Prévision et visualisation de l'affluence dans les transports en commun à l'aide de méthodes d'apprentissage automatique

Florian Toqué

UNIVERSITÉ PARIS-EST

Ecole doctorale MSTIC
Doctorat en codirection
Spécialité : Informatique

Apprentissage automatique appliqué

Prévision et visualisation de l'affluence dans les transports en commun à l'aide de méthodes d'apprentissage automatique

Florian TOQUÉ

Rapporteurs Mounim EL YACOUBI, Professeur, Télécom SudParis
Stéphane BONNEVAY, Maître de conférences - HDR, Université Claude Bernard Lyon 1

Examinateurs Boris MERICSKAY, Maître de conférences, Université Rennes 2
Cristina PRONELLO, Professeure, Sorbonne Universités-UTC

Encadrant Etienne CÔME, Chargé de recherche, IFSTTAR

Directrice Latifa OUKHELLOU, Directrice de recherche, IFSTTAR
Co-Directeur Martin TRÉPANIER, Professeur, Polytechnique Montréal

2 décembre 2019

Florian TOQUE

Prévision et visualisation de l'affluence dans les transports en commun à l'aide de méthodes d'apprentissage automatique

Doctorat en codirection, Informatique, Apprentissage automatique appliqué

Rapporteurs: Mounim EL YACOUBI et Stéphane BONNEVAY

Examinateurs: Boris MERICSKAY et Cristina PRONELLO

Encadrant: Etienne COME

Directeurs: Latifa OUKHELLOU et Martin TREPANIER

UNIVERSITÉ PARIS-EST

Ecole doctorale: MSTIC

Institut de recherche : IFSTTAR

Département/Laboratoire : COSYS/GRETTIA

14-20 Boulevard Newton, 77420 Champs-sur-Marne

Date de soutenance : 2 décembre 2019

Remerciements

Cette thèse s'est déroulée au sein du laboratoire GRETTIA de L'IFSTTAR à Champs-sur-Marne en France et du laboratoire CIRRELT à Montréal au Canada. Sans le soutien et la collaboration de certaines personnes, cette thèse n'aurait pas pu aboutir. Ainsi je tiens à remercier toutes celles et ceux qui ont contribué de près ou de loin à l'avancement de mon travail.

J'ai eu la chance d'être encadré et dirigé par trois personnes d'exception, douées d'une grande intelligence, gentillesse et bienveillance grâce auxquelles j'ai pu m'épanouir et réaliser cette thèse dans les meilleures conditions. Je tiens ainsi à exprimer ma profonde reconnaissance et gratitude à ma directrice de thèse Latifa Oukhellou et mon encadrant Etienne Côme pour leur aide précieuse et leurs nombreux conseils prodigués depuis mon arrivée en stage à l'IFSTTAR jusqu'à la fin de mon doctorat. Je remercie par la même occasion mon directeur de thèse Martin Trépanier pour m'avoir accepté en échange d'un an à Montréal, m'ayant ainsi permis de profiter de l'année la plus enrichissante de ma vie, aussi bien d'un point de vue professionnel que personnel.

Je suis vivement reconnaissant envers la Professeure Cristina Pronello qui m'a fait l'honneur d'être présidente du jury. Mes vifs remerciements sont aussi adressés au Professeur Mounim El Yacoubi et au Maître de conférences Stéphane Bonnevay pour avoir accepté de rapporter mon travail. Je remercie également le Maître de conférences Boris Mericskay qui a accepté d'examiner ce travail et a bien voulu participer au jury de thèse.

Je remercie l'IFSTTAR, Polytechnique Montréal et Thales pour m'avoir accordé le financement nécessaire à la réalisation de cette thèse ainsi que l'université Paris-Est pour la bourse de mobilité internationale octroyée pour mon échange à Montréal.

Je tiens à remercier mes professeurs de l'université Pierre et Marie Curie pour m'avoir donné goût à l'informatique. Je remercie plus particulièrement les directeurs du Master DAC, Ludovic Denoyer et Bernd Amann pour leurs enseignements qui m'ont donné envie de continuer dans le domaine de la Data Science et du Machine Learning. Je remercie également Vincent Guigue pour la qualité de ses enseignements et sa grande capacité à expliquer les concepts clairement qui m'ont permis de fortement apprécier cette branche de l'informatique.

Je tiens à exprimer toute ma gratitude à mes collègues et amis du GRETTIA que j'ai eu la chance de côtoyer tout au long de cette thèse, Anne-Sarah, Moncef, Milad, Khalil, Negin, Mostafa, Florence, ...

Être loin de sa famille et de ses amis peut parfois être difficile, j'adresse donc mes remerciements aux amis que j'ai eu la chance de rencontrer lors de mon échange à Montréal, Mehdi, Marc, Aurélien, Greta, Stéphanie, Julien, Thomas, ... Ainsi qu'à mes amis rencontrés aux cours d'improvisation théâtrale, Jennifer, Dereck, Romane, ...

Enfin, je remercie ma famille et mes amis pour leur soutien indéfectible. Je souhaite que mes parents María et Johann ainsi que mes grands-mères Delfina et Ginette puissent trouver dans la réalisation de cette thèse le fruit de leur éducation, de leur dévouement et de leurs sacrifices. Je tiens également à remercier chaleureusement mes petits frères Yann et Nicolas pour m'avoir soutenu et donné l'envie de montrer l'exemple.

Résumé

Dans le cadre de la lutte contre le réchauffement climatique, plusieurs pays du monde notamment le Canada et certains pays européens dont la France, ont établi des mesures afin de réduire les nuisances environnementales. L'un des axes majeurs abordés par les états concerne le secteur du transport et plus particulièrement le développement des systèmes de transport en commun en vue de réduire l'utilisation de la voiture personnelle et les émissions de gaz à effet de serre. A cette fin, les collectivités concernées visent à mettre en place des systèmes de transports urbains plus accessibles, propres et durables. Dans ce contexte, cette thèse en codirection entre l'Université Paris-Est, l'Institut français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR) et Polytechnique Montréal au Canada, s'attache à analyser la mobilité urbaine au travers de recherches menées sur la prévision et la visualisation de l'affluence des passagers dans les transports en commun à l'aide de méthodes d'apprentissage automatique. Les motivations finales concernent l'amélioration des services de transport proposés aux usagers, tels qu'une meilleure planification de l'offre de transport et une amélioration de l'information voyageur (e.g., proposition d'itinéraire en cas d'événement/incident, information concernant le taux de remplissage des trains à un horaire choisi, etc.). Cette thèse s'inscrit dans un contexte général de valorisation des traces numériques et d'essor du domaine de la science des données (e.g., collecte et stockage des données, développement de méthodes d'apprentissage automatique, etc.). Nous avons notamment comparé différents modèles de prévision tels que des modèles basiques, statistiques, issus de l'apprentissage automatique (Support Vector Regressor, Random Forest, Gradient Boosting, etc.) et issus de l'apprentissage profond (Long-Short Term Memory et Gated Recurrent Unit). Les travaux comportent trois volets principaux à savoir (i) la prévision long terme de l'affluence des passagers à l'aide de données événementielles et de données billettiques, (ii) la prévision court terme de l'affluence des passagers et (iii) la visualisation de l'affluence des passagers dans les transports en commun. Les recherches se basent principalement sur l'utilisation de données billettiques fournies par les opérateurs de transport et ont été menées sur trois cas d'études réels, le réseau de métro et de bus de la ville de Rennes, le réseau ferré et de tramway du quartier d'affaire de la Défense à Paris en France, et le réseau de métro de Montréal, Québec au Canada.

Mots clés : Apprentissage automatique, Apprentissage profond, Transports en commun, Données billettiques, Prévision, Visualisation

Abstract (Résumé en anglais)

As part of the fight against global warming, several countries around the world, including Canada and some European countries, including France, have established measures to reduce greenhouse gas emissions. One of the major areas addressed by the states concerns the transport sector and more particularly the development of public transport to reduce the use of private cars. To this end, the local authorities concerned aim to establish more accessible, clean and sustainable urban transport systems. In this context, this thesis, co-directed by the University of Paris-Est, the french institute of science and technology for transport, development and network (IFSTTAR) and Polytechnique Montréal in Canada, focuses on the analysis of urban mobility through research conducted on the forecasting and visualization of public transport ridership using machine learning methods. The main motivations concern the improvement of transport services offered to passengers such as: better planning of transport supply, improvement of passenger information (e.g., proposed itinerary in the case of an event/incident, information about the crowd in the train at a chosen time, etc.). In order to improve transport operators' knowledge of user travel in urban areas, we are taking advantage of the development of data science (e.g., data collection, development of machine learning methods). In particular, we compared different forecasting models such as basic models, statistical models, machine learning models (Support Vector Regressor, Random Forest, Gradient Boosting, etc.) and deep learning models (recurrent neural networks, namely Long-Short Term Memory and Gated Recurrent Unit). This thesis thus focuses on three main parts: (i) long-term forecasting of passenger demand using event databases, (ii) short-term forecasting of passenger demand and (iii) visualization of passenger demand on public transport. The research is mainly based on the use of ticketing data provided by transport operators and was carried out on three real case study, the metro and bus network of the city of Rennes, the rail and tramway network of "La Défense" business district in Paris, France, and the metro network of Montreal, Quebec in Canada.

Thesis title: Forecasting and visualization of passenger demand in public transport network with machine learning methods

Key words: Machine Learning, Deep Learning, Public Transport, Smart Card Data, Forecasting, Visualization

Table des matières

1 Introduction

1.1 Contexte et motivations

Selon le département des affaires économiques et sociales de l'ONU (DESA), 2,5 milliards de personnes pourraient être ajoutées aux zones urbaines d'ici 2050. L'expansion de la population exigera alors des services et une allocation de ressources supplémentaires. Le DESA a déclaré que «De nombreux pays devront relever des défis pour répondre aux besoins de leurs populations urbaines en croissance, y compris pour le logement, les transports, les systèmes énergétiques et autres infrastructures, ainsi que pour l'emploi et les services de base tels que l'éducation et les soins de santé». D'autre part, le Groupement Intergouvernemental d'Experts sur l'Evolution du Climat (GIEC) réunissant 195 états, a réussi dès 2007 dans un quatrième rapport, à trouver un accord sur les conséquences des changements climatiques. Ce quatrième rapport, confirme le consensus international sur la nécessité d'une action urgente pour limiter les effets du réchauffement climatique fortement lié à l'activité humaine. Dans ce contexte mêlant augmentation de la population dans les zones urbaines et péri-urbaines et augmentation de la production de gaz à effet de serre, il est évident que les pays devront s'adapter et améliorer la planification urbaine ainsi que les services publics de manière plus durable, dans les prochaines années.

Dans cette thèse, nous nous attachons en particulier au secteur du transport qui est une des sources principales de l'augmentation des gaz à effet de serre. En effet, d'après une étude menée par Ritchie et al. [RR17] le secteur du transport est le deuxième secteur émetteur de dioxyde de carbone (CO_2) avec 5,53 milliards de tonnes émises en 2010 derrière le secteur de l'énergie émetteur de 20,33 milliards de tonnes. Plus précisément, les camions, bus et voitures représentent 74% des émissions de CO_2 du secteur du transport. Dans ce contexte, cette thèse porte pour principale ambition d'améliorer les services de transport en commun offerts aux usagers dans le but de favoriser leur utilisation au détriment de la voiture personnelle pour les déplacements quotidiens des citoyens. L'objectif ainsi visé est de réduire l'impact environnementale du transport grâce à une réduction globale des gaz à effet de serre émis lors de nos déplacements.

Nous profitons des avancées technologiques issues de l'essor du domaine de la science des données (e.g., collecte de données, développement de méthodes d'apprentissage automatique) afin d'améliorer la connaissance des opérateurs de transport concernant le déplacement des usagers en zone urbaine. A terme, l'objectif est de ressortir des informations utiles à l'amélioration des services de transport en commun tels que l'offre de transport et l'information fournie aux voyageurs concernant leurs moyens de transport. Plus particulièrement, nous nous attachons à l'étude de la prévision et de la visualisation de l'affluence des passagers au travers de l'utilisation de données billettiques et de données exogènes (e.g., données calendaires, données événementielles, données d'incidents) dans les stations de transport en commun en considérant les trois réseaux de transport des villes de Montréal au Canada, Rennes et Paris en France.

1.1.1 Objectifs sociaux, économiques et environnementaux

La création et l'utilisation des transports en commun peut bénéficier aux usagers mais également au développement des villes qu'ils desservent. En effet, selon différentes études détaillées dans les sections suivantes, l'utilisation des transports en commun est un facteur positif quant à l'amélioration de la santé et la réduction des dépenses des usagers concernant leurs déplacements. De plus, il a été montré dans les travaux de [WR09] que le développement des transports en commun entraîne une augmentation de la croissance économique des villes et aide à lutter contre le réchauffement climatique par le biais d'une réduction des émissions de gaz à effet de serre (en comparaison à l'utilisation de la voiture personnelle).

1.1.1.1. Considérations sociales, sociétales et économiques

Selon les auteurs de [NK19], le monde connaît la plus forte croissance urbaine de l'histoire de l'humanité. Plus de 50% de la population mondiale vit dans les villes et ce chiffre devrait atteindre 70% au cours des vingt prochaines années. Ce phénomène implique une augmentation de la quantité de dioxyde de carbone dans les villes ce qui peut entraîner des problèmes de santé et être un frein à la lutte contre le réchauffement climatique. En effet, d'après les auteurs de [Mit+18; Bai+19], une relation existe entre l'augmentation de la population urbaine et la dynamique de la quantité de CO_2 dans les zones urbaines. Dans ce contexte, il incombe aux décisionnaires publics de développer des systèmes de transports urbains durables et efficaces afin de réduire les émissions de gaz à effet de serre et de veiller à l'épanouissement des individus au sein des villes de demain.

D'autre part, les auteurs [Lit12; Lit13; Sae+14] ont étudié l'impact de l'utilisation des transports en commun sur la santé des usagers. Ces études montrent que les personnes vivant dans des endroits avec un accès aux transports publics utilisent moins la voiture personnelle au profit de modes alternatifs tels que la marche, le vélo ou l'utilisation des transports en commun. Ceci réduit la pollution et les accidents de voiture tout en augmentant la santé physique et mentale des usagers. Par ailleurs, cette étude montre également que cela permet à certains individus d'accéder plus facilement à des services vitaux tels que l'accès à la nourriture ainsi qu'à des services de santé.

D'autres aboutissants positifs résultent du développement des systèmes de transports durables. En effet, selon les auteurs de [WR09], les services de transports publics favorisent la création d'emplois et la croissance économique tout en réduisant l'énergie utilisée et les émissions de gaz à effet de serre.

1.1.1.2. Considérations environnementales

Le réchauffement climatique est le phénomène d'augmentation des températures moyennes observées dans le monde (températures océaniques et de l'air), induit par la quantité d'air piégée à la surface terrestre, du fait des émissions de gaz à effet de serre (e.g., dioxyde de carbone (CO_2), méthane (CH_4), protoxyde d'azote ($N2_O$), etc.). Selon une étude menée par Ritchie et al. [RR17] sur les données issues de l'organisation des nations unies pour l'alimentation et l'agriculture (Food and Agriculture Organization of the United Nations, FAO), le transport est le deuxième secteur émetteur de dioxyde de carbone avec 5,53 milliards de tonnes émises en 2010 (les transports ferrés représentent moins de 3% des émissions de CO_2 et les camions, les bus et les voitures représentent 74% du secteur du transport), derrière le secteur de l'énergie (électricité, construction, chauffage) qui a une émission de CO_2 qui s'élève à 20,33 milliards de tonnes. Réduire l'utilisation de la voiture personnelle au profit des transports en commun semble par conséquent, être un processus important pour réduire l'émission globale de CO_2.

Au niveau mondial, plusieurs pays se sont réunis (COP21, G20, G7) afin d'élaborer des accords permettant de lutter contre le réchauffement climatique. On peut notamment parler de l'accord de Paris entré en vigueur le 4 novembre 2016 et mis en oeuvre par les pays participants de la COP21 (195 pays + Union européenne), lors de la Conférence de Paris sur le climat. Concernant la stratégie de l'Union européenne, la Commission européenne fait la promotion des plans de mobilité urbaine durable qui consistent à mettre en place des systèmes de transports urbains multi-modaux,

qui soient accessibles à tous les usagers et qui répondent aux exigences de durabilité en tenant compte des nécessités de viabilité économique, d'équité sociale et de qualité sanitaire et environnementale. Le Canada quant à lui, a également mis en place un plan climatique visant à réduire les émissions (au Canada le transport représente 25% de ses émissions derrière le secteur de l'industrie qui s'élève à 40%, au Québec le transport est la principale source d'émission avec près de la moitié des émissions devant le secteur de l'industrie). Selon le site du gouvernement canadien, ce pays vise à réduire l'impact du transport grâce à la réalisation d'investissements historiques dans les transports en commun.

Les considérations sociales, économiques et environnementales du développement des transports en commun montrent qu'il s'agit d'un secteur majeur permettant de lutter contre le réchauffement climatique et de répondre aux besoins de mobilité des citoyens dans les zones urbaines et péri-urbaines de demain. Dans ce sens, tout ce qui a été décrit précédemment motive l'étude de la mobilité dans les transports en commun avec l'objectif global d'augmenter leur attractivité. C'est dans ce contexte que se situent les travaux de cette thèse.

1.1.2 Les différents types de données permettant d'analyser la mobilité urbaine

Différents types de données (données personnelles, non personnelles) peuvent être utilisés pour analyser la mobilité urbaine. Sur le territoire de l'Union européenne et au Canada, les données personnelles (informations se rapportant à une personne physique identifiée ou identifiable) sont soumises à des régulations. En effet, sur ces territoires, le traitement de telles données est encadré par le Règlement Général sur la Protection des Données, RGPD. Tout organisme quels que soient sa taille, son pays d'implantation et son activité peut être concerné. En effet, le RGPD s'applique à toute organisation publique et privée qui traite des données personnelles pour son compte ou non, dès lors qu'elle est établie sur ces territoires ou que son activité sociale cible directement des résidents européens ou canadiens.

1.1.2.1. Données utilisées pour l'analyse des déplacements des usagers

De nombreux travaux ont porté sur l'analyse des déplacements des individus en zone urbaine et péri-urbaine et plus particulièrement sur l'exploitation de données numériques pour cette analyse. Plusieurs sources de données existent ayant chacun des avantages et des inconvénients.

Données d'enquêtes L'analyse des déplacements dans les transports en commun s'appuie traditionnellement sur des enquêtes déclaratives [Zmu+13]. On peut notamment citer les différentes enquêtes de déplacements qui existent en France à savoir les Enquêtes Ménage Déplacement (EMD), les Enquêtes Nationale de Transport (ENT), les Enquêtes Globale de Transport (EGT) et les enquêtes Origine-Destination (OD) et au Canada, les Enquêtes Nationale auprès des Ménages (ENM) et les enquêtes OD. Les informations obtenues grâce aux enquêtes sont détaillées, il est par exemple possible d'obtenir des informations socio-économiques et démographiques par individu (e.g., âge, sexe, etc.) ainsi que des informations sur le ménage (taille, structure) ou encore des informations sur le journal de déplacement (e.g., mode, motif, etc.). En revanche, ces enquêtes sont coûteuses et ne permettent pas d'avoir une information exhaustive sur l'ensemble des usagers des transports en commun (échantillon pouvant varier de 5% à 15% du nombre total des ménages pour les enquêtes OD). De plus, ces données ont un biais important dans les réponses (absence/imprécision) et ne sont récoltées que sur un échantillonage temporel large (entre 5 et 10 ans).

Données GPS Ces traces sont mono-mode ou multi-modes, elles permettent d'obtenir la localisation d'un individu ou d'un moyen de déplacement (e.g., bus ,taxi, etc.) de manière continue. Le défi avec ce type de données est de définir le mode de déplacement utilisé par les individus en fonction de plusieurs paramètres tels que leur position et leur vitesse.

Données de téléphonie Ces données collectées par des antennes de téléphonie, sont découpées géographiquement en fonction de la position des antennes. Elles ne permettent pas d'avoir directement l'information sur le mode de déplacement et ne représentent qu'un échantillon réduit de la population (données obtenues par opérateur téléphonique).

Les données de téléphonie peuvent être classées en deux catégories principales, à savoir les statistiques d'appels (Call Detail Records, CDR) et les enregistrements passifs du réseau. Le désavantage des données CDR est que leur fréquence d'échantillonnage peut varier de quelques minutes à plusieurs heures. De plus la localisation des usagers à l'aide de ces données exige une utilisation active du téléphone (appels et sms). Contrairement aux données CDR les données passives permettent quant à elles de localiser les téléphones lorsque ceux ci ne sont pas utilisés. Les enregistrements passifs sont donc plus pratiques que les données CDR pour le suivi des téléphones mobiles car plus fréquents.

Données Bluetooth et Wifi Plusieurs études portant sur l'analyse de la mobilité urbaine à l'aide de données Bluetooth et Wifi ont été réalisées ces dernières années [Lah+14; SKO16; ElT+17]. Ce type de données ne nécessite que de faibles coûts pour être récoltées. En revanche, leur collecte n'est possible que si les appareils des usagers sont connectés aux réseaux ou ont leurs réseaux Bluetooth ou Wifi actifs. Ceci entraîne une collecte partielle des déplacements de passagers dans les transports urbains.

Données billettiques[1] Dès 2004, Bagchi et White [BW04] ont étudié le rôle possible que les données billettiques peuvent jouer dans l'analyse des pratiques de mobilité et ont considéré leur potentiel comme pouvant compléter ou même remplacer les sources de données plus conventionnelles (telles que les données issues d'enquêtes). L'un des problèmes déjà identifié dans ces travaux est celui des données manquantes, telles que les détails personnels des titulaires de carte à puce (e.g., l'âge, le sexe, le revenu, etc. qui sont omis pour des raisons de confidentialité), ainsi que les destinations de déplacement (qui sont rarement recueillis par le système des AFC). L'avantage de ce type de données est qu'elles permettent de collecter la quasi totalité des déplacements des passagers utilisant les transports en commun hormis ceux effectués en situation de fraude. Au vue de ces avantages et de leur mise à disposition de la part des autorités de transport des villes de Montréal, Rennes et Paris, nous avons décidé d'exploiter ce type de données pour nos travaux portant sur l'analyse de l'affluence (prévision et visualisation) dans les transports en commun. Les cartes des différents systèmes de transports étudiés à savoir la carte Opus de Montréal, la carte Korrigo de Rennes métropole et la carte Navigo utilisée pour se déplacer en Ile-de-France sont présentées dans la Figure 1.1.

Fig. 1.1.: Cartes à puce "Opus", "KorriGo" et "Navigo", utilisées par les usagers des systèmes de transport en commun des villes de Montréal au Canada, Rennes et Paris en France.

[1] Données utilisées au cour de cette thèse.

1.1.2.2. Données utilisées pour analyser l'offre des transports urbains

Trois formats de données ont vu le jour ces dernières années pour permettre de stocker, partager et analyser les informations temporelles et géographiques issues de l'offre des transports urbains de manière générique et pérenne. Ces formats sont les suivants.

General Transit Feed Specification, GTFS Ce format a été créé par un employé de Google, Chris Harrelson en 2005. Initialement nommé Google Transit Feed Specification, il a alors été renommé General Transit Feed Specification pour répandre son utilisation plus facilement. Il s'agit d'un format de données permettant de partager les horaires d'un système de transport ainsi que les données géographiques liées à la structure du réseau de transport (e.g., horaire et position des trains, etc.). Il correspond à un ensemble d'au moins 6 fichiers CSV (13 au maximum) définissant chacun une table d'information : opérateur de transport, lignes de transport, déplacements, horaire par station, stations et calendrier.

General Bikeshare Feed Specification, GBFS Sous la direction de la North American Bikeshare Association, le GBFS a été créé par des propriétaires et des exploitants de systèmes de vélo en libre-service du secteur public, privé et sans but lucratif, des développeurs d'applications et des fournisseurs de technologies. Il s'agit du standard de données ouvertes pour les services de vélos en libre-service (VLS) et des services en "free-floating". Le GBFS permet de disposer des flux de données en temps réel dans un format uniforme accessible en ligne publiquement, en mettant l'accent sur la facilité d'accès. Cette spécification a été conçue en gardant à l'esprit les concepts suivants : les données contenues dans le GBFS sont destinées à être utilisées par des individus ayant l'intention de fournir des conseils en temps réel (ou semi temps réel) sur la disponibilité des vélos et sont conçues comme telles.

Mobility Data Specification, MDS Crée en 2018, la MDS est un nouvel ensemble de spécifications de données et d'exigences en matière de partage de données spécialisées pour les scooters électroniques et les bicyclettes en "free-floating". Inspiré par le GTFS et le GBFS, les objectifs de la MDS sont de fournir aux municipalités des interfaces de programmation (API) et des normes de données pour les aider à stocker, comparer et analyser les données sur la mobilité en tant que fournisseur de services.

La MDS aide les villes à collecter et à analyser l'information provenant d'entreprises à but lucratif qui exploitent des scooters et des bicyclettes en "free-floating" sur l'espace public. La MDS est un élément clé de l'infrastructure numérique qui aide les villes et les organismes de réglementation. La MDS est actuellement utilisé dans trois villes des Etats-Unis à savoir, Austin, Los Angeles et Santa Monica. On peut citer l'exemple du Los Angeles Department of Transportation (LADOT) où la MDS a été utilisée pour comprendre comment fonctionne la mobilité en "free-floating". A notre connaissance, ce format de données n'est pas encore utilisé en France et au Canada.

La MDS comprend deux composantes distinctes :

- L'API du fournisseur publiée pour la première fois en mai 2018 pour être mise en œuvre par les fournisseurs de services de mobilité. Lorsqu'une municipalité demande des informations à un fournisseur de services de mobilité, l'API du fournisseur possède une vue historique des opérations dans un format standard.
- L'API de l'agence publiée pour la première fois en avril 2019 pour être mise en œuvre par les organismes de réglementation. La première implémentation a été mise en service en février 2019. Les fournisseurs interrogent l'API de l'agence lorsqu'un événement se produit, comme le début ou la fin d'un déplacement.

1.1.2.3. Autres données permettant d'analyser la mobilité urbaine

Certaines données générées et utilisées par les opérateurs de transport peuvent servir indirectement à l'analyse de la mobilité urbaine. Nous pouvons notamment citer les remontées d'incidents étant survenus sur les réseaux de transport ou encore les données de comptage d'énergie à bord des trains.

Données incidents[2] Dans certains systèmes de transport, dès qu'un incident survient (e.g., accident voyageur, panne électrique, etc.), des informations définissant ce problème sont enregistrées dans une base de données. Ces informations contiennent généralement un descriptif de la cause de l'incident, l'horaire de départ et l'horaire de fin de l'incident ainsi que la station impactée. Ces données qui sont à l'origine,

[2]Données utilisées au cour de cette thèse.

enregistrées pour avoir une trace historique des différents incidents, peuvent également servir à analyser plus en détail leur impact sur l'offre de transport mais aussi sur l'affluence des passagers dans les différentes stations du réseau de transport.

Données de comptage d'énergie à bord des trains Ce type de données correspond, sur le réseau ferré français, à une mesure de la puissance consommée ou réinjectée à la caténaire, de manière géolocalisée, disponible toutes les 5 minutes, 24h sur 24, pour tous les trains. Ces données collectées à l'origine pour des besoins de facturation entre exploitant ferroviaire et fournisseur énergétique peuvent, après pré-traitements et explorations, permettre d'étudier le nombre de passagers utilisant chaque train. En effet, selon les auteurs de [ACO18], il est envisageable d'utiliser ce type de données pour estimer la charge embarquée dans les trains (fret ou passagers) en comparant la puissance absorbée sur des trajets similaires. D'après cette étude, ce type de données peut également servir à d'autres analyses telles que le fait de connaître la position des trains, détecter automatiquement certaines défaillances des infrastructures électriques, etc.

D'autres données reliées au domaine du transport peuvent servir à analyser les flux voyageurs telles que les données issues des requêtes d'itinéraires et des réseaux sociaux. Certaines de ces données peuvent être collectées en récupérant l'information à l'aide d'API, en les demandant directement aux gestionnaires (e.g., société gérant les requêtes d'itinéraires) ou en extrayant le contenu de différents sites web (Web scraping).

Données issues des requêtes d'itinéraires Ces données peuvent en plus d'aider les voyageurs à choisir le chemin le plus optimal pour parvenir à leur destination, permettre aux opérateurs de transport d'estimer la demande des passagers sur les différentes portions du réseau de transport en fonction des itinéraires requêtés par les usagers.

Données issues des réseaux sociaux Certains réseaux sociaux tel que Twitter, peuvent aider les exploitants des systèmes de transport à renseigner leurs usagers concernant différents incidents ayant lieu sur le réseau en quasi temps réel. De la même façon, les usagers utilisant les systèmes de transport en commun peuvent faire savoir leur mécontentement aux travers de messages postés sur ces réseaux mais aussi relayer des informations en temps réel concernant le fonctionnement du système. L'analyse de ces publications peut, par exemple, permettre à posteriori ou en temps réel de prévoir la demande des passagers ou de réagir à un arrêt

momentané de l'offre de transport. L'exploitation de ces données reste néanmoins difficile puisqu'elle nécessite d'utiliser des méthodes à l'état de l'art du traitement automatique du langage (Natural Language Processing, NLP), en plus du fait que ces données puissent être fortement bruitées par des données parasites.

Certaines données externes au domaine du transport telles que les données événementielles (e.g., horaire de concert, rencontres sportives, etc.) et les données météorologiques peuvent également être mobilisées pour enrichir les analyses menées dans ce domaine.

Données événementielles[3] La demande des passagers peut parfois être influencée par des données exogènes telles que les données événementielles. En effet, les événements à forte fréquentation tels que les concerts, spectacles, rencontres sportives, etc., ont un impact certain sur la demande de transport. Ces événements impliquent le plus souvent une augmentation de l'affluence des passagers sur certaines stations/lignes de transport en commun. Comme présenté dans les chapitres 3 et 5, leur utilisation peut servir à mieux prévoir la demande des passagers et à mieux comprendre la mobilité des passagers en cas d'événement.

Données météorologiques De fortes précipitations météorologiques peuvent impacter de manière importante la fréquentation des systèmes de transport en commun (e.g., forte tempête de neige). De la même manière, certains systèmes de transports utilisés en extérieur tels que les vélos en libre service ou les trottinettes en "free-floating" verront leur utilisation diminuer en cas de précipitations. La météo est donc un facteur exogène pouvant impacter l'usage des différents systèmes de mobilité.

1.1.3 Mise à disposition des données entre opérateurs de transport et collectivités

L'émergence du domaine de la science des données et de l'intelligence artificielle a entraîné un fort intérêt pour la collecte et le partage des données comme nous pouvons le voir dans (i) les rapports Jutand [Jut15] et Villani [Vil18] en France et (ii) l'ouverture du site web "portail open data" ainsi que la création de l'institut de valorisation des données (IVADO) au Canada. Si nous prenons le cas de la France, ces deux rapports écrits par Jutand en 2015 et Villani en 2018 portant sur l'ouverture

[3] Données utilisées au cour de cette thèse.

des données et le développement de l'IA en France, visent à faciliter l'accès à des données d'intérêt public (open data), telles que certaines données en lien avec le domaine du transport.

Dans cette même dynamique d'ouverture des données, la mission Etalab fondée en 2011 et lancée au sein de la Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'état français (DINSIC) vise a mener une politique de la donnée en promouvant des actions publiques plus transparentes et collaboratives grâce au numérique. Etalab a ainsi oeuvré pour la mise en place du portail data.gouv.fr permettant d'accéder à des données publiques. Concernant le transport, différentes données sont mises à disposition sur le site web transport.data.gouv.fr, il s'agit notamment de données GTFS ou GBFS de différentes villes en France. Des données de fréquentation fortement agrégées par type de jour et par heure sont également mises à disposition par Ile-de-France Mobilité (organisation en charge des transports en commun en Ile-de-France).

Malgré ces avancées concernant l'ouverture des données, certains freins institutionnels ou commerciaux persistent. Il n'est par exemple pas possible d'accéder à des données individuelles anonymisées concernant les déplacements des usagers, sauf pour des besoins de recherche.

Avec l'arrivée de nouveaux moyens de transport tels que les trottinettes et les vélos en "free-floating", un nouveau format de données a vu le jour, le format MDS. Dans le but de contrôler ces nouveaux moyens de transport, le département des Transports de Los Angeles aux Etats-unis (Los Angeles Department of Transportation, ou LADOT) a obtenu la permission d'accéder à ces données gérées par des sociétés privées. Du côté français, selon la charte de bonne conduite relative à la location de trottinettes électriques, le nouveau Plan Climat Air Énergie territorial (PCAET) [Cli18] guide Paris vers la neutralité carbone 2050 et constitue une mise en œuvre concrète des engagements pris par la France lors de la COP 21. Des objectifs ambitieux sont ainsi visés : zéro véhicule diesel en 2024 et zéro véhicule essence en 2030 à Paris. Il est stipulé dans cette charte [Par19] que certaines données relatives à l'activité des opérateurs doivent être partagées avec la ville de Paris. Nous pouvons citer les trois points suivants directement extraits de la charte:

- "Les opérateurs s'engagent à mettre gracieusement à disposition de la Ville, pour son usage propre, dans le respect de l'application de la réglementation sur la protection des données personnelles, des données sur le déploiement et l'usage du service, nécessaires à la meilleure connaissance des flux et à l'optimisation des espaces de stationnement pour les trottinettes."

- "La Ville de Paris s'engage à conserver ces données confidentielles et à réserver leur usage à l'analyse des phénomènes de flotte libre. La Ville de Paris s'engage à tout mettre en œuvre pour sécuriser l'intégrité de ces données une fois enregistrées au sein de son système d'information."
- "Le détail des données concernées, ainsi que leurs modalités de communication et de protection, sont précisés en collaboration avec les opérateurs par un groupe de travail organisé par la Mairie. Ces éléments font l'objet d'une annexe à la présente Charte."

En conclusion, ces différents types de données soulèvent de nouveaux problèmes (e.g., collecte, partage) quant à l'analyse de la mobilité et l'interaction des différents modes de transport. Néanmoins, ils sont nécessaires à l'élaboration de nouveaux services pouvant bénéficier aux opérateurs de transport et à leurs usagers. Dans ce sens une nouvelle orientation des pouvoirs publics vise à encourager l'ouverture de telles données d'intérêts publics. D'autre part ces données soulèvent un autre problème concernant la protection de la vie privée des usagers, en effet ces données se doivent d'être suffisamment anonymisées pour respecter la vie privée des usagers tout en permettant aux décisionnaires publics de mieux réguler les systèmes de transport urbain.

1.2 Problématiques de recherche et contributions

1.2.1 Objectifs

L'objectif principal de cette thèse est d'augmenter l'attractivité des systèmes de transport en commun. Pour cela, nous nous attachons aux problèmes de prévision et de visualisation de l'affluence des passagers dans les transports en commun. Le but est d'apporter une connaissance détaillée des déplacements des usagers aux opérateurs de transport en commun dans le but de les aider à améliorer les services de transports proposés (e.g., offre, information voyageur, etc.). Le processus complet de prévision contient trois étapes détaillées dans les sections suivantes : le traitement des données, la prévision et la visualisation de l'affluence des passagers.

1.2.2 Traitements des données volumineuses issues de la mobilité dans les transports en commun

Les données billettiques brutes correspondant aux validations de chaque usager pour chaque déplacement, entraînent la création de larges bases de données (plusieurs gigabytes). Une étape de pré-traitements est souvent nécessaire pour traiter ces larges ensembles de données. Afin de réaliser ces pré-traitements (e.g., agrégation du nombre de passagers par quart d'heure), nous avons mis à profit les dernières avancées en matière de méthodes de calculs parallélisés. Ces types de méthodes se basent sur un ensemble de machines réelles ou virtuelles que l'on nomme cluster de machine qui permet de réaliser des opérations sur des données de manière parallélisée sur chacune des machines qui se partagent les opérations a effectuer dans but de réduire le temps de calcul. Dans notre cas, nous avons utilisé la technologie issue du framework Spark créé par Zaharia [Zah+10a] en 2010 pour réaliser les pré-traitements de larges bases de données (données billettiques de Rennes détaillées dans la section 4.2).

1.2.3 Prévision de l'affluence dans les transports en commun

La prévision est un domaine d'investigation majeur en ce qui concerne l'exploitation des données numériques. L'objectif est de développer des modèles de prévision pour prédire le fonctionnement du système urbain [Lap+17; Li+17]. Dans notre cas, nous nous attachons à prévoir l'affluence des passagers dans les systèmes de transport en commun avec une résolution temporelle fine (agrégation du nombre de passagers au quart d'heure). Il est ainsi possible de distinguer deux types d'horizon de prévision, (i) la prévision long terme qui permet d'estimer la demande des passagers plusieurs jours ou mois à l'avance et (ii) la prévision court terme qui permet de prédire la demande des passagers des quelques pas de temps suivants dans un contexte temps réel.

1.2.3.1. Prévision long terme de l'affluence dans les transports en commun

L'un des principaux objectifs des parties prenantes (opérateurs et autorités de transport) est d'adapter le plus précisément possible l'offre de transport à la demande des passagers, quelque soit la période (e.g., période normale, période contenant la présence d'un ou plusieurs événements, période perturbée, journée spéciale, etc.). Selon les opérateurs de transport, un autre objectif est d'anticiper la demande de

titres de transport jetables (cartes à puce non rechargeables) dans le but de faire correspondre la disponibilité des titres à la demande des passagers à une période donnée, en particulier en période d'événements (e.g., concert, rencontre sportive, spectacle, exposition, etc.). C'est notamment le cas à Montréal.

La disponibilité des traces numériques, concernant dans notre cas les données billettiques combinées à la disponibilité d'une base de données événementielles, offre la possibilité de développer des outils de prise de décisions qui peuvent permettre aux opérateurs de transport de mieux comprendre et prévoir la demande des passagers dans les grandes villes. De cette façon, les opérateurs seront en mesure d'améliorer la planification de l'offre de transport. D'autre part, les citoyens pourront profiter d'une tarification spécifique pendant les périodes d'événement, bénéficier d'une réduction du temps d'attente de leur moyen de déplacement (e.g., bus, métro, tramway, etc.) et planifier leurs déplacements de manière plus efficace pour éviter la congestion dans les transports en commun grâce aux informations fournies concernant la demande des passagers prévue.

Pour répondre à cette problématique, nous proposons une mise en forme générique des données contextuelles détaillée dans le chapitre 3. Dans cette étude nous démontrons que l'utilisation conjointe des données événementielles et calendaires permet d'améliorer les résultats en périodes d'événement.

1.2.3.2. Prévision court terme de l'affluence dans les transports en commun

Contrairement aux outils de prévision long terme, les outils de prévision court terme ont pour but de prévoir les flux de passagers, quelques heures ou minutes à l'avance en prenant en compte l'état du réseau de transport en temps réel, ce qui fait de ces modèles des outils pertinents pour résoudre différents problèmes : (i) informer les passagers des situations de congestion, (ii) proposer aux passagers des itinéraires de déplacement adaptés à l'offre de transport et à la demande des passagers et enfin (iii) améliorer l'exploitation du système de transport des lignes affectées par certaines perturbations afin de faire correspondre l'offre de transport à l'affluence des passagers en temps réel.

Dans ce sens, nous nous attachons dans la section 4.2 à pré-traiter et construire une base de données, prévoir des flux de passagers entre stations (prévision de matrices Origines-Destinations) et à analyser en détail ces prévisions dans le but d'aider les opérateurs de transport à adapter l'offre de transport à la demande et plus généralement d'offrir de meilleurs services aux passagers (e.g., meilleur information voyageur, meilleur disponibilité des agents, etc.).

Dans la section, nous étendons les travaux que nous avons initié dans [Toq+16] détaillés à la section 4.2, qui portent sur la prévision court terme de matrices de comptage OD des stations du métro de Rennes Métropole. Nous avons ajouté à l'étude une comparaison de nouveaux modèles et réalisé une analyse plus approfondie des résultats. Dans cette étude nous explorons également la prévision court terme multi pas de temps de la demande des passagers.

1.2.4 Outils de visualisations pour l'analyse de l'affluence

L'analyse des données issues de séries temporelles et des résultats de prévision de séries spatio-temporelles peut s'avérer difficile à réaliser du fait de la grande quantité d'information à traiter. Cependant ces analyses sont primordiales à la compréhension des résultats de prévision et plus en amont, à l'amélioration de ceux-ci. Dans ce sens, l'utilisation d'outils de visualisation s'avère être un choix pertinent pour résoudre ce problème.

Dans ce contexte, nous avons développé deux outils de visualisation dans le but de pouvoir analyser plusieurs ensembles de données de types spatio-temporels, à savoir, des données billettiques, événementielles et d'incidents étant survenus sur le réseau de transport. Ces outils présentés au chapitre 5, permettent principalement d'analyser conjointement les résultats de la prévision de la demande des passagers ainsi que les données contextuelles pouvant impacter l'offre de transport, dans un cadre spatial et temporel. Ce type de visualisation permet ainsi de mieux comprendre pourquoi certaines erreurs sont commises et donne des pistes d'amélioration concernant la création des modèles de prévision. Dans ce sens, ces visualisations se voient être d'excellents outils d'analyse pour les opérateurs de transport et les chercheurs travaillant sur la prévision de séries temporelles impactées par des phénomènes spatio-temporels.

1.2.5 Publications

1.2.5.1. Journal international

[Toq+19] F. Toqué, E. Côme, M. Trépanier, and L. Oukhellou. "Forecasting of the Montreal Subway Smart Card Entry Logs with Event Data". In:(2019). En cours de révision pour la revue Transportmetrica A : Transport Science (Taylor & Francis).

1.2.5.2. Conférences internationales avec actes

[Toq+16] F. Toqué, E. Côme, M. K. El Mahrsi, and L. Oukhellou. "Forecasting dynamic public transport Origin-Destination matrices with long-Short term Memory recurrent neural networks". In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). Nov. 2016, pp. 1071–1076.

[Toq+17] F. Toqué, M. Khouadjia, E. Côme, M. Trépanier, and L. Oukhellou. "Short & long term forecasting of multimodal transport passenger flows with machine learning methods". In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE. 2017, pp. 560–566.

[Toq+18b] F. Toqué, E. Côme, L. Oukhellou, and M. Trépanier. "Short-Term Multi-Step Ahead Forecasting of Railway Passenger Flows During Special Events With Machine Learning Methods". In: CASPT 2018, Conference on Advanced Systems in Public Transport and TransitData 2018. Brisbane, Australia, July 2018, 15p.

1.2.5.3. Conférences sans acte

[Toq+18a] F. Toqué, E. Côme, L. Oukhellou, and M. Trépanier. "Prévision du nombre de passagers entrant dans un réseau de transport multimodal en cas d'évènements atypiques, à l'aide de méthodes d'apprentissage automatique". In:RFTM, 1ères Rencontres Francophones Transport Mobilité. LYON, France, June 2018, 3p.

[Toq19a] F. Toqué, E. Côme, L. Oukhellou, and M. Trépanier. "Short and long term predictions of public transit ridership based on smart card data:the Montreal case". In: Workshop International Network on the Use of Passive Data from Public Transport. Washington DC, USA, Jan. 2019

[Toq19b] F. Toqué, E. Côme, L. Oukhellou, and M. Trépanier. "Visualization tools for space-time series analysis with context awareness: Montreal subway case". In: TransitData 2019, 5th International Workshop and Symposium. Paris, France, July 2019

2 Prévision de la mobilité urbaine : études récentes et méthodologie

2.1 Introduction

Depuis 2004, l'utilisation des données billettiques pour analyser la mobilité dans les transports en commun a reçu une attention considérable de la part des chercheurs. Plus récemment, de nombreuses études portant sur l'analyse de la mobilité urbaine à l'aide des données billettiques se sont attachées au problème de la prévision de la demande des passagers. Le traitement de ces données requiert de les agréger en comptant le nombre de passagers par déplacement Origine-Destination ou en agrégeant le nombre de passagers entrant en un point précis du système de transport en commun (e.g., nombre de passagers entrants ou sortants d'une station), où les matrices Origine-Destination sont plus riches en informations mais sont souvent difficiles à obtenir étant donné que tous les systèmes de billettique n'obligent pas les passagers à valider leur titre de transport en sortie. Dans ce cas de figure, il est néanmoins possible d'estimer ces destinations (voir section 4.2.2.2) même si cela se traduit par une perte d'information. Chacun de ces types de données est utile à la planification, l'exploitation et la gestion des réseaux de transport en commun.

Les objectifs en matière de prévision de l'affluence des passagers peuvent différer en fonction de l'horizon prévisionnel. Nous avons donc distingué deux grands types d'horizon temporel à savoir l'horizon long terme et l'horizon court terme. Pour les prévisions long terme, l'objectif est de prévoir la demande à l'aide des données disponibles bien en avance. Dans ce cas, nous pouvons par exemple utiliser les caractéristiques du type de jour (e.g., vacances, jour férié, fin de semaine, etc.) et des événements prévisibles (e.g., concerts, rencontres sportives, etc.) pour effectuer cette prévision. Ce type de prévision peut être utile pour améliorer la planification de l'offre de transport. En ce qui concerne les prévisions court terme, le processus peut en plus de traiter les données disponibles en avance, prendre en compte les observations sur les flux de passagers récents. Ce type de prévision peut se révéler pertinent en cas de situation atypique. L'objectif principal des opérateurs de transport

est alors d'utiliser la prévision de l'affluence des passagers prévue pour optimiser le fonctionnement du système de transport afin d'adapter l'offre de transport à la demande atypique (e.g., mise en place temporaire de navettes) ou d'informer les passagers de l'incident et de leur proposer d'autres moyens leur permettant d'atteindre leur destination.

Les contributions de cette thèse se placent dans ce contexte. Nous traitons dans la première partie de ce chapitre de l'état de l'art et de la méthodologie de la prévision de la mobilité urbaine au travers de différents exemples de travaux portant sur la prévision long terme et court terme de l'affluence dans différents systèmes de transport (e.g., bus, métro, taxi, etc.). Dans un second temps, nous présentons différentes méthodes de prévision qui seront mobilisées dans les chapitres suivants. Enfin, nous nous attachons à détailler les méthodes d'évaluation et d'optimisation des modèles de prévision.

2.2 Prévision long terme

L'un des premiers objectifs des autorités de transport a été de dimensionner de nouveaux systèmes de transport en commun. Dans ce but, le premier modèle de prévision long terme de la mobilité des individus en zone urbaine, nommé modèle à quatre étapes (four-step model en anglais) [McN07] a été implémenté en 1950. Ce modèle basé sur des données d'enquêtes suit les quatre étapes suivantes :

- La génération des déplacements qui détermine la fréquence des déplacements des points d'origine ou de destination de chaque zone en fonction de différents facteurs, tels que le motif de déplacement, l'utilisation du sol (e.g. zone résidentielle, zone scolaire, etc.), les caractéristiques démographiques des ménages et d'autres facteurs socio-économiques.
- La distribution des déplacements entre origine et destination, en utilisant le plus souvent un modèle nommé "gravity model" [ES90], équivalent à un modèle d'entropie maximale [Kap89].
- Le choix du mode de transport qui est calculé pour chaque déplacement entre une origine et une destination.
- L'affectation d'itinéraire correspondant aux déplacements entre un point d'origine et d'arrivée. Pour cela, le modèle le plus souvent utilisé est le principe d'équilibre des usagers de Wardrop (équivalent à un équilibre de Nash), où chaque conducteur ou groupe d'individus choisit le trajet le plus court.

Certaines études plus récentes portant sur la prévision de la mobilité dans les transports en commun favorisent l'utilisation de données issues de capteurs (e.g., données billettiques, données gps, etc.) capables de quantifier la mobilité des individus de manière beaucoup plus exhaustive que les données d'enquêtes, portant uniquement sur une partie des individus se déplaçant dans les transports en commun. En effet, dans l'étude[T+14], les auteurs utilisent un modèle statistique (ARIMA) pour prévoir le nombre de passagers agrégé par mois, du réseau de bus et de train de la ville de Sydney en Australie. Cette étude se basant sur une agrégation temporelle large des données a permis de démontrer l'efficacité de ce type de méthodes pour la prévision de la demande de passagers plusieurs mois en avance. Il est a noté que l'utilisation de telles données nécessite l'existence d'un réseau de transport déjà existant et permet de prévoir la demande de passagers sur ce réseau uniquement. Au contraire du modèle à quatre étape qui fournit une prévision de la demande de passagers moins précise mais qui peut d'un autre côté, aider à l'élaboration de nouvelles lignes de transport.

Actuellement, les pratiques de la planification de l'offre de transport accordent une attention particulière aux méga-événements (e.g., jeux olympiques, mondial de football, etc.), qui nécessitent un effort important de la part des institutions publiques et privées [PBL03]. Ces événements qui rassemblent d'énormes flux de personnes, ont l'avantage d'être bien définis dans le temps et dans l'espace, il est souvent possible de connaître à l'avance la répartition globale de la demande, étant donné, par exemple, la connaissance du nombre de logements disponibles. Les grands événements quant à eux sont beaucoup moins important en matière d'affluence et de durée (e.g. concerts de musique, rencontres sportives, spectacles, etc.), mais ils sont aussi plus courants et délicats à gérer. Dans la pratique, la majorité des grands événements ne fait pas l'objet d'un traitement ou d'une attention particulière en matière de planification de l'offre de transport, ce qui crée dans la plupart des cas des congestions et de fortes affluences. L'affluence des passagers devient inévitablement plus difficile à prévoir que celle liée à la mobilité quotidienne lors de ce types d'événements [PBL03].

Dans ce contexte, les auteurs de [Kup+13] ont réalisé une modélisation de la demande en période d'événements spéciaux, selon l'approche traditionnelle du modèle à quatre étapes, pour la région métropolitaine de Phoenix dans la région de l'Arizona aux Etats-Unis. Les données recueillies servent à prédire, pour chaque événement, divers indicateurs tels que le nombre de déplacements par mode de transport, les horaires de déplacement, le nombre de déplacements en transport en commun générés par ces événements, etc.

Dans le chapitre 3 nous traitons du problème de prévision long terme avec une résolution fine, à savoir une agrégation du nombre de passagers aux 15 minutes. Dans cette étude, nous nous attachons à prévoir le nombre de passagers entrant dans chacune des stations du réseau de métro de Montréal au Canada, en utilisant des données billettiques et des données événementielles s'étendant sur une période de 3 ans. A notre connaissance, seules quelques ressources liées à la prévision long terme avec une résolution fine (que nous considérons comme étant une agrégation temporelle inférieure ou égale à 2h) sont disponibles dans la littérature, contrairement aux travaux portant sur la prévision court terme. L'étude la plus en lien avec notre travail détaillé dans le chapitre 3 est celle des auteurs de [PRB15]. Dans cette étude, les auteurs ont travaillé sur des approches de prévision long terme en utilisant des données événementielles extraites du web pour prévoir le nombre agrégé de passagers par demi-heure dans 3 stations de métro et 11 arrêts de bus regroupés en 5 zones dans la ville de Singapour. Leur étude a été réalisée sur un ensemble de données s'étendant sur une courte période (total de 16 jours) avec différents modèles d'apprentissage automatique (le modèle ayant obtenu les meilleurs résultats est un réseau de neurones classique, multi-layer perceptron en anglais). Les auteurs ont démontré que l'utilisation d'informations sur les événements, combinées à des données billettiques (nombre de passagers entrants et sortants) peut améliorer la qualité des prévisions de l'affluence des passagers lors d'événements spéciaux.

2.3 Prévision court terme

La prévision court terme des flux de passagers consiste à prévoir quelques pas de temps en avance le nombre de personnes se déplaçant à différents endroits d'un système de transport. Différents modèles de prévision et types d'agrégation ont été utilisés pour répondre à cette problématique, ces agrégations peuvent par exemple correspondre à des périodes très fines (e.g. 2 minutes) ou correspondre à des plages temporelles plus larges (e.g., 1 heure). Les auteurs de [Li+17] ont utilisé des réseaux de neurones nommés Radial Basis Function network (RBF) multi-échelles pour prévoir le nombre de passagers sortant de différentes stations du métro de Pékin à deux pas de temps en avance avec une agrégation de 15 minutes (t+15 et t+30 minutes), en prenant en compte le nombre de passagers entrant dans d'autres stations du métro. Dans cette étude, les auteurs ont effectué une analyse approfondie des résultats obtenus en période d'événements spéciaux. D'autres exemples de prévision des flux de voyageurs dans le métro comprennent les travaux de [RBG16], où les auteurs ont prédit les flux de voyageurs du pas de temps suivant

avec une agrégation extrêmement fine de deux minutes (t+2 minutes). Les auteurs ont utilisé un modèle de réseau bayésien et prédit les flux de passagers (entrée et sortie) à toutes les stations d'une ligne de métro du réseau parisien. Dans l'étude de [Cui+16], les auteurs ont créé un modèle exogène auto-régressif non linéaire flou pour prédire le nombre de passagers entrants au pas de temps suivant (t+1 heure), dans les stations du métro de Shanghai. En plus des prévisions, les auteurs de [Din+16] ont effectué une analyse sur l'importance relative de chaque variable (en pourcentage) utilisée en entrée des modèles de prévision du métro, telles que les activités de transfert par autobus et des caractéristiques temporelles. Ils ont ainsi montré que les principales caractéristiques utilisées par les modèles de prévision court terme sont les observations passées de la demande dans les stations de métro ($\sim 82.0\%$), dans les stations de bus ($\sim 10.4\%$) et les variables correspondant à l'horaire de prévision ($\sim 3.6\%$). Cette étude porte sur la prévision du pas de temps suivant (t+15 minutes) des flux de passagers dans 3 stations du métro de Pékin.

Une recherche plus approfondie des études les plus récentes sur les prévisions court terme dans le domaine des transports révèle que l'aspect spatial et l'aspect temporel sont souvent étudiés dans de tels problèmes de prévision. Par exemple, dans le domaine de la prévision de la demande de covoiturage, une équipe de recherche d'Uber [Lap+17] a étudié les données portant sur la demande de covoiturage d'Uber en se concentrant sur les variables temporelles pour la prévision de jours atypiques (e.g., veille de Noël, jour de l'an, etc.). Les auteurs de [Ke+17] se sont concentrés sur la capture d'informations spatio-temporelles issues du réseau de transport grâce à une approche d'apprentissage profond. Des approches similaires ont été mises en œuvre par les auteurs de [ZZQ17] pour prédire les flux de mobilité dans toute la ville et par [Yao+18] pour prédire la demande de taxis. Des études qui mettent en lumière l'aspect spatio-temporel des prévisions de trafic ont également été menées par [WT16; Che+17] avec une combinaison de modèles de réseaux de neurones à convolution et de réseaux de neurones récurrents ainsi que par [YYZ17] avec un modèle de réseau de neurones à convolution basé sur une structure de graphe.

Afin d'améliorer la précision des modèles de prévision, certaines études ont démontré l'importance des données externes, en particulier les données événementielles. En effet, les événements tels que les concerts, les spectacles et les rencontres sportives sont des sources de perturbation de la demande de mobilité. Selon plusieurs études détaillées ci-après, un des enjeux actuels portent en particulier sur la collecte de données événementielles fiables. Les auteurs de [NHG17] ont développé des approches de prévision court terme pour prévoir les flux agrégés des passagers du métro lors des 4 prochaines heures en utilisant des données issues des réseaux sociaux. Les auteurs se sont concentrés sur la prévision du nombre total de passagers (somme

des entrées et sorties) d'une station de métro du réseau de la ville de New York. Ils ont proposé une méthodologie en deux étapes : la détection d'événements par hashtag (un hashtag permet d'identifier un centre d'intérêt d'un message posté sur le réseau social Twitter) suivie de l'utilisation combinée d'une régression linéaire et d'un modèle de moyenne mobile auto-régressive saisonnière (SARIMA). Des études plus récentes menées par [MRP18; RMP19] portent sur la collecte automatique de données événementielles. Dans ces études, les auteurs s'attachent au problème de la prévision court terme de la demande de taxis dans deux endroits distincts de New York en utilisant des méthodes d'apprentissage profond. La comparaison des modèles a montré que la catégorisation des événements pouvait considérablement aider les modèles de prévision à obtenir de meilleurs résultats.

Comme le montre le tableau 2.1, de nombreuses études portent sur les prévisions court terme des flux de passagers. Les auteurs de ces études exploitent et comparent diverses méthodes et basent leur prévision sur des horizons temporels différents.

Tab. 2.1.: Liste des travaux de prévisions de flux de passagers court terme à l'état de l'art.

Référence	Méthode	Mode	Agrégation	Horizon	Données évén.
[Li+17]	RBF	Métro	15min	1,2	Non
[RBG16]	Bayesienne	Métro	2min	1	Non
[Cui+16]	AR	Métro	1h	1	Non
[Din+16]	MLP	Métro	15min	1	Non
[Lap+17]	LSTM	Taxi	1jour	1	Non
[Ke+17]	C+RNN	Taxi	1h	1	Non
[ZZQ17]	C+RNN	Taxi&Vélo	1h&30min	1&1	Non
[Yao+18]	C+RNN	Taxi	30min	1	Non
[WT16]	C+RNN	Trafic	5min	1	Non
[Che+17]	C+RNN	Trafic	15min	1,2,3,4	Non
[YYZ17]	GCNN	Trafic	15min	1,2,3	Non
[MRP18]	GP	Taxi	1h	1	Oui
[RMP19]	LSTM	Taxi	1jour	1	Oui

Le modèle RBF correspond au modèle Radial Basis Function, un type de réseau de neurones. MLP ou Multi Layer Perceptron correspond à un réseau de neurones classique. LSTM ou Long short-term Memory est un modèle de réseau de neurones récurrent. Nous avons défini la notation C+RNN pour désigner une architecture combinant un réseau de neurones à convolution et un réseau de neurones récurrent. GP correspond au modèle Gaussian Process. Enfin le modèle GCNN correspond à un modèle récent de réseau de neurones qui combine l'utilisation d'un réseau de neurones à convolution pour des données structurées sous forme de graphe.

2.4 Méthodologie

Comme expliqué dans les sections 2.2 et 2.3, de nombreuses méthodes ont été exploitées pour prévoir la mobilité urbaine. Dans cette section, nous détaillons une partie de ces méthodes en faisant référence aux études les ayant exploitées.

2.4.1 Méthodes basiques

Les méthodes basiques servent de référence pour évaluer des modèles de prévision plus avancés. Nous détaillons la méthode moyenne historique et la méthode naïve (LOCF) dans les sections 2.4.1.1 et 2.4.1.2 qui suivent.

2.4.1.1. Moyenne historique (Historical Average, HA)

Le modèle à moyenne historique (HA) est un modèle simple de prévision de série temporelle souvent utilisé pour répondre à des besoins de prévision long terme. Il utilise la moyenne des observations historiques par période (e.g., minute, heure, jour, semaine, mois, etc.) pour prédire les valeurs de la série temporelle aux périodes suivantes.

Par exemple, si l'on se place dans le cas d'une prévision du nombre de passagers entrants par station, agrégés par pas de temps de 15 minutes. Si la période de référence est le type de jour (lundi, mardi, etc.), le modèle à moyenne historique visera à prédire le nombre de passagers entrant dans chaque station en calculant la moyenne des observations historiques par jour (lundi, mardi, etc.) au pas de temps correspondant. Par exemple, la prédiction du pas de temps 8h-8h15 le lundi, correspondra à la moyenne de toutes les valeurs historiques obtenues lors des lundis au pas de temps 8h-8h15.

2.4.1.2. Méthode naïve (Last Observation Carried Forward, LOCF)

Le modèle naïf (LOCF) est une méthode statistique utilisée pour évaluer des modèles de prévision de séries temporelles court terme. Ce modèle prend en entrée la dernière observation de la série et la renvoie comme sortie du modèle. Les articles exploitant des méthodes basiques sont détaillés dans le tableau 2.2.

Tab. 2.2.: Etudes exploitant des méthodes basiques à des fins de prévision de la mobilité urbaine

Méthode	Références
HA	[RBG16; Ke+17; ZZQ17; YYZ17; Yao+18]
LOCF	[RBG16]

HA correspond au modèle moyenne historique et LOCF au modèle de la dernière observation.

2.4.2 Méthodes statistiques

De nombreuses méthodes statistiques ont été créées pour répondre à des problèmes de prévision de séries temporelles. Ces méthodes ont souvent l'avantage d'avoir un temps de calcul rapide ainsi que de bénéficier d'une certaine explicabilité quant aux résultats de prévision contrairement à des méthodes issues de l'apprentissage profond. Il est par exemple possible pour avec les modèles de régression de connaître l'importance des différentes variables d'entrées en étudiant les paramètres associés. Nous détaillons deux méthodes statistiques exploitées dans différentes études portant sur l'analyse de la mobilité à savoir le modèle de régression linéaire Elastic Net et le modèle vecteur autorégressif (VAR) dans les sections 2.4.2.1 et 2.4.2.2 .

2.4.2.1. Régression linéaire (Linear Regression, LR)

La régression linéaire est un modèle statistique qui suppose une relation linéaire entre une variable dépendante et une ou plusieurs variables explicatives (ou variables indépendantes). Afin d'éviter le sur-apprentissage, il est possible d'ajouter au critère un terme de régularisation, de type Lasso (pénalité de type L1), ou de type Tikhonov (pénalité de type L2) ou bien encore une combinaison de ces deux méthodes, nommée Elastic Net.

Elastic Net Cette méthode de régression linéaire a été introduite par [ZH05] et combine linéairement les pénalités L1 et L2 des méthodes de régularisation Lasso et de Tikhonov (Ridge regression en anglais). La méthode Lasso permet de réduire la complexité du modèle en appliquant automatiquement une sélection de variables. La méthode de régularisation de Tikhonov, quant à elle, réduit le nombre de paramètres

pour éviter la multicolinéarité. La méthode Elastic Net minimise la fonction objectif définie dans l'équation 2.1 qui suit:

$$\begin{aligned} \hat{w} = \arg\min_{w} 1/(2 * n_{samples}) * ||y - Xw||^2 \\ +\alpha * l1_{ratio} * \sum_{p=1}^{P} |w_p| \\ +0.5 * \alpha * (1 - l1_{ratio}) * \sum_{p=1}^{P} w_p^2 \end{aligned} \tag{2.1}$$

Où α correspond à un hyperparamètre définissant l'importance donnée aux termes de pénalités L1 et L2, et $l1_{ratio}$ correspond à un hyperparamètre compris entre 0 et 1 permettant de choisir le poids associé à chaque terme de régularisation L1 ou L2 (si ce terme est égale à 1, il s'agira uniquement d'une pénalisation L1).

Si nous nous plaçons dans le cas d'une prévision court terme à sortie univariée (un modèle par série temporelle), nous pouvons définir les entrées et sorties de ce modèle comme suit:

- **Entrées** : Observation de la série temporelle aux n pas de temps précédents (y_{t-n} à y_{t-1})
- **Sortie** : Valeur prédite de la série temporelle au pas de temps suivant à savoir $\hat{y}_t$ avec $\hat{y}_t \in \mathbb{R}$.

Dans le cas d'une prévision multivariée (modèle prenant en compte toutes les séries temporelles en entrée et en sortie) permettant de prendre en compte les relations entre les différentes séries temporelles, la méthode de régression correspond au modèle VAR détaillé dans la section suivante.

2.4.2.2. Vecteur autorégressif (Vector Autoregressive, VAR)

Le modèle vecteur autorégressif, Vector Autoregressive (VAR) en anglais introduit par [Lüt11] est une généralisation multivariée du modèle Autoregressive (AR) qui est une méthode de régression linéaire univariée utilisée pour prévoir la prochaine valeur d'une série temporelle. Le modèle VAR quant à lui a l'avantage de traiter symétriquement l'ensemble des séries temporelles d'entrée afin de prévoir la prochaine valeur de chaque série temporelle en sortie. Il est ainsi capable de prendre en compte les interdépendances linéaires de plusieurs séries temporelles. Le modèle VAR traite

toutes les variables des séries temporelles étudiées de la même manière : chaque variable est régie par une équation expliquant son évolution basée sur ses propres valeurs passées, les valeurs passées des autres variables du modèle et un terme d'erreur. La prévision de la variable y au pas de temps t est calculée à l'aide de l'équation 2.2.

$$\hat{y}_t = w_1 y_{t-1} + ... + w_p y_{t-p} + u_t \tag{2.2}$$

Où $\hat{y}_t \in \mathbb{R}^d$ le vecteur de valeurs prédites, avec d le nombre de séries temporelles, y_{t-1} est le vecteur de valeurs observées au pas de temps $t-1$, w correspond à la matrice de poids correspondant aux paramètres de régression, p correspond au nombre d'observations passées à prendre en compte aussi appelé lag et u représente le terme d'erreur.

Les articles ayant exploité des méthodes statistiques sont détaillés dans le tableau 2.3.

Tab. 2.3.: Etudes portant sur la prévision de la mobilité urbaine qui exploitent des méthodes statistiques

Méthode	Références
LR	[PRB15; Yao+18; MRP18]
MA	[Lap+17; Ke+17]
ARIMA	[Ke+17; ZZQ17; Che+17; YYZ17; Yao+18]
SARIMA	[ZZQ17]
VAR	[ZZQ17]
Dynamic Bayesian Network	[RBG16]

LR correspond au modèle de régression linéaire, MA au modèle Moving Average, ARIMA au modèle Autoregressive integrated moving average, SARIMA au modèle Seasonal ARIMA et VAR au modèle Vector autoregressive.

2.4.3 Méthodes à noyaux

La particularité des méthodes à noyaux réside dans leur capacité à résoudre des problèmes de régression non linéaire en reformalisant le problème de manière linéaire dans un nouvel espace de représentation des données à plus grande dimension. Pour les modèles se basant sur des transformations $\phi(x)$ d'espace de représentation des données non linéaires, la fonction à noyaux est définie par la relation suivante :

$$k(x, x') = \phi(x)^T \phi(x') \tag{2.3}$$

Ainsi tous les calculs sont exprimés comme étant des produits scalaires ce qui évite de travailler directement dans l'espace de représentation transformé [SS01]. Nous

définissons deux modèles basés sur des méthodes à noyaux dans les sections 2.4.3.1 et 2.4.3.2, à savoir les processus Gaussien (GP) et les Support Vector Regressor (SVR) souvent utilisées pour résoudre des problèmes de régression non linéaire.

Le noyau RBF (Radial Basis Function ou squared exponential kernel) est le noyau le plus souvent utilisé par les modèles Gaussian Process et Support Vector Regressor. Ce noyau appliqué sur deux vecteurs $\mathbf{x}$ et $\mathbf{x}'$ représentant les caractéristiques de deux exemples d'un jeu de données, peut être défini comme suit:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{2.4}$$

Où $\|\mathbf{x} - \mathbf{x}'\|^2$ peut être vu comme la distance euclidienne quadratique entre les deux vecteurs avec σ un paramètre déterminant l'importance de chaque exemple du jeu de données.

2.4.3.1. Processus Gaussien (Gaussian Process, GP)

Les auteurs de [Ras03] ont développé un algorithme de régression non linéaire nommé Gaussian Process, une méthode d'apprentissage supervisé qui se base sur des méthodes à noyaux. L'un des avantages de ce modèle réside dans sa capacité à calculer des intervalles de confiance en plus de la prédiction. Avec ce modèle, la prédiction d'un nouveau point x est estimée à partir d'une distribution gaussienne avec la moyenne et la covariance données par les équations 2.5 et 2.6.

$$m(x) = \sum_{n=1}^{N} a_n k(x_n, x) \tag{2.5}$$

$$\sigma(x) = k(x, x) + \alpha^{-1} - k^T C_N^{-1} \mathbf{k} \tag{2.6}$$

Où N est le nombre d'échantillon du jeu de données d'entraînement, α est un hyperparamètre représentant le niveau du bruit, a_n est le $n^{\text{ème}}$ composant de $C_N^{-1} y$, avec C_N^{-1} la matrice de covariance de taille $(N \times N)$, y le vecteur de valeur cible $y = (y_1, ..., y_n)$ et $\mathbf{k}$ est le vecteur $\mathbf{k} = (k(x_1, x), ..., k(x_N, x))$.

Le bruit des valeurs observées est pris en compte et est supposé être issu d'une distribution gaussienne avec une moyenne nulle et un niveau de bruit constant α pour chaque valeur x_n. Ce niveau de bruit contribue à la variance totale de la distribution prédite donnée par l'équation 2.6. Il est important de noter que le

modèle Gaussian Process est non sparse du fait que la fonction du noyau k est évaluée pour toutes les paires possibles des valeurs observées, ce qui entraîne des temps de calcul importants pour les jeux de données volumineux, y compris lors de la prédiction.

2.4.3.2. Support Vector Regressor (SVR)

Support Vector Regressor est un modèle d'apprentissage automatique supervisé, basé sur une méthode à noyaux introduite par [Dru+97]. Pour une simple régression linéaire, l'objectif est de minimiser la fonction objectif régularisée définie par la formule 2.7 :

$$\frac{1}{2}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 + \frac{\lambda}{2}||w||^2 \tag{2.7}$$

Faisant suite à leurs travaux en classification bien connu sur les SVM (séparateurs a vaste marge), Cortes et Vapnik [CV95] ont étudié un problème d'optimisation reprenant les idées des SVM et adapté au problème de la régression. Dans le but d'obtenir une solution sparse, l'erreur quadratique est remplacée par une fonction objectif prenant en compte une marge d'erreur cible ϵ et des variables ressorts ξ (slack variables en anglais), qui permettent de relâcher la marge d'erreur autorisée pour certains exemples plus difficiles à prédire.

Avec cette fonction objectif, l'erreur augmente linéairement avec la distance au-delà de la marge. Le problème d'optimisation est alors défini comme suit:

$$\begin{aligned} \text{Minimiser} \quad & C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) + \frac{1}{2}||w||^2 \\ \text{Sous contraintes} \quad & \begin{cases} y_n \leq f(x_n) + \epsilon + \xi_n \\ y_n \geq f(x_n) - \epsilon - \hat{\xi}_n \\ \xi_n, \hat{\xi}_n \geq 0 \end{cases} \end{aligned} \tag{2.8}$$

Le paramètre C permet de choisir à quel point l'optimisation sera stricte concernant la bonne prévision (ou classification) des échantillons de données d'entrée. Pour une grande valeur de C, l'optimisation choisira un hyperplan à plus petite marge si cet hyperplan prédit correctement tous les points d'entraînement. Inversement, une très petite valeur de C amènera l'optimiseur à rechercher un hyperplan de séparation à marge plus importante, même si cet hyperplan prédit mal plus de points. En

résolvant le problème d'optimisation, on constate que la prédiction d'un nouveau point x_n peut être réalisée en utilisant l'équation suivante :

$$f(x) = \sum_{n=1}^{N} (a_n - \hat{a}_n) k(x, x_n) + b, \tag{2.9}$$

Où a_n et $\hat{a}_n$ ($0 \leq a_n \leq C$ et $0 \leq \hat{a}_n \leq C$) sont les multiplicateurs de Lagrange introduits dans les contraintes du processus d'optimisation. Les conditions de Karus-Kuhn-Tucker (KKT) correspondantes impliquent que $a_n \hat{a}_n = 0$ avec $n \in \{0, ..., N\}$ et que tous les points situées dans la marge répondent à la contrainte $a_n = \hat{a}_n = 0$.

Les articles exploitant les méthodes à noyaux sont détaillés dans le tableau 2.4.

Tab. 2.4.: Etudes exploitant des méthodes à noyaux à des fins de prévision de la mobilité urbaine

Méthode	Références
GP	[PRB15; MRP18; RMP19]
SVR, LSVR	[PRB15; RMP19; YYZ17]
SVM	[Din+16; Li+17]

Le modèle GP correspond au modèle Gaussian Process, SVR au modèle Support Vector Regressor et SVM au Support Vector Machine.

2.4.4 Méthodes à base d'arbres de décision

Les modèles à base d'arbres de décision sont souvent exploités pour répondre à des problèmes de classification ou de régression. Ces modèles ont l'avantage d'avoir une certaine explicabilité grâce par exemple à des indicateurs permettant d'étudier l'importance de chaque variable d'entrée. Avant de détailler le fonctionnement des méthodes de type Forêts aléatoires et Gradient boosting decision trees, il est important de connaître le principe d'un arbre de décision (Classification and regression tree, [Bre+84]).

Un arbre de décision est un arbre binaire où chaque noeud correspond à un test sur une unique variable d'entrée du type $x_p > c$ où c est un seuil. Les exemples du jeu de données d'apprentissage sont répartis dans les deux noeuds fils du noeud suivant le résultat de ce test. La prédiction d'un arbre est faite au niveau de ses feuilles et correspond à la moyenne des individus d'apprentissage de celle-ci. In fine, un arbre définit une partition dyadique de l'espace, et permet d'approcher la fonction cible par des constantes sur chaque élément de la partition. Pour construire l'arbre, un algorithme itératif est utilisé. Celui-ci recherche à chaque étape la variable et le seuil permettant de diminuer le plus possible l'erreur quadratique moyenne commise

par l'arbre sur le jeu de données d'apprentissage (dans le cas de la régression). Le processus est réitéré sur les nouveaux nœuds jusqu'à ce que chaque observation ait la même valeur cible ou que le fractionnement en plusieurs noeuds n'améliore plus la prévision. Les nœuds finaux sont appelés des feuilles. La figure 2.1 donne un exemple d'arbre de décision, ses noeuds sont colorés en bleu et ses feuilles en vert.

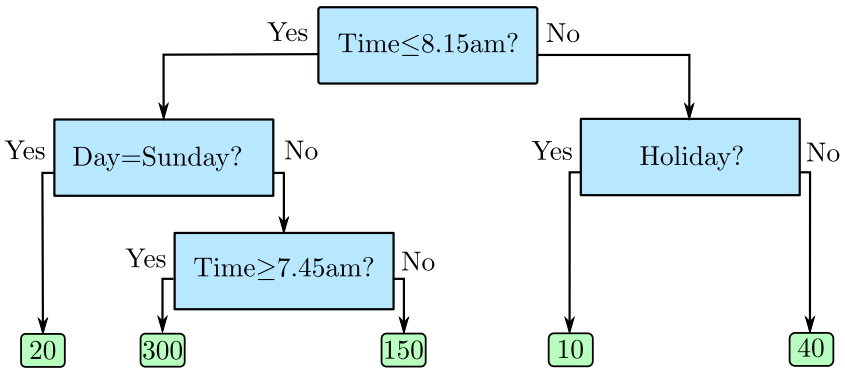

Fig. 2.1.: Schéma d'un arbre de décision dont l'objectif est de prévoir l'affluence des passagers (noeuds en bleu, feuilles en vert).

2.4.4.1. Forêts aléatoires (Random Forest, RF)

Le modèle nommé forêts aléatoires plus connu sous le nom de Random Forest en anglais (RF), est un modèle d'apprentissage automatique bien connu pour résoudre des problèmes de classification ou de régression non linéaires. Le modèle introduit par Breiman [Bre01] est un algorithme d'apprentissage de type bagging qui combine les prédictions de plusieurs arbres de décision (il est ainsi défini comme une forêt). Chaque arbre est entraîné sur différentes parties des données, qui ont été crées en appliquant deux méthodes d'échantillonnage : l'échantillonnage aléatoire avec remplacement, également connu sous le nom de méthode d'agrégation bootstrap aggregating ou bagging, et une sélection aléatoire des caractéristiques, qui est appelée features bagging. Les méthodes de bagging et le calcul de la moyenne des résultats obtenus par les différents arbres rendent les modèles RF plus robustes et précis qu'un simple arbre de décision.

2.4.4.2. Gradient Boosting Decision Trees (GBDT)

Le modèle Gradient Boosting introduit par [Fri01] est un modèle d'apprentissage automatique pour les tâches de régression ou de classification, qui utilise un ensemble de modèles de prévision (dit weak learner en anglais) comme les arbres de décision

dans notre cas pour créer un modèle de prévision robuste. Comme la plupart des autres méthodes ensemblistes, le GBDT construit plusieurs arbres de décision, à la différence près qu'il les construit un à un. Chaque nouvel arbre ayant pour but de corriger les erreurs faites par les arbres déjà formés. Après l'ajout d'un arbre, les poids des données sont réajustés. Les données d'entrée correctement prédites se voient attribuer un poids d'importance plus faible et les exemples mal prédits se voient au contraire attribuer un poids plus important.

Les articles exploitant des méthodes à base d'arbres de décision sont détaillés dans le tableau 2.5.

Tab. 2.5.: Etudes exploitant des méthodes à base d'arbres de décision à des fins de prévision de la mobilité urbaine

Méthode	Références
Regression Trees	[PRB15]
BRT, GBRT, GBDT	[Li+17; Din+16; WT16]
XGBOOST	[Ke+17; Yao+18]
RF	[Din+16]

Les modèles BRT, GBRT et GBDT correspondent aux modèles Boosted Regression Trees, Gradient Boosted Regression Trees et Gradient Boosted Decision Trees. Le modèle XGBOOST est une variante des modèles ensembliste pouvant utiliser des arbres de décision. Enfin le modèle RF correspond au Random Forest.

2.4.5 Méthodes issues de l'apprentissage profond

Les méthodes issues de l'apprentissage profond sont fondées sur le principe des réseaux de neurones et permettent de répondre à des problèmes d'apprentissage supervisé comme la classification et la régression ainsi que d'apprentissage non supervisé tel que le clustering. Ces méthodes sont facilement modulables et permettent de prendre en compte un nombre important de variables en entrée pour répondre à des problèmes de régression non linéaires. Depuis plusieurs années, ce type de modèle est fortement exploité dans les études portant sur l'analyse de la mobilité urbaine et plus principalement dans les études s'attachant à la prévision de la mobilité. Souvent considérés comme des boîtes noires, ces modèles ont un défaut bien connu qui est leur manque d'explicabilité. Pour pallier à ce problème, il est nécessaire d'étudier de manière approfondie leurs résultats de prévision. Nous détaillons trois modèles de réseau de neurones récurrent, à savoir les modèles RNN, LSTM et GRU dans les sections 2.4.5.1, 2.4.5.2 et 2.4.5.3.

2.4.5.1. Réseaux de neurones récurrents, (Recurrent Neural Networks, RNN)

Ces dernières années, les réseaux de neurones récurrents (RNN) ont été appliqués dans différents domaines allant de la traduction automatique à la reconnaissance vocale et la génération de dialogues vocaux par exemple. Ces modèles sont principalement utilisés pour leur capacité à traiter des données historiques couvrant des périodes longues, les rendant plus efficaces pour prévoir des données issues de séries temporelles. Contrairement aux réseaux de neurones classiques, les RNN considèrent que les sorties dépendent des prévisions précédentes (système effectué de manière récurrente sur chaque élément de la séquence). Pour ce faire, ils conservent en "mémoire" les observations précédentes sous la forme d'une couche contenant l'état actuel de la série (couche cachée ou hidden state en anglais) constamment mise à jour. La couche cachée h_t au temps t est calculée sur la base de l'état précédent h_{t-1} et de l'entrée (l'observation) fournie à l'étape courante x_t en utilisant une fonction non linéaire f définie dans l'équation 2.10.

$$h_t = f(Ux_t + Wh_{t-1}) \ , \tag{2.10}$$

où U et W sont des matrices de poids affectées à x_t et h_{t-1} respectivement.

La sortie $\hat{y}_t$ est calculée à partir de cette couche cachée à l'aide d'une fonction g définie dans l'équation 2.11 qui suit :

$$\hat{y}_t = g(Vh_t) \ , \tag{2.11}$$

où V est aussi une matrice de poids.

Le choix d'une fonction appropriée g dépend fortement du type de problème à résoudre (classification ou régression). Différentes fonctions peuvent être utilisées (ReLU, linéaire, softmax, softplus, etc.). Les paramètres à ajuster pendant la phase d'apprentissage des RNN sont les matrices U, V et W.

La structure d'un RNN et d'un RNN sous sa forme déployée est donnée dans la figure 2.2. Cette figure s'inspire de la vulgarisation [Ola15] des réseaux de neurones récurrents menée par Olah en 2015.

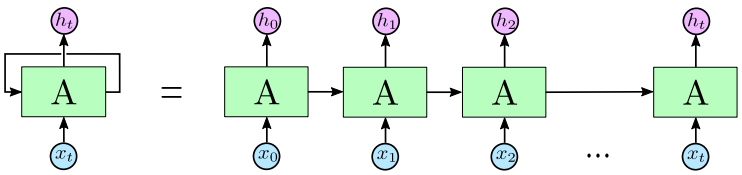

Fig. 2.2.: Schéma d'un RNN et d'un RNN déployé.

2.4.5.2. Long Short-Term Memory, LSTM

En pratique, les RNN souffrent de leur incapacité à mémoriser des informations sur de longues périodes de temps. Pour pallier à ce problème d'autres architectures de réseau de neurones récurrent ont été proposées (e.g. LSTM, GRU, etc.), notamment le Long Short-Term Memory (LSTM) développé par Hochreiter et Schmidhuber [HS97] en 1997. Afin de pouvoir conserver l'information sur de plus longues périodes de temps, le LSTM introduit (en plus de la couche cachée) une cellule mémoire possédant un système de porte (gate en anglais). Ces portes ont pour but de réguler la propagation de l'information afin de contrer le phénomène de la disparition du gradient, vanishing gradient en anglais [PMB13] qui peut survenir pendant l'étape d'apprentissage. Ce phénomène de vanishing gradient est en général responsable du fait que les modèles RNN basiques ne puissent pas apprendre correctement. Les équations du modèles LSTM sont définies dans l'ensemble d'équations 2.12.

Les équations du modèles LSTM sont définies dans l'ensemble d'équation 2.12:

$$\begin{aligned} F_t &= \sigma(b^F + x_t U^F + h_{t-1} W^F) \\ I_t &= \sigma(b^I + x_t U^I + h_{t-1} W^I) \\ O_t &= \sigma(b^O + x_t U^O + h_{t-1} W^O) \\ c_t &= F_t c_{t-1} + I_t tanh(b + x_t U + h_{t-1} W) \\ h_t &= tanh(c_t) O_t \\ \hat{y}_t &= g(b^{\hat{y}} + h_t W^{\hat{y}}) \end{aligned} \tag{2.12}$$

Où F correspond à la "forget gate", la gate d'oubli qui permet de mettre à jour la cellule mémoire c , I et O correspondent aux "gates" d'entrée et de sortie permettant de laisser entrer les informations en entrée et en sortie du LSTM basées sur les données d'entrée x et h_{t-1} qui correspond à l'état caché (ou hidden state en anglais), le vecteur en sortie du LSTM au pas de temps $t-1$. Les vecteurs U et W correspondent aux poids associés aux données d'entrée et aux poids récurrents associés à la couche cachée du pas de temps $t-1$. g correspond à une fonction d'activation permettant d'obtenir $\hat{y}_t$, la prévision ou classification du LSTM au pas de temps t.

Le schéma d'un LSTM est donné en figure 2.3.

2.4.5.3. Gated Recurrent Unit, GRU

Les modèles Gated Recurrent Unit introduits par [Cho+14], sont des mécanismes spéciaux (nommé gate mechanism en anglais) dans les réseaux de neurones récur-

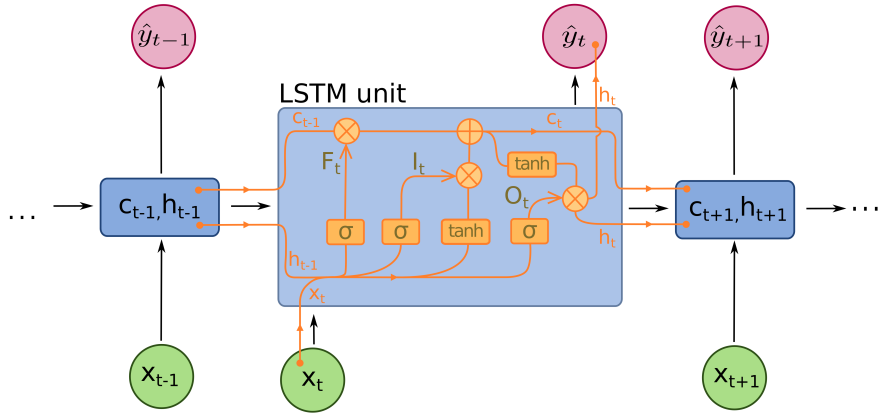

Fig. 2.3.: Schéma d'un Long-Short Term Memory. Les symboles σ et $tanh$ représentent respectivement la fonction sigmoïde et la fonction tangente hyperbolique. Source [Del17b], avec modification du label des sorties $\hat{y}_t$.

rents qui visent à bloquer ou laisser passer l'information qui passe au travers du réseau de neurones. Les modèles GRU, ainsi que les LSTM [HS97], sont des types spéciaux de RNN capables d'apprendre les dépendances temporelles à long terme au moyen d'un mécanisme qui empêche le problème de la disparition du gradient ou vanishing gradient en anglais [PMB13] associé aux modèles RNN. Les GRU sont composés d'un mécanisme de porte ou gate en anglais plus simple que ceux des LSTM ce qui permet un temps de calcul plus rapide. Actuellement, les GRU sont des modèles à l'état de l'art pour les problèmes impliquant l'analyse de séries temporelles (e.g., génération de dialogue vocaux, génération de texte, traduction automatique, etc.). Les équations du modèles GRU sont définies dans l'ensemble d'équation 2.13

$$
\begin{aligned}
Z_t &= \sigma(x_t U^Z + h_{t-1} W^Z) \\
R_t &= \sigma(x_t U^R + h_{t-1} W^R) \\
\hat{h}_t &= \tanh\left(x_t U^h + (R_t * h_{t-1}) W^h\right) \\
h_t &= (1 - Z_t) * h_{t-1} + Z_t * \hat{h}_t \\
\hat{y}_t &= g(b^{\hat{y}} + h_t W^{\hat{y}})
\end{aligned}
\tag{2.13}
$$

Avec R la gate de réinitialisation, Z la gate de mise à jour, W la connexion récurrente entre la couche cachée du pas de temps précédent et actuel, σ une fonction sigmoïde, U la matrice de poids permettant de lier les entrées et la couche cachée actuelle et h_t la couche cachée du modèle GRU au pas de temps t. g correspond à une fonction d'activation permettant d'obtenir $\hat{y}_t$, la prévision ou classification au pas de temps t. Le schéma d'un GRU est donné en figure 2.4.

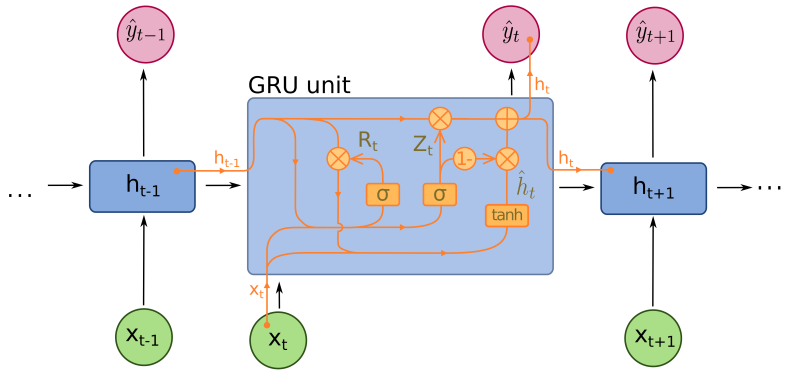

Fig. 2.4.: Schéma d'un Gated Recurrent Unit. Les symboles σ et $tanh$ représentent respectivement la fonction sigmoïde et la fonction tangente hyperbolique. Source [Del17a], avec modification du label des sorties $\hat{y}_t$ et ajout de l'élément $\hat{h}_t$.

Les articles exploitant des méthodes issues de l'apprentissage profond sont détaillés dans le tableau 2.6.

Tab. 2.6.: Etudes exploitant des méthodes issues de l'apprentissage profond à des fins de prévision de la mobilité urbaine

Méthode	Références
NN	[PRB15; Cui+16; Din+16; Ke+17; WT16; Yao+18; WT16; Che+17; YYZ17]
RBF	[Li+17]
LSTM	[Ke+17; WT16; RMP19]
CNN	[Ke+17; RMP19]
CNN+LSTM	[Ke+17; WT16; Che+17; YYZ17; Yao+18]
SAE	[WT16; Che+17]
GC-GRU	[YYZ17]
ST-NN	[ZZQ17]
ST-GCN	[YYZ17]

Le modèle NN, Neural Network, correspond au modèle de réseau de neurones classique, RBF correspond au modèle Radial Basis Function, un type de réseau de neurones. LSTM correspond au modèle Long Short Term Memory un type de réseau de neurones récurrent, CNN correspond aux réseaux de neurones à convolution. Les modèles ST-NN et ST-GCN correspondent aux modèles spatio-temorels : Spatio Temporal Neural Network et Spatio Temporal Graphic Convolutional Network. Le modèle SAE correspond à la méthode Stacked Auto Encoder. Enfin le modèle GC-GRU correspond au modèle Graphic Convolutional Gated Recurrent Unit.

2.5 Processus de prévision de séries temporelles à l'aide d'algorithmes issus de l'apprentissage automatique

La mise en place des méthodes d'apprentissage automatique peut se résumer à deux étapes : (i) la mise en forme des données et l'apprentissage des différents modèles (e.g., modèles de classification, prévision, etc.) et (ii) l'évaluation de ces méthodes. Nous décrivons le processus de sélection des hyperparamètres et l'utilisation d'un jeu de données de validation, utiles à l'apprentissage des modèles issus de l'apprentissage automatique dans la section 2.5.1. Par la suite, nous expliquons le processus d'évaluation des modèles de prévision dans la section 2.5.2. Enfin nous détaillons dans la section 2.5.3 les librairies les plus connues permettant de mettre en forme les données et de créer des algorithmes d'apprentissage automatique.

2.5.1 Hyperparamètres et jeu de données de validation

Le processus commun d'apprentissage des méthodes issues de l'apprentissage supervisé (méthodes de classification et de prévision) correspond au fait de minimiser une fonction de coût sur un jeu de données d'entraînement (train set en anglais). Ceci, dans le but d'obtenir les meilleures performances possibles sur de nouvelles données. Pour estimer ces performances, on utilise classiquement un jeu de données de test (test set en anglais) qui correspond à un ensemble de données différent de celui d'entraînement pour éviter les biais liés au sur-apprentissage des données d'entraînement.

La plupart des algorithmes d'apprentissage automatique ont des hyperparamètres que l'on peut utiliser pour contrôler le comportement des algorithmes. Contrairement aux paramètres, les valeurs des hyperparamètres ne sont pas fixées par l'algorithme d'apprentissage lui-même. Ces hyperparamètres permettent par exemple de contrôler la complexité de la classe de fonctions considérées pour approcher les données. Ceux-ci ne peuvent donc pas être déterminés en étudiant l'erreur obtenue sur le jeu de données d'apprentissage car cela conduirait à toujours choisir la classe de fonction la plus complexe, ce qui peut aboutir à un sur-apprentissage (modèle qui obtient de bonnes performances sur le jeu d'apprentissage mais qui ne réussit pas à bien s'adapter à de nouvelles données). Pour résoudre ce problème, il est courant d'utiliser un ensemble d'exemples de validation (auquel l'algorithme d'apprentissage n'a pas accès) qui n'est utilisé que pour déterminer la meilleure valeur des

hyperparamètres en étudiant les résultats obtenus en fonction des hyperparamètres utilisés sur ces données. Lorsque l'ensemble des valeurs possibles considérées pour les hyperparamètres forment une grille, on parle de procédure grid-search.

Le processus d'apprentissage d'un modèle nommé M peut être détaillé en 5 étapes :

1. Créer un ou plusieurs partitionnements du jeu d'apprentissage en un sous ensemble d'apprentissage et un ensemble de validation
2. Pour chaque sous ensemble d'apprentissage, entraîner le modèle M avec chaque combinaison d'hyperparamètres que nous voulons tester (ce choix est fait manuellement)
3. Comparer les résultats obtenus par chaque combinaison d'hyperparamètres sur les ensembles de validation et sélectionner la combinaison d'hyperparamètres qui permet d'obtenir les meilleures performances
4. Apprendre le modèle M sur l'ensemble du jeu d'apprentissage avec la combinaison d'hyperparamètres sélectionnée
5. Réaliser la prévision (ou classification) sur le jeu de test pour obtenir un résultat final afin de le comparer avec celui obtenu par d'autres modèles

Enfin, notons que diviser l'ensemble du jeu de données en un ensemble d'apprentissage et un ensemble de validation peut s'avérer problématique si l'ensemble de validation contient peu d'exemples. En effet, un ensemble de validation avec peu d'exemple implique une incertitude importante sur l'estimation de l'erreur. Dans ce cas, d'autres procédures permettent de réduire cette incertitude, au prix d'un coût de calcul plus important. Ces procédures sont basées sur l'idée de répéter l'apprentissage et la validation sur différents sous ensembles ou des partitionnements choisis aléatoirement de l'ensemble de données d'apprentissage original. L'une des méthodes les plus courante, est la procédure de validation croisée en k sous ensembles (k-fold cross validation), dans laquelle une partition de l'ensemble de données est formée en le divisant en k sous ensembles qui ne se chevauchent pas. L'erreur de validation peut ensuite être estimée en prenant la moyenne de l'erreur commises sur les k sous ensembles de données par l'algorithme lorsqu'il utilise les données restantes lors de l'apprentissage.

2.5.2 Evaluation des modèles de prévision

Comme nous l'avons vu précédemment, dans le cas d'un problème classique de prévision, il est nécessaire de séparer le jeu de données d'apprentissage de celui de

test. Le jeu d'apprentissage permet aux modèles d'apprendre leurs paramètres et de sélectionner les meilleurs hyperparamètres en divisant ce jeu d'apprentissage en un jeu d'apprentissage et de validation.

2.5.2.1. Découpage du jeu de données pour la prévision de séries temporelles

Dans le cas de la prévision de séries temporelles, il est nécessaire de prendre en considération le fait que les données sont ordonnées. Pour ne pas casser cet ordre, la méthode la plus simple consiste à utiliser les données les plus anciennes comme jeu d'apprentissage et de validation (avec un mélange aléatoire des échantillons, pour ne pas baser l'optimisation des modèles sur une seule période) et d'utiliser les données les plus récentes pour tester les modèles. Selon les auteurs de l'étude [HA15], il existent dans le cas d'une simulation de mise à jour des données d'apprentissage, deux manières pour diviser un jeu de données en une partie d'apprentissage et une partie de test. Les deux méthodes permettant de diviser un jeu de données correspondant à des séries temporelles pour simuler la mise à jour des modèles, sont détaillées dans la figure 2.5.

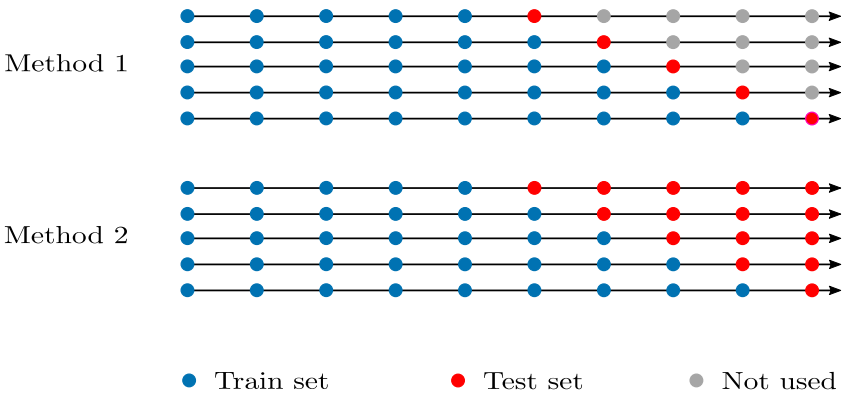

Fig. 2.5.: Méthodes de découpage du jeu de données pour la prévision de séries temporelles

La première méthode qui se rapproche le plus d'un processus réel correspond à créer plusieurs échantillons temporels de différentes tailles en rajoutant une partie plus récente des données dans chaque échantillon. Cette méthode simule la mise à jour récurrente du jeu de données d'entraînement et de test dans le cas où les données seraient disponibles en continu. La deuxième méthode consiste à utiliser l'ensemble du jeu de données en créant plusieurs échantillons d'apprentissage et de test avec une quantité plus ou moins grande de données d'apprentissage. Pour chacune de ces

deux méthodes de découpage du jeu de données, les données du jeu d'apprentissage restent plus anciennes que celles du jeu de test.

2.5.2.2. Métriques d'évaluation

Il existe différentes métriques d'évaluation permettant de comparer et d'évaluer un ou plusieurs modèles d'apprentissage (e.g., classification, prévision). Nous détaillons dans les sections qui suivent les métriques d'évaluation les plus communément utilisées pour l'évaluation de modèles de prévision de séries temporelles. Les équations des différentes métriques d'évaluation (équations 2.14, 2.15, 2.16 et 2.17), se placent dans le cas d'une prévision multivariée (plusieurs séries temporelles), avec S le nombre de séries temporelles, T le nombre de pas de temps étudiées, $\hat{y}_t(s)$ la valeur prévue de la série temporelle s au pas de temps t et y la valeur observée.

Racine carrée de l'erreur quadratique moyenne (Root Mean Square Error, RMSE) Cette erreur est souvent utilisée comme erreur de référence. L'effet de chaque erreur sur la RMSE est proportionnelle à l'importance de l'erreur quadratique ; ainsi, les erreurs plus importantes ont un effet disproportionné sur la RMSE. Par conséquent, la RMSE est sensible aux valeurs aberrantes (outliers). Le calcul de la RMSE peut être exprimé comme suit :

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{s=1}^{S}\sum_{t=1}^{T}(\hat{y}_t(s) - y_t(s))^2}{T \times S}} \tag{2.14}$$

L'erreur quadratique moyenne (Mean Square Error, MSE) Tout comme la RMSE, l'erreur MSE est proportionnelle à l'importance de l'erreur quadratique ; ainsi, les erreurs plus importantes ont un effet disproportionné sur la cette métrique. L'équation de la MSE est la suivante :

$$\mathrm{MSE} = \frac{1}{T * S}\sum_{i=1}^{S}\sum_{t=1}^{T}(\hat{y}_t(s) - y_t(s))^2 \ , \tag{2.15}$$

L'erreur moyenne absolue (Mean Absolute Error, MAE) L'erreur moyenne absolue permet d'avoir une référence plus proche des données concernant l'écart moyen entre les valeurs observées et prédites. Le calcul de la MAE est le suivant :

$$\mathrm{MAE} = \frac{1}{T \times S} \sum_{s=1}^{S} \sum_{t=1}^{T} |\hat{y}_t(s) - y_t(s)| \tag{2.16}$$

Le pourcentage d'erreur moyenne absolue avec seuil (Mean Absolute Percentage Error at v, MAPE@v) La MAPE@v est l'erreur la plus facile à se représenter. En effet, cette erreur correspond au pourcentage d'erreur moyen calculé pour les valeurs observées supérieures ou égales à un seuil v. Le calcul de l'erreur MAPE@v est défini dans l'équation qui suit :

$$\forall \mathrm{y_s(t)} > \mathrm{v},\ \mathrm{MAPE@v} = \frac{100}{T \times S} \times \sum_{s=1}^{S} \sum_{t=1}^{T} \left| \frac{y_t(s) - \hat{y}_t(s)}{y_t(s)} \right| \tag{2.17}$$

2.5.3 Outils de gestion de données et de création d'algorithmes d'apprentissage

Suite à l'essor du domaine de la science des données, de nombreuses librairies informatiques ont vu le jour. Ces outils permettent de mettre en place la gestion de données (section 2.5.3.1) ainsi que la création de différentes méthodes d'apprentissage (section 2.5.3.2, section 2.5.3.4 et section 2.5.3.3). Les librairies les plus connues dans le domaine de la science des données proviennent des langages de programmation Java, Scala, Matlab, R et plus principalement Python.

2.5.3.1. Librairies utilisées pour la gestion et l'analyse de données

La bibliothèque la plus connue et utilisée pour la gestion et l'analyse de données est une librairie Python nommée pandas [McK+10]. Cette bibliothèque permet de facilement analyser des données sous la forme de tableaux et de réaliser des opérations sur ces données telles que des fonctions mathématiques et des fonctions de tri. Spark quant à elle est la bibliothèque la plus utilisée pour la gestion de large base de données, elle peut être utilisée avec les langages Java, Scala, R et Python. Le tableau 2.7 références les librairies les plus exploitées pour la gestion et l'analyse de données.

Tab. 2.7.: Exemple de librairies utilisées pour l'analyse de données

Librairie	Références	Java	Scala	R	Python
Numpy	[VCV11; Oli06]				✓
Scipy	[J+01]				✓
pandas	[McK+10]				✓
dplyr	[Wic+15]			✓	
Spark	[Zah+10b]	✓	✓	✓	✓

2.5.3.2. Librairies de statistiques

Chacun des langages de programmation appartenant à la liste suivante ; Java, Scala, Matlab, R et Python, possède au moins une librairie permettant de mettre en place des algorithmes statistiques. Les bibliothèques les plus connues sont énumérées dans le tableau 2.8. Les langages de programmation les plus utilisés dans le domaine des statistiques sont les langages R et Python.

Tab. 2.8.: Exemple de librairies statistiques

Librairie	Références	Java	Scala	Matlab	R	Python
StatsModels	[SP10]					✓
R Stats Package	[Tea+13]				✓	
Statistics and ML Toolbox	[MM15]			✓		
Smile	[Li16]	✓	✓			
Apache Commons Math	[Mat16]	✓				

2.5.3.3. Librairies d'apprentissage profond

Il est possible d'implémenter des algorithmes d'apprentissage profond avec une grande quantité de langages de programmation. Néanmoins le langage Python est le langage le plus utilisé pour ce type d'étude. Les librairies les plus utilisées dans ce domaine sont les librairies Keras [Cho+15], Tensorflow [Aba+16] et Pytorch [Pas+17]. Le tableau 2.9 référence les bibliothèques d'apprentissage profond les plus utilisées par type de langage de programmation.

Tab. 2.9.: Exemple de librairies spécialisées en apprentissage profond

Librairie	Références	Java	Scala	Matlab	R	Python
Keras	[Cho+15]				✓	✓
Tensorflow	[Aba+16]					✓
Pytorch	[Pas+17]					✓
Deeplearning4j	[Tea16]	✓	✓			
Matconvnet	[VL15]			✓		

2.5.3.4. Librairies d'apprentissage automatique

La librairie la plus exploitée dans le domaine de l'apprentissage automatique est la bibliothèque Python nommée scikit-learn. Cette bibliothèque permet de mettre en place un large éventail de méthodes d'apprentissage automatique en plus de faciliter certains processus tel que l'optimisation d'hyperparamètres. Le tableau 2.10 détaille les librairies d'apprentissage automatique les plus connues pour différents langages de programmation.

Tab. 2.10.: Exemple de librairies spécialisées en apprentissage automatique

Librairie	Références	Java	Scala	Matlab	R	Python
scikit-learn	[Ped+11]					✓
XGBOOST	[CG16]				✓	✓
Caret	[Kuh+08]				✓	
Prophet	[TL18]				✓	✓
Statistics and ML Toolbox	[MM15]			✓		
H2O Sparkling Water	[H2O19]		✓		✓	✓
Smile	[Li16]	✓	✓			
Apache Spark MLib& ML	[Men+16]		✓			
Weka	[Hal+09]	✓				
Deeplearning4j	[Tea16]	✓	✓			
Mallet	[McC02]	✓				

2.6 Conclusion

Dans ce chapitre, nous avons premièrement détaillé différentes études portant sur la prévision long et court terme des flux d'individus utilisant différents moyens de déplacement (e.g., marche, bus, métro, taxi, etc.). En deuxième partie nous avons détaillé certains modèles statistiques ainsi que des modèles issus de l'apprentissage automatique (machine learning) et de l'apprentissage profond (deep learning) souvent utilisés dans les études portant sur la prévision des flux. Nous allons voir dans les chapitres suivants comment nous avons mobilisé ces différentes méthodes statistiques, de machine learning et de deep learning pour la prévision long terme de la demande de passagers à l'aide de données externes telles que des données événementielles, dans le chapitre 3. Nous verrons ensuite, dans le chapitre 4, comment certaines de ces méthodes permettent de réaliser des prévisions court terme de matrices Origine Destination et du nombre de passagers entrant dans les différentes stations d'un réseau de transport en commun.

3 Prévision long terme de l'affluence des passagers avec prise en compte de données événementielles

3.1 Résumé

Les opérateurs de transport en commun ont pour principal objectif d'adapter l'offre de transport à la demande passager dans les réseaux de transport existants et ce, à n'importe quelle période de l'année et de la journée. Dans certains pays, un autre problème mentionné par les opérateurs de transport est la difficulté à adapter la disponibilité des titres de transport spéciaux (jetables) à la demande des passagers. Dans ce contexte, l'estimation et la prévision de l'affluence s'avèrent être des enjeux majeurs pour améliorer les systèmes de transport en commun. La difficulté principale réside dans le fait que l'affluence des passagers dépend de multiples facteurs qui rendent la tâche de prévision difficile. Dans ce chapitre, nous proposons une mise en forme générique des données, permettant l'utilisation de modèles de régression bien connus (moyenne historique, des modèles statistiques et des modèles d'apprentissage automatique) pour la prévision à long terme de la demande des passagers avec une résolution temporelle fine. Plus précisément, nous nous attachons ici à prévoir jusqu'à un an à l'avance le nombre de passagers entrant dans chaque station d'un réseau de transport avec une agrégation temporelle d'un quart d'heure en tenant compte des événements prévus (e.g. concerts, match de hockey, etc.). Pour résoudre le problème de la disponibilité des titres de transport, nous prévoyons la demande des passagers par type de titres de transport en plus de la prévision globale (agrégation de tous les titres). L'évaluation des modèles et de la qualité des prévisions est ensuite menée au travers de deux ensembles de données réelles: un jeu de données billettiques et une base de données événementielles de la ville de Montréal au Canada. Ces ensembles de données couvrent une période totale de trois ans et représentent le nombre de passagers entrant dans chaque station de métro ainsi que plusieurs informations concernant les événements (e.g., horaire de début, horaire de fin, catégorie de l'événement) connues un an à l'avance. Les

meilleurs résultats de prévision sont obtenus avec le modèle de forêts aléatoires (modèle Random Forest).

3.2 Introduction

L'un des principaux objectifs des parties prenantes (opérateurs et autorités de transport) est d'adapter le plus précisément possible l'offre de transport à la demande des passagers, quelque soit la période (e.g., période normale, période perturbée, journée spéciale, période contenant la présence d'un ou plusieurs événements sportifs, culturels, etc.). Pour certains opérateurs de transport, un autre objectif est d'anticiper la demande de titres de transport jetables (cartes à puce non rechargeables) dans le but de faire correspondre le nombre de titres à celui des passagers à une période donnée, en particulier en période d'événements (e.g., concert, rencontre sportive, spectacle, exposition, etc.) lorsque l'affluence est importante.

La disponibilité des traces numériques, concernant dans notre cas les données billettiques combinées à la disponibilité d'une base de données événementielles, offre la possibilité de développer des outils de prise de décisions qui peuvent permettre aux opérateurs de transport de mieux comprendre et de mieux prévoir la demande des passagers dans les grandes villes. De cette façon, les opérateurs seront en mesure d'améliorer le niveau de service de l'offre de transport. D'autre part, les citoyens pourront profiter d'une tarification spécifique pendant les périodes d'événements, bénéficier d'une réduction du temps d'attente de leur moyen de déplacement (e.g., bus, métro, tramway, etc.) et planifier leurs déplacements de manière plus efficace afin d'éviter les périodes de congestion grâce aux informations fournies concernant la demande passager prévue.

Pour répondre à ces questions, nous proposons une mise en forme générique des données contextuelles (événementielles et calendaires), permettant d'utiliser des modèles de régression bien connus pour résoudre la prévision long terme de l'affluence des passagers avec une résolution temporelle fine. Les modèles développés seront ainsi en mesure de prévoir le nombre de passagers entrant dans chacune des 68 stations de métro de la ville de Montréal, au Canada, jusqu'à un an à l'avance en tenant compte des données calendaires et des événements prévus. Nous prévoyons également l'affluence des passagers par type de titres de transport afin de résoudre le problème de l'ajustement du nombre de titres à la demande des passager. La fenêtre temporelle d'agrégation du nombre de passagers a été choisie comme étant égale à 15 minutes, ce qui permet une analyse précise de l'impact des événements

sur la demande des passagers tout en étant un choix "réaliste" pour l'adaptation de l'offre de transport par les opérateurs. Nous comparons plusieurs modèles de prévision bien connus, y compris des modèles simples, statistiques et issus de l'apprentissage automatique. Dans ce contexte, nous analysons l'utilisation de données contextuelles telles que les informations sur le type de jour ainsi qu'une base de données événementielles fournie par la Société de Transport de Montréal (STM) dans le modèle de prévision. Cette méthodologie vise à être reproductible dans le but de pouvoir prévoir la demande passager dans d'autres réseaux de transport du monde (à condition de disposer d'ensembles de données équivalents dans les autres villes, à savoir des données calendaires et événementielles).

Il est important de noter que cette approche ne peut être appliquée que sur des réseaux de transport déjà existants. Pour être efficace, l'approche exige un ensemble de données événementielles historiques qui comprend la fréquence des événements à chaque station comprenant des événements; autrement, le modèle de prévision ne sera pas en mesure de prendre en compte l'information sur les événements à une station qui n'a jamais accueilli d'événement dans la base de données historiques.

Le reste du chapitre est organisé comme suit. L'étude de cas et les données sont présentées dans la section 3.3. La section 3.4 détaille les méthodes de prévision et la mise en forme des données que nous avons développées. La section 3.5.1 décrit les résultats des prévisions sur l'agrégation globale des types de titres de transport, tandis que la section 3.5.2 correspond à une analyse des résultats de prévision par type de titres de transport. Enfin, la section 3.6 présente une conclusion de ces travaux ainsi que quelques pistes de recherches futures.

3.3 Cas d'étude : réseau de métro de Montréal

La prévision de l'affluence des passagers dans un réseau de transport en commun est une tâche difficile. Ceci est principalement dû aux différents facteurs auxquels l'affluence des passagers dépend. Ces facteurs sont néanmoins bien connus et ont été introduit par [ZZQ17] dans une étude portant sur la prévision de flux de personnes dans une ville. Ces facteurs peuvent être définis comme suit. (i) Le facteur temporel, y compris l'intervalle de temps et le type de la journée (e.g., lundi, mardi, etc.), vacances scolaires, jours fériés, etc. (ii) Le facteur spatial tel que le type de zone où la demande est prévue (e.g., zone résidentielle, d'affaires, de commerce, de spectacle, etc.). (iii) Les facteurs qui peuvent être prévus plusieurs heures voire jours en avance tels que les conditions météorologiques, les événements (e.g., sportifs,

culturels, etc.), les grèves des transporteurs et les rénovations. (iv) Enfin, les facteurs imprévisibles tels que les perturbations du réseau de transport pouvant être causées par un problème technique (panne, incendie, etc.), un incident passager ou tout autre facteur qui pourrait avoir un impact important sur l'offre de transport.

3.3.1 Données billettiques

L'ensemble des données réelles utilisées pour évaluer la méthodologie proposée a été fourni par la STM. Nous avons travaillé sur la prévision du nombre de passagers entrant dans chacune des 68 stations composant les 4 lignes de métro de la ville de Montréal. Les données billettiques utilisées ont été agrégées par intervalles de 15 minutes et s'étalent sur une période de trois ans: 2015, 2016 et 2017. Le réseau de métro étudié accueille plus de 670 000 passagers par jour en moyenne. Nous avons également effectué la prévision du nombre de passagers par type de titres de transport. Pour cela, nous avons regroupé le nombre de passagers entrants en fonction de leur type de titres de transport : abonnement mensuel STM (STM monthly pass), abonnement mensuel régional (Regional monthly pass), tickets jetables (Book tickets) et titres occasionnels (Occasional pass).

Dans la Figure 3.1, nous pouvons voir l'utilisation des différents types de titres de transport indiquée en pourcentage au cours de la période 2015-2017. Le type de titres le plus utilisé correspond aux abonnements mensuels de la STM avec environ 140 millions d'entrées par an, ce qui représente 58% de la demande de passagers.

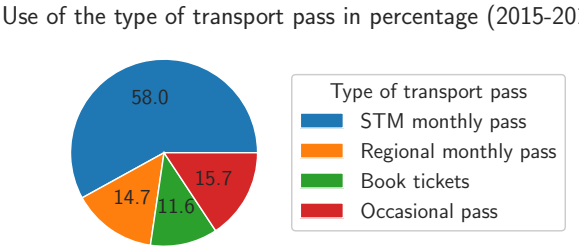

Fig. 3.1.: Pourcentage du nombre de passagers entrant dans le réseau de métro de Montréal par type de titres de transport durant la période 2015-2017.

3.3.2 Données calendaires

La demande passager dépend principalement du type de jour, c'est pourquoi nous avons créé une liste de neuf caractéristiques pouvant le définir avec précision :

- Jour de la semaine (e.g., lundi, mardi, etc.)
- Mois (e.g., janvier, février, etc.)
- Jour férié (e.g., Noël, jour de l'An, etc.)
- 24 décembre
- 31 décembre
- Vacances de Noël
- Vacances universitaires d'été 1 (session intensive, Université de Montréal)
- Vacances universitaires d'été 2 (session ordinaire, Université de Montréal)
- Période de rénovation qui a eu lieu à la station Beaubien pendant 4 mois en 2015

Cette liste de caractéristiques est spécifique à ce réseau de transport et à cette ville, mais elle pourrait facilement être modifiée pour être adaptée à un autre type de réseau de transport et à une autre ville.

3.3.3 Données événementielles

La demande passager dépend également de différents facteurs contextuels, lesquels peuvent être regroupés en trois sous parties, (i) les facteurs non prévisibles tels que les incidents impactant l'offre de transport (e.g., panne, incident passager, etc.), (ii) les facteurs prévisibles à court terme comme par exemple les conditions météorologiques et enfin (iii) les facteurs pouvant être connus plusieurs jours ou mois à l'avance tels que les événements du type rencontre sportive, festival, concert, etc. Il est possible de créer manuellement ou même d'extraire automatiquement une telle base de données événementielles comme nous pouvons le voir dans les travaux développés par [Mor+16; MRP18; RMP19] à condition d'avoir connaissance des sites internet qui possèdent de telles informations. Dans notre cas, ces données événementielles enregistrées manuellement pendant trois années, nous ont été fournies par la STM. Dans cette base de données, chaque événement est caractérisé par un lieu (la station de métro la plus proche), un horaire de début et de fin (le format de l'horaire est "Y-m-d H:M:M:S", environ 80% des horaires de fin d'événement sont disponibles) ainsi qu'une brève description de l'événement rédigée manuellement (la description ne suit pas le même modèle de construction, par exemple, un événement peut avoir différentes descriptions). Nous avons défini manuellement 10 catégories d'événements afin de pouvoir prendre en compte le type d'événement lors de la prévision. La prise en compte de ce type de données s'avère être un véritable défi étant donné qu'elle implique l'utilisation de larges données éparses (matrices creuses) qui peuvent être difficiles à traiter par des modèles de régression. De plus, l'horaire de fin n'est pas disponible pour chaque événement, ce qui rend l'interprétation de l'événement difficile par les modèles de prévision.

Dans la Figure 3.2 nous pouvons observer le nombre d'événements par station et par catégorie. Un événement est considéré s'il correspond à la présence d'un horaire de début dans la base de données (e.g., si le même événement se produit 4 jours de suite et qu'il est représenté par 4 horaires de début dans la base de données, on considérera qu'il existe 4 événements). Nous pouvons voir que la plupart des événements se déroulent à proximité de trois stations : Lucien-L'Allier et Bonaventure proches du Centre Bell qui est une salle omnisport accueillant des rencontres sportives et des concerts, et la station Jean-Drapeau située sur une île connue pour accueillir plusieurs festivals et événements divers.

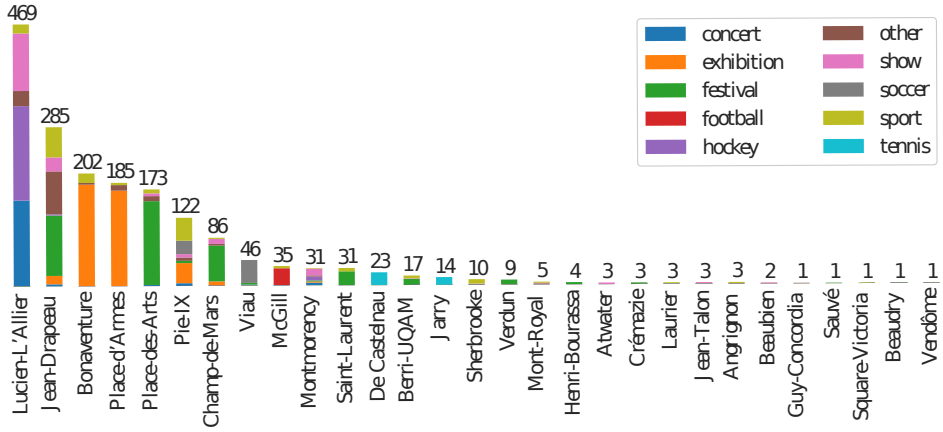

Fig. 3.2.: Nombre d'événements par catégorie et par station accueillant des événements (2015, 2016, 2017).

Dans le but d'obtenir une vue d'ensemble des données billettiques et de voir l'impact du type de jour sur le nombre de passagers entrants, nous avons représenté le nombre de passagers sur trois lundis du même mois (avril 2017) à la station nommée "Lucien-L'Allier" dans la Figure 3.3. Le lundi 3 avril 2017 est représenté par la courbe verte, il correspond à un lundi normal. Nous pouvons observer les pics d'affluence typiques du matin et du soir. Le lundi 10 avril 2017 est représenté en couleur orange, cette journée est considérée comme spéciale car un événement (Concert du groupe Def Leppard qui s'est terminé à 23h) a lieu ce jour-là, proche de cette station. Enfin, le lundi 17 avril 2017, est un jour férié (lundi de Pâques), il est représenté par la courbe de couleur bleue. On observe une diminution de la demande de passagers tout au long de la journée du lundi de Pâques (courbe bleue) par rapport au lundi sans événement (courbe verte). D'autre part, nous pouvons constater une très forte augmentation de l'affluence des passagers à la fin du concert ayant lieu le lundi représenté en orange.

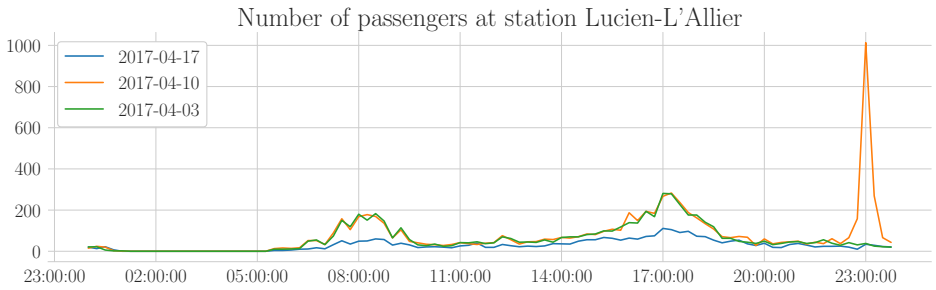

Fig. 3.3.: Nombre de passagers entrant à la station "Lucien L'Allier" lors de trois lundi du mois d'avril 2017.

3.4 Processus de la prévision long terme

Notre objectif est de prévoir le nombre de passagers entrant à chaque instant de la journée dans chaque station d'un réseau de transport jusqu'à un an à l'avance en prenant en compte les événements prévus (e.g., sportifs, culturels, etc.). Nous avons comparé l'utilisation de différents jeux de données en entrée des modèles de prévision ainsi que différents types de modèles de prévision. La section 3.4.1 détaille la mise en forme des données et l'ensemble des données d'entrée comparées. Les méthodes de prévision comparées sont décrites en section 3.4.2. Enfin, les méthodes d'évaluation sont décrites dans la section 3.4.3.

3.4.1 Mise en forme des données pour la prévision de l'affluence

Afin d'évaluer l'importance de chaque ensemble de données contextuelles, nous avons entraîné les modèles de prévision avec quatre groupes de données distincts (D1, D2, D3 et D4). Chacun de ces groupes de données correspond à une concaténation spécifique des 4 ensembles de caractéristiques suivants :

- A : Mois et jour de la semaine, encodage one-hot.
- B : Vacances, 24 et 31 décembre, vacances scolaires de Noël, vacances universitaires 1 et 2, période de rénovation de la station Beaubien. Ces caractéristiques ont été encodées dans des vecteurs one-hot.
- C : Information sur les horaires de l'événement à savoir l'horaire de début, de fin de l'événement ainsi que la période de l'événement sous condition que l'information de fin d'événement soit disponible, en chaque station qui accueille au moins un événement (29 stations). Pour chaque station accueillant un événement, à chaque pas de temps de la journée, nous avons compté le

nombre d'événements qui débutent, terminent ou se déroulent pendant ce pas de temps.

- D : Catégorie de l'événement (10 catégories d'événements). Pour chaque station avec événement, à chaque pas de temps de la journée, nous avons compté le nombre d'événements par catégorie lié à l'horaire de début, de fin ainsi que la période de l'événement si la fin de l'événement est disponible.

Les données d'entrée D1, D2, D3 et D4 des modèles de prévision long terme sont représentées dans le tableau 3.1.

Tab. 3.1.: données d'entrée D1, D2, D3 et D4 des modèles de prévision long terme.

Données	A	B	C	D	Taille du vecteur
D1	✓				$11 + 6 = 17$
D2	✓	✓			$17 + 7 = 24$
D3	✓	✓	✓		$24 + 96 \times 3 \times 29 = 8376$
D4	✓	✓	✓	✓	$8376 + 96 \times 3 \times 29 \times 10 = 91896$

Le jeu de données A correspond au mois et au jour de la semaine, B correspond aux informations calendaires détaillées, C aux horaires des événements et D correspond à l'information de la catégorie de chaque événement.

Dans le but d'analyser l'importance des différentes caractéristiques des données d'entrée par station, nous avons entraîné un modèle de prévision spécifique à chaque station (68 modèles au total). La sortie d'un modèle correspond au nombre de passagers entrant à chaque pas de temps d'une journée. Cela signifie que nous effectuons pour chaque jour de la base de données, une prévision qui correspond à un vecteur de sortie contenant la prévision du nombre de passagers par quart d'heure (la taille du vecteur est égale à 96, le nombre de pas de temps de 15 minutes en 24 heures). Tous les modèles de prévision (un modèle par station) ont les mêmes entrées et sorties que celles décrites dans la Figure 3.4. Cette figure représente un échantillon d'entrée ($x_i \in X$) composé des caractéristiques {A,B} et {C,D} correspondant aux caractéristiques disponibles jusqu'à un an avant les prévisions du jour day_i. Les caractéristiques A et B sont détaillées dans la section 3.3.2 (e.g., jour de la semaine, jour férié, vacances scolaires, etc.). Les caractéristiques C et D sont encodées par pas de temps (96 quarts d'heure par jour) et correspondent aux informations détaillant les événements de la journée.

3.4.2 Méthodologie

Notre objectif est de prévoir la demande passager avec des données calendaires et événementielles disponibles en avance (un an à l'avance dans notre cas) avec une résolution temporelle fine (agrégation par quart d'heure). Dans ce contexte, il n'est

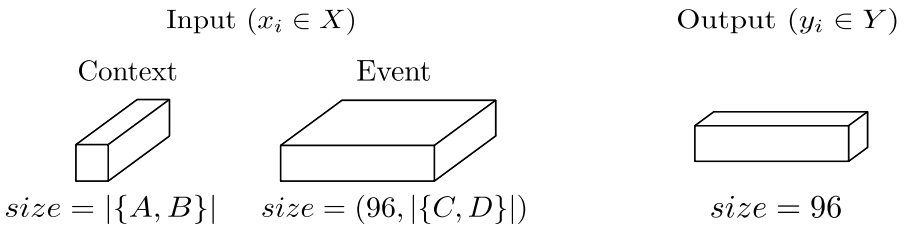

Fig. 3.4.: Exemple d'un échantillon d'entrée ($x_i \in X$) et de sortie ($y_i \in Y$) des modèles de prévision long terme.

pas possible d'utiliser les modèles d'analyses de séries temporelles classiques comme les modèles auto-régressifs (ARIMA, SARIMAX, etc.) en raison du trop grand nombre de variables à prendre en compte lors de l'entraînement et du trop grand nombre de pas de temps de prévision. Nous avons donc comparé différents modèles bien connus qui peuvent être utilisés pour des problèmes de régression. Nous avons comparé un modèle qui utilise la moyenne historique, un modèle de régression linéaire nommé Elastic Net, des modèles d'apprentissage automatique Forêts aléatoires et Gradient Boosting Decision Trees et des modèles basés sur des méthodes à noyaux, Support Vector Regressor et Gaussian Process. Le détail de ces méthodes est donné dans la section 2.4. Afin d'obtenir les meilleures performances de prévision, nous avons optimisé chacun des modèles de prévision. Ce processus d'optimisation ainsi que les hyperparamètres testés sont définis dans la section 3.4.4.2.

Concernant la méthode de régression Elastic Net, nous utilisons en entrée plus d'une variable explicative et nous prédisons plus d'une variable dépendante (total de 96 variables dépendantes), dans ce contexte nous avons utilisé une régression linéaire multivariée. Afin d'éviter le phénomène de colinéarité sujet à perturber l'apprentissage de la méthode de régression linéaire, causé par des caractéristiques catégorielles telles que le jour de la semaine, nous avons restructuré les données en supprimant une des catégories.

3.4.2.1. Prise en compte de la tendance annuelle dans la méthode de prévision

Le principal inconvénient de la prévision basée sur les modèles d'apprentissage automatique est liée au fait que ces modèles ne prennent pas compte de la tendance globale du nombre de passagers d'une année à l'autre. En effet, nous pouvons constater dans la carte de chaleur représentée à la Figure 3.5 le pourcentage de variation entre les années 2015 et 2016 et entre 2016 et 2017 du nombre moyen de passagers par pas de temps et par station (nous ne prenons pas en compte les

stations Beaubien et Rosemont fortement impactées par des rénovations en 2016). 60% des stations présentent un changement d'affluence du même signe (positif ou négatif) entre 2015-2016 et 2016-2017.

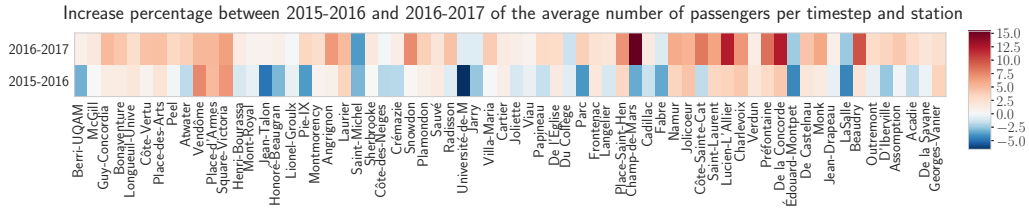

Fig. 3.5.: Facteurs de tendance entre les années 2015-2016 et 2016-2017.

Une première manière de tenir compte de ces tendances de variation dans la prévision consiste à multiplier la demande passager prévue à chaque pas de temps par le facteur de tendance décrit dans l'équation 3.1, obtenu entre 2015 et 2016 (jeu d'entraînement). Nous avons fixé le facteur de tendance des stations Beaubien et Rosemont à 1 afin de ne pas prendre en compte le facteur de tendance de ces stations. Afin d'ajuster la prévision du nombre de passagers par type de titres de transport, nous avons calculé un facteur de tendance spécifique entre les années 2015 et 2016 pour chaque titre de transport.

$$trend\_factor_{2015-2016}(s) = \frac{\frac{1}{T_{2016}} * \sum_{t_1=0}^{T_{2016}} x_{t_1}^{2016}(s)}{\frac{1}{T_{2015}} * \sum_{t_2=0}^{T_{2015}} x_{t_2}^{2015}(s)} \tag{3.1}$$

où,

$$\begin{aligned} t &= \text{Pas de temps de 15 minutes} \\ x &= \text{Nombre de passagers} \\ T_y &= \text{Nombre de pas de temps par année } y\text{, avec } t \in T \\ s &= \text{Station } s \end{aligned}$$

3.4.3 Évaluation

Afin d'évaluer les modèles, nous avons scindé l'ensemble des données en deux parties : (i) l'ensemble des données d'entraînement qui correspond aux années 2015 et 2016 et (ii) l'ensemble de données de test utilisé pour comparer les performances des modèles et qui correspond aux données de l'année 2017. Nous évaluons les résultats obtenus par les différents modèles de prévision avec plusieurs métriques bien connues. Pour mieux comprendre les erreurs, nous avons utilisé trois mesures

d'évaluation des prédictions qui sont l'erreur quadratique moyenne (RMSE), l'erreur médiane absolue (MAE) et l'erreur moyenne en pourcentage pour les valeurs supérieures à v (MAPE@v). Ces métriques d'évaluation sont détaillées dans la section 2.5.2.2.

3.4.4 Développement et optimisation des modèles

Dans cette section, nous détaillons les configurations des différents modèles de prévision. Nous discutons de l'optimisation des hyperparamètres, de la bibliothèque et des ressources utilisées pour construire les modèles.

3.4.4.1. Développement algorithmique

Nous avons utilisé scikit-learn développé par [Ped+11], une célèbre bibliothèque python, pour créer les différents modèles : Régression linéaire Elastic-Net, Processus Gaussien, Random Forest, GBDT et SVR. Nous avons utilisé la classe MultiOuputRegressor de scikit-learn afin d'effectuer des prévisions à sorties multiples avec les modèles SVR et GBDT qui sont originellement des modèles de régression à sortie unique (uni-output).

3.4.4.2. Optimisation

Nous avons effectué un grid search avec 5 ensembles de validation, croisés de manière aléatoire afin d'optimiser tous les modèles statistiques et d'apprentissage automatique. Nous avons fixé le temps de calcul pour l'optimisation de chaque modèle à 2 jours maximum. Nous avons utilisé les hyperparamètres par défaut de scikit-learn pour le modèle GBDT avec les données d'entrée D3 et D4 à cause du temps de calcul trop long. Les hyperparamètres testés sont représentés dans le tableau 3.2. Les expériences ont été menées en parallèle sur 20 cœurs.

3.5 Résultats et analyses des prévisions

Nous comparons dans la section 3.5.1 les résultats de la prévision de l'affluence des passagers avec une agrégation globale des titres de transport et présentons quelques exemples de prévisions comparées aux observations au cours de deux périodes,

Tab. 3.2.: Hyperparamètres testés dans le grid search.

Modèle	Hyperparamètres	Valeurs testées
LR	alpha	0.1, 1, 10
	l1_ratio	0.25, 0.5, 0.75, 1
	normalize	True, False
GP	alpha	0.1, 0.5, 1
	normalyze_y	False, True
RF	n_estimators	100, 150, 200
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 5, 10
	max_features	'auto'
GBDT	n_estimators	100, 150, 200
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 5, 10
	max_features	'auto'
SVR	kernel	'rbf', 'linear'
	gamma	1, 0.1, 0.01, 0.001
	C	0.001, 0.01, 0.1, 1.0, 10

avec et sans événement. Nous mettons en évidence les modèles qui obtiennent les meilleurs résultats de prévision ainsi que l'importance de chaque caractéristique dans la prévision. Dans un contexte plus appliqué au domaine du transport, nous nous attachons à analyser les résultats par station. Enfin, nous détaillons les résultats de prévision obtenus pour chaque type de titres de transport dans la section 3.5.2.

Il est à noter que dans les sections suivantes, les résultats correspondent aux prévisions obtenues par les différents modèles combinées à la prise en compte du facteur de tendance avec la méthode expliquée dans la section 3.4.2.1. La prise en compte du facteur de tendance améliore les résultats d'environ 0,80%.

3.5.1 Résultats de la prévision de l'affluence globale des passagers

3.5.1.1. Analyse des résultats de prévision sur la période globale

Afin d'avoir une compréhension globale des résultats obtenus par les modèles utilisant différents ensembles de caractéristiques comme données d'entrée, nous avons étudié les erreurs agrégées mentionnées dans la section 3.4.3 de toutes les stations. Le tableau 3.3 représente les erreurs RMSE, MAE et MAPE@150 sur les jeux d'entraînement et de test obtenues par les modèles de prévision décrits dans la section 3.4.2. Les modèles sont utilisés pour prévoir le nombre de passagers entrants agrégés par 15 minutes par station. Chaque modèle a été entraîné avec

différentes données d'entrée (D1, D2, D3 et D4) détaillées dans la section 3.4.1. Nous pouvons observer que les meilleurs résultats sont obtenus avec les modèles utilisant la combinaison de toutes les données d'entrée (D4) sauf pour le Processus Gaussien. Ce modèle a sur appris et n'a pas réussi à prendre en compte l'information des données événementielles ainsi que la catégorie des événements, encodée de manière éparse. Le meilleur modèle de prédiction est le modèle RF avec 38,5 et 13,1% d'erreur RMSE et MAPE@150.

Tab. 3.3.: Erreurs sur les jeux de données d'entraînement (train set) et de test (test set) des différents modèles de prévision avec les données d'entrée (D1, D2, D3 et D4).

		Train set (2015-2016)			Test set (2017)		
Données	Modèle	RMSE	MAE	MAPE	RMSE	MAE	MAPE
D1	HA	45,4	18,07	12,7	50,4	21,5	15,3
	LR	49,7	20,5	13,9	52,0	22,5	15,5
	RF	47,0	18,8	13,0	50,4	21,3	15,2
	GP	46,1	18,6	12,9	**50,3**	21,6	15,2
	SVR	56,0	23,7	14,1	57,5	25,4	15,6
	GBDT	48,2	19,4	13,4	51,2	21,8	15,8
D2	HA	32,2	13,0	9,7	44,3	19,2	13,9
	LR	41,2	18,0	12,4	45,0	20,1	14,0
	RF	35,1	14,7	10,7	**41,4**	18,2	13,2
	GP	33,7	14,0	10,3	41,4	18,5	13,4
	SVR	45,7	20,4	12,6	49,2	22,5	14,0
	GBDT	37,6	16,2	11,6	42,3	18,8	13,9
D3	LR	34,6	16,6	11,6	43,7	20,4	14,2
	RF	26,8	12,7	9,3	**39,7**	18,0	13,2
	GP	17,1	7,0	4,9	79,7	36,4	22,7
	SVR	36,8	18,9	12,3	51,1	25,0	16,2
	GBDT	26,4	13,4	9,8	42,8	18,9	14,0
D4	LR	33,8	16,6	11,7	42,6	20,3	14,2
	RF	26,6	12,6	9,3	**38,5**	17,9	13,1
	GP	16,9	6,7	4,9	80,7	37,0	23,1
	SVR	37,0	19,2	12,4	51,1	25,4	16,4
	GBDT	26,1	13,3	9,8	40,8	18,8	14,0

Les différents jeux de données comparés (D1, D2, D3, D4) sont décrits en Section 3.4.1. Les modèles de prévision sont expliqués en Section 3.4.2. Les méthodes d'évaluation RMSE, MAE et MAPE@150 sont détaillées en Section 3.4.3.

Comme nous pouvons le voir dans la Figure 3.6, l'erreur MAPE@v dépend fortement de la valeur seuil v. Par exemple, en considérant le meilleur jeu de données d'entrée D4, le modèle RF a une MAPE@5 (MAPE considérant toutes les observations avec un nombre de passagers supérieur à 5) d'environ 20% et une MAPE@150 de 13%. Nous avons fixé la valeur seuil de la MAPE à 150 afin d'avoir une meilleure estimation de la performance de prévision lorsque la quantité de passagers est suffisamment importante pour avoir un impact sur la demande de titre et l'offre de transport.

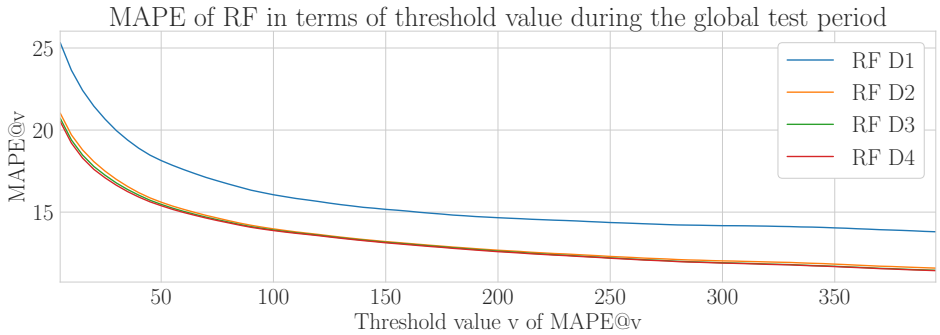

Fig. 3.6.: Erreur MAPE@v du modèle RF avec les données d'entrée (D1, D2, D3, D4) en fonction de la valeur seuil v.

Un exemple d'observation de l'affluence des passagers ainsi que de la prévision du modèle RF à la station Guy-Concordia est donné à la figure 3.7. Nous pouvons remarquer que la demande des passagers de cette station est fortement liée à l'activité des étudiants de l'université Guy-Concordia. En effet, nous observons que le lundi 18 septembre 2017, la demande des passagers semble suivre un schéma régulier avec des pics d'activité correspondant à la fin des cours de l'université. Dans ce cas, nous pouvons voir que le modèle prenant en entrée le jeu de données D4 (courbe violette) réussit à prévoir avec précision le nombre de passagers entrants et réussit à prévoir l'affluence des passagers de manière légèrement plus précise que les modèles utilisant les autres jeux de données en entrée (D1, D2 et D3).

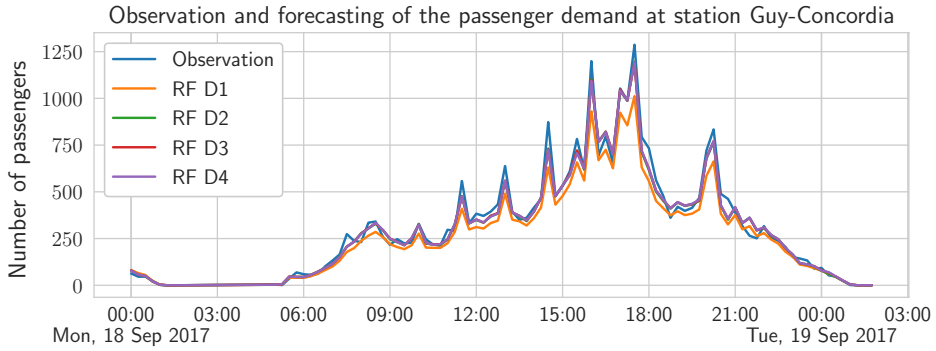

Fig. 3.7.: Observation et prévision de l'affluence des passagers à la station Guy-Concordia le lundi 18 septembre 2017.

3.5.1.2. Analyse des résultats de prévision en période d'événements

Comme certains événements, du type concert, rencontre sportive, spectacle, etc. peuvent avoir un impact sur l'affluence des passagers et par conséquent un impact sur les résultats de prévision, nous avons analysé les résultats du meilleur modèle de

prévision (RF) en considérant les périodes avec et sans événement (voir Tableau 3.4). Nous avons filtré les résultats sur les 17 stations accueillant des événements en 2017 (période de test). Le filtrage de la période des événements est effectué en sélectionnant les couples jour/station avec la présence d'au moins un événement. La période sans événement représente le reste des données (couple jour/station sans aucun événement). Nous observons que le type de données d'entrée a un impact significatif sur l'erreur RMSE lors de la période avec des événements. En effet, l'erreur RMSE diminue légèrement lors de la période sans événement en fonction de l'utilisation des données d'entrée D2 ou D4 (50,7 contre 48,8 de RMSE) contrairement aux résultats obtenus lors de la période d'événements où l'on peut observer une erreur de 153,3 contre 124,7 de RMSE en fonction de l'utilisation des jeux de données D2 et D4. Le modèle prenant en compte le jeu de données D1 est relativement simple pour être comparé de manière pertinente avec le modèle prenant en compte le jeu de données D4 en entrée.

Tab. 3.4.: Erreurs du modèle Random Forest lors de la période de test (2017) en période d'événements et sans événement obtenues sur les 17 stations accueillant des événements en 2017.

	Période sans événement			Période avec événement		
Données	RMSE	MAE	MAPE	RMSE	MAE	MAPE
D1	61,5	28,5	15,0	159,1	47,0	23,6
D2	50,7	24,7	13,2	153,3	43,4	22,2
D3	49,1	24,1	13,2	137,7	43,5	21,4
D4	**48,8**	24,0	13,1	**124,7**	40,7	21,1

Les différents jeux de données comparés (D1, D2, D3, D4) sont décrits en Section 3.4.1. Les modèles de prévision sont expliqués en Section 3.4.2. Les méthodes d'évaluation RMSE, MAE et MAPE@150 sont détaillées en Section 3.4.3.

La Figure 3.8 présente l'erreur MAPE@v des modèles RF en fonction du seuil v pendant la période de test 2017 en présence d'événements. On constate que le classement des modèles de prévision diffère en fonction de la valeur v correspondant au seuil de l'erreur MAPE. Ceci peut s'expliquer par le fait que le calcul de l'erreur MAPE@v est largement impacté par le nombre de passagers observé et la prise en compte des événements par les modèles de prévision. En effet, l'erreur MAPE@v désavantage les modèles qui prévoient un nombre de passagers élevé lorsque le nombre de passagers observé est en réalité faible (e.g., cas d'un modèle prenant en compte la présence d'un événement alors que l'affluence des passagers reste faible). Inversement, l'erreur MAPE avantage les modèles qui prévoient un nombre de passagers faible alors que le nombre de passagers observé est élevé (sous estimation de l'affluence des passagers, souvent réalisée par les modèles ne prenant pas en compte les événements). Afin d'améliorer l'offre de transport et la disponibilité des titres de transport en cas de forte affluence des passagers, il s'avère plus pertinent

d'utiliser un seuil v supérieur à une certaine valeur qui permet d'avantager les modèles qui surestiment la demande des passagers et désavantage ceux qui la sous-estiment. D'après ces données, un seuil égale à 150 semble être un bon compromis pour l'évaluation des modèles.

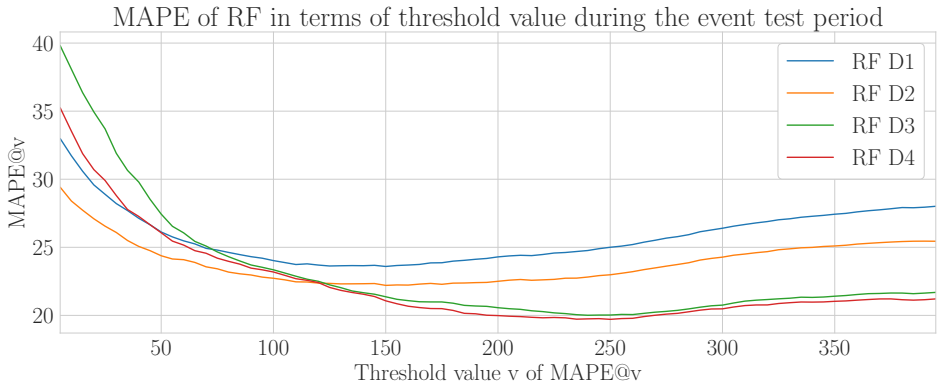

Fig. 3.8.: Erreur MAPE@v des modèles RF avec les données d'entrée (D1, D2, D3, D4) en fonction de la valeur seuil v lors de la période de test (2017) en présence d'événements.

La prise en compte des événements peut s'avérer essentielle pour prévoir avec précision le nombre de passagers entrant dans les stations d'un système de transport en commun. En effet, comme nous pouvons le voir dans la Figure 3.9, les RF utilisant les données d'entrée D1 ou D2 (informations calendaires détaillées) ne réussissent pas à prévoir la forte augmentation de la demande des passagers due à la fin d'un match de hockey à la station Lucien-L'Allier. En revanche, grâce à l'utilisation des informations sur les événements (horaire et catégorie de l'événement) disponibles dans le jeu de données D4, le modèle Random Forest parvient à prévoir avec précision le pic de la demande passager.

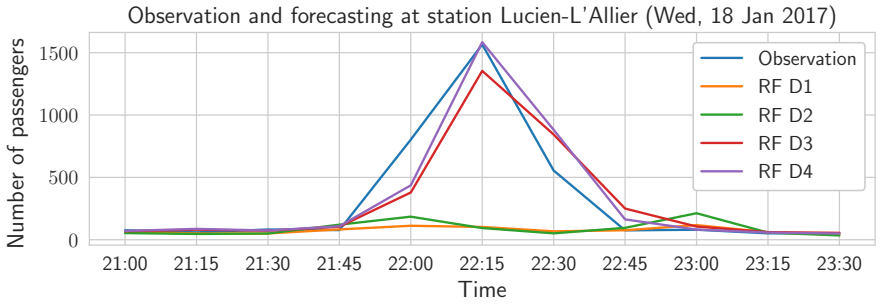

Fig. 3.9.: Observation et prévision du nombre de passagers à la station Lucien-L'Allier le mercredi 18 janvier 2017.

Dans la Figure 3.10 nous pouvons voir l'affluence des passagers observée lors de l'événement "Nuit Blanche" ayant eu lieu la nuit du samedi 4 mars 2017. Lors de cet

événement, les usagers utilisent le réseau de transport en commun de manière très particulière en raison des nombreuses manifestations artistiques qui se produisent toute la nuit dans la zone de la station Place-des-Arts et de l'ouverture du métro pendant toute la nuit. Grâce aux données événementielles D4, le modèle RF a réussi à prévoir la forte augmentation du nombre de passagers à cette station fortement impactée par l'événement.

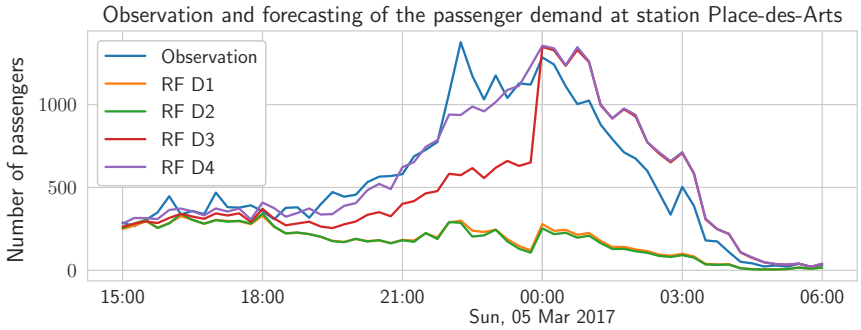

Fig. 3.10.: Observation et prévision du nombre de passagers à la station Place-des-Arts le dimanche 5 mars 2017.

3.5.1.3. Importance des caractéristiques utilisées en entrée des modèles

Certains modèles issus des statistiques ou de l'apprentissage automatique tels que les Random Forest permettent de quantifier l'importance des caractéristiques des données d'entrée. Etant donné qu'un modèle a été appris par station, nous sommes en mesure d'étudier avec précision la contribution de chaque caractéristique pour la prévision de la demande des passagers et ce pour chaque station. Ce type d'analyse s'avère utile pour expliquer et comprendre comment ce type de modèle de prévision fonctionne. Dans la figure 3.11, nous pouvons observer l'importance des caractéristiques du modèle RF avec les données d'entrée D4 pour 3 stations ayant un profil particulier (les stations sont représentées spatialement dans la figure 3.12). L'importance de chaque caractéristique est calculée avec "l'indice de diversité de Gini" appelé "Gini impurity" en anglais [Bre17]. L'importance i des caractéristiques f, notée f_i, est donnée par :

$$f_i = \frac{\sum_{j \in \text{noeuds j qui se divisent sur la caractéristique i}} n_j}{\sum_{j \in \text{ensemble des noeuds}} n_j} \tag{3.2}$$

Avec n_j l'importance du noeud j,

$$n_j = w_j C_j - w_{gauche(j)} C_{gauche(j)} - w_{droit(j)} C_{droit(j)} \tag{3.3}$$

où w_j est le nombre pondéré d'échantillons dans le noeud j, avec C_j l'impureté de ce noeud, et $gauche(j)$ et $droit(j)$ ses noeuds fils respectifs. L'importance de chaque caractéristique est fournie en pourcentage et a été agrégée par catégorie: les informations du type de jour détaillé dans la section 3.3.2, l'horaire des événements (qui correspond à la somme de l'importance des caractéristiques des horaires des événements à savoir début, fin et période de toutes les stations avec événements) et la catégorie de chaque événement disponible pour chaque station hôte d'événement. La caractéristique la plus importante est le jour de la semaine avec respectivement 60,3%, 83,8% et 53,5% d'importance pour les stations Place-des-Arts, Square-Victoria-OACI et Guy-Concordia. On constate également que Place-des-Arts est une station largement impactée par les horaires des événements ainsi que la catégorie de chaque événement (environ 8% et 11% d'importance pour l'horaire et la catégorie des événements), ce qui s'explique par la présence d'un grand nombre d'événements situés près de cette station. Square-Victoria-OACI est une gare située dans un quartier d'affaires, contrairement à la gare Place-des-Arts, nous pouvons voir que les caractéristiques les plus importantes sont différentes et correspondent aux jours fériés et aux vacances de Noël. Enfin, pour ce qui est de la station Guy-Concordia située à proximité de l'université Guy-Concordia, les données les plus importantes correspondent aux jours fériés, aux vacances scolaires de Noël et aux vacances scolaires d'été (partie 1 et 2).

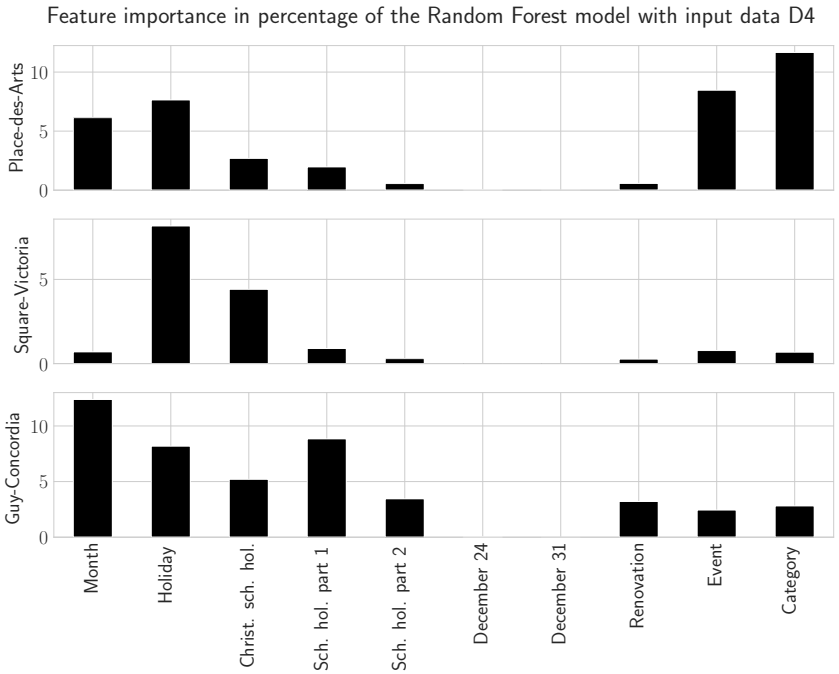

Fig. 3.11.: Importance agrégée des données d'entrée D4 du modèle Random Forest dans les stations Place-des-Arts, Square-Victoria-OACI et Guy-Concordia

3.5.1.4. Analyse des résultats par station

En plus de l'analyse globale (ensemble des stations) détaillée dans la section 3.5.1.1, il est intéressant d'analyser les résultats obtenus pour chaque station étant donné que chacune d'entre elles possède une activité spécifique. Dans la figure 3.12 on peut observer l'erreur MAPE@150 du meilleur modèle de prévision qui est le Random Forest prenant en entrée le jeu de données D4 qui correspond aux informations détaillées sur le jour, l'horaire et la catégorie de l'événement. Nous pouvons voir que le modèle obtient une erreur MAPE@150 supérieure ou égale à 17% dans quelques stations ayant des activités atypiques. Les stations Université de Montréal et Edouard-Montpetit sont situées sur le campus de l'Université de Montréal, ce qui implique une demande de passagers affectée par le calendrier universitaire (MAPE@150 égale à 17% et 20%). L'affluence de la station Lucien-L'Allier est difficile à prévoir (MAPE@150 égale à 20%) car cette station se situe au coeur de la plupart des événements de la ville. Enfin, les nombreux événements qui se sont déroulés à la station Jean-Drapeau située dans une île inhabitée, en font la station la plus difficile à prévoir (50% de MAPE@150).

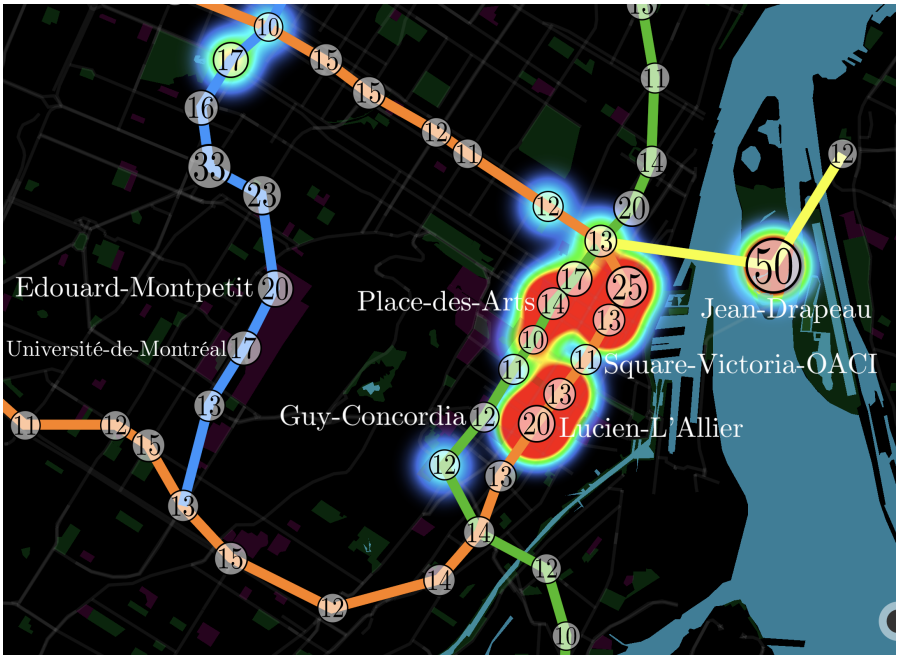

Fig. 3.12.: Erreur MAPE@150 obtenue par station sur la période de test globale (2017) par le modèle Random Forest avec le jeu de données D4.

3.5.2 Résultats de la prévision par type de titres de transport

L'un des objectifs des opérateurs de transport est d'estimer précisément la demande de certains types de titres de transport afin d'adapter leur disponibilité à la demande des passagers. Dans ce contexte, nous voulons étudier les résultats de prévision du modèle Random Forest en nous concentrant sur la prévision de sous-ensembles de données correspondant à l'affluence des passagers par type de titres de transport. Les titres de transport sont regroupés en différentes catégories : Abonnement Mensuel STM (SMP), Abonnement Mensuel régional (RMP), Tickets jetables (BT) et Abonnement Occasionnel (OP).

3.5.2.1. Analyse des prévisions par type de titres de transport lors de la période globale

Comme nous pouvons nous y attendre, les résultats présentés dans le tableau 3.5 montrent que la demande issue des titres occasionnels de transport (titre OP) est la plus difficile à prévoir. Le MAPE@150 est de 27,9% pour ce type de titres qui représente 15,7% de la demande totale de passagers contre 12,2% de MAPE@150 pour le type de titres de transport SMP qui représente 51% de la demande totale. L'utilisation des données d'entrée D4 relatives aux informations des horaires et des catégories d'événements en plus des informations détaillées caractérisant le type de jour sont nécessaires pour obtenir les meilleurs résultats en ce qui concerne la prévision de la demande occasionnelle de passagers titres BT et OP. Ceci est dû à la particularité des titres jetables et occasionnels qui sont naturellement utilisés lors des événements.

Tab. 3.5.: Erreurs des modèles Random Forest utilisant les jeux de données D2 et D4, obtenues par type de titres de transport

		Train set (2015 and 2016)			Test set (2017)		
Titre	Données	RMSE	MAE	MAPE	RMSE	MAE	MAPE
SMP	**D2**	16,4	8,7	9,7	**20,0**	10,7	12,2
	D4	14,1	7,7	8,5	20,1	10,7	12,2
RMP	**D2**	8,3	3,1	9,3	**10,2**	3,7	12,0
	D4	7,5	2,8	8,3	10,3	3,8	12,1
BT	D2	5,6	3,1	16,2	6,7	3,6	20,5
	D4	5,1	2,9	13,2	**6,6**	3,6	19,2
OP	D2	19,5	4,9	28,3	21,4	5,7	30,4
	D4	11,2	4,1	18,9	**17,9**	5,4	27,9

Les différents jeux de données comparés (D2, D4) sont décrits en Section 3.4.1. Les modèles de prévision sont expliqués en Section 3.4.2. Les méthodes d'évaluation RMSE, MAE et MAPE@150 sont détaillées en Section 3.4.3. Les titres de transport sont les suivants: Abonnement Mensuel STM (SMP), Abonnement Mensuel régional (RMP), Tickets jetables (BT) et Abonnement Occasionnel (OP).

3.5.2.2. Analyse des prévisions par type de titres de transport lors de périodes contenant des événements

Les titres de transport mensuels de la STM et les titres de transport mensuels régionaux sont légèrement impactés par les événements. Comme nous pouvons le voir dans le tableau 3.6, la RMSE des 17 stations accueillant des événements passe de 25,0 à 28,0 en période d'événements pour le titre de transport mensuel STM (SMP) et de 12,5 à 13,3 en période d'événements pour le titre mensuel régional. En revanche, les tickets jetables et les titres occasionnels sont fortement influencés par la présence d'événements. En effet, nous observons que le modèle Random Forest obtient les meilleurs résultats pour ces deux types de titres de transport avec les données d'entrée D4, pendant la période avec et sans événement.

Tab. 3.6.: Erreurs des modèles Random Forest obtenues par type de titres de transport en période d'événements et sans événement.

		Période de test sans évén.			Période de test avec évén.		
Titre	Données	RMSE	MAE	MAPE	RMSE	MAE	MAPE
SMP	D2	**25,0**	13,1	11,9	30,5	13,7	18,1
	D4	25,2	13,1	12,0	**28,0**	13,3	16,5
RMP	D2	**12,5**	5,3	11,2	13,5	5,7	11,2
	D4	12,7	5,3	11,4	**13,3**	5,6	11,2
BT	D2	9,0	4,7	17,9	16,2	6,5	51,7
	D4	**8,8**	4,7	17,4	**14,6**	6,4	41,0
OP	D2	21,9	8,6	30,1	114,6	23,7	55,7
	D4	**19,2**	7,8	29,3	**91,0**	22,0	44,0

Les différents jeux de données comparés (D2, D4) sont décrits en Section 3.4.1. Les modèles de prévision sont expliqués en Section 3.4.2. Les méthodes d'évaluation RMSE, MAE et MAPE@150 sont détaillées en Section 3.4.3. Les titres de transport sont les suivants: Abonnement Mensuel STM (SMP), Abonnement Mensuel régional (RMP), Tickets jetables (BT) et Abonnement Occasionnel (OP).

Nous pouvons observer l'impact d'un match de hockey sur la demande des passagers pour chacun des types de titres de transport dans la Figure 3.13. Cet événement est décrit comme commençant à 19h30, mais l'heure de fin n'est pas renseignée. On constate que les Random Forest utilisant les données d'entrée D4 (information sur le type de jour, information sur l'horaire et la catégorie de chaque événement) sont capables de prévoir avec une bonne précision la forte augmentation du nombre de passagers entre 22h00 et 23h00. Le type de titres de transport le plus impacté par l'événement est le titre occasionnel avec une augmentation de 1000 passagers lors du pic d'affluence à 22h15.

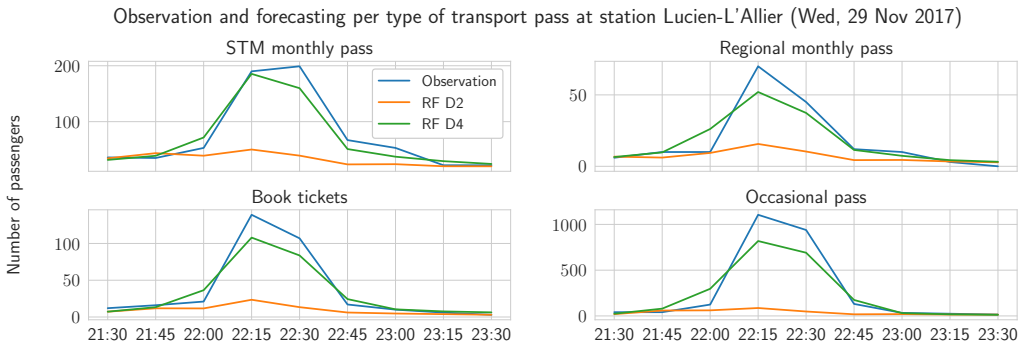

Fig. 3.13.: Observation et prévision de la demande des passagers par titre de transport à la station Lucien-L'Allier en présence d'un événement (match de hockey).

3.6 Conclusion

Dans ce chapitre nous avons étudié l'utilisation des données billettiques, calendaires et événementielles pour prévoir le nombre de passagers entrant dans chaque station d'un réseau de métro avec un horizon prévisionnel long terme (jusqu'à un an à l'avance) et une résolution temporelle fine (agrégation au quart d'heure de l'affluence des passagers). Nous avons effectué les prévisions sur des données réelles (métro de Montréal, Canada) en tenant compte des événements planifiés un an à l'avance tels que des concerts, des matchs de hockey, des festivals, etc. L'objectif opérationnel était double : la prévision à long terme peut être utile aux opérateurs de transport pour adapter l'offre de transport et ajuster la disponibilité des titres de transport à la demande des passagers. Dans ce contexte, nous avons étudié la prévision du nombre de passagers par type de titres de transport ainsi que la prévision de la demande des passagers agrégée sur l'ensemble des titres.

Nous avons proposé une mise en forme générique des données permettant d'utiliser des données contextuelles (données billettiques, calendaires et événementielles) comme entrée de modèles de régression connus : moyenne historique, des modèles statistiques et des modèles issus de l'apprentissage automatique. L'analyse des prévisions globales a prouvé qu'il est possible d'obtenir une bonne précision de prévision à long terme avec une résolution fine même en présence d'événements. Le modèle Random Forest a permis d'obtenir les meilleurs résultats de prévision avec les informations calendaires et événementielles en entrée. Les résultats des prévisions ont mis en évidence l'importance des données relatives aux événements. L'étude a également illustré l'intérêt pour les opérateurs de transport d'utiliser un modèle de régression par station afin de comprendre quelles caractéristiques impactent le plus la demande des passagers par station. Nous avons étudié les résultats liés au réseau

de transport afin de mieux comprendre quelles sont les stations difficiles à prévoir. Nous avons montré que la prévision de la demande passager en fonction de certains types de titres de transport (titres occasionnels et jetables) est plus impactée par les événements et nécessite une prise en compte des données relatives aux événements par les modèles de prévision.

Nous avons également proposé une méthode simple pour prendre en compte la tendance du nombre de passagers par station d'une année sur l'autre. Les résultats de prévision ont démontré l'efficacité de cette méthode combinée à l'utilisation des méthodes d'apprentissage automatique. Néanmoins, des travaux supplémentaires sont nécessaires pour étudier en détail le problème de la prise en compte de la tendance dans la méthode de prévision long terme. Enfin, ces travaux s'avèrent être génériques et peuvent être appliqués à d'autres systèmes de transport en commun dans le monde. La méthodologie de prévision présentée dans cette étude pourrait contribuer à créer des services de mobilité à forte valeur ajoutée pour les citoyens.

4 Prévision court terme de l'affluence des passagers

4.1 Résumé

La prévision de la demande de mobilité s'avère être un problème central pour les autorités et les opérateurs de transport en commun. Plusieurs horizons temporels de prévision peuvent être exploités par ces entités organisatrices des transports, à savoir la prévision long terme détaillée précédemment dans le chapitre 3 et la prévision court terme que nous détaillons dans ce chapitre. Il est à noter que le travail portant sur la prévision court terme reste prospectif au sens où, actuellement, il est rarement possible de récupérer l'ensemble des données billettiques d'un système de transport en commun en temps réel, jeu de données nécessaire à l'application de prévisions court terme. Contrairement à la prévision long terme, où le principal objectif pour les opérateurs de transport en commun est la planification de l'offre de transport, la prévision court terme s'attache à prévoir la fréquentation des usagers en prenant en compte les dépendances temporelles récentes (dans notre cas l'affluence des passagers est prédite lors des 15 prochaines minutes). L'objectif principal de ce type de prévision est de prévoir les flux de passagers à quelques heures ou minutes à l'avance en tenant compte de l'état du réseau de transport en temps réel, ce qui fait de ces modèles des outils pertinents pour résoudre différents problèmes : (i) informer les passagers des situations de congestion, (ii) proposer aux passagers des itinéraires de déplacement adaptés à l'offre de transport et à la demande des passagers et enfin (iii) améliorer l'exploitation du système de transport des lignes affectées par certaines perturbations afin de faire correspondre l'offre de transport à l'affluence des passagers en temps réel (e.g., mise en place de navettes).

Dans ce chapitre nous nous attachons à la prévision de l'affluence des passagers agrégée de deux manières différentes: (i) la prévision de matrices Origine-Destination (OD) représentant le nombre de passagers réalisant des déplacements d'une station d'origine à une station de destination et (ii) le nombre de passagers entrant dans chaque station. La prévision de matrice OD permet d'avoir une connaissance précise du déplacement des usagers, en revanche ce type de prévision peut s'avérer être

difficile à réaliser du fait que les stations de destination ne soient pas toujours collectées par les systèmes de billettique.

Dans ce contexte, nous explorons dans la section 4.2 une approche innovante de prévision court terme des matrices dynamiques Origine-Destination (OD) représentant le nombre de passagers se déplaçant d'une station à une autre dans un réseau de métro. Pour cela nous avons développé un réseau de neurones récurrent de type Long Short-Term Memory (LSTM). Une comparaison avec des approches traditionnelles, basées sur des moyennes historiques et un modèle autorégressif (Vector Autoregressive en anglais, VAR), est effectuée sur un jeu de données réelles. Ce jeu de données correspond à des données billettiques issues du réseau de métro et de bus de Rennes Métropole, France. Les résultats obtenus montrent que l'approche proposée permet de réaliser des prévisions court terme fiables des paires OD (données agrégées au quart d'heure).

Dans un deuxième temps, nous étudions dans la section 4.3 la question de prévision court terme multi pas de temps du nombre de passagers entrant dans chaque station d'un réseau de transport en commun. Nous considérons un modèle d'apprentissage automatique appelé Random Forest (RF) et un modèle plus récent, un réseau de neurones récurrent nommé Gated Recurrent Unit (GRU) issu de l'apprentissage profond. Nous nous concentrons sur 30 stations situées dans le quartier de La Défense, un grand quartier d'affaires bien connu de l'agglomération parisienne. Les résultats sont analysés sur différentes périodes pour illustrer l'efficacité des méthodes de prévision.

Dans ces deux cas nous comparons les méthodes de prévision court terme avec des méthodes de prévision long terme inspirées du chapitre 3, ces méthodes long terme servent ici de référence ou baseline en anglais pour évaluer les modèles de prévision court terme.

4.2 Prévision de matrices OD court terme: cas d'étude Rennes

4.2.1 Introduction

Actuellement, l'offre de transport en commun est déjà, dans une certaine mesure, déterminée par des méthodes de prévision de la demande long terme que les opérateurs ajustent en fonction du type de fréquentation estimé sur la base d'enquêtes

et de données issues de capteurs (par exemple les données billettiques) à partir d'informations calendaires (jours ouvrables et jours ouvrés).

Dans cette section, nous nous attachons au pré-traitement, à la construction de bases de données, à la prévision des flux de passagers entre stations et à l'analyse détaillée de ces prévisions. Plus précisément, la prévision des flux de passagers à laquelle nous nous attachons ici, vise à prévoir à court terme (15 prochaines minutes) le nombre de passagers qui effectueront un déplacement entre chaque station d'origine et station de destination dans un réseau de métro. De nombreuses méthodes statistiques capables de prendre en entrée des données de séries temporelles de type multivarié (plusieurs séries temporelles en entrée) peuvent être utilisées pour réaliser ce type de prévision. Dans notre cas nous utilisons un type de réseau de neurones récurrent, le Long Short-Term Memory (LSTM) connu pour obtenir de meilleures performances de prévisions que d'autres méthodes classiques utilisées en cas d'analyse de séries temporelles (prévision et classification) grâce à sa capacité à prendre en compte des événements passés à long terme. Nous comparons des outils de prévision traditionnels avec cette méthode issue de l'apprentissage profond, tels qu'un modèle basique utilisant la moyenne historique et une approche statistique à savoir un modèle auto-régressif, (Vector Autoregressive en anglais (VAR)) sur un ensemble de données réelles fournies par l'opérateur de transport public de Rennes métropole. Nous expérimentons également l'effet de la prise en compte de données additionnelles correspondant aux matrices OD de systèmes de transports voisins (dans ce cas-ci, le réseau de bus) sur les résultats de prévision des matrices OD du réseau de métro.

Le reste de cette section est organisé comme suit. L'ensemble des données billettiques utilisées tout au long de cette étude est décrit dans la section 4.2.2. Notre approche de la prévision de matrices OD dynamiques du réseau de métro est présentée dans la section 4.2.3. Les résultats expérimentaux sont détaillés et analysés dans la section 4.2.4. Enfin la conclusion de cette étude est présentée dans la section 4.2.5.

4.2.2 Données de Rennes

Dans cette section, nous décrivons l'ensemble des données utilisées pour notre étude ainsi que l'étape de pré-traitement effectuée pour rendre le jeu de données utilisable pour la prévision des matrices OD.

4.2.2.1. Description des données

Dans cette étude, nous utilisons les données billettiques fournies par le Service des Transports en commun de l'Agglomération Rennaise (STAR). Ces données ont été collectées auprès de 70 lignes régulières de bus et 1 ligne de métro desservant Rennes Métropole sur une période de quinze mois (d'avril 2014 à juin 2015). En moyenne, 250000 validations de passagers sont enregistrées par jour. Chaque validation contient un identifiant de passager anonyme (pour les validations effectuées avec des cartes à puce), l'horaire de validation (date et heure à la minute près) et le lieu (station de métro ou arrêt de bus) de la validation, la ligne d'embarquement et le type de tarif. Les destinations de déplacement ne sont pas enregistrées par le système de collecte billettiques (Automatic Fare Collection system en anglais, AFC) car les passagers se doivent de valider leur carte à puce uniquement au moment de l'embarquement. En raison des règlements portant sur la protection de la vie privée, les renseignements personnels sur les titulaires de carte à puce ne sont pas disponibles et les identifiants passagers sont changés régulièrement.

La Figure 4.1 représente une carte de la ville de Rennes ainsi que les 15 stations de métro et arrêt de bus les plus fréquentés. Nous pouvons observer que la ligne de métro est nettement plus fréquentée que le réseau de bus, à l'exception des arrêts de bus situés à proximité de l'Université de Rennes 2 (située au nord-ouest de la ville) et de la station République (située au centre) qui sont utilisés par un nombre plus important de lignes de bus que les autres arrêts et donc plus fréquentés que les autres arrêts.

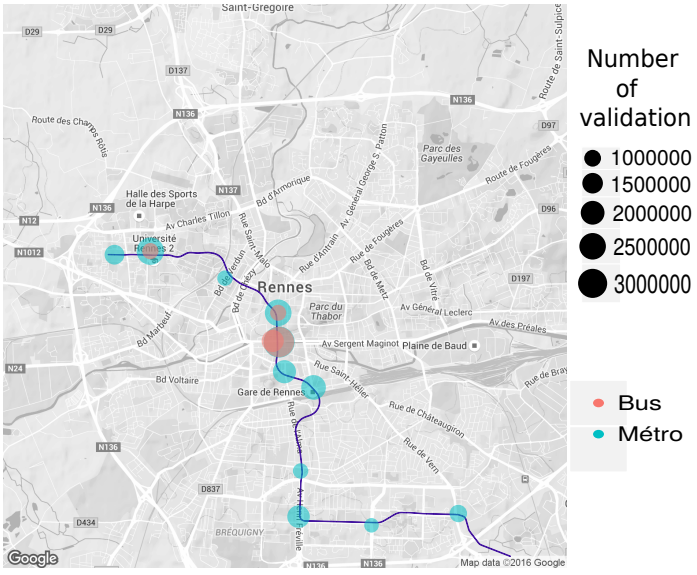

Fig. 4.1.: Carte des 15 stations de métro et arrêt de bus les plus fréquentés de la ville de Rennes.

4.2.2.2. Enrichissement des données billettiques et méthode d'agrégation

Afin de pouvoir prévoir les matrices de comptage OD, nous avons dû compléter les informations concernant l'origine et l'horaire de chaque déplacement (déjà collectées par le système AFC) avec les emplacements de destination. À cette fin, nous appliquons la méthodologie couramment utilisée, proposée par [Bar+02], dans laquelle deux hypothèses sont formulées : (i) le lieu de destination est la station la plus proche (desservie par l'autobus ou le métro) de la station d'origine du déplacement suivant, et (ii) la dernière destination de la journée est la station la plus proche de celle utilisée comme origine du tout premier déplacement de cette même journée.

Étant donné que nous visons à prévoir les matrices OD sur des fenêtres temporelles de 15 minutes, les données enrichies ont ensuite été agrégées en nombre de déplacements par pas de temps de 15 minutes pour chaque paire OD. La figure 4.2 montre le modèle hebdomadaire des paires OD de métro les plus et moins utilisées. Outre la différence significative d'amplitude entre les deux, on peut distinguer deux profils de déplacements distincts entre les jours de semaine et de week-end. Les deux courbes présentent également des variations de fréquence élevées qui reflètent d'importantes fluctuations de la demande des passagers.

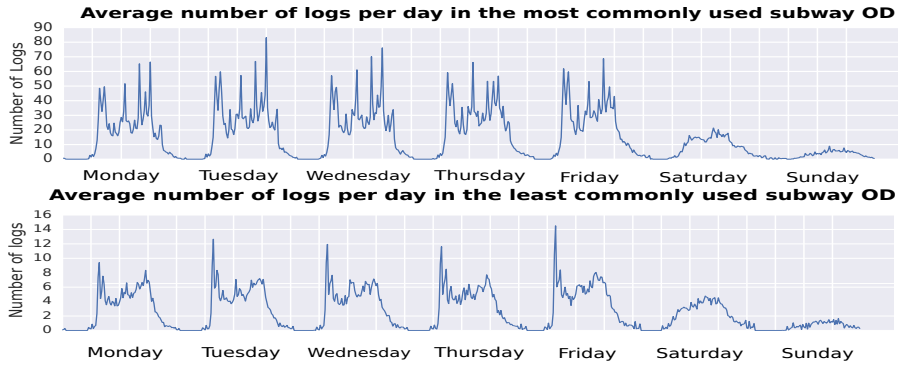

Fig. 4.2.: Nombre de déplacements moyen (par pas de temps de 15 minutes) de la paire OD de métro la plus empruntée (haut) et de la paire OD la moins empruntée (bas).

4.2.3 Prévision des matrices OD avec des LSTM

Dans cette section nous discutons des détails de notre approche pour prédire les matrices de comptage OD.

Les méthodes RNN et LSTM sont détaillées dans les sections 2.4.5.1 et 2.4.5.2. Pour rappel, dans un réseau de neurones récurrent, la couche cachée h_t au temps t

(couche de poids qui correspond aux informations regroupées au temps courant) est calculée sur la base de l'état précédent h_{t-1} et de l'entrée (l'observation) fournie à l'étape courante x_t en utilisant une fonction non linéaire f :

$$h_t = f(Ux_t + Wh_{t-1}) \ , \tag{4.1}$$

où U et W sont des matrices de poids affectées à x_t et h_{t-1} respectivement.

La sortie y_t est calculée à partir de cette couche cachée à l'aide d'une fonction g :

$$y_t = g(Vh_t) \ , \tag{4.2}$$

où V est aussi une matrice de poids.

Notre structure de modèle de prévision contient une couche LSTM, dans laquelle la fonction f (4.1) est la fonction tangente hyperbolique (tanh) et la fonction g (4.2) est une fonction softplus (puisque le problème à résoudre est une régression avec une sortie dans R_+).

Les attributs d'entrée des modèles LSTM correspondent aux paires OD des 300 pas de temps qui précèdent le pas de temps que nous voulons prévoir. Ce que nous appelons sortie du modèle est la matrice complète de comptage OD du pas de temps correspondant aux 15 prochaines minutes pour l'ensemble des paires OD.

4.2.4 Résultats et discussion

Les résultats de prévision de matrices de comptage OD du réseau de métro de Rennes sont présentés et analysés dans cette section. Nous évaluons les modèles LSTM (i.e. avec et sans l'information des matrices de comptage OD issues du réseau de bus) en comparant leurs performances de prévision avec celles obtenues avec deux modèles classiques : un modèle long terme utilisant la moyenne historique et le modèle vecteur autorégressif (VAR). Les formes basiques de ces modèles sont détaillées dans la section 2.4.1.1 et la section 2.4.2.2.

4.2.4.1. Variante du modèle à moyenne historique

Pour effectuer la planification de l'offre d'un réseau de transport public, les opérateurs de transport exploitent souvent le nombre moyen de passagers afin de prévoir les matrices OD. Dans sa forme la plus basique, le modèle à moyenne historique fait

simplement la distinction entre les jours ouvrables (travaillés) et les jours ouvrés (non travaillés). Nous avons mis en place un modèle utilisant la moyenne historique se basant sur des caractéristiques calendaires plus précises qui tiennent compte d'une plus grande diversité de types de jours. Les types de jours suivants ont été pris en compte : Lundi, mardi et jeudi, mercredi, vendredi, samedi hors vacances scolaires (MNH, TTNH, WNH, FNH, FNH, SNH), dimanche et jours fériés (SDO), samedi pendant les vacances scolaires (SH) et les jours ouvrables en période de vacances scolaires (WDH). La prévision est simplement réalisée en calculant la moyenne du comptage de passagers pour chaque paire OD au pas de temps prédit sur l'ensemble des jours disponibles dans les données historiques appartenant au type de jour prédit.

Le modèle implémenté est illustré dans la Figure 4.3.

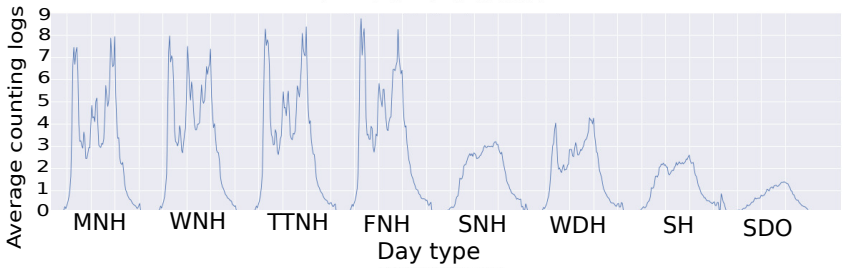

Fig. 4.3.: Illustration des huit types de jours formant le modèle à moyenne historique.

4.2.4.2. Expérimentation

Protocole expérimental Notre objectif est de prévoir la matrice OD du métro au pas de temps suivant en prenant en compte les matrices OD du métro observées précédemment. Nous sommes également intéressés par le fait d'étudier l'apport d'information des matrices OD issues d'autres modes de transport (dans ce cas-ci, les matrices OD issues des comptages de passagers du réseau de bus). Par conséquent, nous avons également réalisé la prévision de matrices OD du métro en nous basant sur des matrices OD observées combinant à la fois les comptages de passagers du bus et du métro. Pour ce second objectif, en raison de la taille du réseau (15 stations de métro et plus de 600 arrêts de bus), il n'est pas possible de prendre en compte la totalité des comptages OD à cause de la trop grande sparsité des données qui rendrait les modèles inefficaces. De plus, ce choix nécessiterait de traiter de grandes quantités de données ce qui rendrait le temps de calcul des modèles beaucoup plus long et pourrait poser des problèmes de mémoire (la matrice complète des OD contient plus de 31000 entrées). Par conséquent, afin d'accélérer l'apprentissage des

modèles, nous avons décidé de ne conserver que les comptages de déplacements des 2000 OD les plus fréquentées de bus (ces 2000 OD couvrent 70% des déplacements de passagers utilisant les bus) en plus des 210 OD issues du métro pour lesquelles nous voulons prévoir le comptage des déplacements.

Après le filtrage des OD réalisé, nous avons divisé les quinze mois de données billettiques enrichies en trois ensembles de données. La première année est divisée aléatoirement en deux ensembles de données : un ensemble d'apprentissage (80% des données, ce qui équivaut à environ 9,5 mois) et un ensemble de validation (l'équivalent de 2,5 mois constituant les 20% restants des données) pour effectuer la sélection des modèles (i.e. régler les hyperparamètres du modèle à des valeurs appropriées). Afin d'évaluer les modèles nous avons utilisé les trois derniers mois de données comme jeu de données de test. Ce cas simule une situation réaliste dans laquelle les données sont disponibles jusqu'à un moment donné, ce qui permet d'entraîner et d'ajuster le modèle de prévision qui pourra être utilisé plus tard afin de prévoir les périodes futures non observées.

Nous utilisons l'erreur quadratique moyenne (MSE) détaillée dans l'équation 2.15 de la section 2.5.2.2 pour mesurer et comparer les performances des quatre modèles (modèle utilisant la moyenne historique, le modèle VAR, le LSTM utilisant les données OD de métro, et le LSTM utilisant les données OD de métro et de bus).

Sélection du modèle LSTM Pour l'étape d'apprentissage, nous avons utilisé une méthode appelée early stopping basée sur les résultats de validation [Pre12]. Cette méthode permet d'éviter le sur-apprentissage en arrêtant l'apprentissage du modèle lorsque l'erreur de validation cesse de diminuer. Nous effectuons l'optimisation des poids du modèle en utilisant l'algorithme d'optimisation ADAM [KB14]. Comme les hyperparamètres de cet algorithme ne nécessitent pas d'optimisation particulière, nous avons focalisé l'optimisation du LSTM sur la taille de la couche cachée: nous avons effectué une optimisation de cet hyperparamètre par grid search (grille contenant les différentes valeurs des hyperparamètres testées lors de l'optimisation) afin de trouver la taille optimale de ce vecteur permettant d'obtenir les meilleurs résultats. Nous avons réalisé cette sélection de modèle de manière distincte entre le LSTM (métro) et le LSTM (métro et bus). Pour le modèle LSTM (métro et bus), l'erreur diminue lorsque la taille de la couche cachée augmente, comme le montre la Figure 4.4. Ceci s'explique par le fait que plus la couche cachée est grande plus le LSTM peut encoder d'informations. La meilleure taille pour le vecteur de la couche cachée est 3000 selon la figure (erreur minimale pour l'ensemble de validation). En

ce qui concerne le LSTM (métro), la taille de la couche cachée permettant d'obtenir les meilleurs résultats est de 900.

Fig. 4.4.: Evolution de l'erreur MSE en fonction de la taille du vecteur de la couche cachée du LSTM (métro et bus).

Sélection du modèle VAR Les résultats du modèle VAR dépendent de la taille de la fenêtre temporelle à prendre en entrée lors de la prévision (cet hyperparamètre s'appelle le lag). Le lag représente le nombre d'observations passées que le modèle prend en compte lors de la prévision. Afin d'optimiser les prévisions du modèle VAR, nous avons entraîné ce modèle avec différentes valeurs de lag et analysé les erreurs de ce modèle sur le jeu de données de validation. D'après la Figure 4.5 la valeur optimale du lag est 10 (erreur minimale sur le jeu de validation, courbe verte).

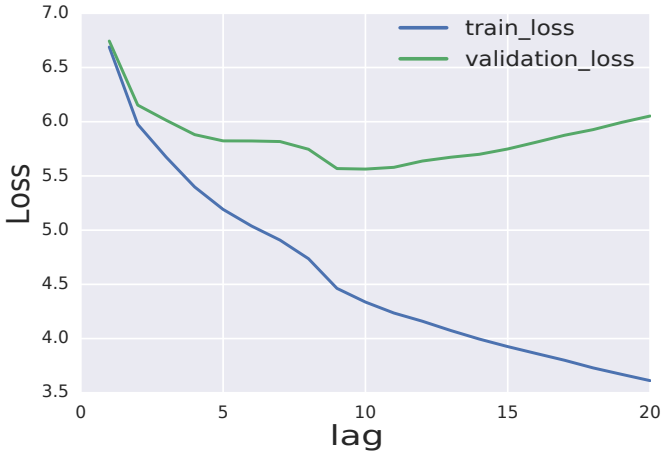

Fig. 4.5.: Evolution des erreurs MSE du modèle VAR sur les jeux de données d'apprentissage et de validation en fonction de la valeur du lag (la meilleur valeur de lag est 10).

4.2.4.3. Résultats

Un exemple de résultat obtenu à l'aide du LSTM (métro et bus) de l'OD la plus fréquentée est présenté dans la Figure 4.6. En général, le modèle est capable de prédire correctement le nombre de déplacements. Cependant, lorsque des changements soudains sont observés, la prédiction est moins précise.

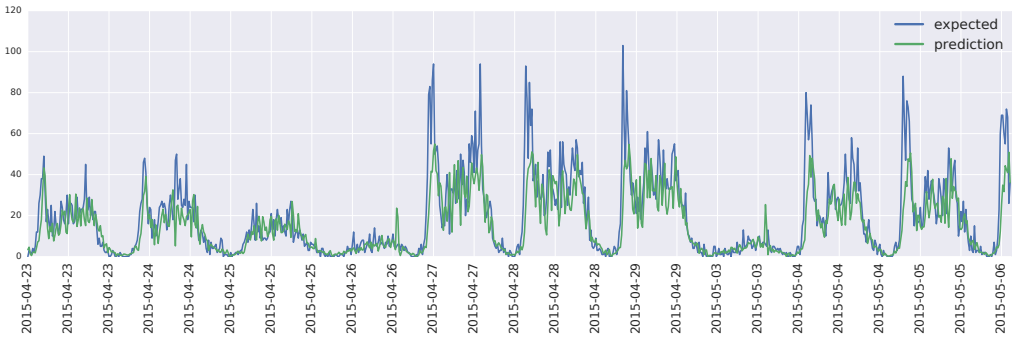

Fig. 4.6.: Observation et prévision du modèle LSTM (métro et bus) de l'OD la plus fréquentée.

L'erreur MSE obtenue par les quatre modèles est présentée dans le tableau 4.1. Nous pouvons observer que contrairement au modèle VAR et aux deux modèles LSTM qui, comme prévu, obtiennent de meilleurs résultats sur l'ensemble de validation que sur l'ensemble de test, le modèle utilisant la moyenne historique obtient de meilleurs résultats sur l'ensemble de test. Cela peut s'expliquer par le fait que (i) l'ensemble de validation présente une variance plus élevée (en raison de la longue période qu'il couvre) et que (ii) les valeurs des comptages OD dans l'ensemble de test étaient par coïncidence plus proches des moyennes calculées à partir de l'ensemble d'apprentissage.

Tab. 4.1.: Erreur MSE obtenue par les différents modèles sur différents jeu de données.

	Apprentissage	Validation	Test
Calendaire	11,6	12,4	8,9
VAR	4,3	5,6	5,9
LSTM (données: métro)	4,0	4,9	5,0
LSTM (données: métro et bus)	**2,7**	**4,5**	**4,7**

Les meilleurs résultats sont obtenus par les modèles LSTM, le modèle LSTM (métro et bus) obtient de meilleurs résultats que le modèle LSTM métro. Cela confirme le fait que la prise en compte de données issues d'autres modes de transport (tel que le bus) permet d'améliorer la performance de prévision pour le réseau de transport d'intérêt

(ici le métro). Cette observation est intuitive puisque ces modes de transport sont reliés et s'impactent mutuellement.

4.2.5 Conclusion

Nous avons proposé une méthodologie basée sur l'apprentissage profond pour prévoir à court terme (15 prochaines minutes) des matrices de comptage OD dynamiques issus de déplacements de passagers dans un transport public (métro). Notre approche est basée sur un réseau de neurones récurrent appelé LSTM, capable de capturer des informations à long terme. La méthodologie proposée a été évaluée à l'aide d'un jeu de données réelles de données billettiques issues du réseau de transport public de Rennes Métropole en France. L'évaluation des performances du LSTM a été réalisée en le comparant à deux approches traditionnelles basées sur un modèle calendaire et un modèle autorégressif (VAR). Nous avons également examiné l'effet de l'ajout des matrices de comptage OD du réseau de transport d'autobus sur le résultat des prévisions des matrices de comptage OD du réseau de métro. Les résultats obtenus sont encourageants : le modèle LSTM qui utilise à la fois les données du réseau de métro et de bus obtient en effet de meilleures performances de prévision que les modèles calendaire, VAR et LSTM utilisant uniquement les données historiques du réseau de métro.

4.3 Prévision court terme de l'affluence : cas d'étude La Défense, Paris

4.3.1 Introduction

Dans cette section, nous étendons les travaux sur la prévision court terme que nous avons initié dans [Toq+16] et qui ont été présentés dans la section 4.2. Cependant, dans cette section nous ne nous intéressons plus à des flux Origine Destination mais au nombre de passagers entrants par station. Nous avons de plus ajouté à l'étude une comparaison de nouveaux modèles et réalisé une analyse plus approfondie des résultats. Nous explorons plusieurs axes de recherche qui peuvent être résumés comme suit.

- D'un point de vue géographique les prévisions réalisées au niveau d'un quartier contribuent à créer des services de mobilité à forte valeur ajoutée pour les citoyens. Nous réalisons ici une étude qui concerne le quartier de La Défense, un grand quartier d'affaires bien connu de l'agglomération parisienne. Nous utilisons un jeu de données réelles d'une période de deux ans pour réaliser la prévision de l'affluence des passagers (nombre de personnes entrant) dans les stations de différents systèmes de transport ferroviaire, à savoir des stations de métro, de Transilien, de RER et de tramway.
- Nous considérons un modèle d'apprentissage automatique appelé Random Forest (RF) et un modèle plus récent, un réseau de neurones récurrent appelé Gated Recurrent Unit (GRU) issu de l'apprentissage profond. Des variantes univariées et multivariées de ces modèles sont développées en plus d'une approche combinée (univariée et multivariée) du modèle GRU. Cette variante est proposée pour considérer les dépendances spatiales entre les différentes séries temporelles.
- Les modèles de prévision court terme sont dans la plupart des cas créés pour prévoir le prochain pas de temps (15 prochaines minutes dans notre cas). Dans cette étude nous comparons les résultats de prévision court terme à plusieurs pas de temps, de T+1 à T+8.
- Une analyse approfondie de la performance des modèles est également effectuée pour évaluer la robustesse des modèles de prévision court terme par rapport à des périodes spécifiques, telles que les jours fériés, les jours spéciaux et certaines périodes contenant des incidents pouvant impacter l'offre de transport (e.g., incendie, accident grave de passager, etc.).

Il est important de noter que ces approches ne permettent pas de prévoir ou d'estimer la demande des passagers sur de nouvelles lignes ou stations, ce qui nécessite des outils et modèles spécifiques déjà étudiés dans le domaine des transports. Une approche bien connue de modélisation des déplacements est le modèle à quatre étapes qui permet d'effectuer des prévisions à long terme de la demande de nouveaux services à l'aide d'analyses de différentes enquêtes de déplacements [McN07].

Le reste de cette partie est organisé comme suit. La section 4.3.2 détaille les méthodologies de prévisions court et long terme développées pour prévoir l'affluence des passagers. Le processus d'expérimentation est détaillé dans la section 4.3.3 qui traite en particulier du jeu de données utilisé ainsi que des méthodes d'évaluation utilisées pour comparer les différents modèles de prévision. La section 4.3.4 décrit les résultats des prévisions. Enfin, la section 4.3.5 présente la conclusion de cette étude portant sur la prévision du nombre de passagers entrant dans les stations du réseau ferré parisien.

4.3.2 Méthodes de prévision

La prévision de l'affluence des passagers à chaque station d'un réseau de transport public est un problème difficile à résoudre, principalement en raison de l'influence de plusieurs facteurs détaillés par [ZZQ17] qui impactent les flux de personnes dans une ville à savoir les facteurs temporels et spatiaux. D'autres facteurs peuvent également impacter la prévision des flux de voyageurs notamment des facteurs directement liés aux systèmes de transport et à la collecte de données. Ces facteurs peuvent être résumés comme suit.

- *Facteurs temporels* y compris l'intervalle de temps et le type de jour (e.g., lundi, mardi, ..., dimanche, jour férié, vacances scolaires, jour de pont, etc.).
- *Facteur spatial* tel que le type de lieu où la station est située (e.g., résidentielle, affaire, commerce, etc.).
- *Facteurs exogènes* tels que la météo et les événements (e.g., concert, spectacle, rencontre sportive, etc.).
- *Influence interne du réseau de transport* qui pourrait être induite par un problème technique (problème de rails, incendie, etc.) ou une grève des transporteurs qui pourrait avoir de graves répercussions sur l'offre de transport.
- *Qualité des données* du système AFC qui collecte le nombre de passagers entrant dans les gares d'un réseau de transport public, agrégé par période de temps, pourrait fournir des données sujettes à différentes interprétations en fonction du contexte. Par exemple, la valeur 0 peut avoir des significations différentes, par exemple : aucun passager, portes ouvertes à cause d'un trop grand nombre

de passagers, portes ouvertes à cause d'un événement spécial (pic de pollution, marche solidaire) ou valeur manquante.

Compte tenu de ces facteurs, nous abordons le problème de la prévision des flux de passagers en considérant des modèles de prévision court et long terme basés sur des méthodes issues de l'apprentissage automatique. Les résultats de prévision sont analysés d'une part de manière globale et dans un second temps, en analysant les résultats de prévision obtenus lors de périodes spéciales, définies à l'aide de bases de données. Ces périodes spéciales correspondent à des jours spéciaux (e.g. jours fériés, vacances, etc.) ainsi qu'à des jours contenant des incidents pouvant impacter l'offre et/ou la demande de transport.

4.3.2.1. Méthodes de prévision long terme

Dans cette section, nous nous attachons à présenter succinctement les méthodes de prévision long terme permettant d'estimer le nombre moyen de passagers entrant dans chaque station d'un système de transport en commun à partir d'une catégorisation préliminaire des types de jours. En effet, l'accent de ce chapitre étant mis sur les méthodes court terme, les méthodes de prévision long terme servent ici comme méthodes de références (baseline en anglais) pour comparer et évaluer les méthodes court terme. Pour rappel, la prévision long terme fait ici référence à la prévision de l'affluence des passagers dans les différentes stations au cours de l'année suivante. Nous avons créé deux modèles long terme à savoir un modèle moyenne historique détaillé en section 2.4.1.1 et un modèle RF LT (Random Forest Long Term) spécialisé pour de la prévision long terme et semblable au modèle RF utilisant les données calendaire défini dans le chapitre 3. Ce modèle est basé sur la méthode des forêts aléatoires détaillée en section 2.4.4.1. Etant donné que l'activité d'une station dépend de différents facteurs temporels, comme le type de jour, nous avons défini les entrées et les sorties de ces modèles long terme comme décrits dans le tableau 4.2.

Tab. 4.2.: Entrées et sorties des modèles de prévision long terme pour la station s au pas de temps t le jour d.

Modèle	Entrées	Sortie
HA	Pas de temps t et jour d représenté par le jour de la semaine	$\hat{y}_t^s$
RF LT	Pas de temps t et jour d représenté par les caractéristiques détaillées dans le paragraphe suivant "Variante long terme du modèle Random Forest (RF LT)"	$\hat{y}_t^s$

Avec $\hat{y}_t^s$, le nombre de passagers prédits au pas de temps t à la station s, avec le pas de temps t encodé par un entier (integer).

Variante long terme du modèle Random Forest (RF LT) est un modèle Random Forest créé pour répondre à des problèmes de prévision long terme. Ce modèle utilise la mise en forme de données détaillée dans la section 3.4.1. Nous construisons cette méthode comme étant un modèle à sorties multiples capable de prévoir toutes les stations en une seule fois, en d'autres termes un seul modèle est créer pour l'ensemble des stations. La sortie de ce modèle correspond au nombre de passagers entrant dans chaque station au pas de temps t du type de jour correspondant au jour prédit. Les entrées de ce modèle sont les suivantes:

- Jour de la semaine (1-7) : Lundi, mardi, etc.
- Mois (1-12) : Janvier, février, etc.
- Jour férié (0-1) : Jour férié dans la région étudiée (pour cette étude, région parisienne).
- Jour de pont (0-1) : Jour de pont pouvant être un lundi ou vendredi se situant entre un jour férié et un jour de week-end.
- Congé scolaire (0-1) : Jour en période de vacances scolaires (zone C).
- Veille de Noël (0-1) : Le 24 décembre.
- Réveillon du Nouvel An (0-1) : 31 décembre.

4.3.2.2. Modèles de prévision court terme

Dans cette étude, la prévision court terme consiste à prévoir le nombre de passagers entrant dans les stations étudiées aux n pas de temps suivants avec n appartenant à $[\![1, 8]\!]$ (les données sont agrégées par intervalle de 15 minutes), à partir des observations passées et des caractéristiques temporelles (type de jour et pas de temps). Des modèles univariés (modèles à entrée et sortie uniques) et multivariés (modèles à entrées et sorties multiples) sont considérés pour prévoir l'affluence des passagers. Nous avons comparé quatre modèles de prévision court terme : un modèle naïf nommé LOCF, un modèle statistique VAR, un modèle issu de l'apprentissage automatique RF et un modèle issu de l'apprentissage profond, le modèle GRU avec différentes variantes (e.g., univariée, multivariée, etc.). Ces modèles sont détaillés dans les sections 2.4.1.2, 2.4.2.2, 2.4.4.1 et 2.4.5.3.

Nous proposons deux variantes de prévision court terme basées sur le modèle RF :

- **RF ST UNI** Le premier type de modèle RF court terme est un modèle univarié (un modèle par station). Ce modèle prend en entrée les n dernières valeurs observées de la série, souvent appelées lag, les informations du type de jour défini avec les caractéristiques décrites dans le paragraphe RF LT de la section 4.3.2.1 et le pas de temps.

- **RF ST MULTI** Le Random Forest court terme multivarié (un seul modèle pour l'ensemble des stations) qui prend en compte la demande des passagers des pas de temps passés de l'ensemble des stations du réseau ferré dans la prévision. Ici, l'objectif est d'obtenir des informations sur la demande passager à la station d'intérêt (prédite) en plus des autres stations. En entrée, le modèle prend les n dernières valeurs observées de chaque station du réseau, le type de jour défini avec les caractéristiques décrites dans le paragraphe RF LT de la section 4.3.2.1 et le pas de temps.

Nous proposons trois architectures basées sur le modèle GRU :

- **GRU UNI** La première version est un modèle univarié (un modèle par série temporelle/station) qui utilise les observations passées d'une série temporelle en plus de l'information sur le pas de temps et le type de jour pour prédire la valeur suivante.
- **GRU MULTI** Modèle qui permet de prendre en compte la dynamique de l'affluence au travers des différentes stations du réseau de transport étudié, nous développons une variante multivariée (un modèle pour l'ensemble des séries) du modèle récurrent. Ce modèle utilise les observations passées de toutes les stations pour prédire la prochaine valeur de toutes les stations.
- **GRU FUSION** Comme les modèles univarié et multivarié fonctionnent différemment selon l'affluence des passagers observée, nous avons créé un modèle qui combine les prévisions des deux modèles. Cette architecture de modèle GRU associe un poids aux prévisions de chaque modèle. Ces poids sont générés par un mécanisme d'attention qui attribut l'importance de chaque modèle (univarié et multivarié) à chaque pas de temps de la prévision. Pour rappel, les équations du modèle GRU sont décrites dans l'ensemble d'équation 4.3. Comme le montrent les équations 4.4 et 4.5, le résultat final de prévision du modèle GRU FUSION correspond à la somme pondérée des prévisions des deux modèles.

$$
\begin{aligned}
Z_t &= \sigma(x_t U^Z + h_{t-1} W^Z) \\
R_t &= \sigma(x_t U^R + h_{t-1} W^R) \\
\hat{h}_t &= \tanh\left(x_t U^h + (R_t * h_{t-1}) W^h\right) \\
h_t &= (1 - Z_t) * h_{t-1} + Z_t * \hat{h}_t \\
\hat{y}_t &= g(b^{\hat{y}} + h_t W^{\hat{y}})
\end{aligned}
\tag{4.3}
$$

Avec R la gate de réinitialisation, Z la gate de mise à jour, W la connexion récurrente entre la couche cachée du pas de temps précédent et actuel, σ une fonction sigmoïde, U la matrice de poids permettant de lier les entrées et la couche cachée actuelle et h_t la couche cachée du modèle GRU au pas de temps t. g correspond à une fonction d'activation permettant d'obtenir $\hat{y}_t$, la prévision ou classification au pas de temps t.

$$\alpha_t(s) = \phi(h_t^{UNI}(s)) \tag{4.4}$$

$$\hat{y}_t^{FUSION}(s) = \alpha_t(s)\hat{y}^{UNI}(s) + (1 - \alpha_t(s))\hat{y}^{MULTI}(s) \tag{4.5}$$

où $\hat{y}_t^{FUSION}(s)$ est la prévision du nombre de passagers à la station s au pas de temps t du modèle GRU FUSION. Les poids $\alpha_t(s)$ (un poids par station) sont obtenus par le mécanisme d'attention représenté par la fonction sigmoïde ϕ sur la couche cachée ($h_t^{UNI}(s)$) du modèle GRU univarié au pas de temps t.

L'approche univariée requiert autant de modèles qu'il y a de stations, alors que l'approche multivariée utilise un modèle pour prévoir le nombre de passagers entrants de l'ensemble des stations.

Les entrées et sorties des modèles court terme sont détaillées dans le tableau 4.3.

Tab. 4.3.: Entrées et sorties des modèles de prévision court terme de la demande passager au pas de temps t le jour d.

Modèle	Entrées	Sortie(s)
HA	Jour (lundi, mardi, etc.) et t	$\hat{y}_t^A$
RF LT	t et τ	$\hat{y}_t^A$
LOCF	y_{t-1}^A	y_{t-1}^A
VAR	$y_{t-n:t-1}^A$	$\hat{y}_t^A$
RF ST UNI	$y_{t-n:t-1}^s$, t et τ	$\hat{y}_t^s$
RF ST MULTI	$y_{t-n:t-1}^A$, t et τ	$\hat{y}_t^A$
GRU UNI	$y_{t-n:t-1}^s$, t et τ	$\hat{y}_t^s$
GRU MULTI	$y_{t-n:t-1}^A$, t et τ	$\hat{y}_t^A$
GRU FUSION	$y_{t-n:t-1}^A$, $y_{t-n:t-1}^s$, t et τ	$\hat{y}_t^A$

Où t est le pas de temps encodé comme un entier entre 1 et 96 et τ représente les caractéristiques de la journée. $\hat{y}_t^A$ représente les valeurs prédites du nombre de passagers à chacune des stations et $\hat{y}_t^s$ le nombre de passagers prévus à la station s au pas de temps t.

4.3.2.3. Prévision multi pas de temps

Les prévisions court terme à plusieurs pas de temps visent à prédire les valeurs de chaque station au pas de temps $t + n$, avec $n > 1$. Dans cette étude, nous prévoyons la demande des passagers jusqu'à 8 pas de temps en avance avec des données agrégées au quart d'heure $\{\hat{y}_{t+n} | n \in [\![1, 8]\!]\}$. La prévision multi pas de temps est une tâche difficile en raison de l'absence d'observations au pas de temps précédent et de l'erreur qui s'accumule sur les différents pas de temps de prévision. Différentes méthodes de prévision multi pas de temps existent. Ici, nous avons adopté la méthode itérative, qui utilise la prévision au pas de temps $t + 1$ comme variable d'entrée pour la prédiction de l'étape suivante. Ainsi, nous pouvons prédire étape par étape la demande des passagers, jusqu'à atteindre le pas de temps ciblé (8 dans notre cas).

4.3.3 Expérimentation

4.3.3.1. Etude de cas

L'ensemble des données utilisées pour évaluer les méthodes proposées a été fourni par l'autorité organisatrice des transports d'Ile-de-France (Ile-de-France Mobilités). Dans cette étude nous nous attachons uniquement aux stations ferrées du quartier de La Défense. Le système de tarification automatique (baptisé Navigo), mis en place en 2001, collecte les validations des passagers en entrée des différents modes de transport, dont le train (métro, RER et Transilien) et le tramway, qui sont gérés par différents opérateurs de transport public. Les informations des différentes stations étudiées sont données dans la section A dans le tableau A.1.

Les données sont composées de validations de passagers en entrée du réseau de transport, agrégées par intervalle de 15 minutes en 2014 et 2015. Le quartier étudié comprend cinq lignes de métro, RER et Transilien qui représentent 17 stations et une ligne de tramway qui contient 13 stations. La ligne de tramway étudiée transporte 35k passagers par jour alors que les stations ferrées étudiées voient passer un flux d'entrée de plus de 215k passagers par jour.

Dans la figure 4.7 nous pouvons observer une semaine de validation issue du jeu de données billettiques (du lundi 23 février 2015 au dimanche 1er mars 2015) des deux stations les plus fréquentées de chaque mode de transport (ferré et tramway). Bien que les validations de la station de tramway contiennent beaucoup de bruit,

ces courbes révèlent néanmoins des tendances temporelles familières quant à l'utilisation des transports en commun. En effet, il est facile de remarquer la différence entre la fréquentation des jours ouvrables et ouvrés, caractérisée par deux pics de fréquentation pendant les jours de semaine (matin et soir) et par une tendance plus diffuse observée pendant les jours de fin de semaine. Le pic de fréquentation à l'heure du déjeuner est également visible pendant les jours ouvrables. Quelques différences de profil peuvent être observées entre les différentes stations, par exemple, le profil de la station ferrée "La Défense Grande Arche" (ID 393) est caractérisé lors des jours ouvrables par un pic de fréquentation plus important en soirée qu'en matinée, ceci s'explique par le fait que cette station se situe au centre du quartier d'affaires.

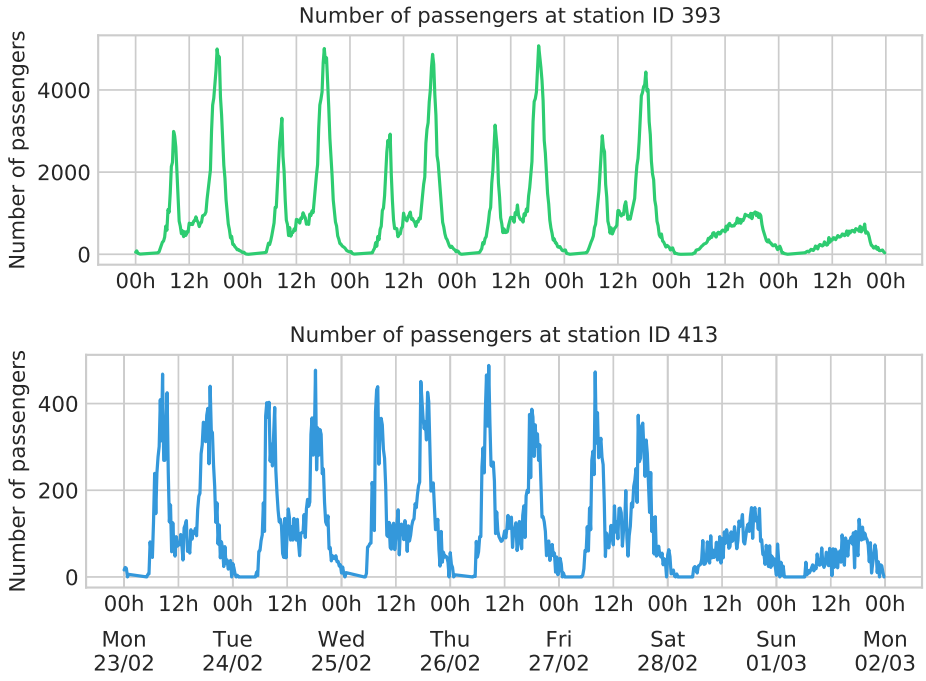

Fig. 4.7.: Profil hebdomadaire de la fréquentation des stations les plus visitées du réseau ferré et de tramway.

4.3.3.2. Méthodes d'évaluation

Nous évaluons les résultats obtenus par les différents modèles de prévision avec trois métriques bien connues: l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE) et le pourcentage moyen de l'erreur absolue calculée au dessus du seuil v (MAPE@v). Ces métriques d'évaluation sont détaillées dans les sections 2.5.2.2.

Comme nous pouvons le voir dans la figure 4.8, nous avons divisé le jeu de données en différentes parties pour évaluer les modèles de prévision court et long terme. Pour la prévision long terme, les modèles ont été entraînés avec les données de l'année 2014 (colorées en bleu) et l'évaluation des modèles a été effectuée sur les données de l'année 2015 (colorées en rouge). Quatre ensembles d'entraînement et de test ont été créés pour évaluer les modèles de prévision court terme. Les erreurs de prévision court terme sur l'année 2015 (ensemble de test global) ont été obtenues en concaténant les prévisions des quatre ensembles de test colorés en rouge. En réalisant ce type d'échantillonage du jeu de donnée, nous nous plaçons dans un cas se rapprochant d'une situation réelle où les modèles pourraient être mis à jour à chaque trimestre. Certains jours ont dû être supprimés de l'ensemble des données de test pour avoir des données d'évaluation pertinentes. Ces jours correspondent à des périodes où le transport était gratuit, phénomène qui entraîne l'ouverture des bornes de validation. Quatre de ces journées correspondent à des pics de pollution qui ont eu lieu en 2014 et six en 2015 (du vendredi 14 au lundi 17 mars 2014, du samedi 21 au lundi 23 mars 2015 et du dimanche 29 au lundi 30 novembre 2015). Les transports étaient également gratuits le dimanche 11 janvier 2015 en raison de la marche républicaine à travers Paris en réaction aux attentats djihadistes des 7, 8 et 9 janvier 2015. Nous avons également supprimé la période du 23 juillet au 24 août 2015 en raison des travaux impactant la ligne principale de l'étude (ligne RER A).

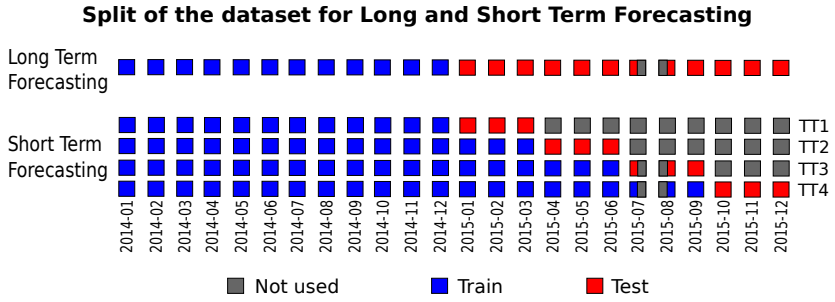

Fig. 4.8.: Échantillonage du jeu de données pour l'évaluation des modèles de prévision.

4.3.3.3. Configuration et développement des modèles de prévision

Dans cette section, nous détaillons les configurations des trois modèles, VAR, RF et GRU. Nous discutons de la taille de la fenêtre temporelle (lag) sélectionnée, de l'optimisation des hyperparamètres ainsi que de la bibliothèque et des ressources utilisées pour développer et entraîner les modèles.

Vector Autoregressive, VAR Nous avons utilisé la bibliothèque python Statsmodels [SP10] pour créer le modèle VAR. La valeur optimale du lag pour ce modèle a été sélectionnée en testant différentes valeurs (grid search) et correspond à 35 pas de temps (8h45 d'observation passée).

Random Forest, RF Nous avons utilisé scikit-learn [Ped+11], une bibliothèque Python, pour développer le processus d'apprentissage et de prévision des différents modèles RF. Le lag sélectionné est de 8 pour les modèles RF, ce qui correspond à 2 heures d'observations passées. Ce lag sélectionné de manière empirique permet d'obtenir de bons résultats de prévision et ne nécessite pas trop de ressources de mémoire et de temps de calcul.

Gated Recurrent Unit, GRU Nous avons utilisé Keras [Cho+15], une bibliothèque d'apprentissage profond, pour développer les différents modèles GRU. Afin de réduire le temps de calcul, nous avons utilisé une implémentation spéciale de la couche GRU, à savoir une couche CUDNN GRU. Cette couche spéciale est optimisée pour effectuer l'apprentissage et la prédiction de manière plus rapide que la couche GRU normale. Après quelques expériences sur cet ensemble de données, il a été démontré que la valeur du lag n'avait pas une grande influence sur la performance des modèles GRU. Ainsi, nous avons fixé la valeur du lag à 100 pas de temps, ce qui représente environ une journée d'observation.

Nous avons optimisé les hyperparamètres des modèles RF et GRU en testant différentes valeurs (grid search) avec la méthode de validation croisée (cross validation method). Les valeurs utilisées pour l'optimisation des modèles sont décrites dans la section A dans les tableaux A.2, A.3 et A.4. Pour plus de détails sur les différents hyperparamètres, il est possible de consulter la documentation des bibliothèques [Ped+11] et [Cho+15]. Les expériences ont été menées sur un ordinateur équipé de 6 cœurs, 32 Go de RAM et une carte graphique NVIDIA GeForce CUDA GTX 1080.

4.3.4 Résultats de prévision

La section 4.3.4.1 présente les résultats des modèles de prévision obtenus sur l'ensemble des jeux de données d'entraînement et de test (2014-2015). Par la suite, une analyse approfondie des résultats de prévision est effectuée dans les sections 4.3.4.2 et 4.3.4.3 avec un accent mis sur l'analyse de sous-ensembles de

données comprenant les jours fériés, les jours spéciaux et les périodes contenant des incidents survenus sur le réseau de transport. Enfin, la section 4.3.4.4 détaille la complexité des modèles de prévision en termes de temps de calcul et de mémoire utilisée.

4.3.4.1. Analyse des résultats sur la période globale d'entraînement et de test

Les résultats des deux modèles long terme et des quatre modèles de prévision court terme détaillés à la section 4.3.2 sont résumés en termes de RMSE dans le tableau 4.4. Selon la méthode d'évaluation expliquée dans la section 4.3.3.2, les performances multi pas de temps pour la période globale d'entraînement et l'ensemble de test sont données dans ce tableau. La grande différence entre l'erreur réalisée lors de la période d'entraînement et l'erreur obtenue en période de test des modèles multivariés (RF ST MULTI et GRU FUSION), démontre que ces modèles ont sur-appris. Le modèle univarié GRU UNI quant à lui parvient à obtenir les meilleurs résultats de prévision sur le jeu de test. En plus de ces analyses, nous pouvons également observer que les modèles GRU (GRU UNI et GRU FUSION) effectuent de meilleures prévisions multi pas de temps que les autres modèles court terme et semblent moins sensibles à la propagation des erreurs entre les différents pas de temps de prévision.

Analyse des résultats de prévision au pas de temps suivant (t+1) par station. En plus de l'analyse globale des performances de prévision, l'analyse de l'erreur de prévision par série temporelle (station) permet de comprendre quelles sont les stations les plus difficiles à prévoir. La figure 4.9 montre les erreurs MAE et MAPE@5 par série temporelle (stations ferrées et tramway) du meilleur modèle long terme RF LT et des modèles de prévision court terme.

L'axe des abscisses des deux graphiques représente les identifiants des stations triées par ordre décroissant du nombre de passagers. Nous pouvons remarquer que les erreurs MAPE@5 augmentent et les erreurs MAE diminuent en fonction du nombre de passagers par station. Cette tendance indique, comme nous pouvions nous y attendre, que les stations à faible fréquentation sont plus difficiles à prévoir que les stations à forte fréquentation (erreur MAPE@5 inférieure à 20% pour les stations les plus fréquentées contre plus de 35% pour les moins fréquentées). L'erreur MAPE@5 du modèle RF long terme montre que les stations les plus irrégulières, en termes de modèles basés sur les données calendaires, sont les stations de tramway avec

Tab. 4.4.: Résultats multi pas de temps des modèles de prévision court et long terme obtenus pendant la période globale d'entraînement et de test en termes de RMSE.

RMSE - Train set (2014) - Global set								
Modèles	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$	$t+8$
HA	87,2	87,2	87,2	87,2	87,2	87,2	87,2	87,2
RF LT	39,2	39,2	39,2	39,2	39,2	39,2	39,2	39,2
LOCF	64,3	104,6	145,1	182,3	217,6	248,8	275,9	299,4
VAR	31,3	38,1	45,7	53,3	61,0	67,7	73,2	77,9
RF ST UNI	20,6	27,1	31,9	35,6	39,2	42,4	45,4	47,9
RF ST MULTI	**13,0**	**18,8**	**23,7**	**28,3**	32,6	36,8	40,7	44,2
GRU UNI	27,2	29,3	31,7	33,6	35,1	36,5	37,6	38,5
GRU MULTI	27,8	30,4	32,5	34,3	35,7	36,9	37,9	38,9
GRU FUSION	25,0	27,0	29,1	30,7	**32,0**	**33,0**	**33,8**	**34,6**
RMSE - Test set (2015) - Global set								
HA	88,0	88,0	88,0	88,0	88,0	88,0	88,0	88,0
RF LT	57,3	57,3	57,3	57,3	57,3	57,3	57,3	57,3
LOCF	65,7	107,5	149,9	188,2	224,7	256,9	285,1	309,4
VAR	34,6	42,4	51,4	59,6	68,4	76,2	82,8	88,6
RF ST UNI	**31,3**	34,4	38,1	40,8	44,1	46,8	49,5	51,5
RF ST MULTI	41,0	42,9	45,4	47,6	50,3	52,4	54,4	56,2
GRU UNI	**31,3**	**33,1**	**36,0**	**38,2**	**40,5**	**42,4**	**44,5**	**46,1**
GRU MULTI	37,7	41,5	44,6	46,8	48,7	50,2	51,6	52,8
GRU FUSION	32,3	34,7	37,7	40,2	42,5	44,3	45,9	47,3

Les modèles de prévision sont décrits dans la section 4.3.2.

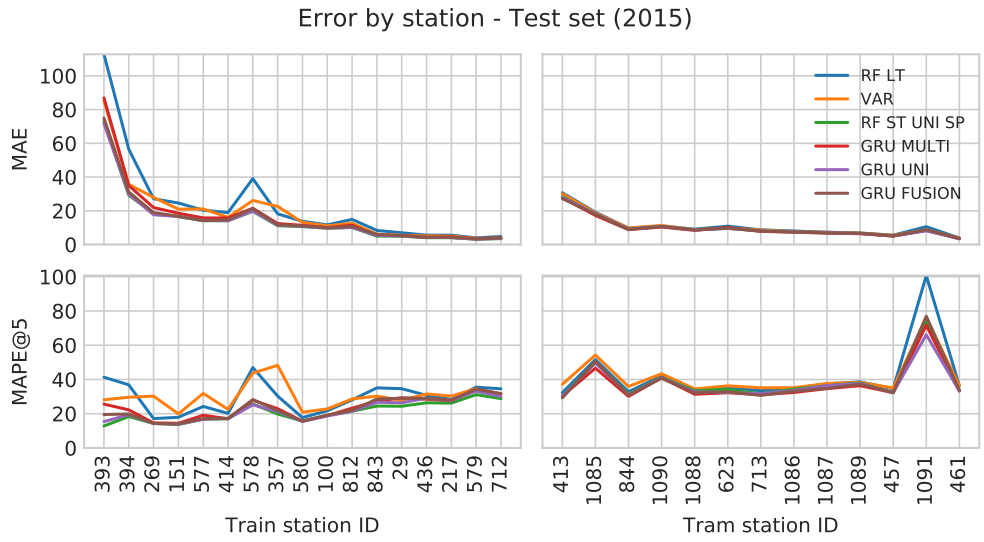

Fig. 4.9.: Erreurs MAE et MAPE@5 des différents modèles de prévision obtenues par station.

les identifiants 1091, 1085 et 1090 et les stations ferrées avec les identifiants 393,

394, 578, 843 et 29. Ces résultats peuvent indiquer que ces stations sont davantage affectées par des facteurs exogènes (conditions météorologiques, événements) ou des facteurs liés au réseau de transport, tels que des problèmes techniques ou des grèves.

Importance des caractéristiques en entrée du modèle RF long terme La figure 4.10 illustre l'importance de chaque caractéristique donnée par le modèle RF long terme pour l'ensemble des stations prédites. L'importance f de chaque caractéristique est calculée avec le facteur appelé "gini importance" [Bre+84]. L'importance de la caractéristique i, désignée par f_i, est donnée par l'équation 3.2 détaillée en section 3.5.1.3.

L'importance de chaque caractéristique est exprimée en pourcentage. Le pas de temps est la caractéristique la plus importante pour la prévision long terme avec un score de 76,60%, suivie du jour avec un score de 12,32% et des jours fériés avec un score de 4,06%. Le facteur le moins important est la veille de Noël.

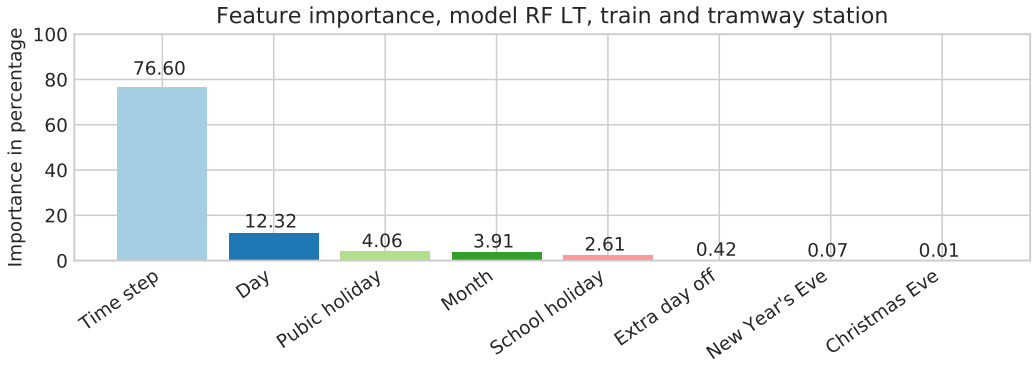

Fig. 4.10.: Importance des caractéristiques des données d'entrée (données calendaires) du modèle RF LT sur l'ensemble des stations de Paris étudiées.

4.3.4.2. Analyse des résultats en période spéciale : jours fériés, réveillon, veille du nouvel an

Les jours fériés, le réveillon de Noël et la veille du nouvel an, sont des jours spéciaux lors desquels l'affluence des passagers peut être considérablement différente de la normale. Nous analysons ici la prévision de l'affluence des passagers agrégée par 15 minutes, calculée à l'aide des modèles de prévision multi pas de temps au pas de temps $t + n$ avec n appartenant à $\{1, 8\}$. Dans le tableau 4.5 nous pouvons observer les erreurs RMSE et MAPE@5 sur un sous-ensemble de test correspondant à ces types de jours spéciaux (12 jours en 2015).

Les meilleurs résultats en termes de RMSE sont obtenus par le modèle GRU univarié jusqu'au pas de temps $t+3$. Nous pouvons remarquer que le modèle RF long terme effectue les meilleures prévisions aux pas de temps $t+n$ avec n entre 4 et 8. Ces résultats peuvent s'expliquer par le fait que les modèles court terme accordent plus d'importance à la valeur observée au pas de temps précédent qu'aux caractéristiques calendaires étant donné que la RMSE explose après plusieurs pas de temps de prévision. En termes de MAPE@5, les résultats multi pas de temps sont moins affectés par la propagation de l'erreur. Dans ce cas les meilleurs modèles sont les modèles multivariés GRU, en raison de leur grande capacité à généraliser, les modèles GRU MULTI et GRU FUSION sont les meilleurs modèles de prévision en termes de prévision multi pas de temps.

Tab. 4.5.: Résultats des modèles de prévision court et long terme obtenus lors des jours spéciaux suivants: jours fériés, réveillon, veille du nouvel an.

	RMSE							
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$	$t+8$
HA	294,6	294,6	294,6	294,6	294,6	294,6	294,6	294,6
RF LT	31,3	31,3	31,3	**31,3**	**31,3**	**31,3**	**31,3**	**31,3**
LOCF	26,7	36,3	45,8	53,7	62,0	68,3	74,0	78,8
VAR	24,6	36,1	47,3	56,8	65,9	74,0	80,9	86,6
RF ST UNI	26,6	37,3	46,0	53,7	62,3	70,6	78,2	84,6
RF ST MULTI	25,7	31,7	39,3	44,4	49,8	53,8	57,9	64,1
GRU UNI	**21,4**	**25,3**	**30,7**	35,5	39,5	42,6	45,2	46,9
GRU MULTI	23,6	29,2	34,7	38,8	41,4	43,1	44,3	45,0
GRU FUSION	21,5	26,6	31,7	35,8	38,4	40,4	42,1	44,2
	MAPE@5							
HA	253,6	253,6	253,6	253,6	253,6	253,6	253,6	253,6
RF LT	39,1	39,1	39,1	39,1	39,1	39,1	39,1	**39,1**
LOCF	43,6	46,7	51,6	56,2	61,6	67,7	73,2	78,6
VAR	38,2	44,8	53,8	63,2	74,0	84,6	94,2	102,5
RF ST UNI	35,5	38,4	41,3	44,5	47,9	50,9	53,9	56,7
RF ST MULTI	33,6	35,2	36,6	38,0	39,8	41,7	43,6	45,8
GRU UNI	33,8	34,5	35,8	36,9	37,9	39,0	40,0	40,7
GRU MULTI	**32,6**	**33,8**	**35,3**	**36,5**	37,7	39,0	40,2	40,9
GRU FUSION	33,2	34,2	35,5	**36,5**	**37,3**	**38,1**	**38,8**	39,4

Les modèles de prévision sont décrits dans la section 4.3.2.

La figure 4.11 montre l'observation et la prévision du meilleur modèle long terme (RF LT), des meilleurs modèles court terme (GRU MULTI et GRU FUSION) en termes de MAPE@5 et du modèle long terme naïf (HA) la veille de Noël à la station 269.

Comme nous pouvons le voir avec la prévision du modèle HA (courbe bleue), cette journée est différente de la normale. Les modèles multivariés court terme réalisent

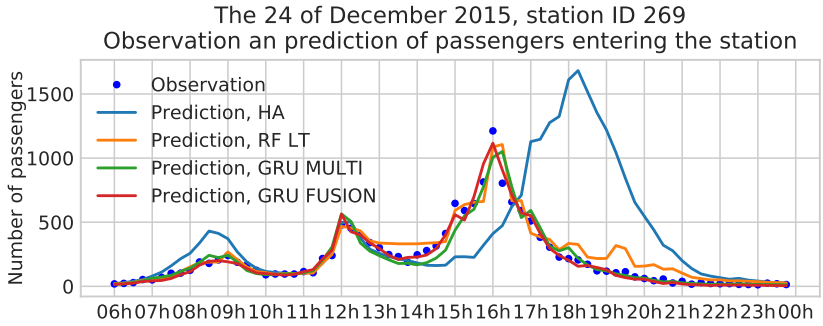

Fig. 4.11.: Observation et prévision de l'affluence des passagers à la station "Esplanade de la Défense", station 269, le jour du réveillon de Noël.

les meilleures prévisions. De plus, nous pouvons observer que le modèle RF LT long terme réalise une prévision pertinente de l'affluence des passagers en ce jour atypique, ce qui prouve que les informations calendaires ont réussi à être prises en compte par le modèle.

4.3.4.3. Analyse des résultats en période spéciale : Période d'incident du réseau de transport

Les incidents qui surviennent sur le réseau de transport peuvent avoir un impact direct sur l'offre de transport par conséquent, la demande des passagers peut également être affectée. Pour quantifier la robustesse des modèles de prévision au cours de ce type de période, nous évaluons leur performance sur une période définie à l'aide d'une base de données d'incidents fournie par l'opérateur de transport d'Ile-de-France (Ile-de-France Mobilités). Pour cela, nous avons extrait de cette base de données les jours possédant au moins un incident sur une des 30 stations étudiées, ce qui nous a permis d'obtenir un sous-ensemble de 45 jours pendant la période de test (2015). Les incidents peuvent être catégorisés comme suit :

- Panne ou problème externe, tel qu'un incendie.
- Rénovation de la station.
- Grève.
- Journées spéciales avec l'ouverture des systèmes AFC.

Les résultats des modèles de prévision court et long terme obtenus pendant la période d'incident sont présentés dans le tableau 4.6. Nous pouvons observer que comme prévu, les modèles long terme ont beaucoup de difficulté à prévoir la fréquentation des usagers pendant la période d'anomalie. Cependant nous pouvons voir que les modèles univariés court terme GRU UNI et RF ST UNI atteignent des niveaux de performance comparables aux performances présentées dans le tableau 4.4 sur la

période d'étude globale. D'autre part nous pouvons voir que les modèles multivariés ne sont pas en mesure de capter le lien spatio-temporel entre les différentes séries temporelles puisque leur performance diminue. En effet, des incidents spécifiques peuvent être caractérisés par un changement d'itinéraire des passagers par rapport à leur itinéraire initial, induisant une forte diminution de l'affluence des passagers à des stations données et une augmentation soudaine de l'affluence à des stations liées par rapport au réseau de transport (e.g., stations proches, stations appartenant à plusieurs lignes de transport, etc.). Les modèles multivariés peuvent, en théorie, apprendre de telles relations entre les stations et produire des résultats similaires à ceux obtenus lors de situations normales, mais en pratique, ces relations sont très difficiles à saisir de par leur rareté au sein de la base d'apprentissage.

Tab. 4.6.: Résultats des modèles court et long terme à plusieurs pas de temps sur la période d'incident, incidents survenus sur le réseau de transport en 2015.

	RMSE							
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$	$t+8$
HA	114,8	114,8	114,8	114,8	114,8	114,8	114,8	114,8
RF LT	97,2	97,2	97,2	97,2	97,2	97,2	97,2	97,2
LOCF	56,7	90,9	125,9	157,4	187,5	214,0	237,1	257,2
VAR	37,1	48,7	61,0	71,9	82,8	92,7	100,4	107,2
RF ST UNI	32,3	38,4	44,9	50,5	55,8	59,5	63,2	66,1
RF ST MULTI	62,9	66,0	71,5	76,5	82,2	86,2	89,1	91,2
GRU UNI	**32,3**	**35,6**	**40,9**	**45,7**	**50,7**	**55,4**	**60,1**	**64,1**
GRU MULTI	52,6	59,8	66,4	71,1	75,2	78,8	81,9	84,6
GRU FUSION	37,6	42,2	48,5	54,2	59,3	63,6	67,5	70,8
	MAPE@5							
HA	57,0	57,0	57,0	57,0	57,0	57,0	57,0	57,0
RF LT	47,7	47,7	47,7	47,7	47,7	47,7	47,7	47,7
LOCF	41,6	48,2	58,4	68,2	79,2	90,2	101,5	112,4
VAR	37,8	44,7	52,9	60,7	69,2	76,7	83,3	88,3
RF ST UNI	**29,9**	**31,2**	**32,6**	34,4	35,6	36,9	38,1	39,2
RF ST MULTI	39,0	40,6	42,2	43,6	45,1	46,3	47,2	48,5
GRU UNI	30,7	31,3	**32,6**	**33,8**	**35,1**	**36,3**	**37,5**	**38,4**
GRU MULTI	35,0	36,6	38,5	39,8	41,2	42,3	43,2	43,8
GRU FUSION	32,4	33,2	34,8	36,0	37,2	38,1	39,0	40,0

Les modèles de prévision sont décrits dans la section 4.3.2.

Pour illustrer le changement d'itinéraire des passagers en cas d'incident, nous pouvons observer dans la Figure 4.12, l'observation et la prévision de l'affluence des passagers dans trois stations différentes. La situation atypique est un incendie qui est survenu à 14h00 le vendredi 9 octobre 2015 à la station 577. La figure 4.12 montre l'affluence des passagers dans trois des stations impactées par l'incident :

- "Nanterre-Préfecture", RER ligne A, identifiant 577

- "La Défense (Grande Arche)", Transilien ligne L, identifiant 394
- "La Défense (Grande Arche)", RER ligne A, identifiant 414

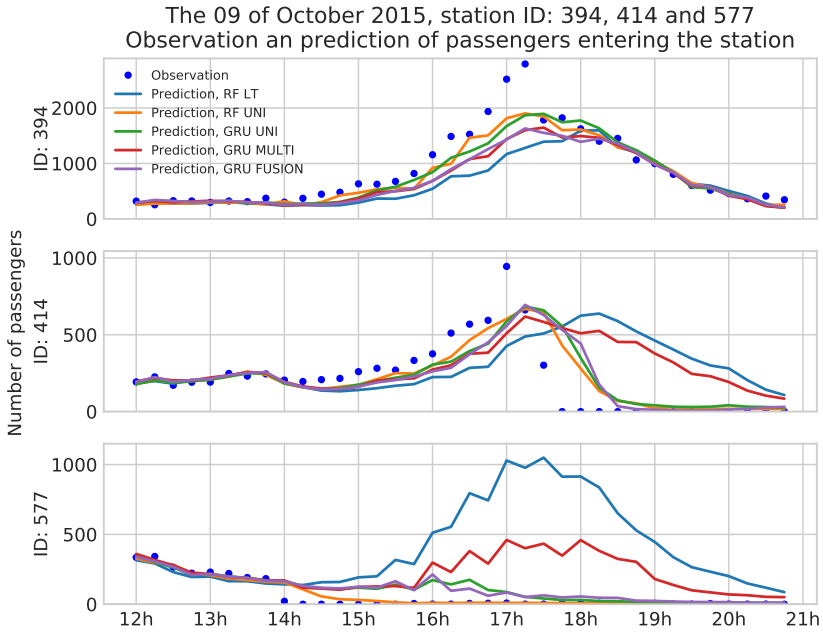

Fig. 4.12.: Observation et prévision des modèles court et long terme le 9 octobre 2015 lors d'un incident technique (incendie) impactant 3 stations.

Nous pouvons observer une forte diminution du nombre de passagers à la station 577 à 14h00 et une augmentation du nombre de passagers aux stations 394 et 414. La demande de transport dans ce cas est en effet différente de l'observation habituelle fournie par le modèle RF long terme, qui correspond à l'affluence des passagers en situation normale. De plus, nous pouvons voir une baisse soudaine du nombre de passagers à 0 entre 17h15 et 17h30 à la station 414 et un retour à la normale de la fréquentation à la station 394. Cette forte diminution est due à l'ouverture des bornes de validation de la station 414 en raison du trop grand nombre de passagers. Si nous examinons de plus près les résultats des prévisions, nous pouvons observer que les modèles court terme univariés semblent produire de meilleures prévisions que celles des modèles court terme multivariés. Cependant, aucun des modèles ne parvient à prévoir avec succès le pic de passagers aux stations 394 et 414 à 17h00. Au vue de cet exemple, nous pouvons dire que ces modèles multivariés ne sont pas en mesure de capter des relations spatio-temporelles aussi complexes entre les séries temporelles. Cela pourrait être dû au fait que seul un petit nombre d'observations de ce type (incident) existe dans la base de données. En effet moins de 10% (27 jours en 2015) des jours appartenant aux données d'apprentissage comportent des incidents qui n'impactent pas obligatoirement la demande, ce qui ne permet pas aux modèles d'apprendre de telles relations atypiques. De plus de

nombreuses configurations de perturbations sont possibles, ce qui rend d'autant plus difficile leur prévision à l'aide de modèles statistiques.

Une représentation de la structure du réseau de transport avec un focus sur les stations et les lignes impactées par l'incident est montrée dans la Figure 4.13.

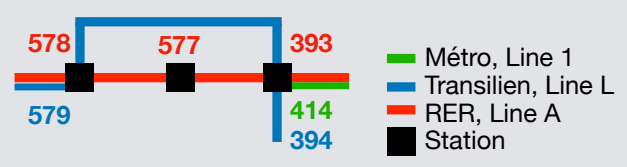

Fig. 4.13.: Structure du réseau de transport avec un focus sur les stations impactées par l'incident du 9 octobre 2015 (incendie).

Le tableau 4.14 montre un classement des meilleurs modèles court terme en fonction du pourcentage du nombre de jours possédant un incident que les modèles arrivent à prédire avec le meilleur score. Les meilleurs modèles de prévision en cas d'incident sont les modèles univariés RF ST UNI et GRU UNI. Nous pouvons observer que les modèles multivariés réalisent la meilleure prévision pour certains jours comportant au moins un incident.

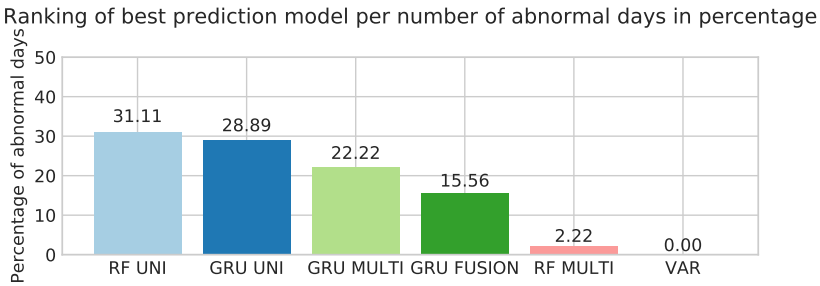

Fig. 4.14.: Classement des modèles de prévision en fonction du nombre de jours avec incident prédit avec le meilleur score.

4.3.4.4. Complexité des modèles de prévision en fonction de leur performance

La performance des modèles en termes de complexité est également intéressante à évaluer. Le tableau 4.7 décrit la complexité des modèles court terme en fonction du nombre de paramètres, de l'erreur RMSE obtenue sur l'ensemble de test global et du temps d'apprentissage des modèles. La complexité des modèles court terme VAR et GRU est calculée comme étant le nombre de leurs paramètres. La complexité des modèles RF est représentée par le nombre total de nœuds dans l'ensemble des arbres. Le tableau 4.7 montre que le meilleur modèle court terme en fonction de la

complexité et du temps d'apprentissage est le modèle VAR, suivi du modèle GRU MULTI.

Tab. 4.7.: Complexité des modèles de prévision court terme en fonction de leur nombre de paramètres, de l'erreur RMSE et de leur temps de calcul.

Modèle	Complexité	RMSE (Test)	Durée de l'apprentissage (s)
VAR	31500	34,6	1,5
RF ST UNI	69623492	31,3	103,3
RF ST MULTI	7399284	41,0	855,0
GRU UNI	2058780	31,3	1267,0
GRU MULTI	546630	37,7	102,0
GRU FUSION	2618970	32,3	1414,0

La figure 4.15 illustre l'erreur RMSE sur l'ensemble de test (2015) et la complexité logarithmique des modèles court terme. Le meilleur modèle en termes de complexité et d'erreur RMSE est le modèle VAR.

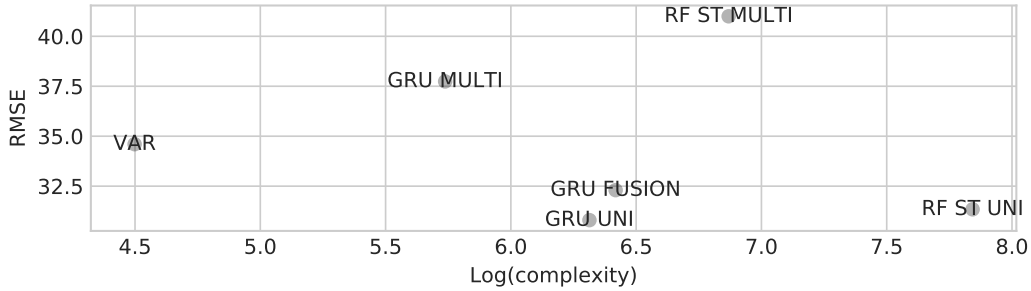

Fig. 4.15.: Erreur RMSE et log complexité des modèles de prévision long terme.

La figure 4.16 représente l'erreur RMSE sur le jeu de test (2015) et le temps d'apprentissage des modèles court terme. Les meilleurs modèles en termes de complexité et de temps d'apprentissage sont les modèles RF ST UNI et VAR.

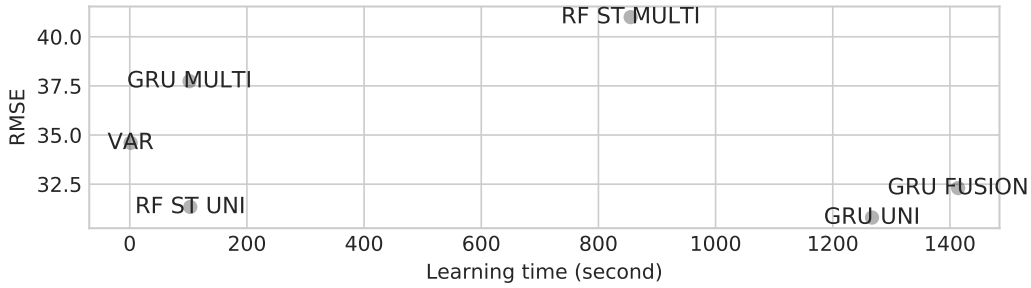

Fig. 4.16.: Erreur RMSE et temps d'apprentissage (secondes) des modèles de prévision court terme.

4.3.5 Conclusion

Dans cette section nous avons étudié la prévision court terme multi pas de temps de l'affluence des passagers en entrée de quelques stations ferrées et de tramway de Paris en utilisant des données billettiques. Ce type de prévision vise à prévoir à plusieurs pas de temps (ici jusqu'à t+8) le nombre de passagers agrégé au quart d'heure, entrant dans chaque station. Contrairement à l'horizon temporel long terme où l'objectif réside dans l'amélioration de la planification de l'offre de transport, les prévisions court terme visent à aider les opérateurs de transport à proposer des itinéraires pour éviter la congestion de passagers ou des incidents ayant lieu dans le réseau de transport et de fournir une meilleure information aux voyageurs dans un contexte temps réel.

Les résultats ont démontré que les méthodes court terme permettent d'obtenir de meilleures performances de prévision face au méthodes long terme. Cependant il est à noter qu'il reste difficile de prévoir l'affluence des passagers lors des périodes d'incident même avec les méthodes court terme (e.g., prévision de l'affluence des passagers suite à un incendie survenu sur le réseau de transport). De manière globale, des résultats fiables ont été obtenus pour les deux modes de transport (ferré et tramway). L'étude de cas décrite dans cet article est particulièrement difficile car elle concerne un grand quartier d'affaires de l'agglomération parisienne (La Défense). Nous avons proposé un modèle Random Forest long terme utilisant des informations calendaires qui surpassent le modèle naïf de prévision long terme, le modèle utilisant la moyenne historique. La prévision court terme à plusieurs pas de temps a été réalisée à l'aide de modèles Random Forest et Gated Recurrent Unit, un réseau de neurones récurrent proposé issu de l'apprentissage profond. Différents modèles de prévision ont été pris en compte, y compris des modèles univariés et multivariés. Les résultats obtenus sur la période globale montrent que le meilleur modèle de prévision court terme multi pas de temps est le modèle GRU univarié. En plus de l'analyse globale des résultats, nous avons effectué une analyse approfondie en nous concentrant sur des périodes atypiques en termes de demande des passagers : (i) les résultats obtenus lors des jours fériés ainsi que d'autres jours spéciaux montrent que les modèles multivariés court terme réalisent de meilleurs résultats que les modèles univariés dans ce type de contexte ; (ii) les résultats obtenus en période d'anomalie du réseau de transport montrent que les modèles univariés court terme sont meilleurs dans ce type de cas atypique et que les modèles multivariés ne permettent pas totalement de saisir les liens spatiaux entre les différentes séries temporelles.

Les travaux futurs devraient porter sur l'amélioration des résultats de prévision court terme. En particulier, il serait intéressant de construire et d'évaluer différentes architectures de réseaux de neurones étant donné que ces modèles ont permis d'obtenir de meilleures performances de prévision sur l'ensemble de test global. De plus, les résultats de prévision montrent que l'aspect spatial du réseau de transport n'est pas bien pris en compte par les modèles. Il pourrait être pertinent d'étudier des modèles permettant de saisir les liens spatiaux entre les séries temporelles, ce qui permettrait une meilleure prise en compte du réseau de transport sous-jacent et, plus particulièrement, du transfert modal qui peut être généré par des incidents non prévisibles. Toutefois, cette analyse nécessite un ensemble de données comportant un nombre important de situations atypiques ayant un impact sur la demande des passagers.

5 Visualisations interactives pour l'analyse de l'affluence des passagers dans les transports en commun

5.1 Résumé

L'analyse des résultats de prévision de séries spatio-temporelles peut s'avérer difficile à réaliser du fait de la grande quantité d'information à traiter. Cependant ces analyses sont primordiales à la compréhension des résultats et plus en amont, à l'amélioration de ceux-ci. Dans ce sens, l'utilisation d'outils de visualisation s'avère être un choix pertinent pour résoudre ce problème.

Nous nous attachons à présenter dans ce chapitre quelques travaux menés sur la visualisation des données de mobilité urbaine ainsi que les ressources technologiques actuelles permettant de créer des visualisations interactives. Nous présentons également deux outils de visualisation que nous avons développés dans le but d'analyser plusieurs ensembles de données spatio-temporelles, à savoir, des données billettiques, événementielles et d'incidents survenus sur le réseau de transport. Ces outils permettent principalement d'analyser conjointement les résultats de la prévision de la demande des passagers ainsi que les données contextuelles pouvant impacter l'offre de transport, dans un cadre spatial et temporel. Ce type de visualisation permet ainsi de mieux comprendre pourquoi certains types d'erreurs sont commis par les modèles prédictifs et comment améliorer ces derniers.

5.2 Introduction

D'un point de vue opérationnel, l'exploration et la visualisation d'une grande quantité de données peuvent aider à construire des outils d'aide à la décision permettant de mieux analyser les systèmes de transport en commun, de mieux répondre aux besoins futurs par une meilleure prévision et de mieux adapter l'offre des transports urbains

aux besoins des citoyens. La création d'outils de visualisation spatio-temporelle est ainsi devenue un secteur important dans lequel de nombreuses startup à forts moyens telles que Uber, n'hésitent pas à investir. Cet intérêt rejoint l'idée que l'un des principaux défis des "transports en commun de demain" réside dans la capacité à utiliser et à maîtriser en avance ou en temps réel, la multitude de données pouvant impacter l'offre et/ou la demande de transport afin d'optimiser le fonctionnement des systèmes de déplacements urbains en conséquence. Comme l'avance les auteurs de [BCO17], un exemple possible de ce contrôle est la combinaison croisée des conditions météorologiques, des événements (e.g., concert, rencontre sportive, etc.) et des services de transport (offre et demande) afin de fournir aux citoyens des services de transport adaptés aux conditions météorologiques et aux événements particuliers. Un autre exemple serait d'adapter l'offre de transport à la demande en temps réel en cas d'incident ayant lieu sur le réseau de transport (e.g., problème technique, accident grave de passager, etc.) à l'aide de scénarios développés grâce à l'analyse de la demande des passagers lors de ce type d'incident.

Comme nous l'avons déjà vu dans les chapitres 3 et 4, l'apprentissage automatique peut jouer un rôle important dans le développement des modèles de prévision. Cependant, l'analyse détaillée de l'affluence des passagers et des erreurs de prévision s'avère être un défi au vue du grand nombre de séries temporelles et d'éléments exogènes à analyser (e.g., météorologie, événement, incident, etc.). Une façon appropriée de s'attaquer à ce problème consiste à combiner la prévision effectuée par des outils d'apprentissage automatique avec une exploration et une analyse humaine aidés par des outils de visualisation interactive. Dans notre cas, nous exploitons les différentes bases de données mises à notre disposition (billettique, événementielle, incident) pour étudier l'impact des situations atypiques sur la demande des passagers (e.g., temps de l'incident, stations impactées, etc.) à l'aide de visualisations capables de regrouper toutes ces informations à différentes échelles temporelles et spatiales. Par ailleurs, ces outils de visualisation permettent également de répondre à plusieurs questions, telles que (i) l'amélioration et l'enrichissement des bases de données incidents et événementielles (ces bases sont souvent remplies manuellement et sont souvent incomplètes) et (ii) le fait d'obtenir une meilleure compréhension des résultats de prévision en détectant de manière plus efficace les périodes et les stations difficiles à prévoir.

Ce chapitre est divisé en trois parties. Nous détaillons dans la première section 5.3, les ressources et les études récentes s'attachant au développement d'outils de visualisation dans le domaine de la mobilité. La section 5.5 traite des outils de visualisation interactive que nous avons développés dans le but d'analyser plus efficacement les prévisions de séries spatio-temporelles en prenant en compte le contexte (e.g., date,

station, présence d'événement et d'incident). Enfin une conclusion et les pistes d'études de ce chapitre seront données en section 5.6.

5.3 État de l'art des méthodes de visualisation de la mobilité urbaine

L'utilisation des données billettiques pour l'analyse de la mobilité dans les transports en commun a fait l'objet d'une attention considérable de la part des chercheurs, comme nous pouvons le constater dès 2011 avec les travaux de [PTM11]. De nombreux graphiques spécialisés ainsi que des outils de visualisation ont été développés au cours de ces dernières années pour analyser en détail la mobilité urbaine. Leur but est d'aider les opérateurs et les décisionnaires de transport à mieux analyser la mobilité dans le but de répondre à des besoins de planification et d'amélioration des services proposés aux usagers. De nombreux exemples de visualisations de différents modes de transport (e.g., bus, métro, vélo en libre service, etc.) étudiés par des chercheurs ou des startups sont détaillés dans les sections suivantes. Nous pouvons différencier trois types de visualisation, à savoir, (i) les outils d'analyse de l'offre de transport en commun détaillés en section 5.3.1, (ii) les outils d'analyse de la demande de mobilité décrite par la fréquentation des réseaux de transport en chaque station ou ligne, détaillés en section 5.3.2 et (iii) les outils mêlant visualisation de l'offre et de la demande, décrits en section 5.3.3.

5.3.1 Outils d'analyse de l'offre de transport en commun

Certains outils de visualisation sont dédiés à analyser l'offre de transport à savoir les horaires de passage des différents moyens de déplacement (e.g., métro, train, tramway, bus, etc.). L'objectif applicatif de ce type de visualisation vise à mieux comprendre l'offre de transport en vue d'optimiser la fréquence des passages des différents moyens de transport.

On peut notamment citer les auteurs de [SBR15] qui ont développé une application web qui utilise des graphiques de type "marey graph" (graphiques temps-distance), crées par Marey [Hra+11], permettant d'analyser les déplacements des métros de la ville de New York en temps réel. L'application a été conçue en interne à l'aide des ressources de données existantes (la source de données temps réel du GTFS sur les

arrivées prévues des trains) et d'outils open-source. Un exemple de graphe de Marey est illustré dans la figure 5.1.

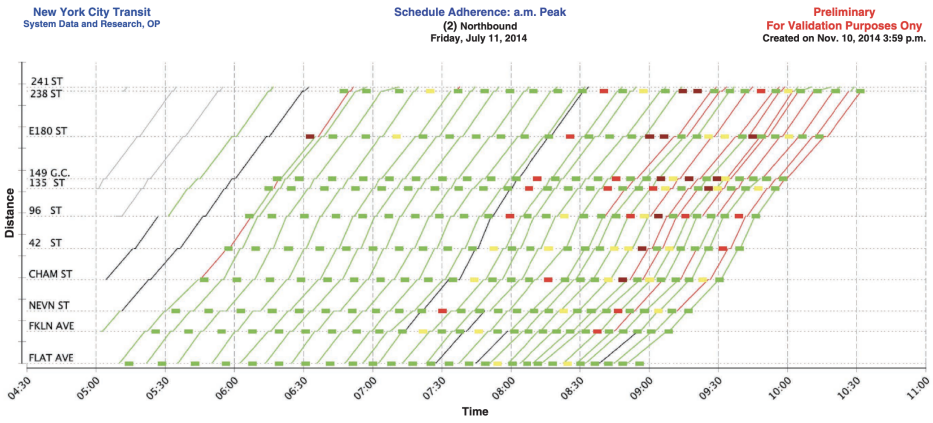

Fig. 5.1.: Variante d'un graphique de type marey graph détaillant le retard sur l'horaire de passage prévu des trains en plus de l'horaire d'arrivée des trains dans certaines stations du métro de New York City aux Etats-Unis. Source [SBR15].

Dans le même ordre d'idée, l'étude [AOT16] présente BusViz, une visualisation interactive qui aide les exploitants de services d'autobus et les opérateurs de transport en commun à prendre de meilleures décisions en utilisant une grande quantité de données observées pour analyser et visualiser la performance des opérations d'autobus de la ville de Singapour. Les auteurs utilisent notamment des graphes de Marey pour analyser l'offre de bus. A l'aide de cette application les utilisateurs peuvent également ajouter des services d'autobus à une carte interactive pour afficher la direction et le trajet physique de chaque service d'autobus.

5.3.2 Outils d'analyse de la demande des passagers

Dans le but d'adapter l'offre de transport à la demande des passagers et de proposer de meilleurs services aux usagers (e.g., tarification, information voyageur, etc.), certains chercheurs et industriels ont développé des outils de visualisation permettant d'analyser l'affluence et les déplacements des passagers pour en extraire de l'information.

Dans sa thèse, [Gor12] a travaillé sur la visualisation des origines de déplacements quotidiens des usagers du réseau de transport de Londres, collectées à l'aide de la carte à puce "Oyster card". D'une manière plus générale, les membres du laboratoire Senseable City Laboratory du MIT ont réalisé un outil de visualisation [KSR12] dans lequel les utilisateurs peuvent passer d'un mode de visualisation à l'autre pour obtenir différentes perspectives sur le même ensemble de données. Les auteurs ont

expérimenté cette visualisation sur les données billettiques issues du réseau de bus de Singapour.

En mixant une méthode d'apprentissage automatique et une visualisation cartographique détaillée (e.g., détails de l'utilisation du territoire, tels que le nombre d'habitant, le taux d'emploi moyen par habitant, etc.), les auteurs de [EL14] ont développé une interface graphique permettant de visualiser les résultats d'une classification des stations de vélo en libre-service de la ville de Paris en fonction de leur profil d'utilisation. Sur le même type de données, les auteurs de [CMO14] ont développé une visualisation innovante permettant d'analyser de manière interactive les ODs les plus fréquentées par les utilisateurs du système de vélo en libre-service de Paris à l'aide de carte de chaleur. Un exemple de cette visualisation est donné dans la figure 5.2.

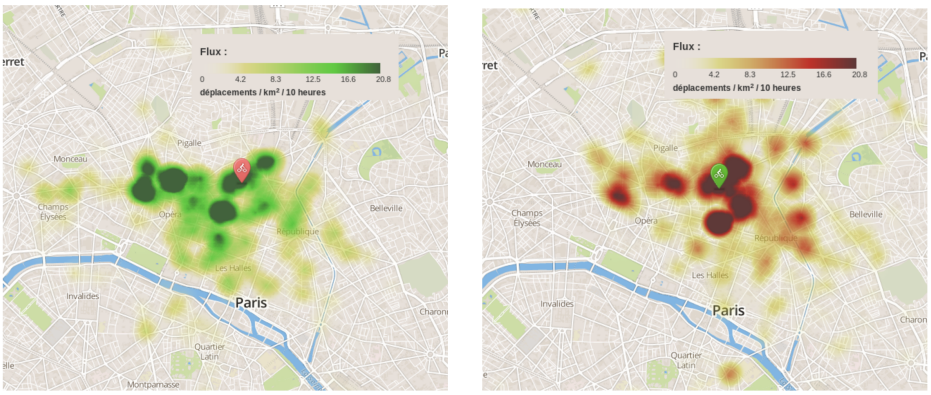

Fig. 5.2.: Visualisation des ODs les plus fréquentées pas les utilisateurs du système de vélo en libre-service de Paris en France. Source [CMO14].

Concernant le réseau de bus, des cartes de flux conditionnels (ou flow-comaps) développées par [TRC14] permettent de comparer visuellement la demande des passagers des Bus à Haut Niveau de Service (BHNS) ou Bus Rapid Transit (BRT) en anglais et autres trajets de bus en fonction des événements calendaires (par exemple, une journée de travail, un week-end, un congé scolaire et un jour férié). Leur étude porte sur l'analyse des données billettiques (entrées et sorties) des utilisateurs du réseau de BHNS de la ville de Brisbane en Australie.

Dans [Yok+14; Ito+14] les auteurs ont concentré leur travail sur la visualisation des flux de passagers (origine-destination) et les données des médias sociaux (tweets) à l'aide de rubans orientés sur une carte et d'une carte de chaleur. L'objectif est ici d'analyser facilement le changement de comportement des passagers (choix d'itinéraire) en cas de plainte des passagers visant les services de transport.

5.3.3 Outils de visualisation de l'offre de transport et de l'affluence des passagers

Certains outils de visualisation ont le double objectif d'analyser l'offre de transport et la demande des passagers. On peut notamment citer les auteurs de [TV08] qui dès 2008, ont créé un site intranet pour fournir des statistiques quotidiennes sur les opérations de transport. Les opérateurs de transport peuvent grâce à cet outil, obtenir le profil de charge de n'importe quel parcours d'autobus exploité, à n'importe quelle date, et utiliser cette information pour améliorer les services fournis. Les limites du système de billettiques comprennent néanmoins le manque de données exhaustives de tous les passagers, en particulier les jours d'événements spéciaux, ainsi que le manque d'information en temps réel (les statistiques ne sont disponibles que le lendemain).

Les auteurs de [BD14] ont développé un outil de visualisation en ligne permettant d'analyser les horaires d'arrivée du métro de Boston. Ils ont également élaboré une carte de chaleur qui montre le nombre moyen de personnes qui entrent et sortent des stations des différentes lignes du réseau de métro à chaque heure sur une période d'un mois. Enfin, ils ont mis au point un outil permettant d'analyser la durée de trajet entre différentes OD et le temps d'attente entre deux trains pour chaque jour de la semaine.

Plus récemment, les auteurs de [Gir+16] ont proposé une interface web pour analyser l'utilisation des bus et les profils de charge des lignes empruntées par les usagers dans des zones spécifiques du réseau.

5.3.4 Outils développés par des startups

Au cours de ces dernières années, plusieurs startups ont travaillé sur la réalisation d'outils de visualisation permettant d'analyser des éléments de la mobilité urbaine tels que les déplacements des taxis et des individus dans une ville. Uber a notamment crée deux outils de visualisation spatiale et interactive détaillés dans la section 5.4.1 (Kepler.gl et deck.gl). Deux exemples de visualisations réalisées avec deck.gl sont données dans la figure 5.3.

La startup sharedstreets.io a développé un outil open source permettant d'analyser les données de mobilité dans le but d'apporter une connaissance approfondie des différents éléments caractérisant la mobilité des personnes dans une ville (e.g., nombre de scooters en libre service disponibles, durée d'un trajet moyen, etc.).

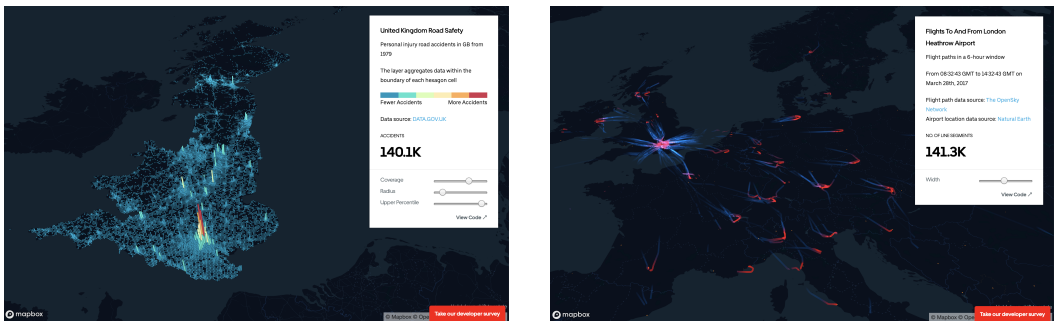

Fig. 5.3.: Exemples de visualisations cartographiques réalisées à l'aide de l'outil deck.gl (hexagone et ligne). Sources: [Ubea; Ubeb]

La société remix est également connue pour ses outils de visualisation interactive permettant d'analyser des données de mobilité spatio-temporelles. Ses outils rendent notamment possible la visualisation de flux d'individus dans un réseau de transport. Un de leurs outils permet également de simuler la création de nouvelles lignes de transport dans une ville à l'aide d'une interface graphique.

Nous avons détaillé dans cette section une liste de travaux portant sur les outils d'analyse de l'offre de transport en commun, les outils d'analyse de la demande de mobilité décrite par la fréquentation des réseaux de transport en chaque station ou ligne, les outils mêlant visualisation de l'offre et de la demande, ainsi que les outils de visualisation de la mobilité urbaine, développés au sein de startups au cours de ces dernières années. Les pistes de recherche concernant la visualisation des offres de transport ainsi que de la mobilité des individus au sein des transports en commun visent à prendre en compte de nouvelles sources de données qui n'ont pas encore été exploitées. Par exemple, nous pouvons envisager que la prise en compte de données exogènes telles que la météo, la présence d'événements et d'incidents dans un système globale de visualisation sera un intérêt majeur de l'analyse de l'affluence des passagers dans les transports en commun. En effet, la prise en compte de ces données peut permettre une meilleure analyse de la mobilité des individus étant donné que la mobilité peut se voir impactée par ce type de facteurs externes.

5.4 Ressources permettant la création de visualisations spatio-temporelles

Les services de cartographie correspondent le plus souvent à des cartes interactives qui constituent la principale interface utilisateur pour la plupart des applications des Systèmes d'Information Géographique (SIG). Ces services s'avèrent très utiles

pour afficher des systèmes de transport en commun et comprendre la dynamique spatiale des réseaux de transport. Différents exemples d'interfaces de programmation d'application (API pour application programming interface) utiles pour afficher des fonds de carte sont détaillés dans la section 5.4.1. Certains de ces services de cartographie permettent également de rajouter des éléments pour analyser des jeux de données.

Dans le but de créer des visualisations spatio-temporelles spécialisées, il est possible de mêler l'utilisation de ces services de cartographie à des outils de visualisation interactive à l'aide de différentes bibliothèques informatiques. Nous avons détaillé quelques librairies récentes permettant de créer des visualisations interactives dans la section 5.4.2.

5.4.1 Services de cartographie personnalisable

Il existe différentes API (Application Programming Interface) permettant de réaliser des cartographies personnalisables. L'API de cartographie la plus connue est celle de Google Maps, cette API a été créée par Google et lancée en 2004 aux Etats-Unis et au Canada et en 2006 en France.

Technologies basées sur le framework WebGL Certaines API de cartographie sont basées sur le framework WebGl crée en 2011. Cet outil permet de gérer, de créer et d'afficher des éléments graphiques complexes en 3D, dans la plupart des navigateurs modernes, du côté client. Le principal avantage de ce framework réside dans sa capacité à appeler le pilote OpenGL ES du système d'exploitation du client dans le but d'accélérer les calculs nécessaires à l'affichage des différents éléments graphiques de type WebGL, en exploitant si possible l'accélération matérielle du ou des processeurs graphiques du terminal.

Plusieurs technologies de visualisations spatiales interactives sont basées sur ce framework :

- **Mapbox GL JS** est une bibliothèque JavaScript open source qui utilise le framework WebGL pour afficher des cartes interactives à partir de tuiles vectorielles et de style Mapbox. Nous avons utilisé cette bibliothèque ainsi que des fonds de carte Mapbox pour créer la visualisation spatiale définie en section 5.5.3.

- **deck.gl** est un framework WebGL développé par UBER dans le but d'explorer et de visualiser des ensembles de données à large échelle. Ce framework peut être utilisé avec Mapbox GL.
- **Kepler.gl** est un framework développé par UBER permettant l'exploration de larges ensembles de données géolocalisées. Développé en s'appuyant sur les technologies Mapbox GL et deck.gl, kepler.gl peut afficher des millions de points représentant des milliers de déplacements et effectuer des agrégations spatiales à la volée.

5.4.2 Librairies de visualisation interactive

L'émergence de la science des données a engendré la création de plusieurs librairies de visualisation permettant de mieux analyser les données.

- **Tableau** est un outil souvent utilisé pour de la visualisation de données. L'avantage principal de cet outil est sa facilité d'utilisation. Son principal problème est sa faible flexibilité et son coût.
- **Leaflet** est la principale bibliothèque JavaScript open-source pour les cartes interactives. Cette librairie est directement utilisable depuis le langage R et depuis Python à l'aide de la librairie Folium. Elle est capable de fonctionner avec différents services de tuiles (e.g., Mapbox, Mapzen, OpenStreetMap, etc.).
- **Plotly.js** est une bibliothèque JavaScript open-source permettant de créer des graphiques et des dashboard interactifs.
- **Dash** est un framework Python et R open-source permettant la création d'applications analytiques et interactives.
- **D3.js** (ou D3 pour Data-Driven Documents) créée par Michael Bostock [BOH11] est une bibliothèque graphique JavaScript open-source qui permet l'affichage interactif de données variées. Nous avons utilisé cette bibliothèque dans chacun de nos outils de visualisation. L'avantage de cette bibliothèque est qu'elle permet de créer une grande multitude d'éléments interactifs.

5.5 Outils de visualisation pour l'analyse de la prévision de séries spatio-temporelles

Dans le but d'analyser les prévisions de séries spatio-temporelles détaillées dans les chapitres 3 et 4, nous avons développé deux outils génériques de visualisation

utiles pour ce type d'analyse : un outil pour visualiser temporellement les erreurs de prévision et un outil pour analyser les erreurs en se focalisant sur la structure du réseau de transport étudié. Les analyses peuvent ainsi être regroupées en deux étapes, à savoir (i) repérer les jours difficiles à prévoir de manière efficace à l'aide de la visualisation temporelle (e.g., filtrage des jours avec événement/incident, sélection des fortes erreurs de prévision facilité par la vue globale de nombreux couples station-jour) et (ii) accès de la vue temporelle à la vue spatiale d'un simple clique pour analyser les erreurs directement sur le réseau de transport.

5.5.1 Cas d'étude

Nous avons exploité trois ensembles de données pour analyser la prévision de la demande des passagers à chaque station du réseau de métro de Montréal, Québec, Canada. Ces trois ensembles de données ont été fournis par la Société de Transport de Montréal (STM) et correspondent aux jeux de données suivants: (i) les données billettiques (validation en entrée uniquement) avec une agrégation temporelle au quart d'heure, détaillées en section 3.3.1, (ii) une base de données événementielles (e.g., concert, spectacle, match de hockey, etc.) détaillée en section 3.3.3 et (iii) une base de données d'incidents survenus sur le réseau de métro de la ville de Montréal, détaillée ci-après.

Données des incidents survenus sur le réseau de métro de Montréal Ce jeu de données s'étend sur une période de 3 ans (2015-2017) et correspond aux enregistrements manuels des incidents survenus sur le réseau de métro. Ces enregistrements ont été réalisés par les opérateurs de transport. Chaque incident est défini par un horaire de début et de fin, une cause définissant l'incident (e.g., blocage de porte, accident grave de passager, panne électrique, etc.) ainsi que l'identifiant de la station où a eu lieu l'incident.

5.5.2 Visualisation temporelle

Afin d'analyser temporellement les résultats des prévisions sur l'ensemble des stations, nous avons développé une carte de chaleur permettant de mettre en évidence les périodes et les stations difficiles à prévoir. Pour cela, nous présentons dans une carte de chaleur interactive les résidus de prévision avec deux résolutions temporelles, à savoir la résolution par année permettant d'avoir une vue globale des résidus sur plusieurs jours et mois et une résolution par jour permettant d'analyser

en détail les résidus à chaque pas de temps de la journée étudiée. Chacune des deux résolutions temporelles disponibles (jour et année) permet d'observer les résidus de prévision avec un choix de normalisation sélectionnable dans l'interface graphique. Plusieurs méthodes de normalisation ont été développées pour analyser sous différents angles les résidus (e.g., périodes, stations, couple station-temps difficiles à prévoir). En effet, sans ces différents types de normalisation, l'interprétation des résidus peut être incomplète dans le cas où la visualisation ne met pas en évidences les informations importantes telles que les périodes ou les stations difficiles à prédire.

Concernant la résolution temporelle à la journée, les choix de normalisation sont définis dans les équations 5.1, 5.2, 5.3 et 5.4. Dans ces équations, le terme rn correspond à un résidu normalisé.

La normalisation 5.1 utilise la valeur absolue maximale de tous les pas de temps de l'année notés $\mathcal{T}_y$, observée sur l'ensemble des stations $\mathcal{S}$. Cette méthode permet de repérer plus simplement les couples station-pas de temps difficiles à prédire (parmi tous les pas de temps de l'année).

$$\forall s \in \mathcal{S}, \forall t \in \mathcal{T}_y, rn_t^s = r_t^s / max(|r_{\mathcal{T}_y}^{\mathcal{S}}|) \tag{5.1}$$

La normalisation 5.2 utilise la valeur absolue maximale observée sur l'ensemble des stations $\mathcal{S}$ et sur l'ensemble des pas de temps de la journée analysée (l'ensemble des pas de temps d'une journée est noté $\mathcal{T}_d$). Cette méthode permet de repérer plus aisément les couples station-pas de temps difficiles à prédire (parmi tous les pas de temps de la journée étudiée).

$$\forall s \in \mathcal{S}, \forall t \in \mathcal{T}_d, rn_t^s = r_t^s / max(|r_{\mathcal{T}_d}^{\mathcal{S}}|) \tag{5.2}$$

La normalisation 5.3 utilise la valeur absolue maximale de l'ensemble des pas de temps de l'année noté $\mathcal{T}_y$ observée par station. Cette méthode permet de repérer quels sont les pas de temps difficiles à prévoir par station parmi l'ensemble des pas de temps de l'année.

$$\forall s \in \mathcal{S}, \forall t \in \mathcal{T}_y, rn_t^s = r_t^s / max(|r_{\mathcal{T}_y}^{s}|) \tag{5.3}$$

La normalisation 5.4 utilise la valeur absolue maximale de la journée analysée par station. Cette méthode permet de mettre en évidence les pas de temps difficiles à prévoir par station parmi l'ensemble des pas de temps de la journée étudiée.

$$\forall s \in \mathcal{S}, \forall t \in \mathcal{T}_d, rn_t^s = r_t^s / max(|r_{\mathcal{T}_d}^s|) \tag{5.4}$$

La résolution temporelle à l'année permet quant à elle d'analyser les résidus de prévision, de manière agrégée par couple station-jour à l'aide de la méthode décrite en équation 5.5.

$$\forall t \in D, ra_d^s = \begin{cases} max(r_t^s), & if\ max(r_t^s) >= |min(r_t^s)| \\ min(r_t^s), & otherwise \end{cases} \tag{5.5}$$

avec ra_d^s le résidu agrégé de la station s, le jour d et t un pas de temps appartenant à l'ensemble des pas de temps D du jour d.

Ces résidus de prévision agrégés peuvent également être normalisés à l'aide de trois méthodes qui permettent de plus facilement mettre en évidence des stations ou des périodes difficiles à prédire pour les modèles de prévision. La première méthode consiste à normaliser les résidus à l'aide de la valeur maximale absolue observée sur l'ensemble des stations par jour. Elle permet dans ce cas, de mettre en évidence pour chaque journée de l'année, les stations difficiles à prédire. La deuxième méthode normalise les résidus avec la valeur maximale absolue observée sur l'ensemble des jours par station. Nous pouvons à l'aide de cette normalisation, repérer plus facilement les stations les plus difficiles à prévoir. Enfin la dernière méthode utilise la valeur maximale absolue observée sur l'ensemble des jours de l'année étudiée à toutes les stations. Cette méthode permet de repérer les couples station-jour difficiles à prédire.

Un exemple de cet outil de visualisation avec la résolution temporelle à l'année est illustré à la Figure 5.4. Plusieurs interactions homme/machine sont réalisables et sont représentées par des indications numérotées en bleu sur la Figure:

1. Permet de choisir le modèle de prévision à analyser.
2. Sélection de la méthode de normalisation.
3. Choix des données externes à visualiser (en vert sur la carte de chaleur) à savoir les événements (e.g., concert, spectacle, etc.) ou les incidents (e.g., incendie, panne, accident grave de passagers, etc.).
4. Filtre les couples station-jour en fonction de la présence de données externes.

5. Sélection de la date (ici l'année).
6. Sélection de la résolution temporelle, visualisation par année ou jour.
7. Carte de chaleur permettant d'analyser les résidus de prévision normalisés entre -1 et 1. Une cellule de couleur bleue signifie que l'observation du nombre de passagers correspond à une valeur inférieure à celle de la prévision. Les noms des stations sont affichés sur l'axe des ordonnées et les dates sur l'axe des abscisses.

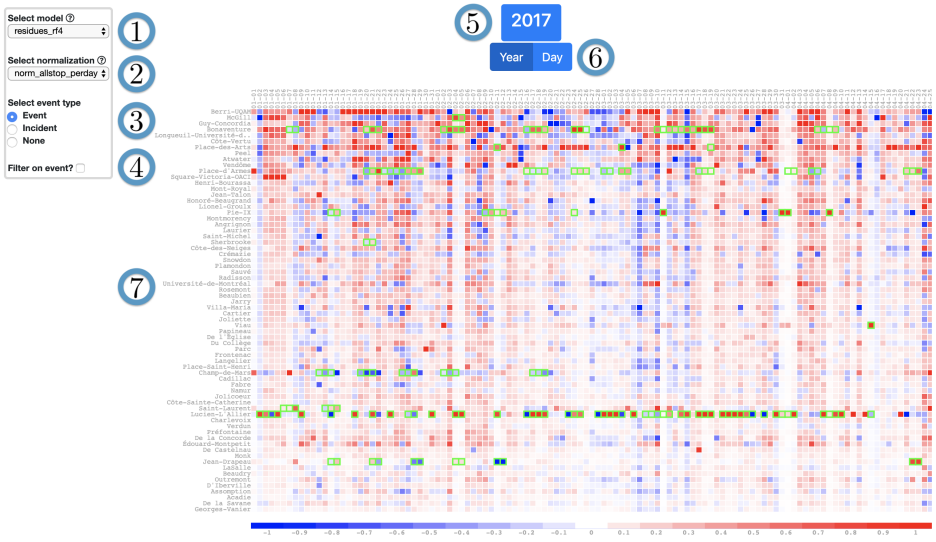

Fig. 5.4.: Visualisation temporelle des résidus de prévision à l'aide d'une carte de chaleur interactive, développée avec la librairie D3.js.

5.5.3 Visualisation spatiale

Dans le but d'analyser spatialement les résidus de prévision et d'analyser en détail la dynamique de l'entrée des passagers dans les différentes stations du réseau de métro à chaque pas de temps d'une journée, nous avons développé une visualisation qui se base sur une carte interactive (zooms spatiaux et choix de l'horaire à afficher). Pour cela, nous avons représenté chaque station par une bulle de couleur correspondant à un résidu normalisé avec une des méthodes détaillées en section 5.5.2. Cette visualisation permet également de repérer les stations qui sont proches d'un événement ou d'un incident enregistré dans les base de données (repérable par la couleur du contour de la bulle).

Cette visualisation qui permet de mieux comprendre la dynamique spatiale de l'affluence des passagers est détaillée dans la Figure 5.5.

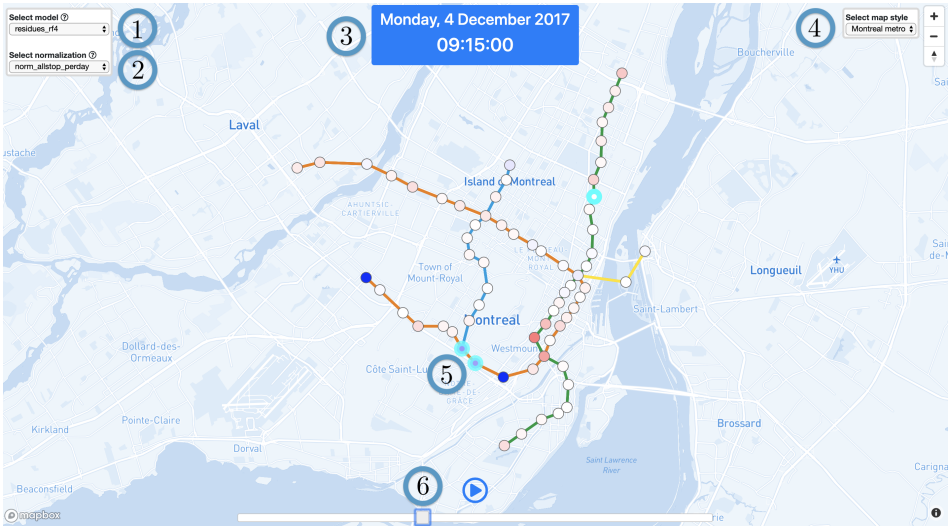

Fig. 5.5.: Outil de visualisation spatial des résidus de prévision, développée avec D3.js et le service de cartographie Mapbox GL JS.

Plusieurs interactions sont possibles à l'aide de cet outil de visualisation. Ces interactions peuvent être localisées sur la figure à l'aide des indications numérotées de 1 à 6 sur la Figure :

1. Choix du modèle de prévision.
2. Sélection de la méthode de normalisation des résidus de prévision.
3. Sélection de la date (jour).
4. Choix du style du fond de carte.
5. Réseau de transport représenté par les stations de métro et les différentes lignes de métro colorées en orange, bleue, vert et jaune. Certaines stations sont encerclées en bleu clair, cela signifie qu'un incident à eu lieu à ces stations lors des précédents pas de temps (fixés à 8 dans cet exemple) de la journée.
6. "Slider" permettant de choisir le pas de temps (9h15 dans la figure). Bouton permettant de lancer une animation temporelle des différents pas de temps de la journée.

Depuis chacune des visualisations (temporelle et spatiale), il est possible d'obtenir une visualisation détaillée de la journée à une station donnée. Pour cela il suffit de cliquer sur une cellule station-jour sur la visualisation temporelle et sur une station si l'on interagit avec la visualisation spatiale. Un exemple de visualisation détaillée de la journée du samedi 28 octobre 2017 est donné dans la Figure 5.6. Les trois indications numérotées correspondent à :

1. La liste et le détail des événements et incidents. Ici, nous pouvons voir qu'une rencontre de hockey a débuté à cette station à 19h00.
2. L'option d'accéder à la visualisation spatiale de cette journée.
3. Un graphique permettant d'analyser l'observation du nombre de passagers par pas de temps à la station analysée (en bleu) et les prévisions des différents modèles. Ici nous pouvons voir que c'est le modèle pred_rf4 en rouge (modèle Random Forest numéro 4) qui réussit à prévoir la forte augmentation du nombre de passagers suite à la fin de la rencontre de hockey. Pour plus d'informations concernant ce modèle, sa méthodologie est donnée en section 3.4.2 et le détail des données d'entrée de ce modèle (D4) est donné en section 3.4.1).

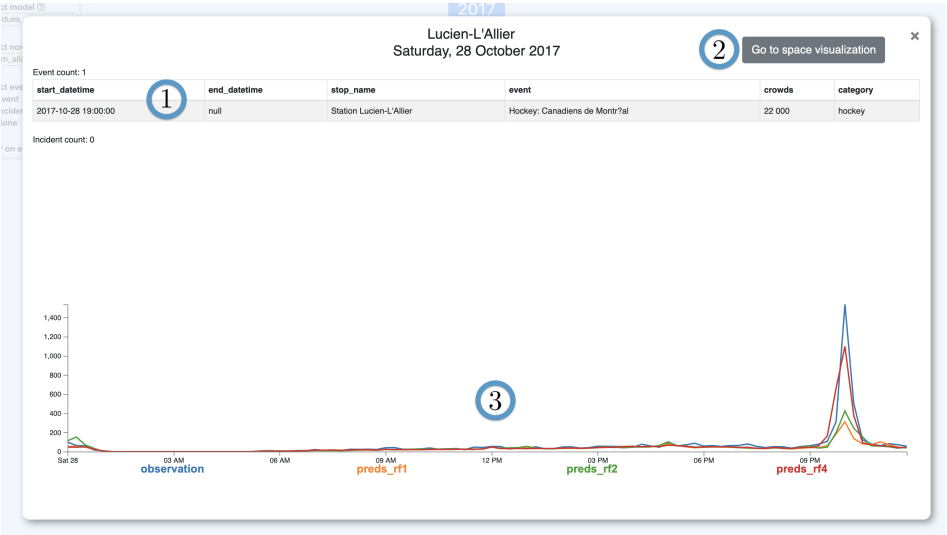

Fig. 5.6.: Informations détaillées de la journée du 28 octobre 2017 à la station Lucien-L'Allier.

5.5.4 Cas d'usage

Nous avons utilisé les trois jeux de données fournis par la Société de Transport de Montréal détaillés en section 5.5.1 pour tester les outils de visualisation. Dans le but de comparer différents modèles de prévision, nous avons exploité les résidus des modèles Random Forest détaillés en section 3.4.2. Ces modèles permettent de prévoir jusqu'à un an en avance le nombre de passagers agrégés au quart d'heure entrant dans chaque station d'un réseau de transport. Les modèles présentés ici utilisent des données d'entrée différentes, chacun de ces ensembles de données (D1, D2, D3 et D4) sont détaillés dans la section 3.4.1. Le modèle RF1 utilise les données calendaires basiques à savoir le jour et le mois, le modèle RF2 utilise les

données calendaires avancées (e.g., vacances scolaires, jours fériés, ponts, etc.), et le modèle RF4 utilise les données calendaires avancées et les informations concernant la base de données événementielles à savoir les horaires et la catégorie de chaque événement (e.g., concert, match de hockey, soccer, tennis, etc.).

Il est important de noter que ce type d'analyse (visualisation des résidus de prévision) permet de mettre en évidence la capacité des modèles de prévision long terme à faire ressortir des phénomènes anormaux des données observées (détection d'anomalie). En effet, comme ces modèles se basent sur la moyenne des valeurs observées dans la base de données historique, un fort écart entre l'observation réelle et la prévision de ces modèles (résidus avec une forte valeur négative ou positive) permet d'indiquer une affluence anormale des passagers survenue à cause d'une situation ayant impactée le réseau de transport.

Nous avons choisi de mettre en évidence l'utilité de ce type de visualisation à l'aide de trois exemples, la section 5.5.4.1 détaille une comparaison des résultats entre deux modèles de prévision, la section 5.5.4.2 permet d'analyser les résultats de prévision en cas d'événement (concert) et la section 5.5.4.3 permet d'analyser un exemple d'incident étant survenu sur le réseau de transport.

5.5.4.1. Analyse de deux modèles utilisant différentes variables d'entrée

Nous analysons ici les résidus du modèle Random Forest RF1 qui utilise uniquement les données calendaires basiques comme caractéristiques d'entrée (jour et mois). Nous pouvons observer dans la Figure 5.7, les résidus normalisés en fonction de l'axe des abscisses sur l'année de test, à savoir l'année 2017. Comme nous pouvons le voir sur la figure, les périodes 1, 2 et 3 correspondent à des jours avec des fortes valeurs de résidus. Ceci s'explique par le fait que ces périodes correspondent à des jours fériés ou à des jours de vacances scolaires, phénomène qui entraîne une diminution du nombre de passagers sur la plupart des stations.

La figure 5.8 montre les résidus du modèle RF2 sur la période étudiée précédemment dans la Figure 5.7. Nous pouvons voir que les valeurs des résidus obtenues sur les périodes 1, 2 et 3 ont nettement diminué comparées à celles obtenues par le modèle RF1. Cet exemple prouve que le modèle RF2 a réussi à prendre en compte les caractéristiques calendaires avancées lors de la prévision, à savoir les informations concernant les jours fériés, les vacances scolaires, etc.

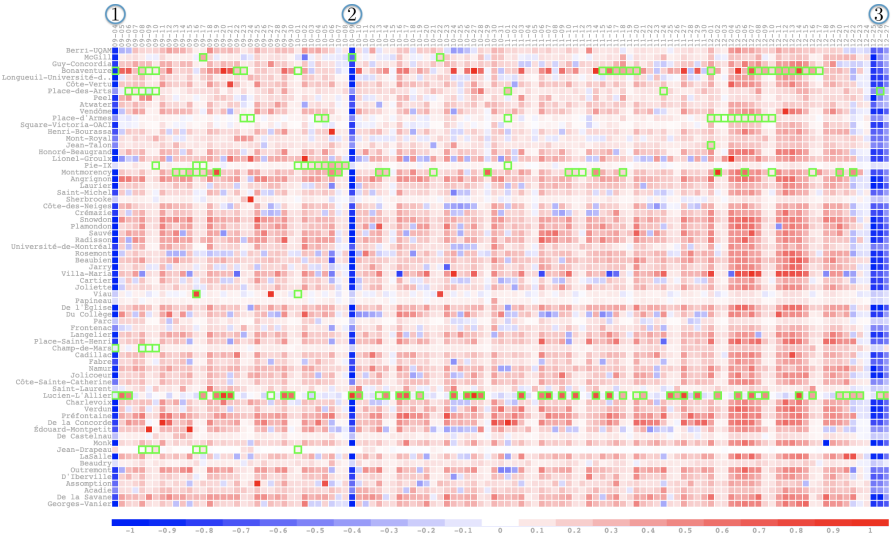

Fig. 5.7.: Visualisation temporelle des résidus du modèle de prévision RF1 de l'année 2017.

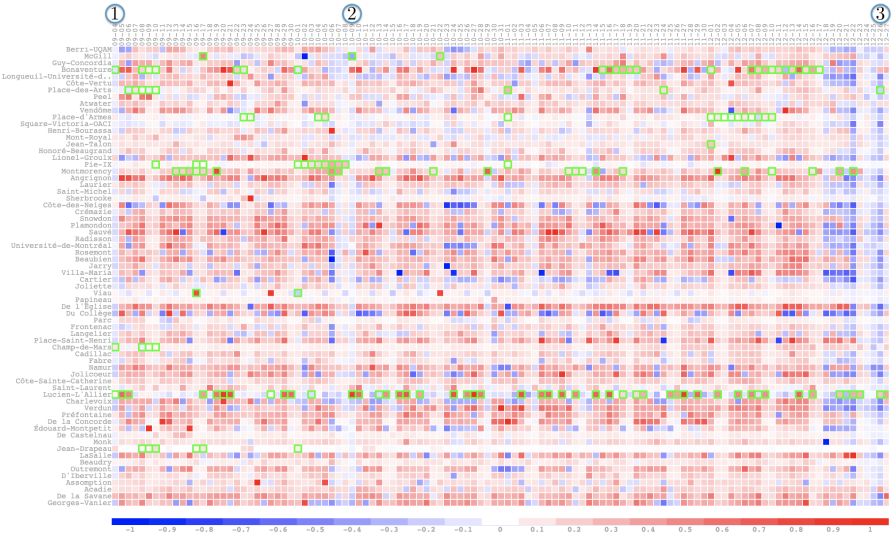

Fig. 5.8.: Visualisation temporelle des résidus du modèle de prévision RF2 de l'année 2017.

5.5.4.2. Analyse de l'impact d'un événement (concert de Bruno Mars) sur l'affluence des passagers

La Figure 5.9 détaille les informations disponibles du mardi 29 août 2017 à la station Lucien-L'allier. Au niveau de l'indication numéro 1, nous pouvons voir qu'un concert de Bruno Mars a débuté à 20h00 et s'est terminé à 22h45 ce jour-là. Le graphe indiqué par le numéro 2 nous permet quant à lui d'analyser les observations du nombre de passagers (courbe bleue) ainsi que les prévisions des différents modèles, RF1 (orange), RF2 (vert) et RF4 (rouge). Cette visualisation nous permet de facilement comprendre la cause de la forte hausse de l'affluence de passagers aux

alentours de 22h45. En effet, il ne fait aucun doute que l'origine de l'augmentation de la demande des passagers correspond à l'entrée d'une partie des spectateurs du concert de Bruno Mars à la station Lucien-L'allier. A l'aide de ce type de visualisation, il nous est également possible de dire que le modèle RF4 a réussi à prédire la forte hausse de la demande des passagers.

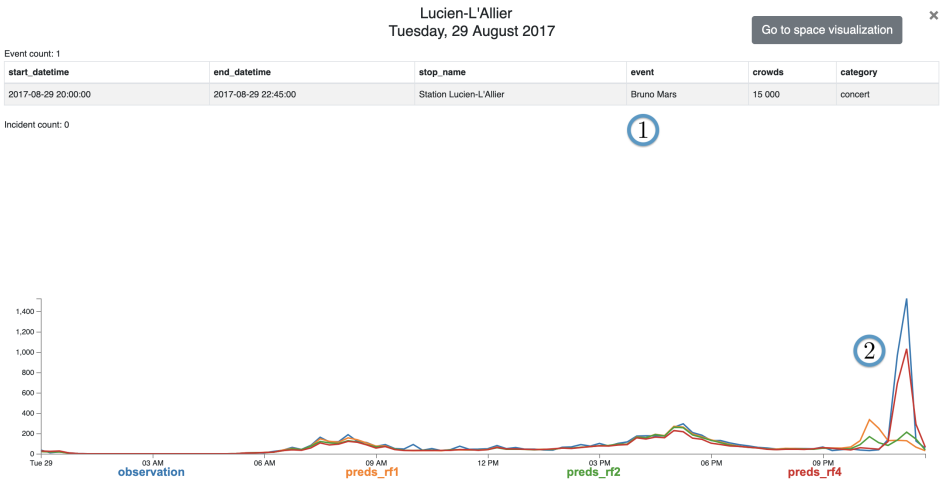

Fig. 5.9.: Informations détaillées de la journée du 29 août à la station Lucien-L'Allier.

Afin d'analyser la répercussion de cette fin de concert sur l'ensemble du réseau de transport, nous avons ouvert la visualisation temporelle à la journée du 29 août 2017, disponible dans la Figure 5.10. Nous pouvons observer au niveau de l'indication numéro 2, le pas de temps auquel le concert a débuté (rectangle vert) ainsi que la hausse des résidus qui a eu lieu à partir de 23h00, un quart d'heure après que le concert se soit terminé. Par ailleurs, cette visualisation nous permet également d'observer que le nombre de passagers a subi une forte hausse à une autre station, la station Bonaventure, qui est une station localisée non loin de la station Lucien-L'allier (station la plus proche de l'événement). Nous pouvons donc déduire à l'aide de cette visualisation que ce type d'événement impact deux stations et que le temps de retour à la normale de l'affluence des passagers est d'environ 45 min après la fin de l'événement.

5.5.4.3. Analyse de l'impact d'un incident important sur le réseau de transport

L'ensemble des données d'incident permet à l'aide de ces outils de visualisation, de repérer efficacement et d'expliquer des phénomènes ayant un fort impact sur l'affluence des passagers. Dans la Figure 5.11 nous pouvons voir l'exemple d'un

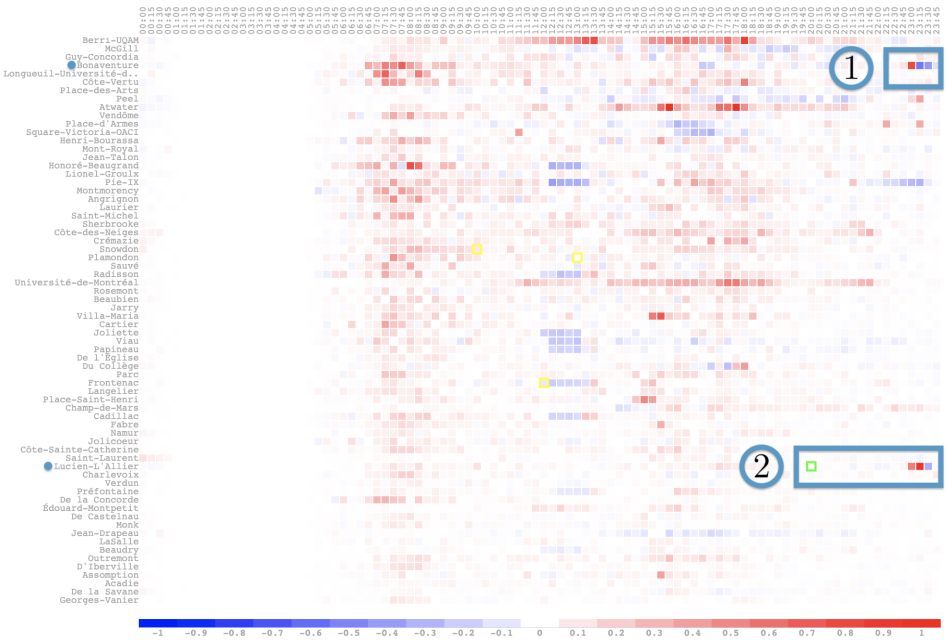

Fig. 5.10.: Visualisation temporelle des résidus du modèle de prévision RF4, la journée du 29 août 2017.

grave incident de passager ayant eu lieu à la station Atwater le mercredi 15 février 2017 à 15h00. Nous observons une forte diminution du nombre de passagers entrant à la station impactée, à la suite de l'incident. Comme nous pouvions nous y attendre, aucun modèle de prévision long terme n'a pu prévoir l'impact de cet incident sur le nombre de passagers observés.

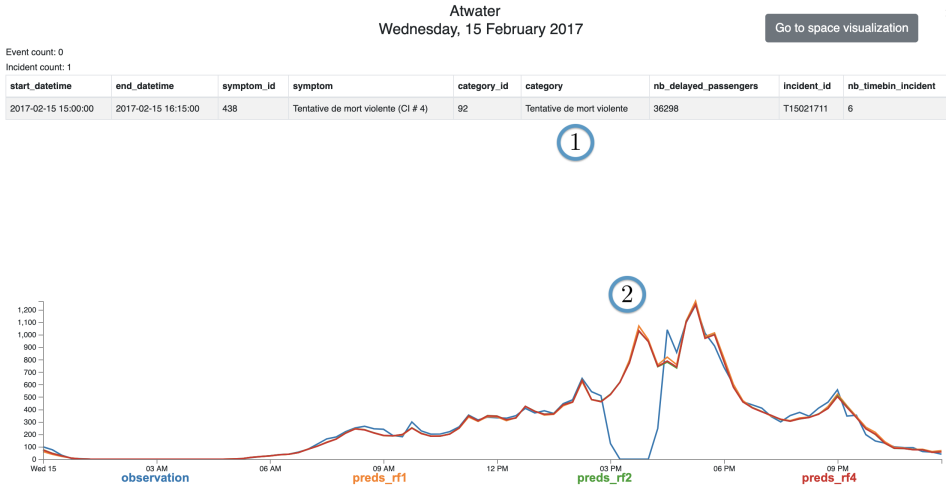

Fig. 5.11.: Informations détaillées de la journée du 15 février 2017 à la station Atwater.

En revanche, comme le montre la Figure 5.12 la visualisation temporelle des résidus des modèles de prévision long terme permet de mettre en évidence la forte per-

turbation survenue sur le réseau de transport. En effet, nous pouvons repérer les différentes stations du réseau de métro impactées, en localisant les couples station-pas de temps avec une forte valeur de résidu. Nous pouvons ainsi voir les stations réagissant de manière directe à cet incident (entre 15h00 et 15h15), ici la réaction de la dynamique des passagers se traduit par une baisse de la fréquentation en entrée de certaines stations (couleur bleue). Un quart d'heure après l'incident, une hausse de l'affluence des passagers est visible sur différentes stations du métro (couleur rouge). Avec ce type de visualisation nous ne pouvons qu'imaginer la topologie du réseau de métro et déduire que certaines stations ont été fermées à la suite de l'incident, ce qui a provoqué le changement d'itinéraire de certains passagers vers les stations non bloquées les plus proches.

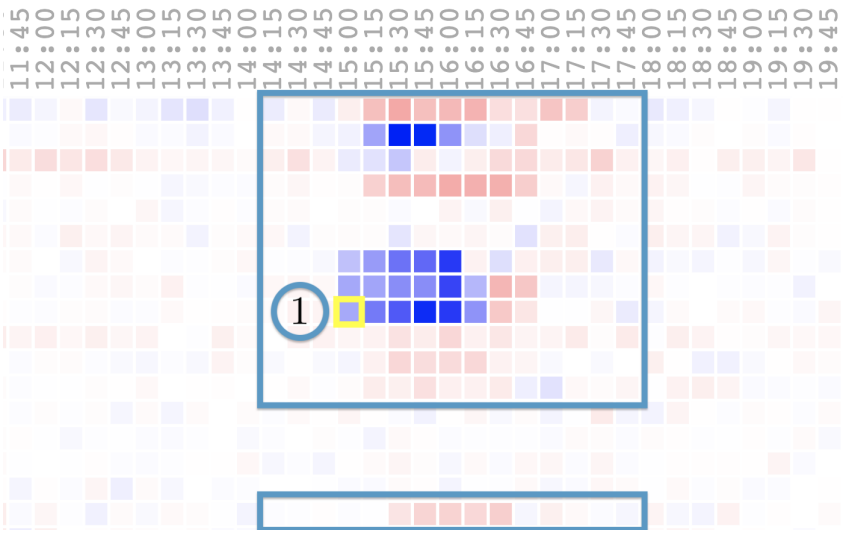

Fig. 5.12.: Zoom sur la visualisation temporelle de quelques stations impactées par un incident le 15 février 2017.

A l'aide de la visualisation spatiale présentée dans les Figures 5.13, 5.14, 5.15 et 5.16, il est plus aisé de comprendre la dynamique de l'affluence des passagers observée pendant et après cet incident, lors des pas de temps 15h00, 15h15, 15h30 et 15h45. Nous constatons que l'incident est survenu à la station Atwater repérable par un cercle bleu clair sur la ligne verte. Une forte baisse de la fréquentation sur la plupart des stations de la ligne verte, a lieu pendant et après l'événement. En revanche, certaines stations de la ligne orange (ligne proche de la ligne verte) notamment les stations appartenant aux deux lignes sont affectées 15 minutes après l'incident par un nombre plus important de passagers. Ceci conforte notre hypothèse et nous permet d'affirmer que les passagers se sont déplacés sur la ligne orange dans le but de continuer leurs déplacements.

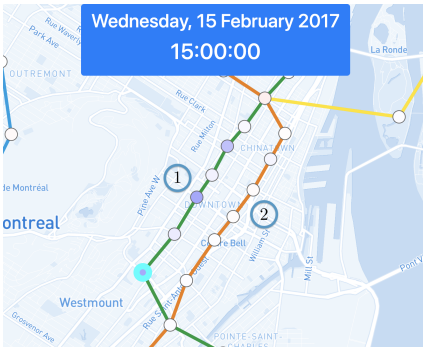

Fig. 5.13.: Visualisation spatiale du 15/02/2017 à 15h00.

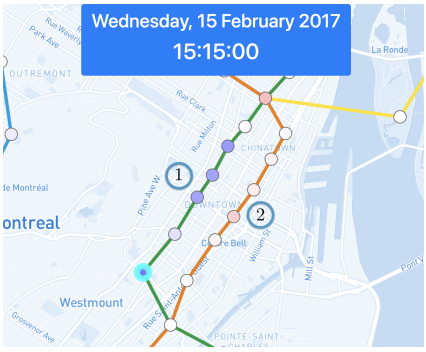

Fig. 5.14.: Visualisation spatiale du 15/02/2017 à 15h15.

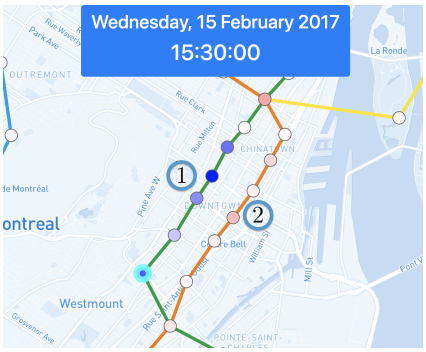

Fig. 5.15.: Visualisation spatiale du 15/02/2017 à 15h30.

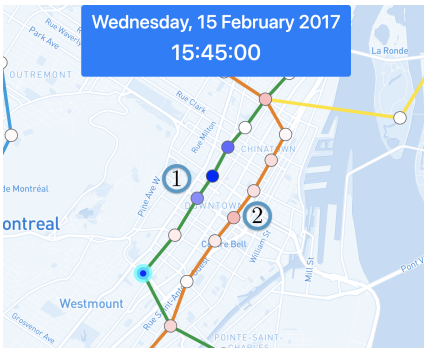

Fig. 5.16.: Visualisation spatiale du 15/02/2017 à 15h45.

5.6 Conclusion

Le développement des méthodes d'analyse de données a conduit à l'essor de la collecte d'ensembles de données par une multitude de capteurs urbains. Ces données multi-sources peuvent être mobilisées pour analyser en profondeur les systèmes urbains. Dans notre cas, nous utilisons des données billettiques, événementielles et une base de données d'incidents fournies par la Société de Transport de Montréal (STM), en vue d'améliorer les connaissances que nous avons sur les déplacements des usagers dans le système de transport en commun.

Dans ce travail, nous avons associé des méthodes de prévision de l'apprentissage automatique à des outils de visualisation de données dans le but de mieux comprendre la dynamique de l'affluence des passagers dans ce réseau de transport. Pour cela, nous avons créé deux outils de visualisation interactive permettant d'analyser de

manière spatiale et temporelle les résidus des prévisions de l'affluence des passagers en présence d'événements.

Au travers de différentes expérimentations sur un jeu de données réelles, il nous a été possible de démontrer la forte utilité de ce type d'outils de visualisation open source pour l'ensemble de la communauté du transport (chercheurs et opérateurs). D'un point de vue plus global, ces outils peuvent également être utiles pour les personnes travaillant sur de l'analyse de larges ensembles de séries spatio-temporelles qui requiert la prise en compte de données externes.

Les différents avantages de ce type d'outils sont les suivants: repérer plus simplement les périodes et les séries temporelles difficiles à prévoir, comprendre pourquoi certaines erreurs ont été commises, analyser spatialement la dynamique des résidus de prévision au travers du réseau de transport et enfin améliorer les bases de données historiques et les modèles de prévision au travers d'une connaissance des données plus détaillée.

Bibliographie

[Aba+16] Martin Abadi, Paul Barham, Jianmin Chen, et al. “Tensorflow: A system for large-scale machine learning”. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 2016, pp. 265–283 (cit. on p. 41).

[ACO18] Patrice Aknin, Etienne Côme, and Latifa Oukhellou. “L’ouverture des données, une opportunité pour la recherche sur les transports et la mobilité”. In: *Transports urbains* 132.1 (2018), pp. 21–27 (cit. on p. 9).

[AOT16] Afian Anwar, Amedeo Odoni, and Nelson Toh. “Busviz: Big data for bus fleets”. In: *Transportation Research Record* 2544.1 (2016), pp. 102–109 (cit. on p. 102).

[BW04] M. Bagchi and P. R. White. “What role for smart card data from bus systems”. In: *Proceedings of the Institution of Civil Engineers. Municipal Engineer. March Issue ME1*. Vol. 157. 2004, pp. 39–46 (cit. on p. 6).

[Bai+19] Yuping Bai, Xiangzheng Deng, John Gibson, Zhe Zhao, and He Xu. “How does urbanization affect residential CO2 emissions? An analysis on urban agglomerations of China”. In: *Journal of cleaner production* 209 (2019), pp. 876–885 (cit. on p. 2).

[Bar+02] J.J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda. “Origin and destination estimation in New York City with automated fare system data”. In: *Transportation Research Record* 1817 (2002), pp. 183–187 (cit. on p. 71).

[BD14] Mike Barry and Brian Card Visualizing MBTA Data. “An interactive exploration of Boston’s subway system http://mbtaviz.github.io”. In: *Retrieved: Jan* (2014) (cit. on p. 104).

[BCO17] Pierre Borgnat, Etienne Côme, and Latifa Oukhellou. “Processing, mining and visualizing massive urban data.” In: *ESANN*. 2017 (cit. on p. 100).

[BOH11] M. Bostock, V. Ogievetsky, and J. Heer. “D^3 Data-Driven Documents”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2301–2309 (cit. on p. 107).

[Bre17] Leo Breiman. *Classification and regression trees*. Routledge, 2017 (cit. on p. 59).

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 30).

[Bre+84] Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. “Classification and regression trees”. In: (1984) (cit. on pp. 29, 90).

[CG16] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM. 2016, pp. 785–794 (cit. on p. 42).

[Che+17] Xingyi Cheng, Ruiqing Zhang, Jie Zhou, and Wei Xu. “DeepTransport: Learning Spatial-Temporal Dependency for Traffic Condition Forecasting”. In: *arXiv preprint arXiv:1709.09585* (2017) (cit. on pp. 21, 22, 26, 35).

[Cho+14] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In: *arXiv preprint arXiv:1406.1078* (2014) (cit. on p. 33).

[Cho+15] François Chollet et al. *Keras*. `https://github.com/keras-team/keras`. 2015 (cit. on pp. 41, 87).

[Cli18] Agence Parisienne du Climat. *Le Plan Climat Air Energie : une stratégie globale pour le climat à Paris*. `https://www.apc-paris.com/plan-climat`. 2018 (cit. on p. 11).

[CMO14] Etienne Côme, K Mohamed, and Latifa Oukhellou. “Cartographie interactive de matrices Origines/Destinations”. In: (2014) (cit. on p. 103).

[CV95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297 (cit. on p. 28).

[Cui+16] Chunsheng Cui, Hongfei Jia, Liping Huang, and Xiaopeng Zhang. “Fuzzy multivariate NARX model for passenger entrance flow prediction in the Shanghai subway system”. In: *Journal of Intelligent & Fuzzy Systems* 31.6 (2016), pp. 3047–3054 (cit. on pp. 21, 22, 35).

[Del17a] François Deloche. *A diagram for a one-unit Gated Recurrent Unit (GRU). From bottom to top : input state, hidden state, output state. Gates are sigmoïds or hyperbolic tangents. Other operators : element-wise plus and multiplication. Weights are not displayed. Inspired from Understanding LSTM, Blog of C. Olah*. `https://upload.wikimedia.org/wikipedia/commons/5/5f/Gated_Recurrent_Unit.svg`. File licensed under `https://en.wikipedia.org/wiki/en:Creative_Commons` and `https://creativecommons.org/licenses/by-sa/4.0/deed.en`. 2017 (cit. on p. 35).

[Del17b] François Deloche. *A diagram for a one-unit Long Short-Term Memory (LSTM). From bottom to top : input state, hidden state and cell state, output state. Gates are sigmoïds or hyperbolic tangents. Other operators : element-wise plus and multiplication. Weights are not displayed. Inspired from Understanding LSTM, Blog of C. Olah*. `https://upload.wikimedia.org/wikipedia/commons/6/63/Long_Short-Term_Memory.svg`. File licensed under `https://en.wikipedia.org/wiki/en:Creative_Commons` and `https://creativecommons.org/licenses/by-sa/4.0/deed.en`. 2017 (cit. on p. 34).

[Din+16] Chuan Ding, Donggen Wang, Xiaolei Ma, and Haiying Li. “Predicting Short-Term Subway Ridership and Prioritizing Its Influential Factors Using Gradient Boosting Decision Trees”. In: *Sustainability* 8.11 (2016), p. 1100 (cit. on pp. 21, 22, 29, 31, 35).

[Dru+97] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. “Support vector regression machines”. In: *Advances in neural information processing systems*. 1997, pp. 155–161 (cit. on p. 28).

[ES90] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*. Vol. 3. Vsp, 1990 (cit. on p. 18).

[EL14] Côme Etienne and Oukhellou Latifa. “Model-based count series clustering for bike sharing system usage mining: a case study with the Vélib’system of Paris”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.3 (2014), p. 39 (cit. on p. 103).

[Fri01] Jerome H Friedman. “Greedy function approximation: a gradient boosting machine”. In: *Annals of statistics* (2001), pp. 1189–1232 (cit. on p. 30).

[Gir+16] Antoine Giraud, Martin Trépanier, Catherine Morency, and Félix Légaré. *Data fusion of APC, smart card and GTFS to visualize public transit use*. Tech. rep. CIRRELT, Centre interuniversitaire de recherche sur les réseaux d’entreprise, la logistique et le transport, 2016 (cit. on p. 104).

[Gor12] Jason Benjamin Gordon. “Intermodal passenger flows on London’s public transport network: automated inference of full passenger journeys using fare-transaction and vehicle-location data”. PhD thesis. Massachusetts Institute of Technology, 2012 (cit. on p. 102).

[H2O19] H2O.ai. *H2O*. H2O version 3.26.0.3. Sept. 2019 (cit. on p. 42).

[Hal+09] Mark Hall, Eibe Frank, Geoffrey Holmes, et al. “The WEKA data mining software: an update”. In: *ACM SIGKDD explorations newsletter* 11.1 (2009), pp. 10–18 (cit. on p. 42).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (cit. on pp. 33, 34).

[Hra+11] Rob Hranac, Jaimyoung Kwon, Mark Bachmann, and Karl Petty. *Using marey graphs to visualize transit loading and schedule adherence*. Tech. rep. 2011 (cit. on p. 101).

[HA15] Rob J Hyndman and G Athanasopoulos. “Measuring forecast accuracy”. In: *Business Forecasting: Practical Problems and Solutions*. John Wiley & Sons, 2015, pp. 177–184 (cit. on p. 38).

[Ito+14] Masahiko Itoh, Daisaku Yokoyama, Masashi Toyoda, et al. “Visual fusion of mega-city big data: an application to traffic and tweets data analysis of metro passengers”. In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE. 2014, pp. 431–440 (cit. on p. 103).

[J+01] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. [Online; accessed <today>]. 2001 (cit. on p. 41).

[Jut15] Francis Jutand. *Ouverture des données de transport*. `https://www.ladocumentationfrancaise.fr/rapports-publics/154000182/`. [Online; accessed 14-August-2019]. 2015 (cit. on p. 10).

[Kap89] Jagat Narain Kapur. *Maximum-entropy models in science and engineering*. John Wiley & Sons, 1989 (cit. on p. 18).

[Ke+17] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun Michael Chen. "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach". In: *Transportation Research Part C: Emerging Technologies* 85 (2017), pp. 591–608 (cit. on pp. 21, 22, 24, 26, 31, 35).

[KB14] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014) (cit. on p. 74).

[KSR12] Kristian Kloeckl, Oliver Senn, and Carlo Ratti. "Enabling the real-time city: LIVE Singapore!" In: *Journal of Urban Technology* 19.2 (2012), pp. 89–112 (cit. on p. 102).

[Kuh+08] Max Kuhn et al. "Building predictive models in R using the caret package". In: *Journal of statistical software* 28.5 (2008), pp. 1–26 (cit. on p. 42).

[Kup+13] Arun Kuppam, Rachel Copperman, Jason Lemp, et al. "Special events travel surveys and model development". In: *Transportation Letters* 5.2 (2013), pp. 67–82 (cit. on p. 19).

[Lah+14] Pierre-Antoine Laharotte, Romain Billot, Etienne Côme, et al. "Spatiotemporal analysis of bluetooth data: Application to a large urban network". In: *IEEE Transactions on Intelligent Transportation Systems* 16.3 (2014), pp. 1439–1448 (cit. on p. 6).

[Lap+17] Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. "Time-series extreme event forecasting with neural networks at uber". In: *International Conference on Machine Learning*. 2017 (cit. on pp. 13, 21, 22, 26).

[Li16] H Li. *Smile-Statistical Machine Intelligence & Learning Engine*. 2016 (cit. on pp. 41, 42).

[Li+17] Yang Li, Xudong Wang, Shuo Sun, Xiaolei Ma, and Guangquan Lu. "Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks". In: *Transportation Research Part C: Emerging Technologies* 77 (2017), pp. 306–328 (cit. on pp. 13, 20, 22, 29, 31, 35).

[Lit12] Todd Litman. *Evaluating public transportation health benefits*. Victoria Transport Policy Institute, 2012 (cit. on p. 3).

[Lit13] Todd Litman. "Transportation and public health". In: *Annual review of public health* 34 (2013), pp. 217–233 (cit. on p. 3).

[Lüt11] Helmut Lütkepohl. *Vector autoregressive models*. Springer, 2011 (cit. on p. 25).

[MRP18] Ioulia Markou, Filipe Rodrigues, and Francisco C Pereira. "Real-Time Taxi Demand Prediction using data from the web". In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2018, pp. 1664–1671 (cit. on pp. 22, 26, 29, 47).

[MM15] Wendy L Martinez and Angel R Martinez. *Computational statistics handbook with MATLAB*. Chapman and Hall/CRC, 2015 (cit. on pp. 41, 42).

[Mat16] Commons Math. “The apache commons mathematics library”. In: *Apache Commons* (2016) (cit. on p. 41).

[McC02] Andrew Kachites McCallum. *Mallet: A machine learning for language toolkit (2002)*. 2002 (cit. on p. 42).

[McK+10] Wes McKinney et al. “Data structures for statistical computing in python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX. 2010, pp. 51–56 (cit. on pp. 40, 41).

[McN07] Michael G McNally. “The four-step model”. In: *Handbook of Transport Modelling: 2nd Edition*. Emerald Group Publishing Limited, 2007, pp. 35–53 (cit. on pp. 18, 79).

[Men+16] Xiangrui Meng, Joseph Bradley, Burak Yavuz, et al. “Mllib: Machine learning in apache spark”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1235–1241 (cit. on p. 42).

[Mit+18] Logan E Mitchell, John C Lin, David R Bowling, et al. “Long-term urban carbon dioxide observations reveal spatial and temporal dynamics related to urban characteristics and growth”. In: *Proceedings of the National Academy of Sciences* 115.12 (2018), pp. 2912–2917 (cit. on p. 2).

[Mor+16] Luís Moreira-Matias, João Gama, Michel Ferreira, João Mendes-Moreira, and Luis Damas. “Time-evolving OD matrix estimation using high-speed GPS data streams”. In: *Expert systems with Applications* 44 (2016), pp. 275–288 (cit. on p. 47).

[NHG17] Ming Ni, Qing He, and Jing Gao. “Forecasting the subway passenger flow under event occurrences with social media”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.6 (2017), pp. 1623–1632 (cit. on p. 21).

[NK19] Mark Nieuwenhuijsen and Haneen Khreis. “Urban and Transport planning, environment and health”. In: *Integrating Human Health into Urban and Transport Planning*. Springer, 2019, pp. 3–16 (cit. on p. 2).

[Ola15] Christopher Olah. “Understanding lstm networks”. In: (2015) (cit. on p. 32).

[Oli06] Travis E Oliphant. *A guide to NumPy*. Vol. 1. Trelgol Publishing USA, 2006 (cit. on p. 41).

[Par19] Mairie de Paris. *Charte de bonne conduite relativeà la location de trottinettes électriques en libre-service*. `https://www.api-site.paris.fr/paris/public/2019%2F4%2FCharte%20trottinettes-VDEFmodif.pdf`. 2019 (cit. on p. 11).

[PMB13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. “On the difficulty of training recurrent neural networks”. In: *International Conference on Machine Learning*. 2013, pp. 1310–1318 (cit. on pp. 33, 34).

[Pas+17] Adam Paszke, Sam Gross, Soumith Chintala, et al. “Automatic Differentiation in PyTorch”. In: *NIPS Autodiff Workshop*. 2017 (cit. on p. 41).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 42, 53, 87).

[PTM11] Marie-Pier Pelletier, Martin Trépanier, and Catherine Morency. “Smart card data use in public transit: A literature review”. In: *Transportation Research Part C: Emerging Technologies* 19.4 (2011), pp. 557–568 (cit. on p. 101).

[PRB15] Francisco C Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. “Using data from the web to predict public transport arrivals under special events scenarios”. In: *Journal of Intelligent Transportation Systems* 19.3 (2015), pp. 273–288 (cit. on pp. 20, 26, 29, 31, 35).

[PBL03] F Potier, P Bovy, and C Liaudat. “Big events: planning, mobility management”. In: *European Transport Conference*. 2003 (cit. on p. 19).

[Pre12] Lutz Prechelt. “Early stopping—but when?” In: *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 53–67 (cit. on p. 74).

[Ras03] Carl Edward Rasmussen. “Gaussian processes in machine learning”. In: *Summer School on Machine Learning*. Springer. 2003, pp. 63–71 (cit. on p. 27).

[RR17] Hannah Ritchie and Max Roser. “CO_2 and Greenhouse Gas Emissions”. In: *Our World in Data* (2017) (cit. on pp. 1, 3).

[RMP19] Filipe Rodrigues, Ioulia Markou, and Francisco C Pereira. “Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach”. In: *Information Fusion* 49 (2019), pp. 120–129 (cit. on pp. 22, 29, 35, 47).

[RBG16] Jérémy Roos, Stephane Bonnevay, and Gérald Gavin. “Short-Term Urban Rail Passenger Flow Forecasting: A Dynamic Bayesian Network Approach”. In: *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*. IEEE. 2016, pp. 1034–1039 (cit. on pp. 20, 22, 24, 26).

[Sae+14] Brian E Saelens, Anne Vernez Moudon, Bumjoon Kang, Philip M Hurvitz, and Chuan Zhou. “Relation between higher physical activity and public transit use”. In: *American journal of public health* 104.5 (2014), pp. 854–859 (cit. on p. 3).

[SS01] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001 (cit. on p. 26).

[SP10] Skipper Seabold and Josef Perktold. “Statsmodels: Econometric and statistical modeling with python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 57. Scipy. 2010, p. 61 (cit. on pp. 41, 87).

[SKO16] Neveen Shlayan, Abdullah Kurkcu, and Kaan Ozbay. “Exploring pedestrian Bluetooth and WiFi detection at public transportation terminals”. In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2016, pp. 229–234 (cit. on p. 6).

[SBR15] Boris Suchkov, Mikhail Boguslavsky, and Alla Reddy. “Development of a real-time stringlines tool to visualize subway operations and manage service at New York City transit”. In: *Transportation Research Record* 2538.1 (2015), pp. 19–29 (cit. on pp. 101, 102).

[TRC14] Sui Tao, David Rohde, and Jonathan Corcoran. “Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap”. In: *Journal of Transport Geography* 41 (2014), pp. 21–36 (cit. on p. 103).

[ElT+17] Samy El-Tawab, Raymond Oram, Michael Garcia, Chris Johns, and B Brian Park. “Data analysis of transit systems using low-cost IoT technology”. In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE. 2017, pp. 497–502 (cit. on p. 6).

[TL18] Sean J Taylor and Benjamin Letham. “Forecasting at scale”. In: *The American Statistician* 72.1 (2018), pp. 37–45 (cit. on p. 42).

[Tea16] Eclipse Deeplearning4j Development Team. “DL4J: Deep Learning for Java”. In: (2016) (cit. on pp. 41, 42).

[Tea+13] R Core Team et al. “R: A language and environment for statistical computing”. In: (2013) (cit. on p. 41).

[Toq19a] Toqué, Florian and Côme, Etienne and Oukhellou, Latifa and Trépanier, Martin. “Short and long term predictions of public transit ridership based on smart card data: the Montreal case”. In: *Workshop International Network on the Use of Passive Data from Public Transport*. Washington DC, USA, Jan. 2019 (cit. on p. 16).

[Toq19b] Toqué, Florian and Côme, Etienne and Oukhellou, Latifa and Trépanier, Martin. “Visualization tools for space-time series analysis with context awareness: Montreal subway case”. In: *TransitData 2019, 5th International Workshop and Symposium*. Paris, France, July 2019 (cit. on p. 16).

[Toq+16] F. Toqué, E. Côme, M. K. El Mahrsi, and L. Oukhellou. “Forecasting dynamic public transport Origin-Destination matrices with long-Short term Memory recurrent neural networks”. In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. Nov. 2016, pp. 1071–1076 (cit. on pp. 15, 16, 78).

[Toq+18a] Florian Toqué, Etienne Côme, Latifa Oukhellou, and Martin Trépanier. “Prévision du nombre de passagers entrant dans un réseau de transport multimodal en cas d’évènements atypiques, à l’aide de méthodes d’apprentissage automatique”. In: *RFTM, 1ères Rencontres Francophones Transport Mobilité*. LYON, France, June 2018, 3p (cit. on p. 16).

[Toq+18b] Florian Toqué, Etienne Côme, Latifa Oukhellou, and Martin Trépanier. “Short-Term Multi-Step Ahead Forecasting of Railway Passenger Flows During Special Events With Machine Learning Methods”. In: *CASPT 2018, Conference on Advanced Systems in Public Transport and TransitData 2018*. Brisbane, Australia, July 2018, 15p (cit. on p. 16).

[Toq+19] Florian Toqué, Etienne Côme, Martin Trépanier, and Latifa Oukhellou. “Forecasting of the Montreal Subway Smart Card Entry Logs with Event Data”. In: (2019). En cours de soumission (cit. on p. 15).

[Toq+17] Florian Toqué, Mostepha Khouadjia, Etienne Côme, Martin Trépanier, and Latifa Oukhellou. “Short & long term forecasting of multimodal transport passenger flows with machine learning methods”. In: *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2017, pp. 560–566 (cit. on p. 16).

[TV08] Martin Trépanier and Francois Vassiviere. “Democratized smartcard data for transit operator”. In: *15th World Congress on Intelligent Transport Systems and ITS America’s 2008 Annual MeetingITS AmericaERTICOITS JapanTransCore*. 2008 (cit. on p. 104).

[T+14] Chi-Hong Patrick Tsai, Corinne Mulley, Geoffrey Clifton, et al. “Forecasting public transport demand for the Sydney greater metropolitan area: A comparison of univariate and multivariate methods”. In: *Road & Transport Research: A Journal of Australian and New Zealand Research and Practice* 23.1 (2014), p. 51 (cit. on p. 19).

[Ubea] Uber. *Example of deck.gl visualisation, hexagon layer*. `https://deck.gl/#/examples/core-layers/hexagon-layer`. Accessed: 2019-09-26 (cit. on p. 105).

[Ubeb] Uber. *Example of deck.gl visualisation, line layer*. `https://deck.gl/#/examples/core-layers/line-layer`. Accessed: 2019-09-26 (cit. on p. 105).

[VCV11] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. “The NumPy array: a structure for efficient numerical computation”. In: *Computing in Science & Engineering* 13.2 (2011), p. 22 (cit. on p. 41).

[VL15] Andrea Vedaldi and Karel Lenc. “Matconvnet: Convolutional neural networks for matlab”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM. 2015, pp. 689–692 (cit. on p. 41).

[Vil18] Cédric Villani. *Mission Villani sur l’IA - AI for humanity*. `https://www.aiforhumanity.fr/`. [Online; accessed 14-August-2019]. 2018 (cit. on p. 10).

[WR09] Glen Weisbrod and Arlee Reno. *Economic impact of public transportation investment*. Citeseer, 2009 (cit. on pp. 2, 3).

[Wic+15] Hadley Wickham, Romain Francois, Lionel Henry, Kirill Müller, et al. “dplyr: A grammar of data manipulation”. In: *R package version 0.4* 3 (2015) (cit. on p. 41).

[WT16] Yuankai Wu and Huachun Tan. “Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework”. In: *arXiv preprint arXiv:1612.01022* (2016) (cit. on pp. 21, 22, 31, 35).

[Yao+18] Huaxiu Yao, Fei Wu, Jintao Ke, et al. “Deep multi-view spatial-temporal network for taxi demand prediction”. In: *arXiv preprint arXiv:1802.08714* (2018) (cit. on pp. 21, 22, 24, 26, 31, 35).

[Yok+14] Daisaku Yokoyama, Masahiko Itoh, Masashi Toyoda, et al. “A framework for large-scale train trip record analysis and its application to passengers’ flow prediction after train accidents”. In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2014, pp. 533–544 (cit. on p. 103).

[YYZ17] Bing Yu, Haoteng Yin, and Zhanxing Zhu. “Spatio-temporal Graph Convolutional Neural Network: A Deep Learning Framework for Traffic Forecasting”. In: *arXiv preprint arXiv:1709.04875* (2017) (cit. on pp. 21, 22, 24, 26, 29, 35).

[Zah+10a] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. “Spark: Cluster computing with working sets.” In: *HotCloud* 10.10-10 (2010), p. 95 (cit. on p. 13).

[Zah+10b] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. “Spark: Cluster computing with working sets.” In: *HotCloud* 10.10-10 (2010), p. 95 (cit. on p. 41).

[ZZQ17] Junbo Zhang, Yu Zheng, and Dekang Qi. “Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction.” In: *AAAI*. 2017, pp. 1655–1661 (cit. on pp. 21, 22, 24, 26, 35, 45, 79).

[Zmu+13] Johanna Zmud, Martin Lee-Gosselin, Marcela Munizaga, and Juan Antonio Carrasco. *Transport survey methods: Best practice for decision making*. Emerald Group Publishing Limited, 2013 (cit. on p. 5).

[ZH05] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320 (cit. on p. 24).

Liste des figures

Liste des tables

Annexes

A

A.1 Tables

Tab. A.1.: Informations des stations étudiées.

ID	Name	Line	Type
393	LA DEFENSE-GRANDE ARCHE	A	RER
394	LA DEFENSE-GRANDE ARCHE	L	Transilien
269	ESPLANADE DE LA DEFENSE	1	Metro
151	CHARLES DE GAULLE ETOILE	A	RER
577	NANTERRE-PREFECTURE	A	RER
414	LA DEFENSE-GRANDE ARCHE	1	Metro
578	NANTERRE-UNIVERSITE	A	RER
357	HOUILLES-CARRIERES-SUR-SEINE	A	RER
580	NANTERRE-VILLE	A	RER
100	BOULOGNE-PONT DE SAINT CLOUD	10	Metro
812	SAINT-CLOUD	L	Transilien
843	SURESNES-MONT VALERIEN	L	Transilien
29	ASNIERES	J	Transilien
436	LE VAL-D'OR	L	Transilien
217	COURBEVOIE	L	Transilien
579	NANTERRE-UNIVERSITE	L	Transilien
712	PUTEAUX	L	Transilien
413	LA DEFENSE-GRANDE ARCHE	T2	Tramway
1085	PONT DE BEZONS	T2	Tramway
844	SURESNES-LONGCHAMP	T2	Tramway
1090	VICTOR BASCH	T2	Tramway
1088	CHARLEBOURG	T2	Tramway
623	PARC DE SAINT-CLOUD	T2	Tramway
713	PUTEAUX	T2	Tramway
1086	FAUBOURG DE L ARCHE	T2	Tramway
1087	LES FAUVELLES	T2	Tramway
1089	JACQUELINE AURIOL	T2	Tramway
457	LES COTEAUX	T2	Tramway
1091	PARC PIERRE LAGRAVERE	T2	Tramway
461	LES MILONS	T2	Tramway

Tab. A.2.: Hyperparamètres du modèle VAR.

Parameters	Range	Best
Norm	MinMax (0,1)	-
Lag	$x \in \mathbb{N} \vert 0 < x < 51$	**35**

Tab. A.3.: Hyperparamètres des modèles RF.

Parameters	RF LT	RF ST MULTI	RF ST UNI
Norm	MinMax (0,1)	MinMax (0,1)	-
Lag	-	8	8
N estimators	[100,150,**200**]	[100,150,**200**]	[100,150,**200**]
Max features	[**auto**, sqrt]	[**auto**, sqrt]	[**auto**, sqrt]
Min samples leaf	[**1**, 5, 10]	[**1**, 5, 10]	[1, **5**, 10]
Min samples split	[2, 5, **10**]	[**2**, 5, 10]	[**2**, 5, 10]

The optimization of the hyperparameters of the Random Forest models is performed by grid search with a cross-validation with 5 different splits (80% training set, 20% validation set).

Tab. A.4.: Hyperparamètres des modèles GRU.

Parameters	GRU MULTI	GRU UNI	GRU FUSION
Norm	divide by the mean	divide by the mean	divide by the mean
Lag	100	100	100
Drop	[0,0.0001,**0.001**]	**0.001**	**0.001**
Forget	[0.5, **0.8**, 1]	**0.8**	**0.8**
N neurones MULTI	[200,**350**,500]	-	**350**
N neurones UNI	-	[**100**,200,300]	**100**
N epochs	[15,**20**,25,30]	[15,20,25,**30**]	**30**

The optimization of the hyper parameters of the GRU models have been performed by grid search with a cross-validation with 3 different split (80% train set, 20% validation set).

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.