple, Amita et al. (2015) n'étudient que deux trajets uniques, connectant les zones commerciales et résidentielles de Delhi, en Inde. De ce point de vue, l'étude avec Sherbrooke a une réelle valeur ajoutée potentielle, grâce au vaste échantillon de données disponible et au réseau d'application étendu à travers la ville.

Ensuite, tous ces articles cherchent à utiliser des modèles algorithmiques et d'apprentissage machine pour prédire les différents retards. Pour les alimenter, on voit que des variables cohérentes sont toujours nécessaires. Il apparaît alors que le choix des variables d'entrée à considérer est crucial pour l'étude menée avec la ville de Sherbrooke.

Les méthodes utilisées sont variables, mais sont souvent complexes. Certes, Amita et al. (2015) utilisent une régression multilinéaire mais dans l'ensemble ; réseaux de neurones, filtres de Kalman ou encore machines à vecteur de support sont privilégiés. Pour le présent projet, il serait pourtant intéressant de commencer par des méthodes plus simples qui peuvent déjà être efficaces, et suffisantes pour une implantation opérationnelle. Ensuite seulement seront considérées des méthodes plus poussées.

Pour les résultats des études enfin, trois grandes idées se dégagent. Premièrement, la précision de la prédiction semble augmenter avec le nombre de données historiques en entrée (Fan Jiang, 2017 ; Yin et al., 2016). De plus, la précision semble diminuer avec la durée de l'horizon de prédiction (Sun et al., 2016). Finalement, le réseau de neurones semble être la méthode la plus efficace (Amita et al., 2015 ; Yin et al., 2016).

Ces conclusions seront prises en compte dans la suite, au cours des différentes étapes de l'étude.

CHAPITRE 2 DONNÉES

2.1 La structure des arrêts

Il a été vu précédemment que dix-huit lignes sont portées à l'étude. Pour prédire les retards aux différents arrêts, il faut avant toute chose connaître l'organisation spatiale de ces derniers.

Figure 2.1 : Structure du réseau d'arrêts des autobus de Sherbrooke

Cette carte montre la localisation des différents arrêts de la ville en fonction de leur latitude et longitude. Chaque point correspond à une position géographique d'un arrêt. Un bleu plus foncé signifie qu'il y a une plus forte densité d'arrêts : souvent deux arrêts, dans les directions opposées, qui sont proches l'un de l'autre. Cette visualisation globale permet déjà de dessiner le réseau routier de la ville. On y retrouve notamment les grandes artères Nord/Sud et Est/Ouest qui la quadrillent. L'étape suivante est de collecter les données opérationnelles correspondant aux temps de passage à ces arrêts.

2.2 Les données opérationnelles : données SAE

La Société de Transport de Sherbrooke offre l'accès à ses données opérationnelles, appelées données SAE. Elles seront utilisées comme base de travail, pour analyser la circulation actuelle d'une part, et construire un modèle en conséquence d'autre part. Comme constaté dans la revue de littérature, l'accès à une telle quantité d'informations opérationnelles est rare : cet avantage sera exploité dans le cadre de l'étude. Toutes les trente secondes, ou à chaque arrivée/départ d'arrêt, un signal est envoyé par l'autobus. Ici, il est considéré que les données SAE sont prises lors de l'arrivée à un arrêt. Une donnée SAE équivaut donc à un arrêt donné pour un trajet donné d'un jour donné. Par ce signal, le bus transmet plusieurs informations. Les principales sont son identification, sa position géographique, l'heure, l'arrêt où il est rendu et son retard réel à l'instant correspondant.

Les données sont transmises sous la forme suivante :

	vehicle_id	latitude	longitude	time	real.delay	speed	bearing	route_id	trip_id	stop_id	sequence	day
1	62104	45.34884	-71.99117	2018-03-12 05:51:13	0	8	50	17S	477346	965	2	2018-03-12
2	62104	45.34958	-71.99007	2018-03-12 05:52:00	1	9	46	17S	477346	967	3	2018-03-12
3	62104	45.35106	-71.98895	2018-03-12 05:52:13	0	9	28	17S	477346	2271	4	2018-03-12
4	62104	45.35485	-71.98799	2018-03-12 05:53:05	0	8	272	17S	477346	1209	5	2018-03-12
5	62104	45.35489	-71.99406	2018-03-12 05:54:02	0	8	271	17S	477346	1211	6	2018-03-12
6	62104	45.35651	-71.99413	2018-03-12 05:54:30	1	9	359	17S	477346	1513	7	2018-03-12

Figure 2.3 : Structure des données SAE

Il convient de remarquer que les valeurs de retard réel sont arrondies à l'unité. Ainsi, toutes les études de retard et d'adhérence horaire, qui seront faites par la suite, seront aussi arrondies à l'unité. Les arrondis se font par défaut : un retard de trente secondes est arrondi à zéro minute, un retard de quatre-vingt-dix secondes est arrondi à une minute. Il n'est pas possible d'avoir une précision supérieure mais, de toute façon, cela n'est pas primordial au niveau opérationnel.

Quatre semaines de données seront utilisées pour le projet : il s'agit de la période du samedi 10 mars au vendredi 6 avril 2018 inclus. Cela représente un peu plus d'un million de données SAE. L'accent est mis sur ces quatre semaines pour alléger le traitement et les temps de calcul, il sera ensuite étudié si ces données peuvent être généralisées à une année entière.

Un autre type de données utile est le GTFS (General Transit Feed Specification). Ce sont les données informatiques standard régissant toute la planification du réseau de transport en commun d'une ville, pour un trimestre donné. Les heures d'arrêts planifiées pour chaque trajet à chaque arrêt seront les principales données utilisées dans l'étude. C'est par rapport à cette référence que les retards peuvent être définis. Enfin, il est décidé d'emblée de séparer ces données en deux

parties : la semaine et le weekend. Le GTFS montre en effet que l'agencement des trajets est différent dans les deux cas. De plus, l'expérience des spécialistes suggère que les comportements de circulation sont également différents selon plusieurs points de vue : il n'y a pas de pointe le weekend notamment.

2.3 Modèle de prédiction simple

À partir des données SAE et du GTFS, la première étape est d'établir un modèle de prédiction simple. Les données SAE ne fournissent que le retard réel au moment de la mesure, par comparaison entre l'heure d'arrivée réelle à l'arrêt et l'heure d'arrivée prévue par le GTFS. Or, à l'arrêt 1 d'un trajet, il faut trouver un moyen pour prédire le retard à l'arrêt 7 par exemple. Le modèle de prédiction simple imaginé doit permettre d'y répondre de la manière la plus intuitive possible. Pour cela, le retard réel actuel est simplement propagé à tous les arrêts suivants. Un affinage est ensuite effectué en tenant compte des contraintes du GTFS. Voici un exemple :

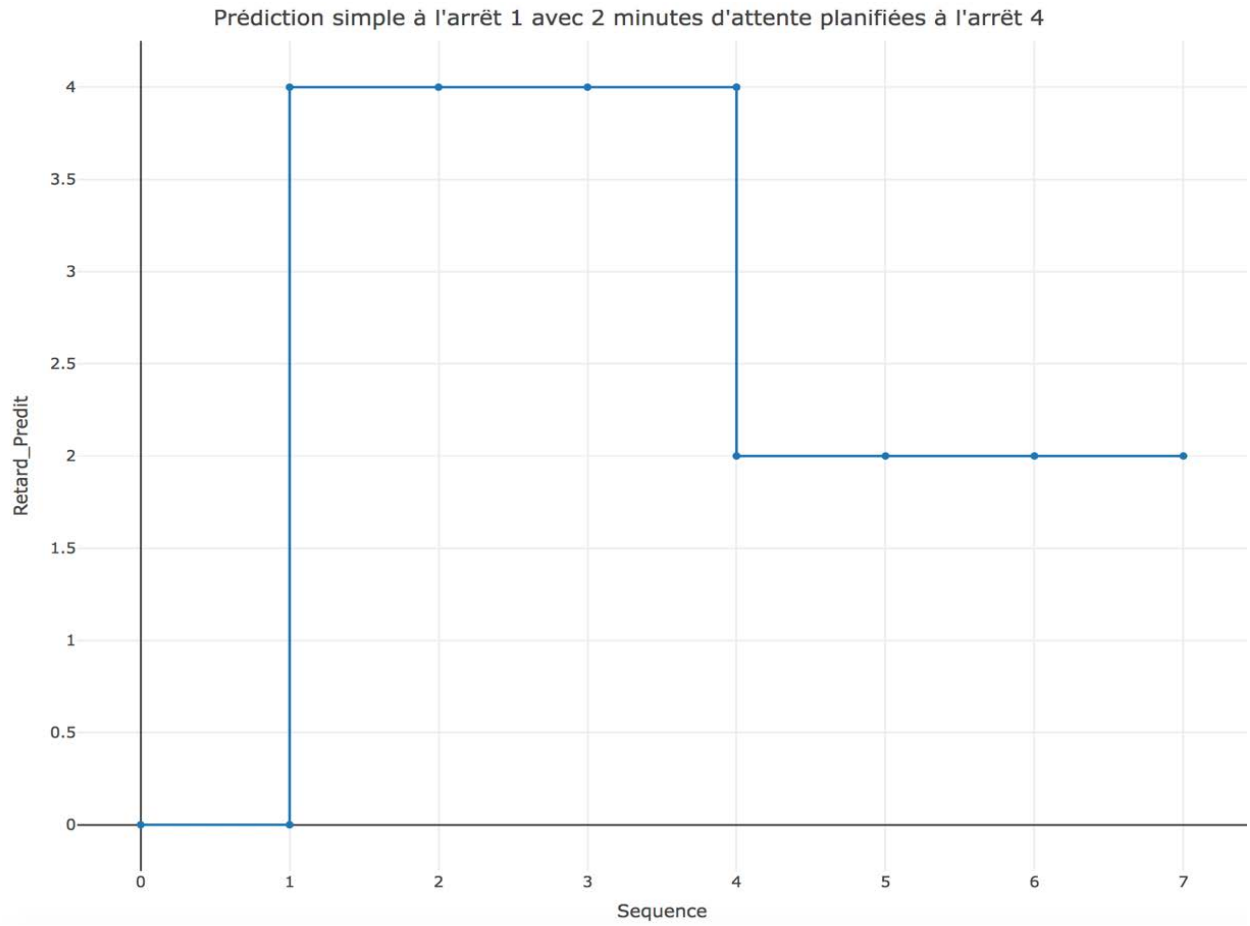

Figure 2.4 : Propagation du retard par prédiction simple, en tenant compte du GTFS

En abscisse, il y a le numéro de séquence de l'arrêt d'un trajet (2 équivaut au deuxième arrêt du trajet, 3 au troisième, etc …). En ordonnée, c'est le retard prédit. Ici, il est supposé que le bus est à l'arrêt 1 avec un retard réel de quatre minutes par rapport au GTFS. Il est aussi supposé que le GTFS prévoit un temps de battement (d'arrêt) de deux minutes à l'arrêt 4. À l'arrêt 1, le modèle de prédiction simple prédit le retard pour chacun des arrêts suivants du trajet. Il va propager le retard de quatre minutes sur tous les arrêts suivants en tenant compte de la contrainte du temps de battement prévue par le GTFS. Puisque le chauffeur sait qu'il est en retard, il n'utilisera pas ce temps de battement et continuera sa route immédiatement à l'arrêt 4. Ainsi, le retard prédit est de quatre minutes pour les arrêts 2,3 et 4. Il est ensuite de deux minutes pour tous les arrêts suivants. Dans la suite du rapport, ces nouvelles prédictions seront appelées les données PRED (pour prédiction).

	stop_id	wait.time	sequence	pred.delay	trip_id	vehicle_id	current_time	arrival_time	route_id	day
1	965	0	2	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:50:42	17S	2018-03-12
2	967	0	3	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:50:55	17S	2018-03-12
3	2271	0	4	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:51:15	17S	2018-03-12
4	1209	0	5	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:52:11	17S	2018-03-12
5	1211	0	6	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:53:04	17S	2018-03-12
6	1513	0	7	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:53:26	17S	2018-03-12

Figure 2.5 : Structure des données PRED

Dans le cas de la figure 2.4, le trajet 477346 (trip_id) est à l'arrêt 2 (sequence) à 5h51min13s (current_time). Il aurait dû y être, selon le GTFS, à 5h50min42s (arrival_time). Le retard réel à l'arrêt 2 est compris entre zéro et une minute, donc est arrondi à zéro minute. Ainsi, le retard prédit depuis l'arrêt 2 vers les arrêts suivants est de zéro minute (colonne des pred.delay), puisqu'il n'y a pas de contraintes opérationnelles prévues par le GTFS ici.

2.4 Combinaison des données SAE et PRED : introduction de ΔR et ΔT

L'étape suivante effectuée est de combiner les données SAE et PRED. Deux nouvelles notions sont définies pour cela : l'arrêt de visée et l'arrêt cible. L'exemple précédent va permettre d'illustrer cela : le bus est à l'arrêt 1 et on cherche à prédire le retard à l'arrêt 7. Dans ce cas, l'arrêt 1 est appelé l'arrêt de visée : c'est l'arrêt à partir duquel la prédiction est faite. L'arrêt 7 est lui nommé l'arrêt cible : c'est l'arrêt pour lequel la prédiction est faite.

À partir de là, les deux variables principales de notre étude peuvent être définies : ***ΔR = Retard réel à l'arrêt cible – Retard prédit pour l'arrêt cible depuis l'arrêt de visée ; ΔT =heure d'arrivée effective à l'arrêt cible –heure d'arrivée effective à l'arrêt de visée.***

L'heure d'arrivée effective à l'arrêt de visée correspond à l'heure où est faite la prédiction pour l'arrêt cible depuis l'arrêt de visée. Ainsi, ΔR symbolise l'erreur de la prédiction. ΔT représente l'horizon temporel de la prédiction. Voici la structure de données obtenue :

	stop_id	wait.time	sequence	pred.delay	trip_id	vehicle_id	current_time	arrival_time	route_id	day	latitude	longitude
1	965	0	2	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:50:42	17S	2018-03-12	45.34884	-71.99117
2	967	0	3	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:50:55	17S	2018-03-12	45.34958	-71.99007
3	2271	0	4	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:51:15	17S	2018-03-12	45.35106	-71.98895
4	1209	0	5	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:52:11	17S	2018-03-12	45.35485	-71.98799
5	1211	0	6	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:53:04	17S	2018-03-12	45.35489	-71.99406
6	1513	0	7	0	477346	62104	2018-03-12 05:51:13	2018-03-12 05:53:26	17S	2018-03-12	45.35651	-71.99413

	time	real.delay	speed	bearing	deltaR	deltaT	hour	week_day	period
1	2018-03-12 05:51:13	0	8	50	0	0	5	2	1
2	2018-03-12 05:52:00	1	9	46	1	47	5	2	1
3	2018-03-12 05:52:13	0	9	28	0	60	5	2	1
4	2018-03-12 05:53:05	0	8	272	0	112	5	2	1
5	2018-03-12 05:54:02	0	8	271	0	169	5	2	1
6	2018-03-12 05:54:30	1	9	359	1	197	5	2	1

Figure 2.6 : Création des variables ΔR et ΔT

ΔR correspond à l'adhérence horaire du modèle de prédiction. L'objectif choisi est de maximiser le pourcentage de ΔR dans l'intervalle [-1 ; +3]. Cet intervalle répond aux exigences opérationnelles de l'adhérence horaire. Une adhérence horaire positive est préférée puisque cela signifie que le bus est en retard, ce qui est mieux qu'un bus en avance. En effet, le pire scénario survient lorsqu'un usager arrive à l'heure à la station mais que le bus y est déjà passé.

2.5 Performance de ΔR selon différents horizons, point de départ de l'étude

Le premier pas de la démarche consiste à analyser les performances du modèle de prédiction simple. Après tout, il est intuitif et devrait procurer de bons résultats. En considérant l'ensemble des données combinées SAE et PRED de semaine, il y a 44 332 841 données. Il faut remarquer qu'il y a beaucoup plus de données combinées que de données SAE simples. Cela est logique puisqu'à chaque arrêt, donc chaque mesure SAE, une prédiction est effectuée pour tous les arrêts ultérieurs. En tenant compte de tous les horizons temporels, le pourcentage de ΔR dans l'intervalle [-1 ; +3] est de 87%. La performance du modèle est donc de 87%.

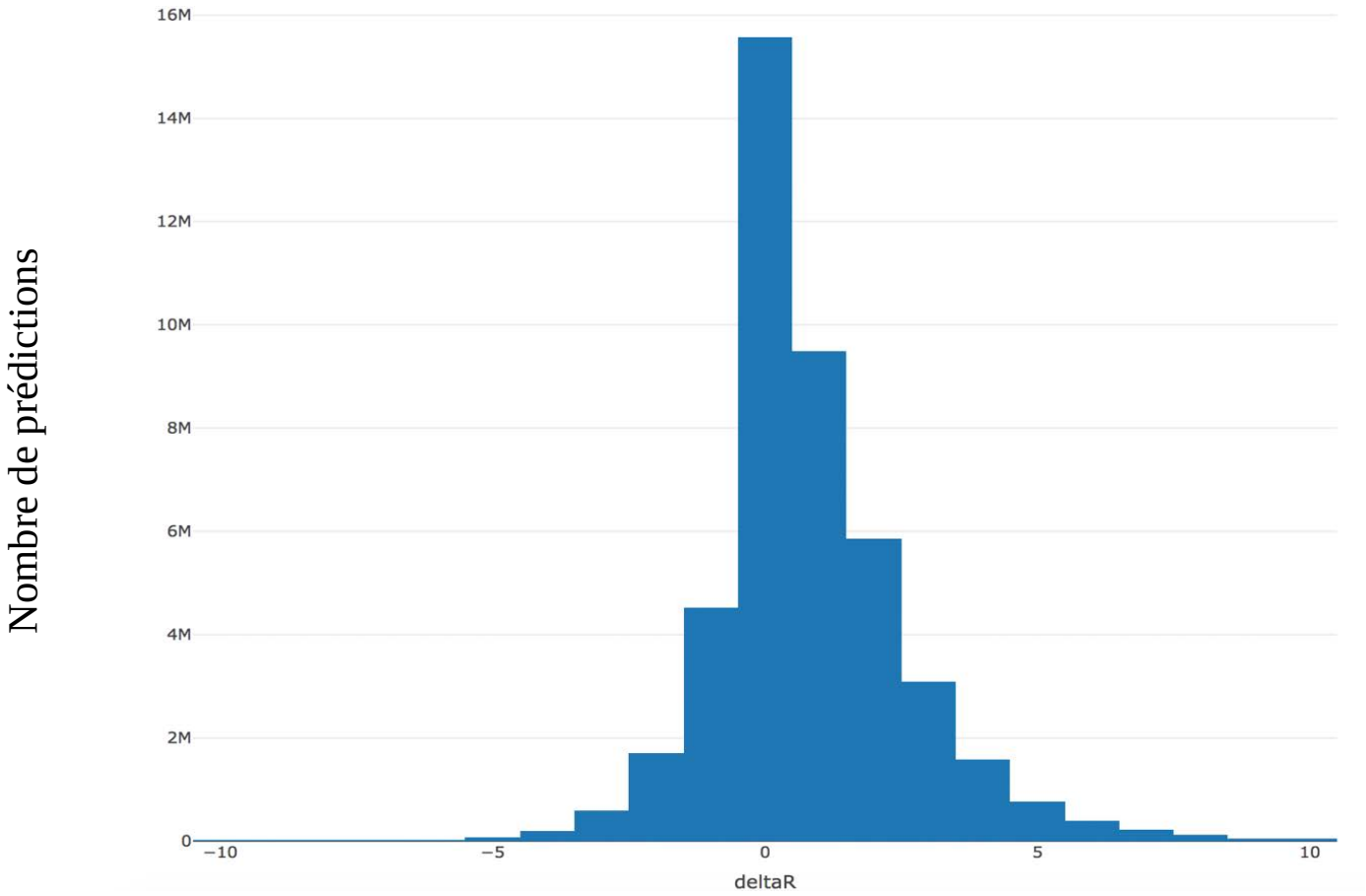

Figure 2.7 : Histogramme de performance du modèle de prédiction simple, pour tous horizons et pour les quatre semaines d’étude, en jour de semaine

En filtrant selon l’horizon ΔT, des variations sont obtenues. Il est constaté que plus l’horizon augmente, plus la fiabilité de la prédiction diminue. Par exemple, une performance moyenne de 91% est obtenue pour les horizons de zéro à cinq minutes, et de 87% pour ceux de dix à quinze minutes. Avec une segmentation supplémentaire selon l’heure de la journée, on obtient le graphe suivant :

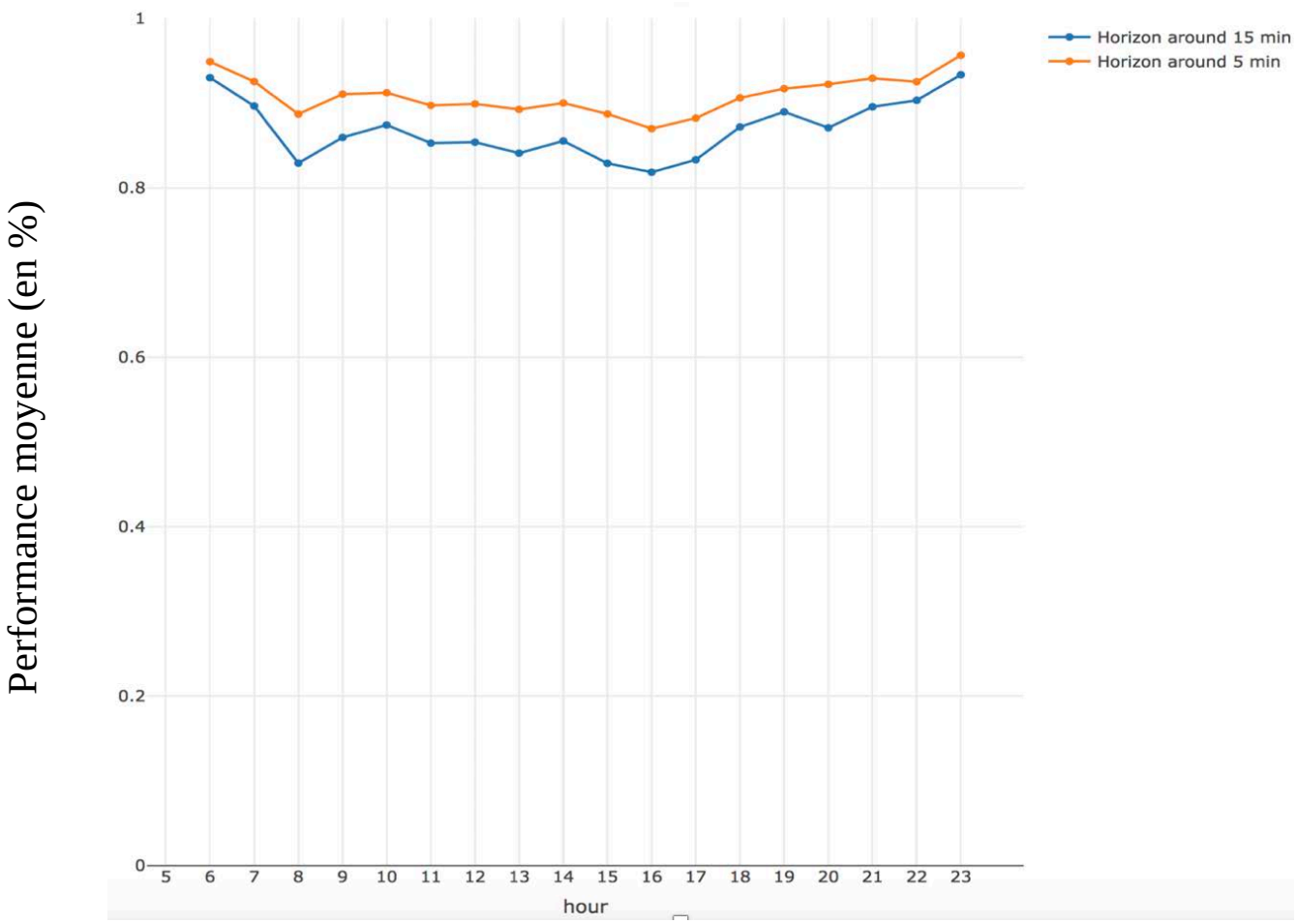

Figure 2.8 : Performance globale du modèle de prédiction simple, par heures et horizons

En orange, on retrouve la courbe pour les horizons entre zéro et cinq minutes (« horizon around 5 min »). En bleu, c'est celle pour les horizons entre dix et quinze minutes (« horizon around 15 min »). Il faut également noter que les heures sont arrondies par défaut dans l'étude : les prédictions effectuées entre 8h00 et 8h59 sont comptabilisées pour la huitième heure par exemple.

Le constat est, comme expliqué plus haut, que la prédiction d'horizon zéro/cinq minutes est toujours meilleure que celle d'horizon dix/quinze minutes. De plus, les prédictions sont en moyenne moins bonnes en période de pointe (8h et 16/17h), et ce quel que soit l'horizon. En continuant à augmenter l'horizon, le comportement par heure reste similaire mais la performance moyenne diminue petit à petit.

Ainsi, la prédiction simple est assez bonne dans l'ensemble, mais elle connaît plus de difficultés dans les périodes de perturbations et pour les horizons temporels importants. C'est pourquoi d'autres méthodes doivent être explorées afin d'améliorer le modèle.

2.6 Définition de ΔR2, nouvelle variable d'étude

Il a été vu précédemment que la prédiction simple prenait en compte les caractéristiques du GTFS, comme les temps de battement par exemple. Pour créer un modèle efficace, il est nécessaire de tenir compte de ces caractéristiques et il n'y a pas moyen de les prédire sans utiliser le GTFS. C'est pourquoi, pour établir un nouveau modèle, il est décidé de repartir des valeurs obtenues avec la prédiction simple, et de les ajuster grâce à des termes correctifs. Ainsi, le travail ne repart pas de zéro et la mainmise sur les particularités du GTFS est conservée.

Une nouvelle variable d'étude est donc définie, elle sera nommée ΔR2 : ***ΔR2 = ΔR - ΔRestimé.*** Le nouvel objectif est de maximiser le pourcentage de ΔR2 dans l'intervalle [-1 ; +3]. Le but est d'estimer la valeur de ΔR à l'aide d'une méthode différente, pour ensuite retrancher cette estimation à la valeur initiale de ΔR et atteindre une valeur de ΔR2 la plus proche possible de zéro. Autrement dit, il est cherché à prédire l'erreur du modèle de prédiction simple pour pouvoir la corriger, et obtenir un nouveau modèle le plus performant possible.

CHAPITRE 3 MÉTHODOLOGIE

3.1 Variables de prédiction utilisées

3.1.1 Présentation de l'idée

Pour prédire ΔR, des variables de prédiction sont nécessaires. De nombreuses variables existent et les plus adéquates doivent être déterminées. Un modèle de prédiction généralisable est recherché, c'est-à-dire que si une ligne de bus est rajoutée au réseau ou si la même étude est faite dans une ville différente, le modèle devra être adaptable. Pour répondre à cette exigence, et après plusieurs études qui sont détaillées dans le chapitre 5, il a été décidé de conserver trois catégories de variables : les catégories de tronçons, les distances inter-arrêts et les retards en temps réel.

3.1.2 Découpage des tronçons en trente-huit catégories

Un tronçon est défini comme une séquence d'arrêts successifs (stop_id) desservie par n'importe quel trajet du réseau, pendant une période donnée de la journée. En comptabilisant toutes les lignes et trajets de semaine, cela donne un total de 1 661 tronçons. L'idée est ensuite de les catégoriser. Deux caractéristiques de chaque tronçon sont principalement intéressantes : le retard réel moyen, ainsi que le ΔRmoyen (performance moyenne de la prédiction simple). Ces deux variables seront donc utilisées pour caractériser les tronçons.

Quatre périodes dans la journée sont également distinguées :

- Période 1 : de 6h à 9h (pointe du matin)
- Période 2 : de 9h à 15h (après-midi)
- Période 3 : de 15h à 18h (pointe du soir)
- Période 4 : de 18h à 00h (soirée)

Ces distinctions sont faites car il est supposé que les comportements d'un même tronçon varient, selon les pointes/hors-pointes notamment. À partir des résultats obtenus suite à la classification, trente-huit catégories sont définies. Ces catégories sont choisies arbitrairement, en quadrillant l'espace selon les valeurs du ΔRmoyen et du retard réel moyen. Un numéro est attribué à chacune de ces catégories, de 1 à 38.

L'idée de cette catégorisation est de caractériser chaque tronçon selon sa performance vis-à-vis du modèle de prédiction simple. Il convient de se rappeler que plus le ΔRmoyen est proche de zéro, plus la prédiction pour un tronçon donné est bonne. Plus le ΔRmoyen est élevé, plus le tronçon a tendance à créer plus de retard que prévu. Plus le ΔRmoyen est faible (de valeur négative), plus le tronçon a tendance à créer plus d'avance que prévu. La valeur du retard réel moyen informe sur la prise de retard réelle provoquée par le tronçon. Pour l'étude, si les retards existent mais sont bien prédits, cela n'est pas si grave. Ainsi, les zones qui ont un retard réel moyen élevé (entre 3 et 7 minutes), mais avec un ΔRmoyen très bon (entre -0,5 et 0,5 minute), sont considérées comme bonnes. Cette catégorisation procure les premières variables importantes pour prédire ΔR.

3.1.3 Distance de Haversine inter-arrêts

Dans un second temps, il est supposé que la distance entre deux arrêts influe directement sur la fiabilité de la prédiction simple. Il est donc décidé d'introduire une variable de distance pour chaque couple arrêt de visée/arrêt cible. La distance utilisée est la distance de Haversine, qui est une distance à vol d'oiseau. L'hypothèse faite est que prendre une distance à vol d'oiseau ne change pas fondamentalement la distance réelle qui serait plus de type Manhattan à priori, sachant que les rues sont souvent à angle droit. Cela fait du sens puisque, pour un couple arrêt de visée/arrêt cible, la distance est calculée en faisant la somme des distances d'Haversine de chaque tronçon faisant partie du couple. Or, pour un tronçon simple, la distance à vol d'oiseau est très proche de la distance réelle.

Par exemple, pour quatre trajets consécutifs d'une même ligne, les variables de prédiction pour chaque couple visée/cible afficheront une distance identique. Cela est normal puisque c'est le même couple d'arrêts. En revanche, les catégories changent si les périodes de la journée sont différentes.

3.1.4 Retard réel à l'arrêt de visée

La dernière variable utilisée est le retard réel à l'arrêt de visée. Par exemple, si le bus est à l'arrêt 3 et prédit le retard pour l'arrêt 10, la variable du retard réel à l'arrêt 3 est prise en compte. En effet, en temps réel de condition opérationnelle, il n'y a pas accès aux mesures du retard suivantes donc on ne connaît pas encore le retard réel à l'arrêt cible 10. Il faut par conséquent prendre la dernière

mesure de retard disponible à chaque instant, c’est-à-dire le retard réel à l’arrêt de visée. L’heure à laquelle est effectuée la prédiction sera également considérée comme une variable d’entrée.

En résumé, les variables de prédiction utilisées dans l’étude sont : les catégories des tronçons (38 variables, chacune égale au nombre de tronçons de la catégorie correspondante sur l’intervalle arrêt de visée/arrêt cible considéré par la prédiction), la distance inter-arrêts, le retard réel à l’arrêt de visée et l’heure (arrondie par défaut) de la prédiction.

3.2 Méthodes potentielles envisagées

Dans cette sous-partie sont présentées trois méthodes qui ont été étudiées mais qui n'ont finalement pas été choisies : la régression linéaire, les arbres de décision et les réseaux de neurones. Dans le cadre de l'étude, on rappelle que le logiciel OpenSource R est utilisé.

3.2.1 Régression linéaire

La première idée suivie pendant le travail est la méthode qui semble la plus simple : la régression linéaire. En effet, si cette méthode peut donner des résultats satisfaisants, elle est intéressante puisque très facile à mettre en œuvre. Un « catalogue » spécifique de variables est alors établi : il servira d'entrée au modèle. Les quatorze variables suivantes sont utilisées :

v1 = ΔR du jour précédent pour le même trip_id
v2 = Retard réel à l'arrêt de visée
v3 = Moyenne des trois retards précédant l'arrêt de visée
v4 = Différence retard réel arrêt de visée - retard réel précédent
v5 = Différence retard réel arrêt cible - retard réel arrêt visée pour le jour précédent
v6 = Différence retard réel arrêt cible - retard réel arrêt visée pour le trip_id précédent, sur une même période
v7 = Indicateur météo « TPRH » : | Température - Point de rosée | / Humidité relative
v8 = Heure de la journée
v9 = Vitesse
v10 = Angle
v11 = Période de la journée
v12 = Jour de la semaine
v13 = Température
v14 = Catégorie de temps

Après simulation, les trois variables les plus importantes semblent être v1 : le ΔR du jour précédent pour le même trip_id ; v5 : la différence entre le retard réel de l'arrêt cible pour le jour précédent et le retard réel de l'arrêt de visée pour le jour précédent ; et v14 : la catégorie de temps (normal, glace, neige, brouillard, pluie) qui sera présentée plus en détail au chapitre 5. Ci-dessous se trouve une représentation graphique de la relation qui existe entre le ΔR et le ΔR du jour précédent pour un même trip_id (ce qui correspond à la variable v1). L'exemple pris est celui du trajet 477662.

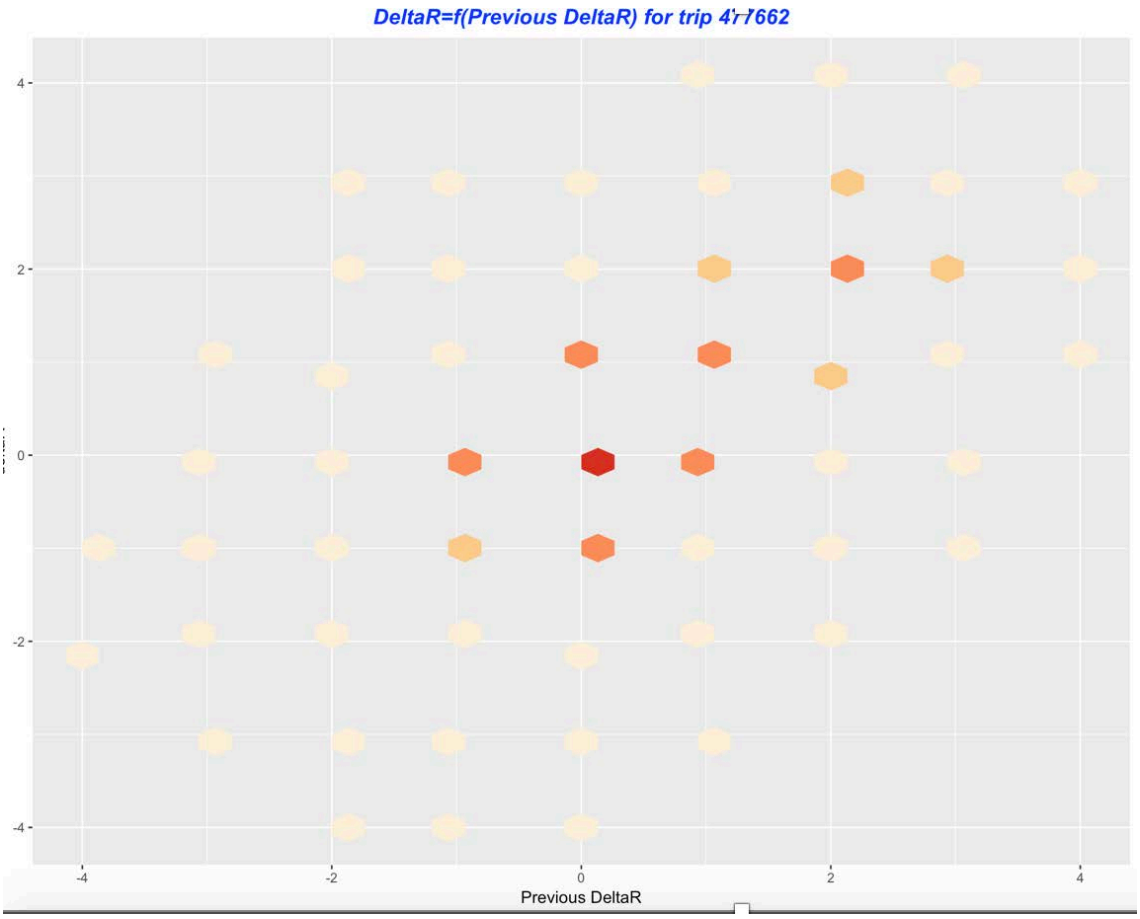

Figure 3.1 : Corrélation linéaire entre ΔR et v1, pour l'exemple du trajet 477662

Cette représentation en graphes hexagonaux montre qu'il y a une forte corrélation, ce qui est rassurant. Cependant, hormis les variables v1, v5 et v14, les autres variables ne démontrent pas une corrélation évidente avec ΔR. Les résultats de performance amenés par la régression linéaire seront présentés au chapitre 4, section 1.1.

3.2.2 Arbres de décision

La seconde méthode est celle des arbres de décision. La méthode se divise en deux parties : l'entraînement sur un échantillon donné, puis le test sur un échantillon distinct. Après entraînement sur les données d'une certaine ligne, pour des horizons de dix à quinze minutes, l'arbre suivant est obtenu. La ligne étudiée a été choisie pour sa grande longueur, sa fréquence élevée, et sa bonne représentativité des caractéristiques globales du réseau. Par souci de confidentialité, cette ligne sera

appelée la ligne A. Cet arbre est ensuite appliqué à l'échantillon test, et des prédictions pour ΔR2 en sont ainsi obtenues.

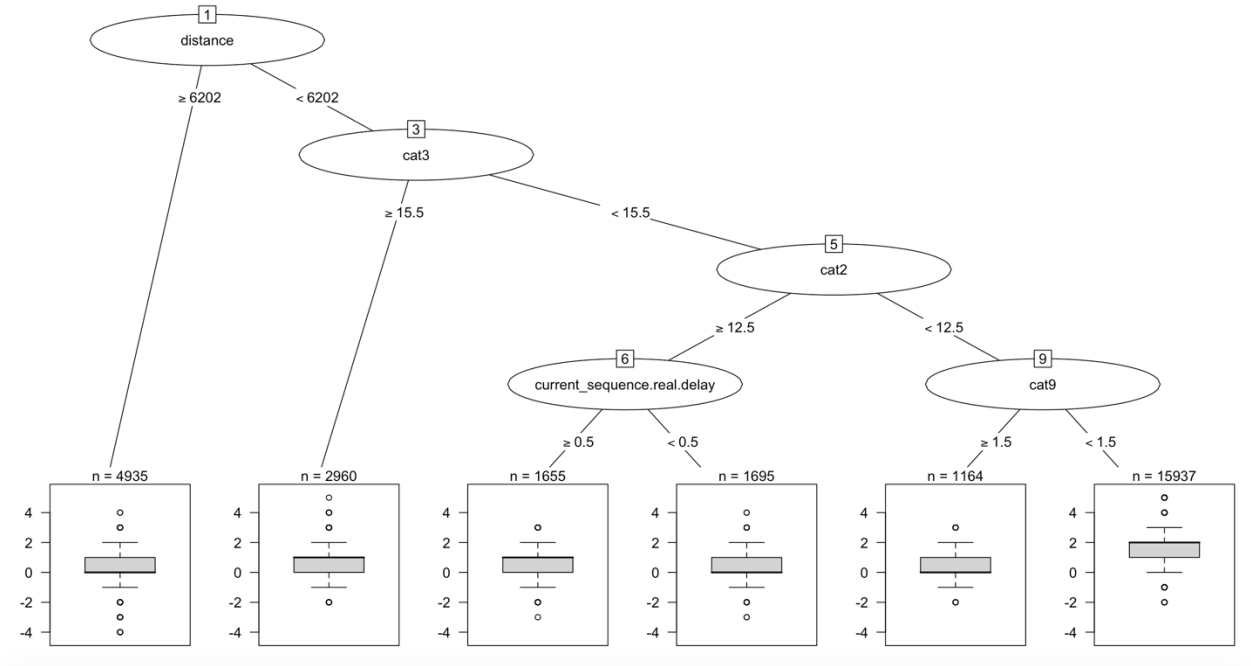

Figure 3.2 : Exemple d'arbre de décision généré, pour la ligne A

Les résultats de performance amenés par les arbres de décision seront présentés au chapitre 4, section 1.2.

3.2.3 Réseaux de neurones

La troisième et dernière méthode testée est celle des réseaux de neurones. Cette méthode est sans doute la plus complexe, elle contient énormément de paramètres sur lesquels il est possible de jouer. Pour simplifier, les paramètres utilisés par défaut par le logiciel R sont conservés. La figure 3.3 en page suivante est tirée d'une simulation sur ce même logiciel R. Un réseau de neurones à trois couches y est modélisé. Celui-ci cherche à prédire ΔR à partir de treize variables d'entrée de notre étude. Le package « neuralnet » est utilisé. On peut déjà remarquer que la structure semble extrêmement lourde avec une grande quantité de paramètres à estimer pour le modèle. Cette densité ne semble à priori pas idéale pour une utilisation opérationnelle flexible.

Les résultats de performance des réseaux de neurones seront eux présentés plus en détail au chapitre 4, section 1.3.

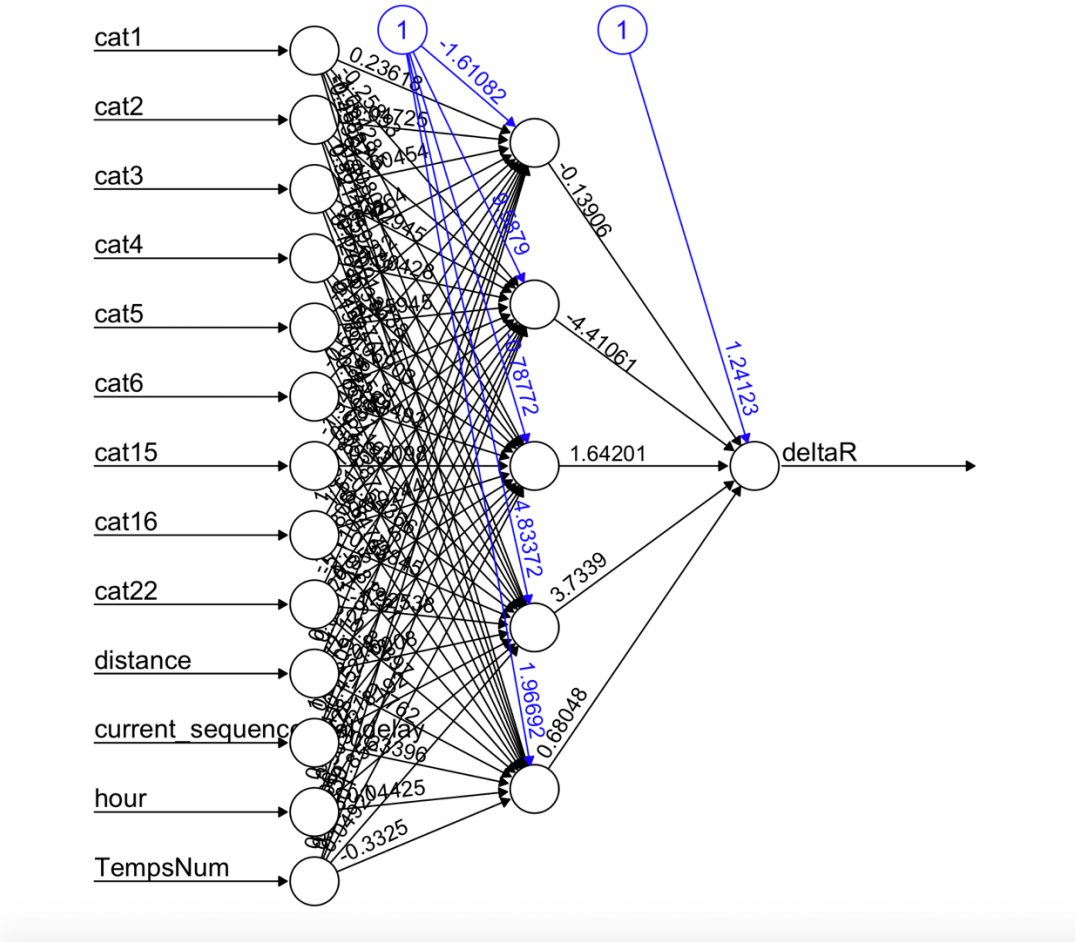

Figure 3.3 : Paramétrage d'un réseau de neurones à treize entrées et trois couches

3.3 Méthode utilisée, le *random forest*

3.3.1 Présentation de la méthode

La méthode de *random forest* qui est finalement choisie pour le modèle de prédiction est celle implémentée dans le package « randomforest » du logiciel R. Par définition, un *random forest* a besoin d'être entraîné sur un échantillon, avant d'être testé sur un autre. Les données seront donc à chaque fois divisées en deux parties : un échantillon « train » et un échantillon « test ». Plusieurs paramètres sont utilisables, mais un seul principal sera considéré : le nombre d'arbres de décision aléatoires à utiliser pour l'entraînement. Il est choisi d'en utiliser dix. Selon les simulations effectuées, plus on augmente le nombre d'arbres et meilleur est le résultat, mais les performances n'augmentent pas significativement en comparaison du temps de traitement. Voici les variations de performances selon le nombre d'arbres, pour une nouvelle ligne (que nous appellerons B), tous horizons temporels inclus.

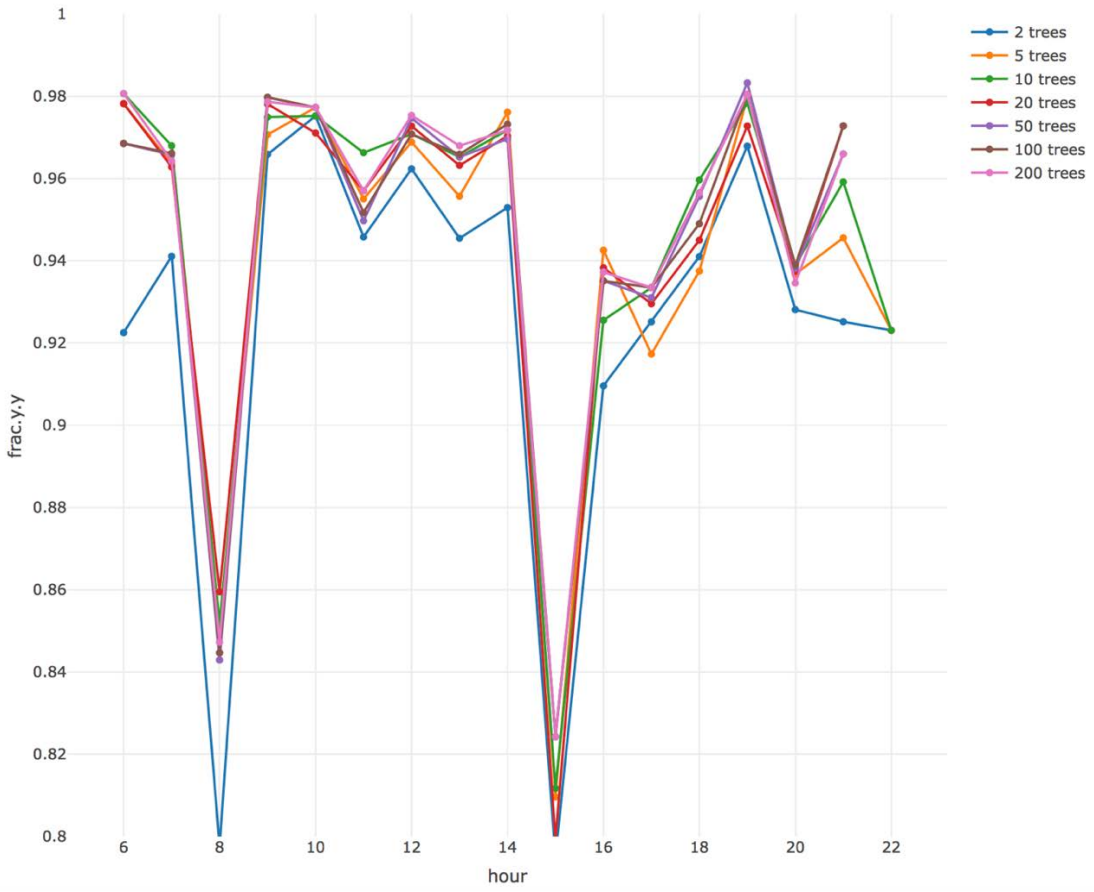

Figure 3.4 : Performances *random forest* de la ligne B, selon le nombre d'arbres utilisés

3.3.2 Deux modèles différents : la semaine et le weekend

Pour la suite, le travail va continuer à être divisé en deux parties : l'étude de la semaine (du lundi au vendredi inclus) d'une part, et l'étude du weekend (les samedis et dimanches) d'autre part. Cela vient du fait que les conditions opérationnelles sont très différentes entre ces deux périodes.

Les données de semaine couvrent quatre semaines de l'année 2018 : la semaine du 12 au 16 mars (entraînement), la semaine du 19 au 23 mars (entraînement), la semaine du 26 au 30 mars (test) et la semaine du 2 au 6 avril (test). Les données du weekend couvrent cinq weekends : le weekend du 10 et 11 mars (entraînement), le weekend du 17 et 18 mars (entraînement), le weekend du 24 et 25 mars (entraînement), le weekend du 31 mars et 1er avril (test) et le weekend du 7 avril et 8 avril (test). Comme précédemment, chaque journée est également divisée en quatre périodes, encore une fois cela provient des observations opérationnelles : de grosses variations existent entre les périodes de pointe et hors-pointe notamment. Les jours sont donc chacun divisés en quatre périodes, telles qu'elles ont été définies au chapitre 3. Des tables de catégories sont alors construites pour chacun des tronçons, selon la période de la journée. Un tronçon est caractérisé par son numéro d'arrêt (stop_id) d'origine et son numéro d'arrêt (stop_id) de destination. Les numéros de séquence (sequence) ne peuvent plus être utilisés ici, puisque cette catégorisation doit fonctionner quelle que soit la ligne d'autobus.

Entre la semaine et le weekend, les distances d'Haversine restent les mêmes, ce qui est logique. Cependant, on constate que les catégories affectées sont différentes. Les catégories changent et peuvent même devenir des NA (ce qui veut dire que les tronçons ne sont pas desservis durant cette période de la journée). Ces différences dans le comportement des tronçons confortent dans l'idée de faire une distinction entre un modèle de semaine et un de weekend.

3.3.3 Cas des jours fériés

Une autre problématique est l'appréhension des jours fériés. Dans la période de temps étudiée ici, cela concerne le vendredi 30 mars (vendredi Saint) et le lundi 2 avril (lundi de Pâques). Après une analyse focalisée sur ces deux jours, il est constaté qu'ils semblent mieux s'adapter à la modélisation de semaine plutôt qu'à celle de weekend. Dans la suite, les jours fériés seront ainsi considérés comme des jours de semaine. Cette hypothèse sera cependant à vérifier avec plus de données sur plus de jours fériés.

3.3.4 Importance relative des variables

Il est maintenant possible de commencer à utiliser les *random forests* puisque les variables d'entrées nécessaires sont à disposition. R nous permet de savoir quelles sont les variables les plus prépondérantes parmi celles utilisées en entrée du modèle. Voici par exemple l'importance relative des variables, pour une autre ligne (appelée C par souci de confidentialité), avec dix arbres utilisés dans le modèle de *random forest*.

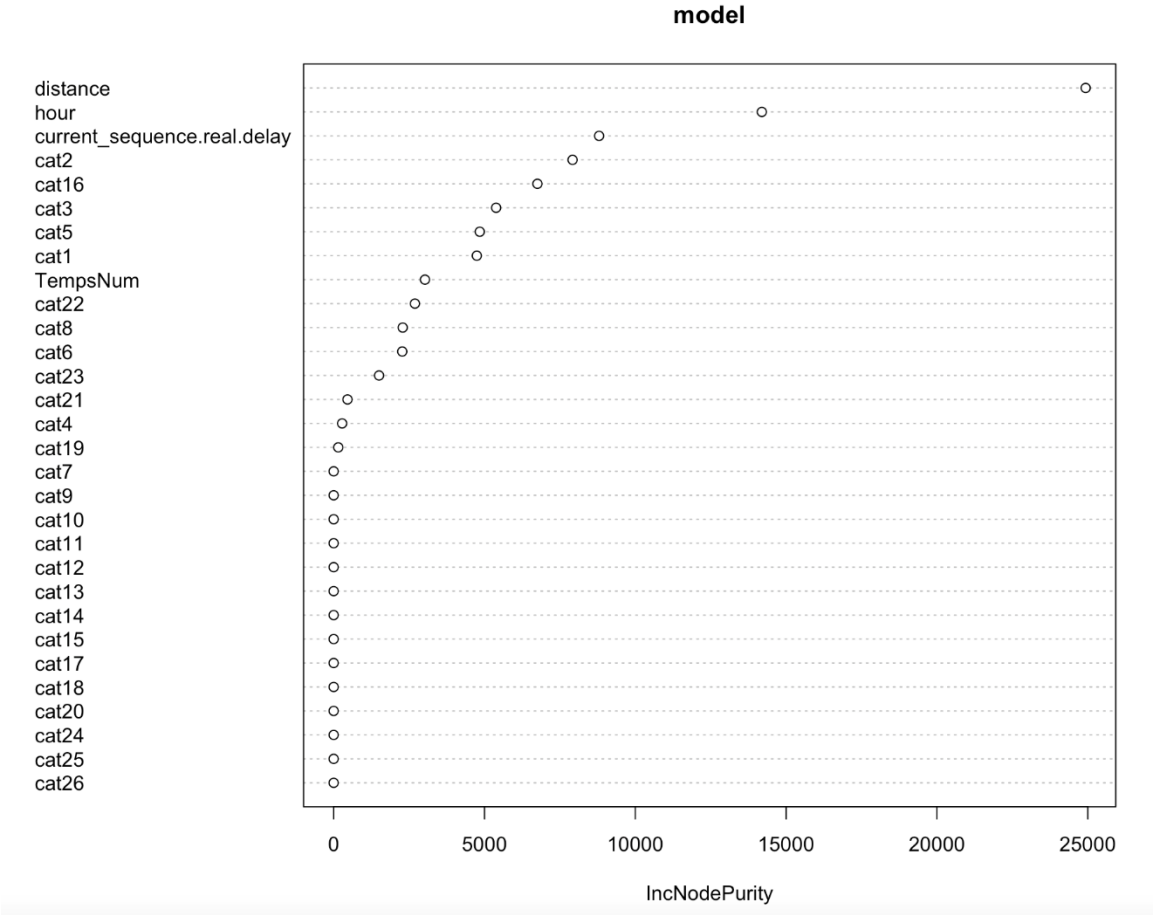

Figure 3.5 : Importance relative des variables pour le modèle *random forest* à dix arbres de la ligne C

Cette figure 3.5 indique que la distance, l'heure de la journée et le retard réel à l'arrêt de visée sont les trois variables, dans cet ordre, ayant les plus d'impact sur la prédiction. On constate d'ailleurs que ce comportement se reproduit pour la grande majorité des lignes.

CHAPITRE 4 RÉSULTATS

Cette partie du rapport rassemble les résultats de l'étude. Les différentes méthodes présentées au chapitre précédent seront d'abord examinées successivement. Dans un second temps, l'accent sera porté sur la problématique de l'implantation du modèle choisi en conditions opérationnelles, pour la Société de Transport de Sherbrooke.

4.1 Performances comparées des différentes méthodes

4.1.1 Régression linéaire

Comme exposé précédemment dans ce rapport, la régression linéaire a l'avantage d'être la méthode la plus simple et intuitive. Par conséquent, elle se calcule facilement et serait très efficace en termes de temps de traitement, si on décidait de l'implanter. En entrée, les quatorze variables définies en 3.2.1 sont utilisées.

Le premier constat est que la régression linéaire donne des résultats globalement similaires au modèle de prédiction simple. La figure 4.1 est plutôt éloquente à ce sujet.

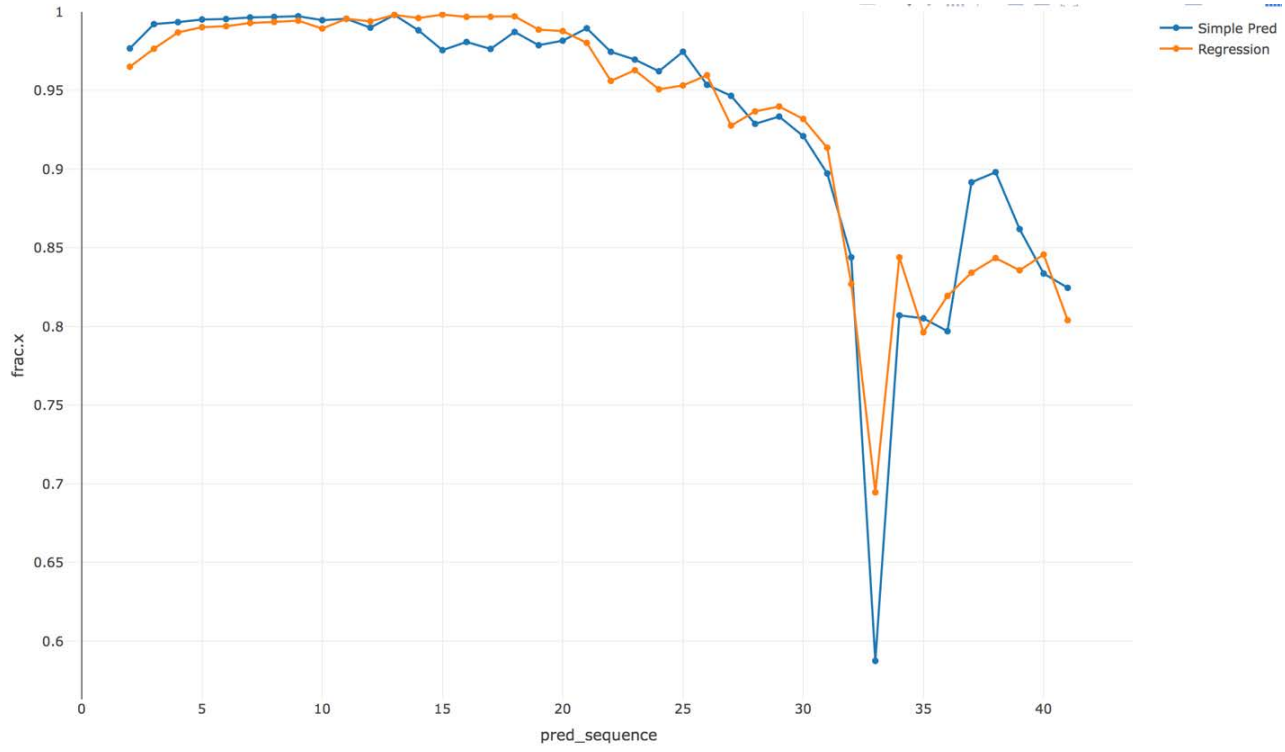

Figure 4.1 : Comparaison des performances pour le voyage 478112 (direction 1) de la ligne A

On rappelle que l'indicateur ΔR2, défini au chapitre 2, est notre indicateur d'intérêt. On veut maximiser le pourcentage de sa fraction dans l'intervalle [-1 ; 3]. En comparant ce pourcentage avec celui de la prédiction simple ΔR, l'amélioration de notre modèle peut être déterminée. C'est pourquoi, dans la suite, les courbes de performances des différentes méthodes vont être successivement superposées à celles de la prédiction simple, afin de pouvoir procéder à des comparaisons.

Dans cette figure 4.1, un voyage précis a été choisi : le voyage 478112 de la ligne A. Un numéro de voyage correspond à un trajet quotidien précis à un horaire précis dans une direction donnée. Il y a également une distinction opérée entre la semaine et le weekend. Autrement dit, le voyage 478112 correspond à un même trajet quotidien de la ligne A, chaque jour de semaine, dans la même direction (la direction 1 ici) et à la même heure. Toutes les données correspondantes dans l'échantillon sont alors considérées, pour tous les horizons. De là, on compare les performances des deux modèles – prédiction simple et régression linéaire – en fonction du numéro de séquence de l'arrêt. La courbe de prédiction simple est bleue, tandis que celle de régression linéaire est orange. Ici, la courbe de régression épouse celle de prédiction simple : cela est rassurant puisque cela renforce l'idée que la prédiction simple est une bonne estimation de départ. De plus, cet exemple a été choisi car il illustre le point fort de la régression linéaire. La station numéro 33 apparaît ici comme très problématique pour la prédiction. En réalité, cette station est une des stations centrales du réseau. Les horaires GTFS pour cette station sont « falsifiés » par les horairistes pour permettre une régulation opérationnelle des flux d'autobus. Les détails sont assez techniques mais il convient surtout d'en analyser les impacts sur le modèle. Pour la prédiction simple, la fraction dans l'intervalle [-1 ; +3] chute à 58%. Celle de la régression chute aussi, mais seulement à 69%. Dans ce cas, la régression semble amortir les fluctuations brutales de performance obtenues avec la prédiction simple.

Dans la figure 4.2 de la page suivante, tous les voyages de la ligne A, dans cette même direction 1, sont cette fois considérés. On retrouve le même problème à la station centrale évoquée plus haut. Mais cette fois, les stations précédentes sont aussi impactées. La performance de la prédiction simple diminue progressivement, depuis la station 26 jusqu'à la station 33. La performance de la régression, elle, reste stable jusqu'à la station 31 avant de commencer à ressentir la fluctuation. Le point fort de la régression est donc confirmé : elle permet d'atténuer les défauts de la prédiction simple. Cependant, elle ne permet pas d'améliorer significativement ses performances dans le cas

général. Cela se voit ici du départ à la station 25 puis de la station 34 au terminus, où la régression est même moins précise que la prédiction simple.

Ainsi, d'autres méthodes doivent être approfondies, à commencer par les arbres de décision.

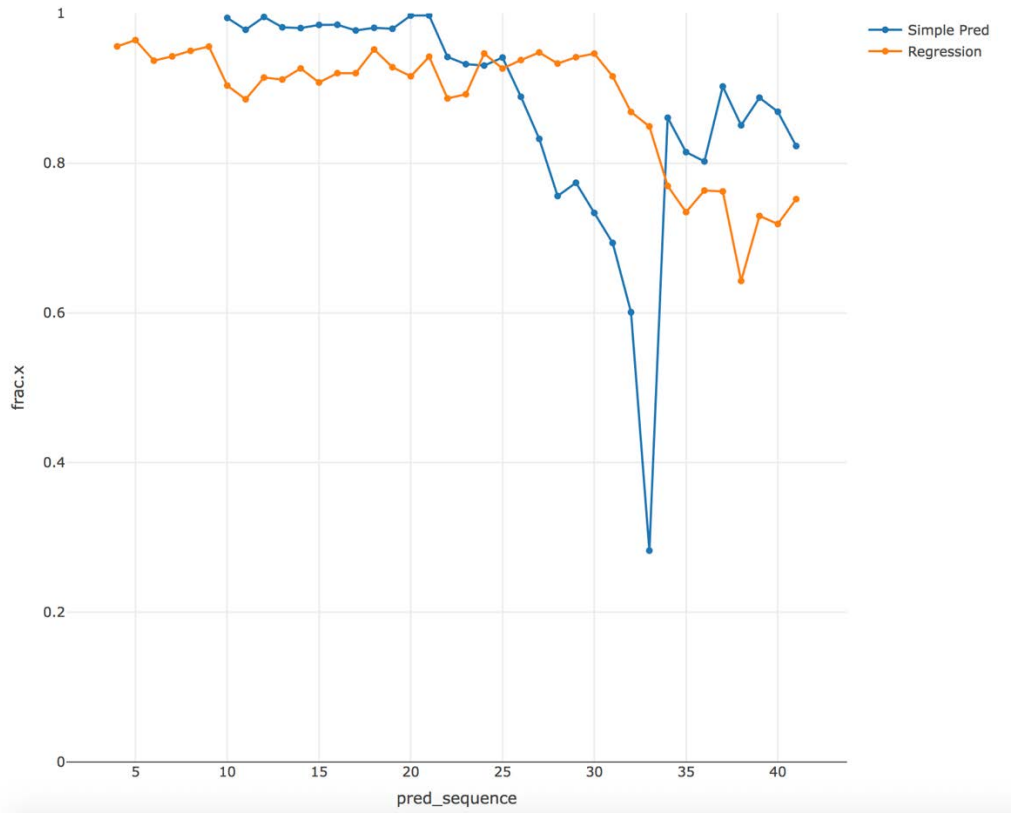

Figure 4.2 : Comparaison des performances pour tous les voyages de la ligne A, direction 1

4.1.2 Arbres de décision

La méthode des arbres de décision a été présentée précédemment, en section 3.2.2. L'exemple de la ligne A est repris pour comparer les performances de cette méthode avec celles de prédiction simple et de régression linéaire. La ligne A est considérée car c'est une ligne longue, complexe et achalandée. Cette fois, l'échantillon analysé comporte tous les voyages de la ligne A, dans les deux directions (direction 0 et direction 1). De plus, on regarde les performances par heure de la

journée, et non plus par numéro d'arrêt. Les résultats sont présentés en figure 4.3. La prédiction simple est en bleu, la régression en vert et l'arbre de décision en orange.

La prédiction simple connaît quelques chutes, notamment durant les pointes du matin (autour de 8/9h) et du soir (autour de 16h). Les heures creuses du midi semblent aussi être touchées. On remarque une nouvelle fois que la régression linéaire n'est pas très éloignée de la prédiction simple.

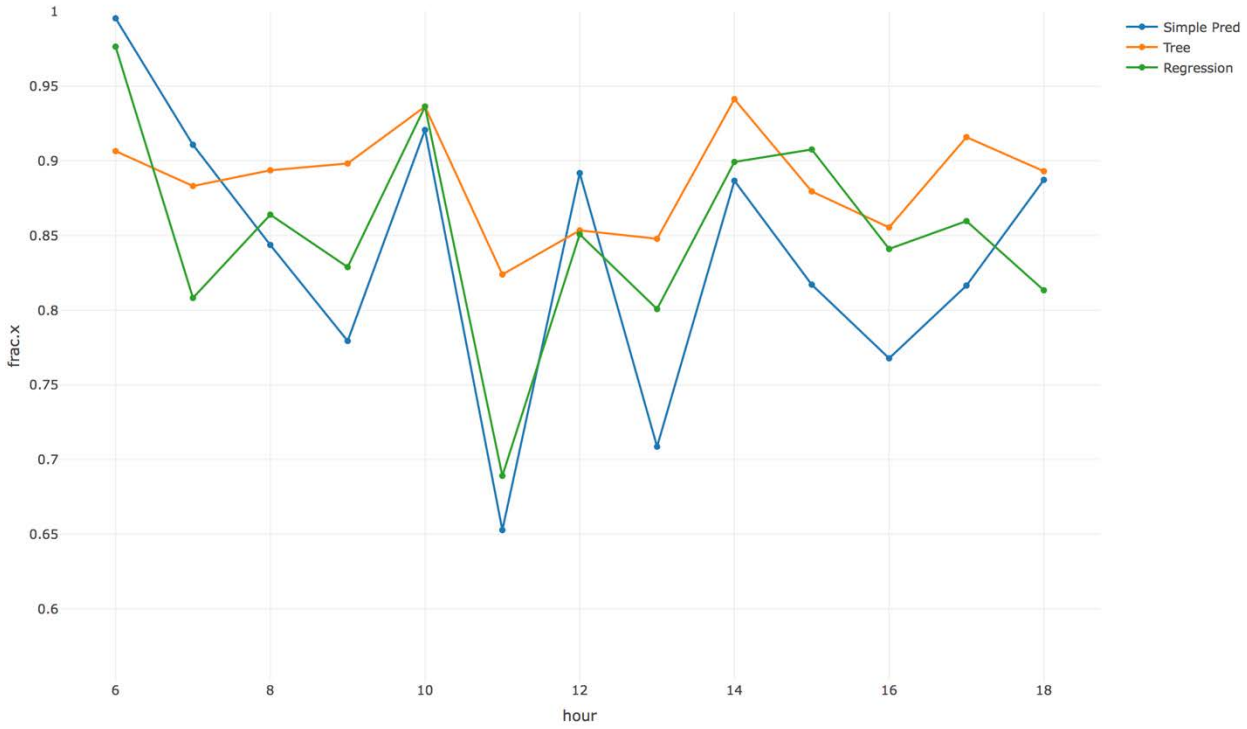

Figure 3 : Comparaisons des performances avec l'arbre de décision, pour tous les voyages de la ligne A

Pourtant, ici, l'arbre de décision semble donner de meilleurs résultats à quasiment toutes les heures de la journée, avec une moyenne de fraction de $\Delta R2$ dans l'intervalle [-1 ; +3] autour de 90%. Il semble aussi assez stable, ce qui est primordial pour être utilisé en opérationnel. Puisque l'arbre de décision semble être intéressant, il existe une méthode spécifique qui se base sur plusieurs arbres de décision combinés les uns avec les autres, comme expliqué en partie 3.3. Cette méthode est le *random forest*, ou forêt aléatoire en français.

4.1.3 Random forests

Il convient à présent de montrer pourquoi il a été décidé de privilégier la méthode des *random forests*, comme exposé précédemment dans la partie 3.3. Ci-dessous, on peut voir les performances des différents modèles de prédictions pour tous les voyages de cinq lignes réunies, par heure de la journée, avec des horizons compris entre dix et quinze minutes. Ces lignes sont respectivement les lignes A, B, C, D et E. Les lignes A, B et C sont les mêmes que celles étudiées précédemment dans ce rapport. Les lignes D et E sont deux autres lignes du réseau. Encore une fois, on les nomme ainsi pour respecter la confidentialité. Toutes ces lignes ont été choisies car elles ont beaucoup d'arrêts, sont achalandées et ont plusieurs points potentiellement problématiques. Entraînement et test sont faits sur l'ensemble des cinq lignes en même temps ici.

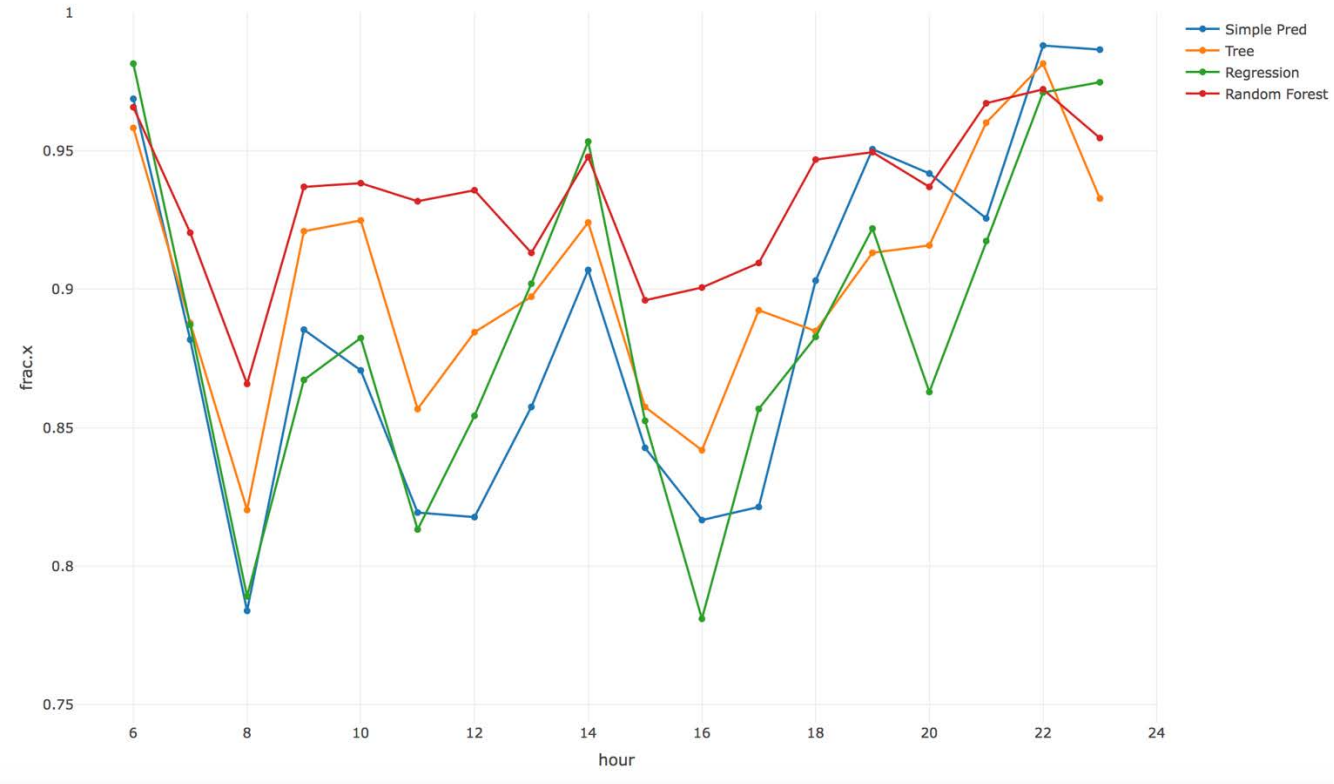

Figure 4 : Comparaison des performances de prédiction pour cinq lignes réunies, avec horizons de dix à quinze minutes : les lignes A, B, C, D et E

La prédiction simple (en bleu) n'est pas très bonne, surtout en périodes de pointe du matin (8h), de pointe du soir (16/17h) et du midi (11/12h). Cela fait du sens puisqu'on s'attend à ce que les périodes de pointe entraînent plus de difficultés à prédire les retards. La régression linéaire (en vert)

n'apporte pas de différence significative globale par rapport à la prédiction simple. En revanche, l'arbre de décision (en orange) est lui meilleur à toutes les heures de la journée, excepté de 18h à 20h puis en toute fin de soirée. Et surtout, c'est la courbe du *random forest* (en rouge) qui est intéressante. Dans cet exemple, elle a toujours une ordonnée bien supérieure à celle de la prédiction simple. À part à 8h (donc pour les prédictions effectuées entre 8h00 et 8h59), la fraction de $\Delta R2$ dans [-1 ; +3] est toujours supérieure ou égale à 90% (seulement légèrement en-dessous de 0,9 pour 15h). Le comportement de la courbe de *random forest* suit des variations assez similaires avec celle de l'arbre de décision, mais en améliorant toujours la performance. Cela confirme l'idée que les *random forests* sont en quelque sorte des « arbres de décisions améliorés ». En effet, on rappelle qu'ils sont construits à partir de plusieurs arbres de décision combinés, dix dans notre cas comme expliqué en 3.3.1. Enfin, puisque cette analyse est effectuée sur la totalité de cinq lignes, c'est-à-dire sur un grand échantillon de données, on peut supposer que la meilleure performance des forêts aléatoires est un phénomène globalisé et récurrent.

L'analyse des résultats est également réalisée par ligne, direction et heure de la journée. Voici en premier exemple les résultats pour la ligne C, direction 1, par heure de la journée, en semaine :

Figure 5 : Performances des modèles par heure, en semaine, pour la ligne C direction 1

La performance de la prédiction simple connaît une nouvelle fois des creux pendant les heures de pointe (9h et 17h). Pourtant, le *random forest* ne semble pas affecté par ces périodes de perturbation. Et de manière générale, le *random forest* semble encore être très souvent plus performant que la prédiction simple. La figure suivante reprend la même analyse, mais pour la ligne D direction 1 cette fois.

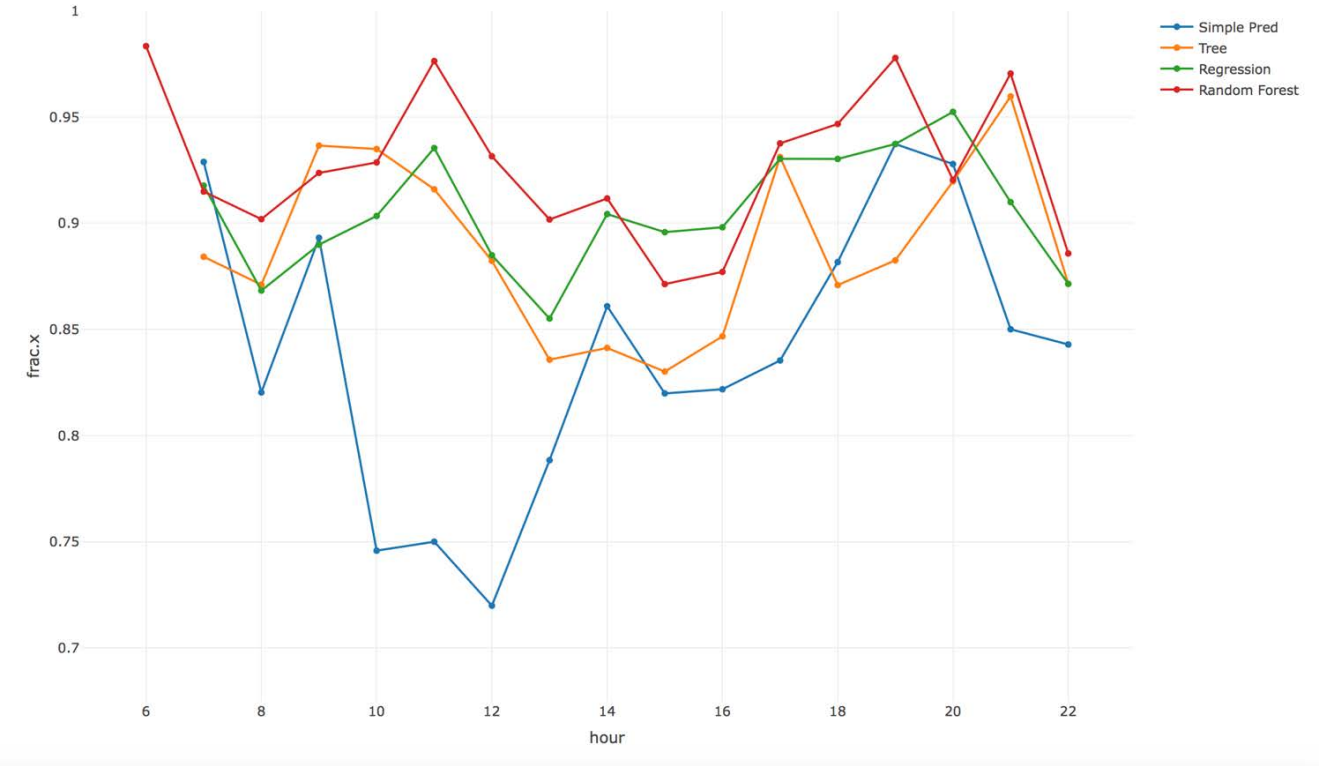

Figure 6 : Performances des modèles par heure, en semaine, pour la ligne D direction 1

On regarde à présent les mêmes performances, mais par journée cette fois. Les lignes C et D sont toujours analysées, ainsi que la A, mais dans les deux directions maintenant. La prédiction est réalisée sur l'échantillon test de 10 jours, correspondant aux deux dernières semaines de notre base de données. Les résultats sont présentés dans les trois figures qui suivent. Le 30 mars semble être un jour particulier. Cela s'explique par deux éléments : c'est un jour férié début d'un long weekend d'une part (Vendredi Saint), et il coïncide avec une forte tempête de neige d'autre part. Pour la quasi-totalité des jours, le *random forest* est encore meilleur que la prédiction simple (courbe rouge contre courbe bleue).

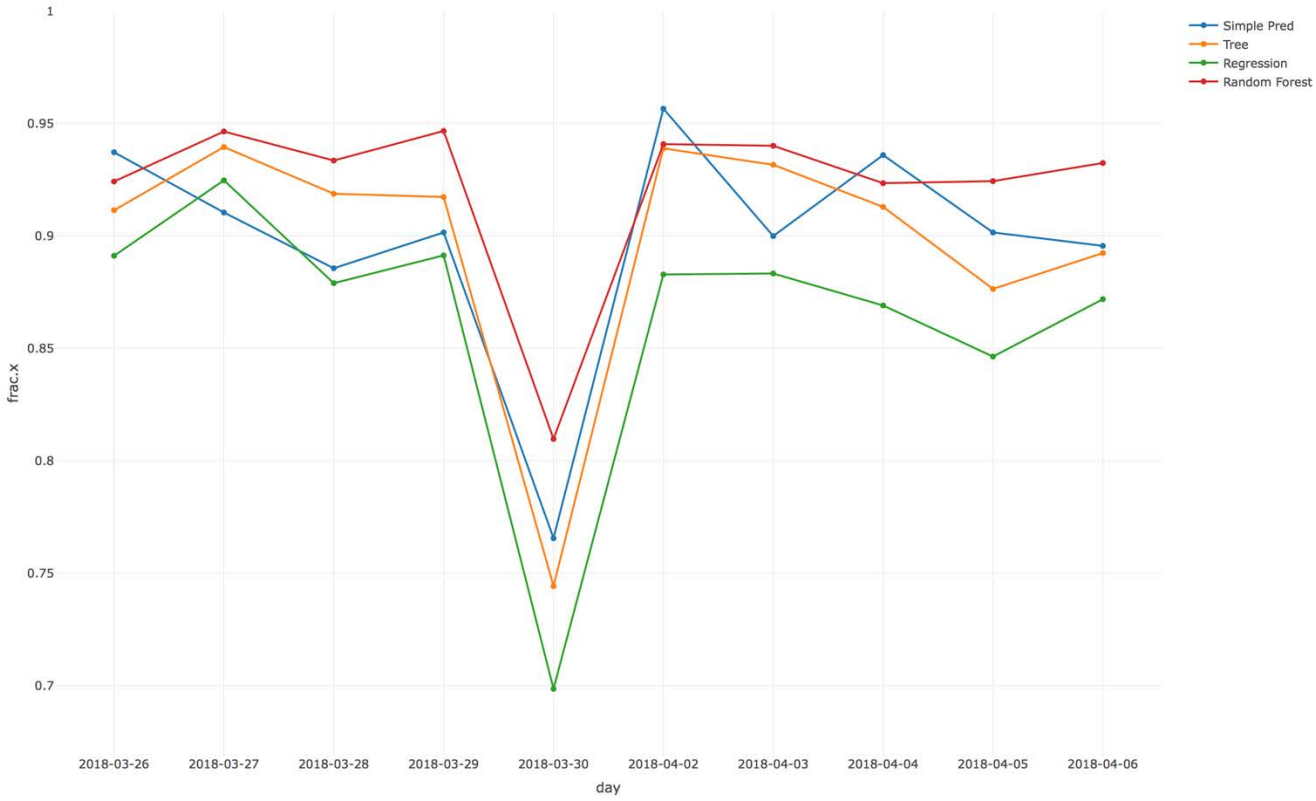

Figure 7 : Performance des modèles par jour, en semaine, pour la ligne C

Encore une fois, ces observations semblent se généraliser à l'ensemble du réseau, puisque les comportements sont globalement similaires entre les différentes lignes. Les figures 4.8 et 4.9 suivantes présentent ainsi la même analyse, mais respectivement sur les lignes D et A cette fois. Les tendances sont les mêmes : le 30 mars est problématique et le *random forest* est le plus performant.

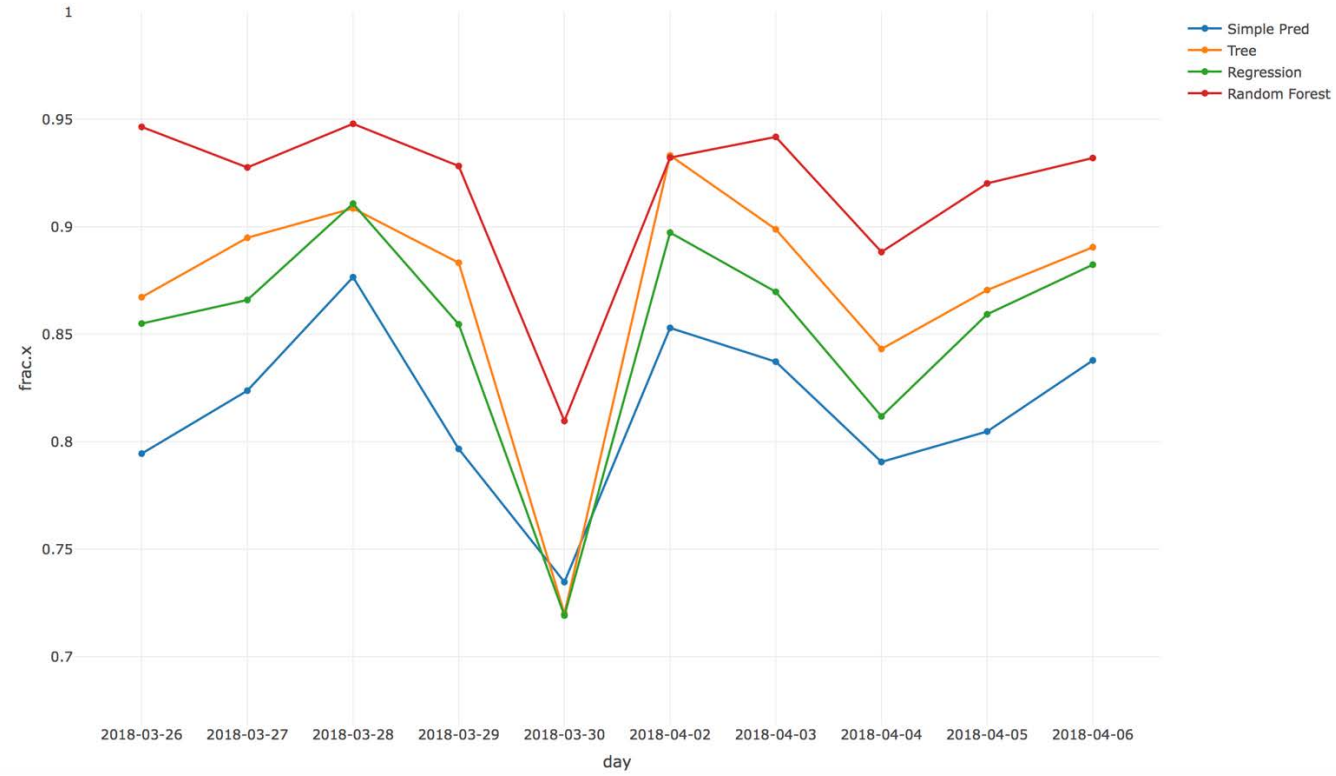

Figure 8 : Performance des modèles par jour, en semaine, pour la ligne D

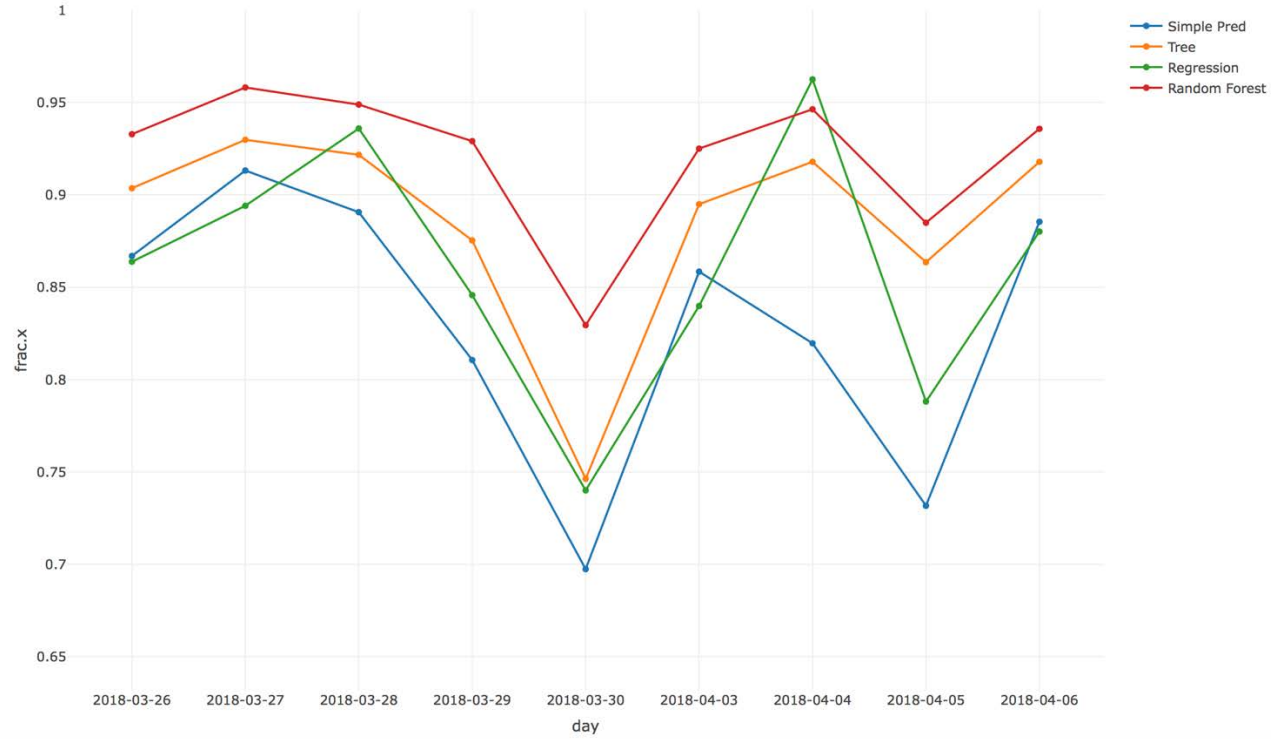

Figure 9 : Performance des modèles par jour, en semaine, pour la ligne A

L'étude focalisée sur ce jour problématique du 30 mars propose des observations à souligner.

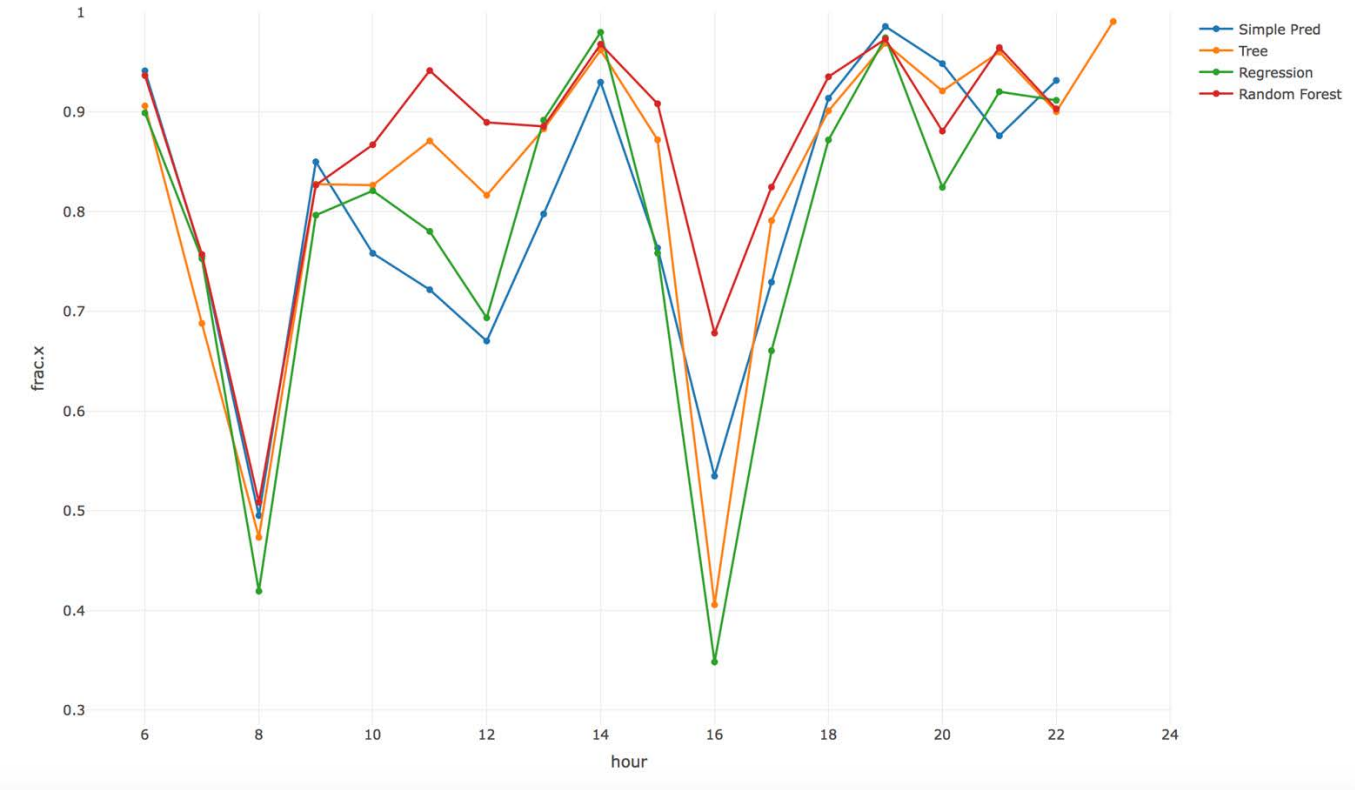

Figure 10 : Étude détaillée du 30 mars, pour les cinq lignes combinées (A, B, C, D, E)

En ce jour où les conditions exceptionnelles ont fait que les prédictions ont été peu précises, ce graphe 4.10 montre que ce sont les pointes du matin (8h) et du soir (16h) qui ont été violemment touchées, et ce pour toutes les méthodes de prédiction. La superposition des deux très fortes contraintes que sont les pointes et les événements spéciaux ont fait chuter les performances. La prédiction simple y a une fraction dans [-1 ; +3] d'environ 50%. La régression et l'arbre de décision sont même pires. Pourtant le *random forest* réagit là encore plus favorablement en atténuant la chute. Il est équivalent à la prédiction simple en pointe du matin, et il est meilleur en pointe du soir avec un pourcentage de performance frôlant les 70%.

Enfin, l'attention est portée sur le weekend. Avec le modèle d'entraînement caractéristique au weekend qui est utilisé, on retrouve aussi les mêmes constats, même si seulement six journées d'entraînement et quatre de test sont à disposition. Augmenter la taille des échantillons donnerait plus de certitude. La figure 4.11 présente les performances de la ligne C, par jour de weekend.

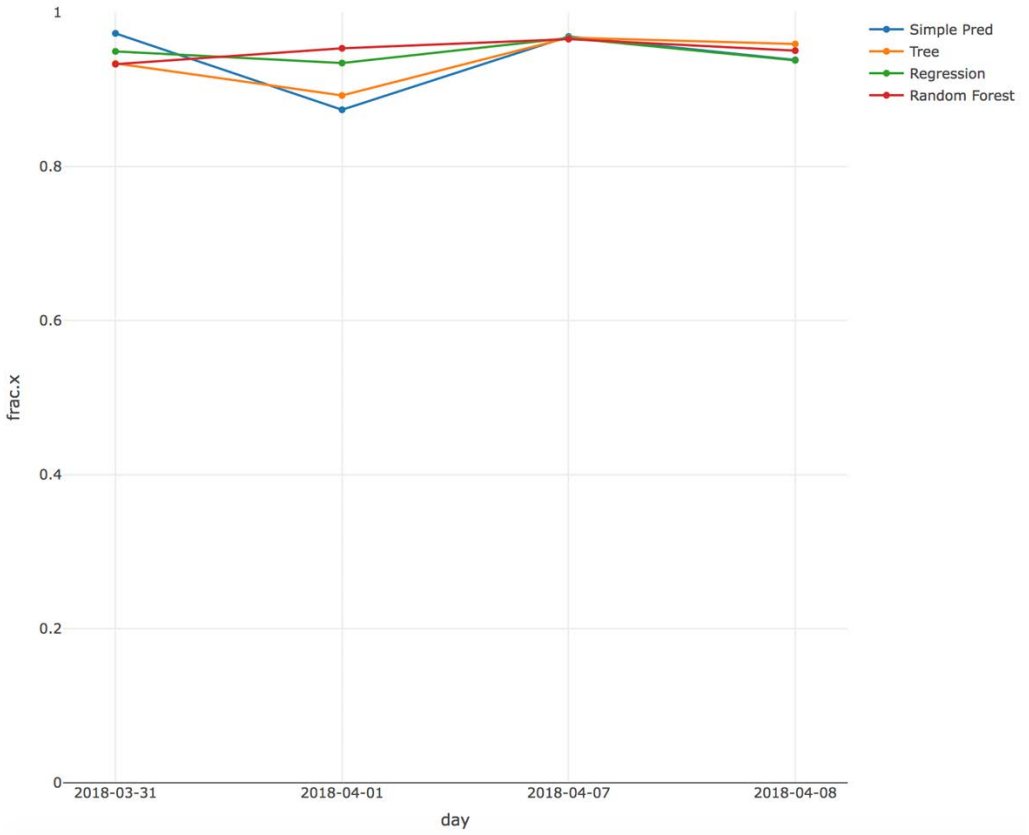

Figure 11 : Performance des modèles par jour, en weekend, pour la ligne C

Ainsi, le modèle du *random forest* semble adapté à la problématique de l'étude puisqu'il améliore le modèle de prédiction simple, dans une grande majorité des cas. De plus, il est stable et n'est pas excessivement chronophage en termes de temps de traitement, ce qui est primordial en vue de l'implantation opérationnelle visée.

C'est pourquoi cette méthode a été sélectionnée pour la construction du modèle. La fonction d'implantation qui l'utilise sera présentée plus en détail en partie 4.2.

4.1.4 Réseaux de neurones

La dernière méthode d'apprentissage machine envisagée pour améliorer la prédiction simple est celle des réseaux de neurones. La revue de littérature, au premier chapitre, a montré que les réseaux sont souvent considérés dans la théorie car ils engendrent de bons résultats. Cependant, ils ont deux principaux inconvénients. Le premier est l'instabilité. Il est difficile d'avoir de la reproductibilité dans les résultats. Même avec un échantillon d'entraînement fixe, des résultats de prédiction sensiblement différents peuvent être obtenus pour le même échantillon test. Le second est la complexité, déjà soulignée en section 3.2.3. De très nombreux paramètres peuvent être modifiés, le temps de traitement est long, est le réseau de neurones est souvent considéré comme une « boîte noire ». En effet, il donne des résultats comme par magie sans qu'on sache réellement comment il opère, contrairement aux régressions, arbres de décision et *random forests* où les étapes de traitement sont plus compréhensibles.

Dans le graphe 4.12 ci-dessous, on peut voir les performances des différentes prédictions pour tous les voyages de la ligne A, avec des horizons compris entre dix et quinze minutes. Les cinq méthodes d'apprentissage machine de l'étude y sont superposées. Les réseaux de neurones, en violet, procurent ici la performance qui semble la moins bonne, après avoir utilisé les paramètres par défaut du package « neuralnet ». Néanmoins, cela est difficilement interprétable puisque si l'on réitère la simulation, les performances obtenues ne seront pas forcément les mêmes du fait de l'instabilité de la méthode. Devant ces inconvénients rédhibitoires à une mise en opération robuste et flexible à court terme, cette méthode des réseaux de neurones est laissée de côté pour le moment. Une meilleure compréhension de ses paramètres pourrait peut-être permettre dans le futur de parvenir à la maîtriser plus efficacement.

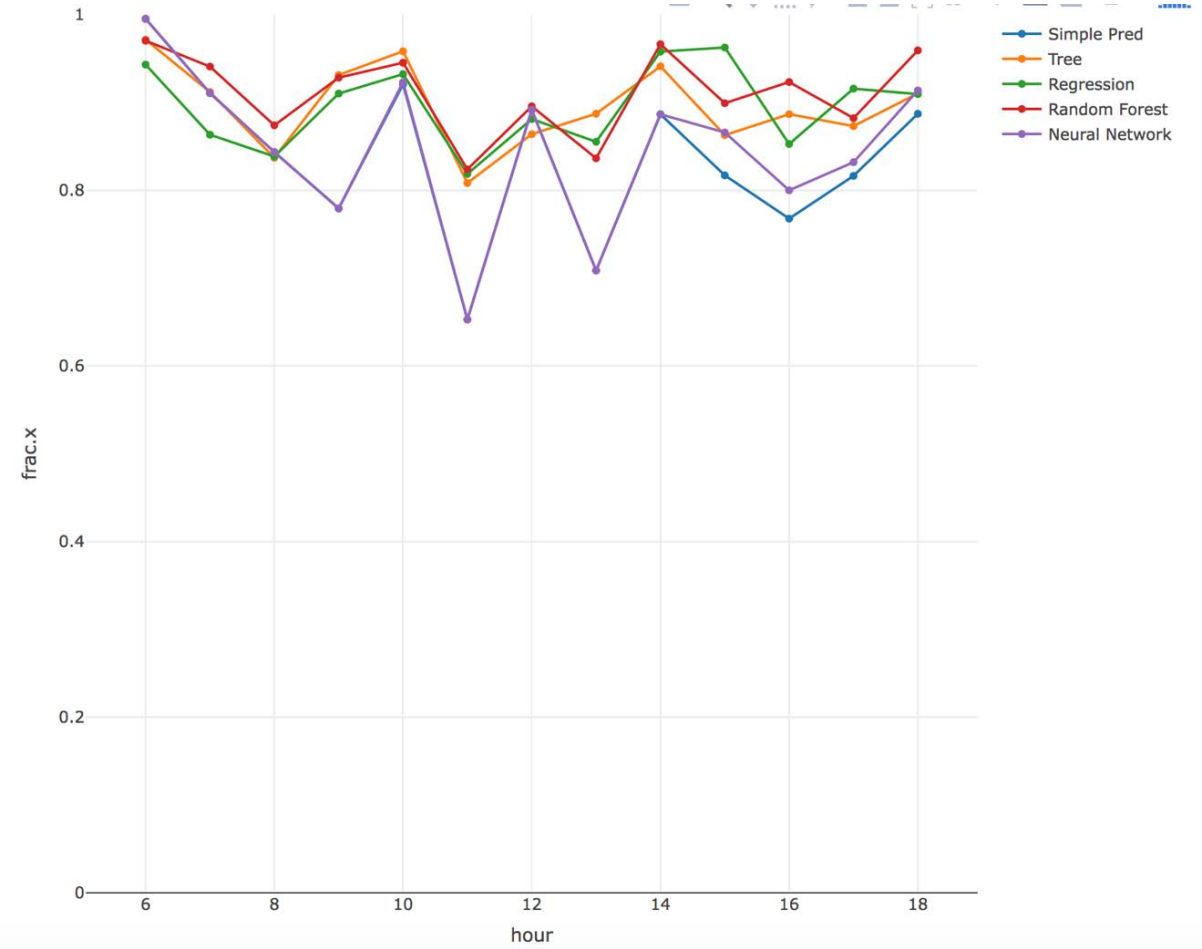

Figure 12 : Comparaison des performances des cinq méthodes étudiées, pour la ligne A, avec des horizons de dix à quinze minutes

En fin de compte, c'est donc bien la méthode de *random forest* qui est privilégiée, au vu des conclusions obtenues à partir des résultats exposés dans cette partie.

L'étape suivante est de procéder à l'implantation d'une fonction algorithmique efficiente, utilisant cette technique de forêt aléatoire, dans le système opérationnel de la Société de Transport de Sherbrooke. C'est en effet ce but qui motive l'étude depuis le début.

4.2 Implantation du modèle en conditions opérationnelles

On rappelle que le but principal poursuivi est d'obtenir une solution opérationnelle qui fonctionne en temps réel. La STS doit pouvoir l'utiliser pour informer les usagers à tout moment sur l'état de circulation des autobus. C'est pourquoi l'étape suivante est d'implémenter une fonction globale simple et facilement manipulable qui englobe tout le travail effectué précédemment.

4.2.1 Fonction implémentée

La fonction implémentée prend en entrée les arguments suivants :

- les données SAE
- les données PRED
- le GTFS
- la table de catégories des tronçons (présentées au chapitre 3)
- les modèles de *random forest* entraînés sur les données d'entraînement (ceux-ci sont entraînés indépendamment en amont pour gagner du temps lors de l'application de la fonction)

En sortie, la fonction créée renvoie un fichier CSV de la forme suivante, avec les fractions ΔR pour la prédiction simple et ΔR2 pour le *random forest*, pour une ligne ou un numéro de trajet au choix, par jour et par heure :

day	hour	frac_simplePred	frac_randomForest
18-03-26	6	1	1
18-03-26	7	1	1
18-03-26	8	0.94	0.98
18-03-26	9	0.93	1
18-03-26	10	1	1
18-03-26	11	0.7	1
18-03-26	12	0.77	0.88
18-03-26	13	1	0.94
18-03-26	14	0.98	1
18-03-26	15	0.86	0.94
18-03-26	16	0.85	0.9
18-03-26	17	0.88	0.85
18-03-26	18	1	1
18-03-26	19	1	0.87
18-03-26	20	1	0.98
18-03-26	21	1	0.99
18-03-26	22	0.97	0.96

Figure 4.13 : Fichier CSV en sortie de la fonction opérationnelle

4.2.2 Temps de process

Le temps que prend la fonction pour tourner est un élément important à considérer puisqu'on veut l'utiliser de manière continue en temps réel. Deux étapes du programme se révèlent être les plus chronophages :

- *l'affectation des catégories à une ligne de données*

Le temps de traitement est proportionnel au nombre de lignes de données. Par exemple, 10 000 lignes de données sont traitées en 6 minutes. Et, pour donner un ordre d'idée, si l'on prend toutes les données pour deux semaines ayant un horizon entre 10 et 15 minutes, on a environ 1 250 000 données. Donc pour tourner sur toutes ces données, cela prend 750 minutes soit 12 heures 30.

- *l'entraînement des modèles de random forests*

La méthode du *random forest* se compose de deux parties : l'entraînement et la prédiction. Si la prédiction est très rapide, la phase d'entraînement prend plus de temps. Surtout, il faut noter que cette phase d'entraînement prend un temps non proportionnel. Voici un exemple, en partant du principe que l'on veuille prédire la direction 1 de la ligne A. Il faut alors choisir sur quel échantillon de données le *random forest* doit être entraîné. Si l'entraînement se fait sur toutes les données de la ligne A, il y a 30 000 données et cela prend 5 secondes. Cependant, s'il se fait sur toutes les données de toutes les lignes, il y a 772 458 données et cela prend 30 minutes. Le choix de l'échantillon affecte donc grandement la rapidité de l'entraînement des *random forests*. Dans tous les cas, il est choisi de sortir cette étape de la fonction, puisque celle-ci peut être faite en amont avec un GTFS donné. Elle sera par la suite réutilisable autant que souhaité dans la fonction.

4.2.3 Pistes d'optimisation

Pour l'affectation des catégories à une ligne de données, chaque donnée est actuellement traitée séparément. Une optimisation peut être faite en regroupant les combinaisons arrêt de visée / arrêt cible identiques. Cependant, pour l'entraînement des modèles du *random forest*, le temps de process peut difficilement être amélioré puisqu'un package prédéfini de R est utilisé, et non une fonction codée manuellement. Le choix de l'échantillon d'entraînement est donc capital pour optimiser le ratio efficacité/rapidité.

CHAPITRE 5 AXES DE RÉFLEXION COMPLÉMENTAIRES

5.1 Segmentation géographique selon le réseau routier

L'idée utilisée dans cette étude a été de caractériser les tronçons d'arrêts en fonction de leur comportement vis-à-vis du retard réel et de la performance de prédiction simple. Cela permet d'être généralisable : si une nouvelle ligne est créée dans le réseau de Sherbrooke, on peut lui appliquer le même modèle sans tout re-paramétrer. De même, si l'on veut exporter le modèle à une autre ville complètement différente, il suffit de recalculer les catégories de tous les tronçons d'arrêts, selon le même principe, pour leur appliquer ensuite la même méthode. Les tronçons sont dans ce cas définis comme des couples de stop_id consécutifs.

Cependant, la méthode de segmentation suivante est également envisageable :

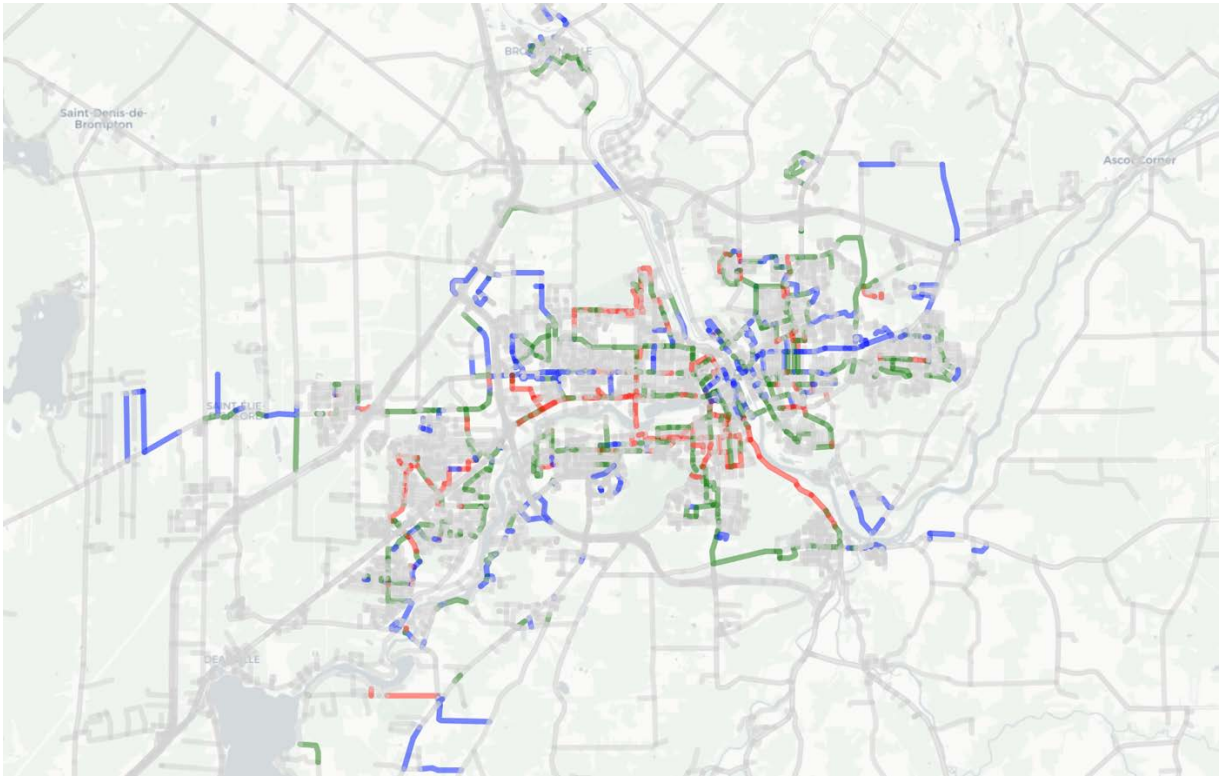

Figure 5.1 : Caractérisation du réseau routier de Sherbrooke en fonction du retard moyen

Ici, on utilise les données de segmentation diffusées par Données Québec pour caractériser le réseau routier géographique de Sherbrooke. Un tronçon y est défini comme un segment de rue entre deux intersections successives. La toile grise de la carte représente le réseau routier de Sherbrooke.

Ensuite, chaque arrêt est affecté à un tronçon. En général, il y a une paire d'arrêts par tronçon (un arrêt pour chaque direction de ligne). Si plusieurs lignes différentes empruntent le même tronçon, les données de toutes les lignes sont prises en compte et une moyenne est effectuée. Puis une caractérisation graphique de ces tronçons peut être faite selon la moyenne de la variable de notre choix. Dans le premier exemple ci-dessus, on fait la moyenne du retard réel pour chaque tronçon. Si le tronçon est gris, cela signifie qu'il ne comporte pas d'arrêt du réseau d'autobus de Sherbrooke. S'il est bleu, le retard moyen réel est strictement inférieur à 1 minute. S'il est vert, il est compris entre 1 et 2 minutes. Enfin, s'il est rouge, il est supérieur ou égal à 2 minutes. Une vision très générale est obtenue mais ces données peuvent ensuite être adaptées en ciblant plus précisément certains horizons, certains jours ou certaines périodes de la journée.

Et cette même carte peut être dressée pour la variable d'intérêt : ΔRmoyen.

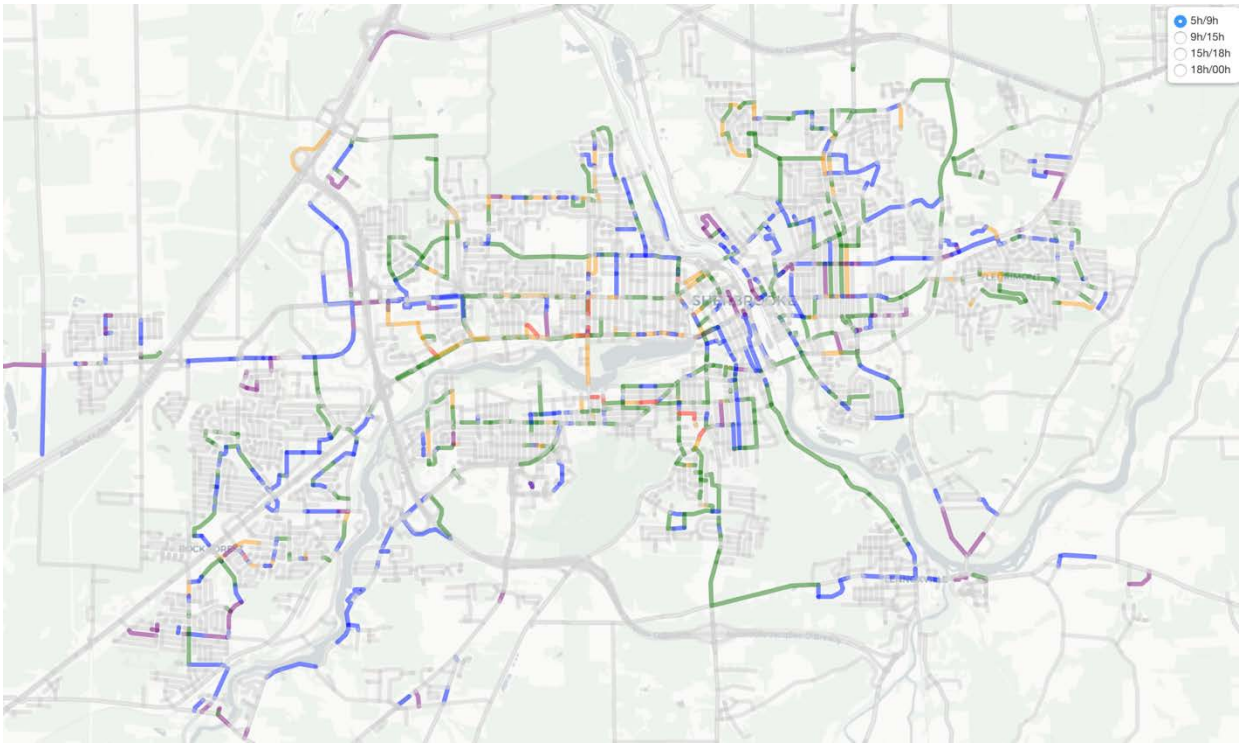

Figure 5.2 : Caractérisation du réseau routier de Sherbrooke en fonction du ΔRmoyen, pour les mardis/mercredis/jeudis, en période de pointe matinale (5h/9h)

Les mardis/mercredis/jeudis sont ciblés ici. Ils ont des comportements semblables entre eux, même s'ils peuvent différer des lundis et vendredis (du fait des ponts et des weekends prolongés affectant ces derniers). De plus, on a choisi de segmenter par période de la journée.

Les divisions par périodes habituelles sont affichées dans la fenêtre en haut à droite de l'image. En cliquant sur la période souhaitée, l'aperçu graphique correspondant apparaît. Le code couleur est le suivant selon les valeurs du ΔRmoyen : violet pour une valeur entre -1 et 0, bleu entre 0 et 1, vert entre 1 et 2, orange entre 2 et 3 et rouge en dehors de l'intervalle [-1 ; 3]. De manière visuelle, on peut donc repérer les endroits où la prédiction est mauvaise, c'est-à-dire les zones rouges où le ΔRmoyen est en dehors de l'intervalle d'acceptation [-1 ; +3]. Ces erreurs semblent se produire principalement sur les grands axes. La rue King, le boulevard Jacques-Cartier et le boulevard de Portland sont notamment problématiques.

Cette méthode semble intéressante et intuitive. Cependant, elle a deux principaux défauts en l'état. Elle ne permet pas de faire la distinction entre deux directions différentes pour un même tronçon. Si, par exemple, il y a un ralentissement dans un sens de circulation mais que l'autre sens reste fluide, il ne sera pas possible de faire ressentir le contraste dans les prédictions. De plus, un autre inconvénient est que cette approche est dépendante au réseau routier de la ville de Sherbrooke. Il n'est pas envisageable d'appliquer le même modèle à une autre ville puisque le nouveau réseau routier sera complètement différent et ne sera pas forcément accessible, comme ici grâce à Données Québec.

5.2 Conditions météorologiques

Un facteur important devant être considéré dans l'étude est la météo. On suppose que les conditions météorologiques peuvent être un critère d'influence important sur la circulation. Cela est d'autant plus vrai au Québec, où les conditions sont très variables, notamment en période hivernale où la neige impacte à priori de façon non négligeable la fiabilité de la performance opérationnelle. Après réflexion, il est supposé que la neige, la glace et le brouillard sont parmi les plus impactants. On cherche à savoir si l'on peut améliorer le modèle en se servant de ces variables climatiques. Pour tenter de vérifier ces hypothèses, les données météorologiques fournies par le gouvernement du Québec sont utilisées. Elles prennent la forme suivante :

	Date	Hour	Temperature	Point_de_Rosee	Humidite_Rel	Dir_Vent	Vit_Vent	Visibilite	Pressure	TPRH	Temps
1	2018-03-01	1	2.5	1.7	94	NA	5	16.1	97.81	2.48	ND
2	2018-03-01	2	3.6	2.0	89	24	11	16.1	97.85	3.58	Pluie
3	2018-03-01	3	1.2	0.4	94	29	26	8.1	97.94	1.20	Brouillard
4	2018-03-01	4	0.1	0.0	99	28	11	1.0	98.40	0.10	Brouillard
5	2018-03-01	5	0.3	0.4	99	28	21	2.8	98.13	0.30	Brouillard
6	2018-03-01	6	0.8	-1.1	98	28	18	2.4	98.21	0.81	Brouillard

Figure 5.3 : Structure des données météorologiques disponibles

Pour chaque journée, une ligne de données par heure est disponible. Les valeurs correspondantes peuvent donc être affectées à toutes les prédictions effectuées durant la même heure de la même journée. On considère que le vent et la pression atmosphérique n'ont pas d'impact majeur, tandis que la visibilité est liée au point de rosée et à l'humidité. En effet, le point de rosée est la température la plus basse, pour un couple (pression; humidité) donné, à partir de laquelle la vapeur d'eau de l'air se condense en eau liquide. Les variables de température, point de rosée, et humidité relative sont considérées comme les plus importantes. Plutôt que d'avoir trois variables, il est choisi de les fusionner en une seule : *le « TPRH »,* défini selon *|Température – Point de Rosée| / Humidité relative*. On part du principe que, plus l'humidité est élevée, plus la visibilité est faible et plus la chaussée est glissante. Il en va de même lorsque la température est proche du point de rosée (température de rosée), donc lorsque la valeur du numérateur du TPRH est faible. Ainsi, il est supposé que plus le TPRH est faible, plus les conditions climatiques nuisent à la fiabilité du service.

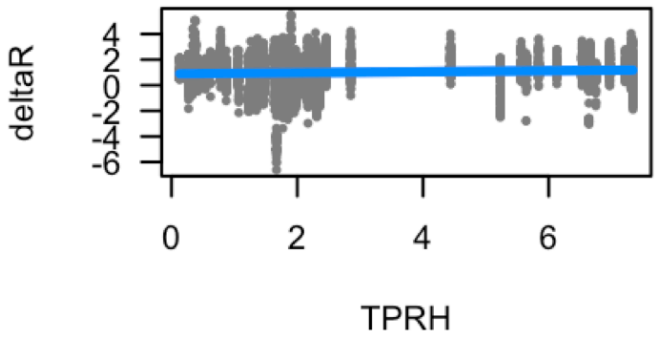

Figure 5.4 : Corrélation entre ΔR et TPRH, pour la ligne F

Malheureusement, les résultats de corrélation ne sont pas très concluants : les valeurs du TPRH semblent échelonnées. Cela est dû au fait que les données météorologiques accessibles ne sont pas très précises, ne sont disponibles seulement qu'heure par heure, et prennent souvent les mêmes valeurs. Dans l'état actuel, il est donc choisi de mettre de côté l'étude du TPRH. La focalisation de l'étude sera alors recentrée sur la variable catégorielle « Temps » (dernière colonne de la figure 5.3). C'est une variable, déterminée par le gouvernement du Québec, qui annonce le type de temps qu'il fait. On divise celle-ci en cinq catégories : Temps « normal » (0), Glace (1), Neige (2), Brouillard (3) et Pluie (4). Cette variable sera appelée « TempsNum ».

Le graphique suivant présente la distribution du ΔR en fonction de ces cinq catégories, sous forme de diagramme en « boîte à moustaches ».

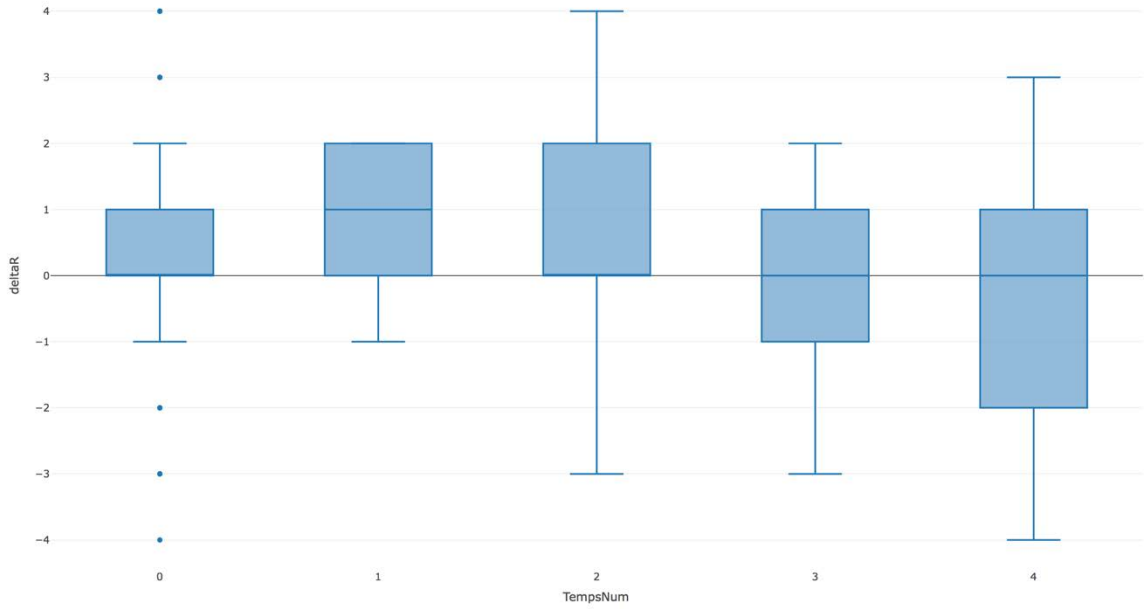

Figure 5.5 : Diagramme "boîte à moustaches" de ΔR en fonction du temps météorologique

Il se voit qu'il existe des différences entre les catégories. Les précipitations (pluie et neige) entraînent une plus forte variabilité de ΔR. Pour le temps « normal », il n'y a pas de comportement particulier, ce qui semble logique. Pour la glace enfin, ΔR est en moyenne positif. Cela signifie que le retard réel est supérieur au retard prédit, donc que le chauffeur accumule du retard. Cela se comprend puisque, si la route est gelée, le chauffeur va ralentir par sécurité. Ainsi, cette variable est intéressante mais elle possède deux principales limites qui font qu'elle n'a pas été gardée dans le modèle final. D'abord, elle dépend de la méthode d'affectation du gouvernement du Québec qui n'est pas connue. Mais surtout, cette information n'est pas accessible en temps réel, et ne peut donc pas être utilisée dans la fonction opérationnelle actuelle de l'étude. La prise en considération de la météo est néanmoins à garder en tête pour améliorer le modèle dans le futur.

5.3 Données d’achalandage

Les données d’achalandage en temps réel seraient également intéressantes à appréhender, telles que le nombre de montants/descendants à chaque arrêt. La prise en compte du temps arrêté en station à cause des flux de voyageurs est certainement influente sur les retards, donc attrayante pour l’étude, mais les données à ce sujet ne sont pas disponibles présentement pour la ville de Sherbrooke.

CHAPITRE 6 CONCLUSION ET RECOMMANDATIONS

Au cours de ce projet, plusieurs pistes auront été explorées pour chercher à établir un modèle de prédiction des retards des autobus urbains de la ville de Sherbrooke, au Québec.

Après une phase de découverte et de mise en forme des données brutes fournies par la Société de Transport de Sherbrooke, les variables explicatives les plus adéquates ont été recherchées, dans le but de les utiliser en entrée du modèle. Plusieurs méthodes d'apprentissage machine ont ensuite été essayées, pour finalement choisir celle de la forêt aléatoire *(random forest)* qui paraît être la plus robuste. Certains axes de réflexion supplémentaires ont été étudiés puis écartés pour servir le but premier : la mise en opération de la méthode et l'application en temps réel par la STS. La fonction du modèle a été créée de sorte à ce qu'elle soit la plus pratique et flexible possible. Elle peut ainsi être adaptée à toute modification du réseau de transport en commun de Sherbrooke, voire même à une ville complètement différente. Néanmoins, le modèle a besoin d'être entraîné, ce qui est limitant car chronophage, surtout si l'on souhaite le ré-entraîner régulièrement. La fonction créée est présentement en cours d'implantation au sein de la Société de Transport de Sherbrooke, et les résultats effectifs sont attendus. Si les conclusions satisfont la STS, l'étape de transition entre le système de prédiction actuel et celui proposé dans cette étude sera un virage délicat à négocier.

Il est tout à fait recommandé que ce travail soit poursuivi car de nombreux défis peuvent toujours être attaqués. Le modèle est encore améliorable avec de nouvelles variables et méthodes, dont certaines sont déjà évoquées dans ce rapport. D'autres variables à considérer pour approfondir peuvent être par exemple : les données de circulation pour tenir compte des congestions en temps réel, les données d'achalandage pour caractériser les temps de montée/descente aux arrêts (à partir des cartes à puces dans les villes qui en sont équipées, par exemple), ou encore les données d'événements spéciaux pouvant influer temporairement et localement sur le trafic des autobus (travaux, feux tricolores en panne, festivals, etc …). Enfin, cette thématique peut être renforcée grâce à des alliances avec d'autres groupes de recherche, dont un très actif en Floride, dans le but de développer ce projet à plus grande échelle et à l'international.

LISTE DES RÉFÉRENCES

Amita, Johar, Jain Sukhvir Singh, et Garg Pradeep Kumar. « Prediction of bus travel time using artificial neural network ». *International Journal for Traffic and Transport Engineering 5*, n° 4 (décembre 2015): 410-24. https://doi.org/10.7708/ijtte.2015.5(4).06.

Jiang, Fan. « Bus Transit Time Prediction Using GPS Data with Artificial Neural Networks », s. d., 27.

Jithendra. H. K, Naga Ravi Kanth Devarapalli, et M.S.R.I.T. « Predicting Bus Arrival Time Based on Traffic Modelling and Real-Time Delay ». *International Journal of Engineering Research And* V4, n° 06 (17 juin 2015). https://doi.org/10.17577/IJERTV4IS060512.

Ma, Xiaoliang, Fadi Al Khoury, et Junchen Jin. « Prediction of Arterial Travel Time Considering Delay in Vehicle Re-Identification ». *Transportation Research Procedia* 22 (2017): 625-34. https://doi.org/10.1016/j.trpro.2017.03.056.

Sun, Fangzhou, Yao Pan, Jules White, et Abhishek Dubey. « Real-Time and Predictive Analytics for Smart Public Transportation Decision Support System ». In *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 1-8. St Louis, MO, USA: IEEE, 2016. https://doi.org/10.1109/SMARTCOMP.2016.7501714.

Yin, Tingting, Gang Zhong, Jian Zhang, Shanglu He, et Bin Ran. « A Prediction Model of Bus Arrival Time at Stops with Multi-Routes ». *Transportation Research Procedia* 25 (2017): 4623-36. https://doi.org/10.1016/j.trpro.2017.05

Plan stratégique organisationnel 2025, publié par la Société de Transport de Montréal et adopté par son conseil d'administration le 8 juin 2017. https://www.stm.info/sites/default/files/pso-2025.pdf