Analyse et prédiction des flux piétons dans un pôle de transport multimodal à partir de données multi-sources

Paul de Nailly

UNIVERSITÉ GUSTAVE EIFFEL

Ecole doctorale mathématiques et sciences et technologies de l'information et de la communication

Thèse présentée pour obtenir le grade universitaire de Docteur

Discipline : Informatique

Paul DE NAILLY

Analyse et prédiction des flux piétons dans un pôle de transport multimodal à partir de données multi-sources

Soutenance le 23/01/2023 devant le jury composé de :

Catherine MORENCY, Professeure titulaire, Ecole Polytechnique de Montréal	Rapporteure
Julien CHIQUET, Directeur de Recherche à l'INRAE, UMR MIA-Paris	Rapporteur
Michel BIERLAIRE, Professeur, Ecole Polytechnique Fédérale de Lausanne	Examinateur
Paul HONEINE, Professeur d'université, Université de Rouen	Examinateur
Latifa OUKHELLOU, Directrice de recherche, Université Gustave Eiffel	Directrice de thèse
Allou SAMÉ, Directeur de recherche, Université Gustave Eiffel	Co-Directeur de thèse
Etienne CÔME, Chargé de recherche, Université Gustave Eiffel	Encadrant
Jacques FERRIERE, Régie Autonome des Transports Parisiens	Invité

Remerciements

Cette thèse s'est déroulée dans le cadre d'un accord CIFRE entre le laboratoire GRETTIA de l'Université Gustave Eiffel et la Régie Autonome des Transports Parisiens (RATP). Je souhaite remercier ici l'ensemble des personnes qui m'ont aidé dans l'avancement de ce travail tant sur le plan technique, que relationnel et personnel.

Cette thèse a pu aboutir en grande partie grâce à l'encadrement dont j'ai pu bénéficier tout au long de ces trois années. Je remercie tout d'abord ma directrice de thèse Latifa Oukhellou qui a su me motiver à me lancer dans cette thèse à l'issue de mon stage de Master 2 puis qui a réussi à jongler entre bienveillance et rigueur dans le travail, m'aidant à devenir une personne plus confiante et épanouie dans mon travail. Mon co-directeur Allou Samé et mon encadrant au GRETTIA Etienne Côme ont su me pousser à être curieux et force de proposition dans les travaux engagés, tout en gardant une grande proximité avec moi en dehors du travail. Je les remercie sincèrement pour tout cela. Je remercie également mon encadrant aux études générales de la RATP Jacques Ferriere qui a su me soutenir dans mon travail et au-delà, à travers des idées innovantes et un soutien régulier, là aussi tout en conservant une proximité et une disponibilité tout du long.

Je suis vivement reconnaissant envers les personnes ayant relu et rapporté mon travail de thèse. La professeure Catherine Morency et le directeur de recherche Julien Chiquet m'ont ainsi fait l'honneur de rapporter sur mes travaux. Les professeurs Michel Bierlaire et Paul Honeine ont également pris du temps pour examiner mon travail. Toutes ces personnes ont été d'une grande bienveillance dans leur retour sur ce travail, cela s'est fait ressentir dans les échanges riches que nous avons pu avoir lors de la soutenance.

Je tiens à remercier mes collègues et amis du GRETTIA et des études générales de la RATP. Les doctorants du GRETTIA Thomas, Benoit, Louise, Rodolphe, Cyril, Pascal, Demeng, Khadidja, Milad et Florian et post-doctorants Paul et Hugues. Les collègues à la RATP Yasmine, Nelly, Renaud, Gilles, Marie, Pauline, Rémy, Charles, Antoine, Rachid, Morgane, Felipe, William, Laurent, Nathalie, Julien, Julie, Nathanael, Daniel et tous les autres, ainsi que mon stagiaire Ibrahima.

Le soutien de personnes proches externes au lieu de travail a été lui aussi fondamental. Je remercie donc avec beaucoup de tendresse ma copine Aurélie qui m'a soutenu et porté sur bien des moments difficiles ainsi que sa famille qui m'a accueilli. Je ne peux oublier mon colocataire Samuel qui a su faire preuve de bien des ruses pour m'aider à sortir la

tête du travail. J'ai trouvé beaucoup de réconfort auprès de mes amis Alexis, Pierre C., Salma, Adrien, Bastien, Kaelan, Pierre B., Nicolas, Marine, Thibaut, Milène, Charles, Julia, Alexandra, Daphné, Maeva, Pierre E., Guillaume, Enora et Marine qui m'ont tous motivé à continuer à sortir faire du sport, aller à la danse ou simplement se retrouver. Le soutien de ma famille a lui aussi été fondamental. Je remercie chaleureusement mes parents Philippe et Laurence qui, à travers leur éducation et leur profonde bienveillance m'ont porté jusqu'ici. Je remercie de plus mes soeurs Hélène et Julie qui ont travaillé pour garder un contact avec moi même lorsque le travail m'isolait. Nos relations ont évolué positivement pour cela et je leur en suis reconnaissant. Je remercie aussi mes cousins Clémence et Benoit.

Résumé

Au sein des grandes métropoles, les quartiers d'affaires sont des pôles attracteurs majeurs qui concentrent les activités ; attirant chaque jour, *via* des systèmes de transports en commun, une population nombreuse. La bonne compréhension de la dynamique des flux piétons dans les espaces de transport de ces quartiers est un sujet de première importance, notamment afin d'éviter les situations de très forte affluence mal gérées. Cette thèse est appliquée au cas particulier du quartier d'affaire de La Défense dans l'ouest parisien.

Dans ce contexte, la thèse s'attache à développer des méthodes de traitements de données multi-sources de mobilité afin de synthétiser et comprendre les données fortement bruitées des comptages piétons en de multiples points des espaces de transport ; puis de prévoir l'affluence à court terme dans ces mêmes espaces. Ces deux axes de travail ont vocation à enrichir l'information voyageurs à destination des usagers des transports collectifs mais peuvent également servir aux opérateurs de transport pour une régulation « à la demande » de l'offre de transport.

Le premier chapitre se concentre sur la mise en place d'un modèle linéaire dynamique de décomposition afin de comprendre comment les variations de séries temporelles de comptages piétons se traduisent dans les différentes composantes cachées du modèle, chacune liée à un élément de contexte (tendance, saisonnalitée(s), impact de variables contextuelles, ...). L'accent est mis sur la décomposition comparée des séries de comptage de flux entrants vers deux lignes de transport massivement empruntées dans le quartier d'affaires.

Le deuxième chapitre propose une approche de clustering à base d'apprentissage statistique afin de synthétiser les données de fréquentation multivariées, surdispersées et corrélées de l'ensemble du pôle de transport, en lien avec du contexte (variables calendaires et d'événementiel) et au sein de catégories facilement interprétables. L'approche permet de détecter des périodes de temps aux dynamiques de déplacements homogènes et de leur associer des profils de déplacement caractéristiques. Des modèles de mélange basés sur des distributions « somme et partages » et Poisson log-normal sont développés et comparés sur la base de leur capacité à bien modéliser les données et à détecter des périodes homogènes les plus continues possibles.

Le troisième chapitre s'attache à la mise en place de modèles de prédiction probabilistes des flux voyageurs avec des méthodes basées sur l'apprentissage profond. La force de ces modèles réside dans leur capacité à modéliser l'incertitude, particulièrement adaptée dans

le domaine des transports, en s'appuyant sur une abstraction des données contextuelles et en faisant l'hypothèse de distributions en sortie. Nous proposons pour cela un modèle basé sur les distributions « sommes et partages » et le comparons à d'autres modèles issus de l'état de l'art à la fois sur des données ouvertes disponibles et sur les données collectées dans les espaces de transport du quartier de La Défense.

Mots clés : Séries temporelles multivariées, Comptages, Modèles de décomposition, Clustering, Modèles de mélange, Prévision, Apprentissage profond, Transports en commun, Pôles de transport.

Abstract

Within large cities, business districts are major attractors that concentrate activities ; attracting a large population every day, thanks to the public transport system. A good understanding of the dynamics of pedestrian flows in the transport areas of these districts, is a subject of primary importance. Especially in order to avoid poorly managed situations of very high affluence. This thesis is applied to the particular case of the La Défense business district in western Paris.

In this context, the thesis focuses on developing methods for processing multi-source mobility data in order to synthesize and understand the highly noisy data from pedestrian counts at multiple points in the transportation areas, and then to predict short-term traffic in these same areas. These two lines of work aim to enrich passenger information for public transport users, but can also be used by transport operators to regulate transport supply on demand.

The first chapter focuses on the implementation of a dynamic linear model for decomposition in order to understand how the variations of pedestrian count time series are translated in the different hidden components of the model, each one linked to a contextual element (trend, seasonality, impact of contextual variables, ...). The focus is on the comparative decomposition of the series of counts of incoming flows to two massively used transport lines in the business district.

The second chapter proposes a clustering approach based on statistical learning in order to synthesize multivariate, overdispersed and correlated pedestrian flow data of the whole transport hub, in relation with the context (calendar and event variables), within easily interpretable categories. The approach allows to detect time periods with homogeneous travel dynamics and to associate them with characteristic travel profiles. Mixture models based on $\ll$ sum and shares $\gg$ and Poisson log-normal distributions are developed and compared on the basis of their ability to model the data well and to detect homogeneous time periods as continuously as possible.

The third chapter focuses on the implementation of probabilistic prediction models of passenger flows with methods based on deep learning. The strength of these models lies in their ability to model uncertainty, which is particularly adapted to the transportation domain, by relying on an abstraction of contextual data and by assuming output distributions. To this end, we propose a model based on the distributions $\ll$ sums and shares $\gg$ and

compare it to other models from the state of the art, both on available open data and on data collected in the transport areas of the La Défense district.

Keywords : Multivariate time series, Counts, Decomposition models, Clustering, Mixture models, Forecasting, Deep learning, Public transports, Transport hubs.

Table des matières

Liste des notations

Général

T	Nombre total de tranches de temps dans les séries temporelles
P	Nombre total de lieux de comptages des flux piétons dans le pôle de La Défense
J	Nombre total de jours dans les séries temporelles
H	Nombre de tranches de temps par jour
$\mathbf{y}$	Vecteur de comptages, pouvant être indicé de deux manières : 1. $\mathbf{y} = \{y_{t,p}\}_{t\in\{1,\dots,T\},p\in\{1,\dots,P\}}$ avec $y_{t,p}$ un comptage effectué à une tranche de temps t et un lieu de comptage p 2. $\mathbf{y} = \{y_{j,h,p}\}_{j\in\{1,\dots,J\},h\in\{1,\dots,H\,p\in\{1,\dots,P\}}$ avec $y_{j,h,p}$ un comptage effectué à la tranche de temps h d'un jour j et un lieu de comptage p
θ	Paramètres de distribution d'une loi
$p(y;\theta)$	Densité de probabilité avec les paramètres θ pour une variable y
$M(a_1,\dots,a_n)$	Médiane de l'ensemble des données $\{a_1,\dots,a_n\}$
$\bar{\sigma}(a_1,\dots,a_n)$	Écart type de l'ensemble des données $\{a_1,\dots,a_n\}$
$\mathbb{E}(Y;\theta)$	Espérance de la variable aléatoire Y associée à une distribution de paramètres θ
$\mathbb{V}(Y;\theta)$	Variance de la variable aléatoire Y associée à une distribution de paramètres θ
$\mathrm{Cov}(Y,Y';\theta)$	Covariance de Y et Y' associée à une distribution de paramètres θ
$L_Y(\theta)$	Log-vraisemblance de Y associée à une distribution de paramètres θ

Chapitre 1

Introduction

1.1 Contexte

L'urbanisation est un phénomène qui implique l'augmentation du nombre d'habitants dans les villes, et dont découle un certain nombre de problématiques dans la gestion des espaces urbains. D'après le rapport *Perspectives sur l'urbanisation mondiale : révision 2018* (p10) émis par les Nations Unies [Uni19], 68% de la population mondiale devrait se concentrer dans les zones urbaines en 2050. L'accélération de cette urbanisation, combinée à la croissance de la population mondiale, ajouterait ainsi 2,5 milliards de personnes supplémentaires dans les villes. Dans ce contexte d'urbanisation rapide, il est indispensable de mettre en place un cadre de développement urbain intelligent, afin de répondre aux grands enjeux en matière de mobilité, de travail, d'éducation, de soins ou encore d'énergie. Les grandes métropoles présentent le plus souvent une configuration de type « cœur et périphéries ». Les périphéries comprennent des zones résidentielles et d'activités secondaires, tandis que les coeurs concentrent les quartiers d'affaires, les lieux de divertissement, les bâtiments administratifs, ainsi que les principaux nœuds de transport (urbains et interurbains). Les auteurs de [DA11] définissent les quartiers d'affaires (« central business districts », CBDs) comme des zones de concentration massive des activités, ou encore de polarisation du capital et des activités économiques et financières dans les villes. Ces quartiers peuvent constituer le centre même des villes comme les quartiers financiers (« financial downtown ») de Toronto, de Chicago ou de New-York. Dans les villes européennes, le quartier d'affaires est généralement séparé du centre-ville, mais reste un pôle attractif majeur. On pense ici au Bankenviertel à Francfort, Canary Wharf à Londres, Moskva-City à Moscou ou encore La Défense à Paris. Les quartiers de loisirs (« recreational business district », RBDs) sont une autre composante indispensable des villes [LT03], dans laquelle les visiteurs viennent pour les loisirs, le tourisme et la consommation. Avec le développement du tourisme urbain, les quartiers d'affaires ont progressivement attiré une population autre que celle des travailleurs, et se sont par conséquent de plus en plus étendus à des fonctions

de loisirs en incluant des commerces, des restaurants, des cinémas ou encore des salles de concert [Zhu+15]. Toutes ces activités (travail, consommation et loisir) attirent quotidiennement un flux très important de voyageurs qui, pour beaucoup, utilisent les transports publics (85% des voyageurs pour le quartier de La Défense à Paris [Kro+14], 62% pour le quartier financier de Pittsburgh [BM83]). Les transports en commun présentent beaucoup d'avantages ; de manière non exhaustive, ils permettent d'accroître l'accessibilité des zones d'activités aux personnes sans voiture, ils réduisent la congestion et sont moins émetteurs en gaz à effet de serre.

Les grands quartiers d'affaires sont aujourd'hui remis en question sur plusieurs aspects, pour répondre aux enjeux du développement durable ou aux conséquences de la crise sanitaire de Covid19 notamment. Ils doivent rester des quartiers agréables, malgré la densité très forte de population s'y rendant quotidiennement. De plus, avec les conséquences de la crise sanitaire de Covid19, ils doivent se réinventer face aux nouvelles habitudes de déplacement conséquentes au télétravail et aux nouvelles formes de consommation (achats sur internet). Un récent article [Gui22], met en lumière les difficultés actuelles rencontrées par le quartier de La Défense pour attirer les jeunes actifs. La congestion, l'environnement excessivement minéral ainsi que les nouvelles habitudes de télétravail prises pendant la crise sanitaire contribuent à diminuer l'attractivité de ce centre économique.

Cette thèse s'inscrit dans ce contexte où le quartier d'affaires, couplé au transport public permettant aux voyageurs de s'y rendre, doit évoluer afin de s'adapter à ces nouveaux enjeux de développement durable et sociétaux. La multiplication des sources de données numériques (billettique, capteurs, etc.) et le développement de la modélisation statistique aident à mieux comprendre le déplacement des voyageurs dans ces espaces très concentrés en emplois et activités. L'objectif est, pour le quartier d'affaires et l'opérateur de transport, d'être en capacité de comprendre plus finement les flux voyageurs dans les espaces de transport, afin de mieux les appréhender et les anticiper pour contribuer à rendre le quartier plus agréable.

1.1.1 Les opérateurs de transports en commun : enjeux et responsabilités

Les transports en commun sont au cœur même du fonctionnement des zones urbaines, dans la mesure où ils permettent de déplacer les usagers entre leur lieu de résidence et leur lieu de travail, d'étude ou de loisir. Le développement des transports en commun améliorerait même la croissance économique des villes [WR09]. Au-delà de ces aspects, les transports en commun aident les usagers à réduire leurs dépenses liées aux déplacements et aident à lutter contre le réchauffement climatique. Les enjeux associés aux transports en commun étant de première importance, les opérateurs de transport détiennent par conséquent une grande responsabilité quant à la gestion des villes. Il est indispensable pour les opérateurs de transports en commun de connaître finement les dynamiques de déplacements au sein des réseaux afin de :

1. Cibler des anomalies de fonctionnement, et mieux les prévenir,

2. Développer de manière cohérente le réseau,
3. Anticiper à plus ou moins long terme les futurs déplacements des usagers, de manière à adapter l'offre de transport en conséquence.

Cette thèse s'est effectuée au sein de l'opérateur des transports en commun parisien RATP (Régie Autonome des Transports Parisiens). Le groupe RATP exploite huit modes de transport dans quatorze pays, mais son coeur d'activité reste l'Ile-de-France. Au sein de cette région, la RATP exploite 16 lignes de métro (302 stations pour 206 km de lignes), 8 lignes de tramway (105 km), une partie des lignes de bus (4 775 bus en exploitation), ainsi qu'une partie des lignes A et B du réseau express régional ou RER (66 gares pour 117 km). La RATP transporte ainsi plus de 3,3 milliards de passagers par an en Ile-de-France, ce qui en fait l'un des réseaux les plus fréquentés au monde.

1.1.2 Les espaces souterrains comme pôles d'échanges

Dans le contexte d'une ville plus durable, l'idée que les grands quartiers d'affaires devraient se réinventer gagne en popularité. Le quartier d'affaire devrait ainsi se développer, en jonglant entre une augmentation de sa densité et une meilleur accessibilité en transports, une réappropriation par les piétons et une mise en avant des espaces verts. La construction de nouvelles infrastructures est une problématique majeure des grands quartiers d'affaires tant l'espace est dense. Les travaux de [Pen+20] présentent les espaces souterrains (« urbain underground spaces », UUS) comme une solution viable ayant fait ses preuves dans de nombreux quartiers d'affaires dans le monde, parmi lesquels le quartier de La Défense à Paris (figure 1.1).

FIGURE 1.1 – Les espaces souterrains du quartier de La Défense peuvent être très chargés, notamment pendant les heures de pointe.

Il a notamment été établi que les espaces souterrains ont un impact significatif sur le développement urbain [QPW17]. Dans le cadre des transports urbains, les espaces souterrains accroissent la capacité de transport dans les villes. Ils sont ainsi utilisés pour optimiser

les systèmes de transport dans les grands quartiers d'affaires, notamment parce qu'ils permettent le passage d'infrastructures ferroviaires lourdes dont les avantages sont nombreux (sécurité, ponctualité, fiabilité et faibles émissions). Comme mentionné dans les travaux de [Pen+20 ; Cui+13], le passage d'un métro est un facteur majeur de développement des espaces souterrains des quartiers d'affaires car, en plus d'être un transport de masse, le métro peut être connecté aux tours de bureaux et aux zones commerciales. Des hubs de transports multimodaux souterrains ont naturellement été mis en place dans les grands quartiers d'affaires, car présentant une structure de réseau piéton très efficace pour faire le lien entre les lignes de transport, les zones de bureau et les zones commerciales. Ces travaux de thèse prennent l'exemple du quartier de La Défense, situé à l'ouest de Paris, dont les éléments techniques sont présentés dans la section 1.2.

1.1.3 Le lissage des heures de pointe

Beaucoup de systèmes de transports en commun doivent faire face à la gestion des heures de pointe, notamment celles du matin. Il s'agit de périodes au cours desquelles la demande peut dépasser la capacité de service du réseau, formant un goulot d'étranglement. Cette demande est principalement engendrée par les trajets domicile-travail (et domicile-études) qui se concentrent aux mêmes heures. Une heure de pointe mal gérée peut facilement entraîner des perturbations et une diminution de la qualité de service. Des programmes de gestion de la demande lors des heures de pointe ont été menés à travers le monde, avec l'aide de mesures comme la communication par les entreprises auprès de leurs employés, la tarification supplémentaire aux heures de pointe ou le rabais des tarifs aux heures creuses. Des villes comme Londres, Taipei, Washington DC ou Melbourne ont expérimenté la différenciation des prix entre périodes de pointe et périodes creuses. L'étude de [LWL10] a révélé que les usagers sont sensibles à la différenciation des prix, dans le cas de Taipei au moins. La ville de Melbourne a poussé le concept en rendant le trajet gratuit vers le quartier d'affaires central pour tout usager complétant son voyage avant 7 heure le matin. L'étude de [Cur10] montre que 23% des voyages hors heure de pointe étaient effectués par des usagers ayant décalé leurs habitudes temporelles de déplacement. Dans le travail de [YT18], les auteurs proposent une expérimentation où l'usager est « récompensé » d'un voyage gratuit hors pointe s'il a voyagé et payé plus d'un certain nombre de fois pendant la période de pointe.

A l'origine, cette thèse s'inscrit dans le cadre d'une expérimentation de lissage des heures de pointe menée depuis 2019 dans le quartier de La Défense, en partenariat avec la Région Île-de-France, Île-de-France Mobilités, Paris La Défense, la RATP, la SNCF et quatorze entreprises du territoire. Il est estimé qu'environ 100 000 personnes accèdent au site de La Défense entre 8h30 et 9h30, pendant l'hyperpointe. Dans ce contexte, la Région Ile-de-France, Ile-de-France Mobilités, l'établissement public local Paris La Défense et les opérateurs de transports (RATP et SNCF) se sont engagés à mettre en place des solutions alternatives permettant de lisser l'heure de pointe du matin, et ainsi contribuer à la désaturation des transports en commun empruntés quotidiennement par les salariés

des entreprises (RER A, Transilien, ligne 1 du métro, tramway 2, bus...). En contrepartie, les entreprises participant à l'expérimentation se sont engagées à réduire la part de leurs salariés utilisant les transports en commun à l'arrivée sur La Défense, à l'heure de pointe du matin, chaque jour ouvré et en particulier le mardi et le jeudi. Un travail de suivi des flux de voyageurs et de leur évolution suite au lancement de l'expérimentation est nécessaire afin d'en évaluer ses effets. Ce rôle a été assigné à la RATP. Dans ce cadre, elle a équipé le pôle de transport avec un dispositif de comptage de flux, dont les données sont depuis analysées (par la RATP) notamment pour mesurer les effets de l'expérimentation. Ces données, détaillées dans la section 2.2.1, nourrissent les travaux menés dans cette thèse.

1.1.4 La survenue de la pandémie de Covid19

Cette thèse a été menée dans le contexte particulier de la pandémie de Covid19. Au delà de l'impact notable que cette pandémie a eu sur la manière de mener des travaux de thèse, et des adaptations qui en ont découlé, les données utiles aux travaux ont aussi été modifiées en profondeur. Le quartier de La Défense, pôle tertiaire majeur où la majorité des métiers sont adaptables au télétravail, s'est vidé de ses employés pendant de longues périodes en 2020 et 2021. Cette situation inédite, dans laquelle la préoccupation du lissage des heures de pointe était mise de côté, a d'abord été vue comme un problème dans le cadre de cette thèse. La pandémie a finalement été un terrain d'expérimentation nouveau, car apportant des cas d'étude non anticipés, comme l'apparition d'habitudes de mobilité différentes, avec le télétravail notamment. Les différents confinements et couvre-feux ont été autant de périodes particulières dans nos données, mais ont également eu un impact sur les habitudes des usagers, même lorsque la pandémie commençait à s'atténuer. Dans les travaux de [Tho+21] sur l'impact de la pandémie sur les déplacements quotidiens, les auteurs précisent que l'utilisation des transports publics pourrait prendre un certain temps avant de retrouver les niveaux d'utilisation pré-Covid19, mais il se peut aussi que ce retour à la normale se stabilise à un niveau plus bas. En France, un ensemble de restrictions ont été mises en place pour lutter contre la propagation de la pandémie, les plus importantes ont été :

— Les trois confinements en Mars-Mai 2020, Novembre-Décembre 2020 et Avril 2021. Ces trois confinements n'avaient pas tous les mêmes niveaux de restrictions, mais ont chacun eu un impact notable sur la mobilité des personnes.

— Les couvre-feux qui, de Décembre 2020 à début Juin 2021 ont imposé à la population un retour au domicile avant 20 heures, voire 18 heures à certaines périodes.

Ces restrictions ont eu un impact considérable sur les mobilités des personnes, avec des périodes aux fréquentations nulles (télétravail à 100% obligatoire), fortement diminuées (télétravail obligatoire une partie de la semaine) ou décalées (couvre-feux). La pandémie de Covid19 a diminué drastiquement l'affluence dans les transports en commun, le travail de [GC21] qui rassemble un ensemble de résultats sur ce thème énonce que la baisse pendant

les confinements a été de l'ordre de 80%-90% dans les grandes villes de Chine et des Etats-Unis. Au Royaume-Uni, la baisse aurait été de 70%.

La pandémie de Covid19 aura donc sa place dans tous nos travaux, en tant que sujet d'étude ou bien en tant qu'élément de contexte à prendre en compte.

1.2 Eléments techniques autour du pôle de La Défense

Le quartier de La Défense est situé dans un bassin d'emplois dynamique à l'ouest de Paris. Le quartier est avant tout un quartier d'affaires, regroupant des immeubles de grande hauteur englobant trois millions de mètres carrés de bureaux (figure 1.2).

FIGURE 1.2 – Le quartier d'affaires de La Défense, à l'ouest de Paris.
Source : Image importée depuis 500px.

Quelque 2 500 entreprises et sièges de grandes multinationales attirent quotidiennement une population de 180 000 salariés. Au-delà du quartier d'affaires, La Défense est aussi un pôle commercial de premier plan, notamment avec le centre commercial « Westfield Les Quatre Temps ». D'autres structures, commerciales ou non, attirent quotidiennement un grand nombre de visiteurs autres que les salariés, on pense par exemple au Centre des Nouvelles Industries et Technologies (ou CNIT, un autre centre commercial et de congrès) ou à la Grande Arche (monument, expositions). Ces centres d'attractions sont complétés par de nombreux commerces, restaurants et cinémas dans le quartier. Le quartier est également un quartier d'habitation comptant 20 000 habitants. On peut mentionner que le quartier abrite un pôle universitaire (Léonard-de-Vinci) ainsi que quatre écoles de commerce, le tout attirant 45 000 étudiants. Enfin, un autre pôle attracteur situé à proximité du quartier est le stade de La Défense Arena. Cette immense salle, la plus grande d'Europe avec une capacité de 40 000 places, est le théâtre d'événements majeurs comme des concerts ou

des rencontres sportives. Ces événements, limités dans le temps, attirent un surplus très important d'usagers transitant par le pôle de transport.

Une étude réalisée pour l'établissement public Paris La Défense en 2006 montrait que 90% des salariés qui se rendaient à La Défense utilisaient les transports en commun. Cette donnée est valable pour les salariés uniquement, mais témoigne d'une utilisation massive des transports en commun pour se rendre dans le quartier. Les usagers arrivant dans le quartier de La Défense passent par le pôle de transport situé sous le quartier, appelé « Coeur Transport ». Sa dénomination dans le réseau de transports parisiens est « La Défense Grande Arche ». Les voyageurs peuvent choisir parmi de nombreuses lignes, ferrées ou non, pour se rendre à La Défense. La ligne A du réseau express régional (RER A), qui dessert Paris ainsi que de nombreuses banlieues Est et Ouest, passe ainsi au coeur du pôle de transport. La ligne 1 du métro parisien dessert également le quartier depuis Paris. Deux lignes de trains régionaux (Transilien L et U) sont également accordées au pôle de transport. La ligne 2 du tramway d'Ile-de-France correspond elle aussi avec les autres lignes à La Défense. Enfin, 16 lignes de bus desservent le quartier, la plupart s'arrêtant dans deux gares routières situées à proximité des lignes ferrées. Les lignes de RER A et de Métro 1 sont connues comme complémentaires et concurrentes, car elles partagent en partie des itinéraires similaires. Les usagers de La Défense empruntant le RER A peuvent se rendre dans les banlieues ouest ou est *via* Paris, tandis que ceux empruntant le métro 1 se rendent essentiellement à Paris, avec une desserte de la ville plus fine que le RER A. La figure 1.3 illustre les tracés et les stations des deux lignes de transport, et montre que les deux lignes sont parallèles lors de la traversée de Paris.

En plus de la station « La Défense Grande Arche », deux autres stations de moindre importance sont situées dans le périmètre du quartier d'affaires :

— **Esplanade**
Il s'agit d'une station de métro située entre le pôle et la Seine. La station est desservie par le métro 1.

— **Nanterre-Préfecture**
Cette station, située à l'ouest du pôle, est desservie par la ligne A du RER.

Nos travaux se concentreront essentiellement sur les données issues de la station « la Défense Grande Arche », au coeur du quartier d'affaires, et véritable pôle de transport alignant un nombre très important de lignes, ferrées ou non.

1. Les données sont disponibles sous la licence Open Database. Fond de carte et données issues de OpenStreetMap et OpenStreetMap Foundation.

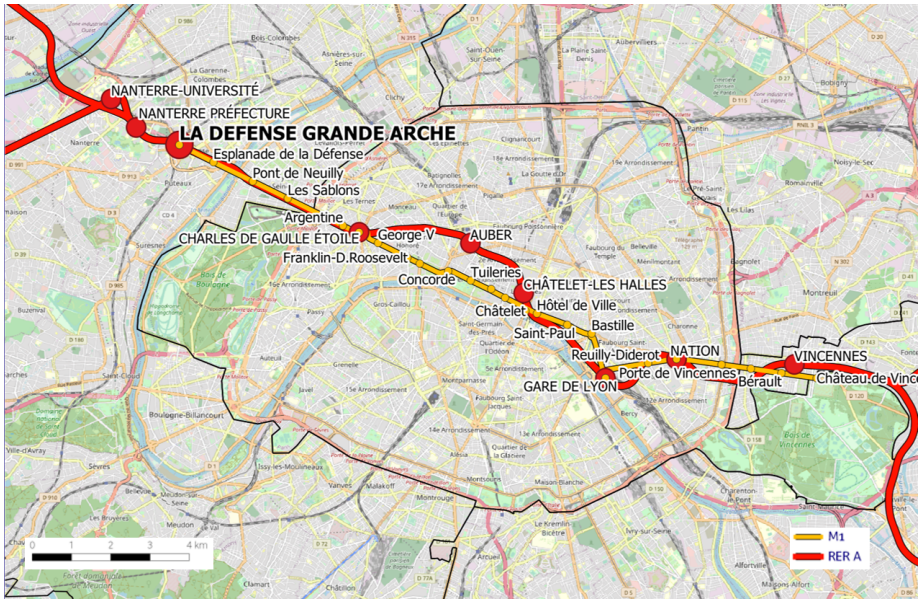


FIGURE 1.3 – Tracés et stations de la ligne 1 du métro parisien et du RER A.
Source : RATP/EDT 2021. © OpenStreetMap contributors [1]. Ile de France Mobilité 2020, last update : 2021/04/21

1.3 Motivations et contributions

1.3.1 Objectifs

Les paragraphes précédents ont mis en évidence la question de la gestion de l'affluence dans les espaces de transport denses, dans un contexte qui évolue. Nous proposons d'apporter à travers cette thèse des méthodes utiles, pour les gestionnaires d'espaces publics, à la compréhension et à l'anticipation de l'affluence dans les espaces de transport. Ce type de travail peut aider à l'amélioration de la qualité de service en indiquant aux opérateurs de transports où et quand allouer de la ressource (ex. augmenter la fréquence des trains ou placer des agents aux bons endroits en gare). Dans cette optique, nos contributions se déclinent en trois volets distincts : la décomposition des séries temporelles de fréquentation, la catégorisation des flux passagers en profils types de mobilités, et la prévision des affluences à venir. Les deux premiers volets permettent d'obtenir des visions théoriques d'utilisations de l'espace de transport, conséquentes à des périodes données ou des événements distincts. Ces deux volets sont destinés à être utilisés par des opérateurs de transports ou des gestionnaires des espaces afin qu'ils puissent définir par exemple où et quand allouer de la ressource (nombre de trains, agents en stations, etc.). Le troisième volet est destiné à l'information voyageurs en plus des deux acteurs précédemment énoncés. Il permet notamment aux voyageurs d'anticiper quels accès seront les plus chargés et quand, en conséquence de

certains événements. Ce type de travail pourrait être repris pour des recherches appliquées à d'autres espaces de transport que celui de La Défense. Dans les sections suivantes, nous détaillons les trois volets de la thèse, en y ajoutant l'exploration des données.

1.3.2 Explorer des données de mobilité en lien avec du contexte

Les données numériques (ex. billettique, WiFi) permettent des analyses dynamiques à des niveaux de précision géographique et temporelle fins, et offrent aux acteurs le potentiel de gérer efficacement leurs ressources. Cependant, les données numériques prises isolément sont partielles et biaisées, et leur capacité à saisir des phénomènes complexes et interreliés est encore réduite. Dans cette première partie de thèse, nous explorerons l'ensemble des données disponibles autour du pôle de La Défense comme les flux voyageurs, les données calendaire, l'exploitation des trains ou encore l'événementiel. Leur combinaison peut permettre de tirer parti des atouts de chaque type de données : pertinence, représentativité et fiabilité à des niveaux spatio-temporels fins. Ce travail concerne ainsi l'enrichissement, le traitement et la visualisation des données, et permet de mieux les comprendre en vue de les intégrer dans les différents travaux de la thèse.

1.3.3 Quantifier l'impact des variables contextuelles sur les mobilités

Le premier volet de cette thèse est un travail qui vise à explorer l'aspect temporel des données de fréquentation. Croiser les comptages piétons avec du contexte est une analyse qu'il est possible de faire simplement, lorsque l'étape de qualification des données a été menée. Cependant la compréhension de l'impact de différents éléments de contexte sur les flux piétons peut être difficile lorsque ces derniers s'accumulent. Il peut alors être délicat de faire un lien direct entre un élément de contexte et son impact sur les déplacements des piétons. Ce travail part du constat que le bruit dans les séries temporelles peut être en partie expliqué via un modèle de décomposition qui permet de mettre en valeur les différentes composantes cachées internes de ces séries. L'objectif de notre travail sera d'appliquer un de ces modèles qui nous permettra de comprendre comment les variations de séries temporelles de comptages piétons se traduisent dans ces composantes cachées, liées à divers éléments de contexte. Nous concentrerons l'étude sur un cas de décomposition comparée des séries de comptages de flux entrants vers les lignes de RER A et de ligne 1 du métro à La Défense. Ce cas d'étude permettra de mettre en lumière comment les éléments de contexte influencent les usagers sur l'utilisation préférentielle de l'une ou l'autre des lignes. Les données journalières de billettique collectées sur le long terme (9 années) ainsi que des élements de contexte comme les grèves ou les travaux de maintenance seront utilisées dans cette analyse. Nous nous appuierons sur les modèles linéaires dynamiques qui, à travers une structure de type espace-état, permettent d'intégrer de manière flexible une grande variété de facteurs, calendaires ou non, dans la modélisation. Nous n'avons, dans nos recherches, pas trouvé d'application des modèles de décomposition pour l'analyse comparée

des mobilités dans les transports en commun. Ce travail est une porte ouverte à ce type d'application qui peut s'avérer utile à tout gestionnaire de transports publics souhaitant comprendre plus finement la dynamique temporelle des flux piétons dans les espaces de transports.

1.3.4 Détecter des périodes aux mêmes dynamiques de mobilité

Un pôle de transport multimodal, comme celui de La Défense, est un lieu d'étude complexe pour les données de comptages, car il s'y croise quotidiennement un grand nombre de flux de passagers. L'extraction d'informations à partir de ces données, souvent bruitées, est une tâche difficile. Il peut donc s'avérer délicat pour le gestionnaire en charge des espaces d'avoir une vision claire des dynamiques de déplacements piétons, pourtant nécessaire pour la planification des services de transport ou la gestion des flux piétons. Une approche de clustering s'avère ainsi être une approche indispensable lorsque l'on souhaite mener une étude des dynamiques de déplacements. Cette approche permet en effet de synthétiser des données de fréquentation en lien avec du contexte, au sein de catégories facilement interprétables. L'objectif est d'utiliser des approches de clustering pour détecter des périodes de temps aux dynamiques de déplacements homogènes d'une part et de leur associer des profils de déplacements caractéristiques d'autre part. Pour cela, nous développerons des méthodes à base d'apprentissage statistique permettant de détecter des catégories de profils liées à des segments de temps homogènes, mais aussi les changements sur ces profils liés à du contexte. En terme méthodologique, nous nous appuierons sur des modèles probabilistes de mélange pour intégrer les aspects catégorisation et modélisation des événements. Les modèles de régression spécifiques aux segments mis en place prennent en compte les corrélations entre les séries, la surdispersion, ainsi que l'impact des facteurs exogènes. À cette fin, nous mettons en place et comparons deux modèles de mélange adaptés à notre problématique : le modèle de mélange de Poisson log-normal et le modèle de mélange de distributions « sommes et partages ». Ce chapitre contribue à enrichir les modèles de mélange en y ajoutant une modélisation inspirée des modèles « sommes et partages », issus de la littérature. Il s'agit d'un premier travail où ce type de modèle est comparé à un modèle de Poisson log-normal ; les forces et faiblesses de chaque méthodologie sont ainsi explorées. D'un point de vue opérationnel, ce type de travail peut fournir à tout gestionnaire d'espaces publics une vision synthétique des déplacements, afin de mieux les interpréter.

1.3.5 Prédire les futures mobilités

La prévision de l'affluence dans les espaces de transport est un sujet de recherche majeur pouvant servir à enrichir l'information voyageurs à destination des usagers des transports collectifs, qui peuvent ainsi mieux planifier leur déplacement. Elle peut également servir aux opérateurs de transport pour une régulation « à la demande » de l'offre de transport. Beaucoup de travaux de prédiction de la demande des usagers se concentrent sur des

prédictions moyennes et ne peuvent donc pas être utilisées pour l'analyse de l'incertitude. Cette incertitude est pourtant particulièrement adaptée dans le domaine des transports, où le risque d'une forte affluence mal gérée est à éviter. Nous proposons dans ce chapitre la mise en place de modèles de prédiction probabilistes des flux voyageurs dans le quartier de La Défense avec des méthodes basées sur l'apprentissage profond. Ces modèles sont capables de modéliser l'incertitude en s'appuyant sur une abstraction des données contextuelles et en faisant l'hypothèse de distributions en sortie. Nous mettons en place pour cela un nouveau modèle de prédiction probabiliste basé sur l'apprentissage profond, venant enrichir la littérature associée à ces modèles. Notre modèle apprend une représentation latente des données en entrée avec l'aide d'un réseau de neurone récurrent, puis la traduit en prévisions de flux passagers en plusieurs points avec la modélisation « somme et partages », rencontrée dans le chapitre précédent. Nous comparons ce modèle avec d'autres modèles issus de l'état de l'art sur des données en open source et sur notre cas d'étude du pôle de La Défense. Notre modèle semble trouver un avantage dans certaines situations où les données présentent des régularités.

1.3.6 Publications

Journal international

— de Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Merad-Boudia, Y. "What can we learn from 9 years of ticketing data at a major transport hub ? A structural time series decomposition". In : Transportmetrica A : Transport Science, 1-25 (2021).

— de Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Merad-Boudia, Y. "Multivariate count time series segmentation with « sums and shares » and Poisson lognormal mixture models. A comparative study using pedestrian flows within a multimodal transport hub." In : Soumis, en cours d'évaluation pour la revue Advances and Data Analytics and Classification (2022).

— de Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Merad-Boudia, Y. "DeepNegPol : A multivariate probabilistic time series forecasting with « sums and shares » distributions." In : En préparation, à soumettre à la revue IEEE Transactions on Intelligent Transportation Systems (2022).

Conférences avec présentations

— de Nailly, P., Oukhellou, L.,Côme, E., Samé, A., Ferriere, J., Darrort, N. "A combined use of sensor count data and smart card data to study the flows of a multimodal transport hub". In : 6th TransitData Workshop and Symposium, Toronto, Canada, August 2020.

— de Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Darrort, N. "Using sensor count data to study the spatio-temporal flows within a multimodal transport hub". In : European Transport Conference, Milan, Italy, September 2020.

— de Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Merad-Boudia, Y. "Comparison between sums and share and Poisson log-normal mixture models : a case study using pedestrian flows within a multimodal transport hub". In : 11th Triennal Symposium on Transportation Analysis, Mauritius Island, June 2022.

— de Nailly, P., Côme, E., Samé, A., Oukhellou, L., Ferriere, J., Merad-Boudia, Y. "Applying a new "sums and shares" mixture model to pedestrian flows : the "La Défense" hub case study". In : European Transport Conference, Milan, Italy, September 2022.

Chapitre 2

Exploration de données de mobilité multi-sources

2.1 Introduction

Le développement durable des territoires est un enjeu qui semble aujourd'hui de plus en plus prégnant, dans un souci de gestion efficace des ressources et des espaces partagés. Les acteurs qui gèrent les territoires voient dans les données numériques un potentiel d'analyse dynamique, à des niveaux élevés de précision géographique et temporelle, nécessaire à une gestion efficace et durable du territoire. Les données numériques présentent des avantages certains par rapport aux données d'enquêtes traditionnellement utilisées pour l'observation et l'analyse de la mobilité dans les espaces urbains. Les données numériques peuvent être acquises quasi continuellement, et présentent une finesse spatiale et temporelle.

En revanche, prises isolément, les données numériques restent limitées, car incapables de saisir des phénomènes complexes et interdépendants. Leur combinaison à d'autres données classiques (données socio-économiques, données calendaires et d'événementiel, etc.) et/ou numériques (GPS, WiFi, téléphonie, etc.) permettrait de tirer parti des atouts de chaque type de données. L'exploitation d'une grande quantité de données doit passer par la résolution d'un certain nombre de problèmes afin d'être pertinente. Cela peut concerner leur enrichissement, leur traitement ou leur visualisation. Dans le domaine du transport en commun, les études passent régulièrement par l'utilisation de données numériques issues de la billettique. La donnée billettique prend sa source au niveau des lignes de contrôle présentes aux différents points d'accès des stations de métro ou de RER. Chaque passage de personne à une ligne de contrôle est enregistré avec le temps, le lieu et le sens de passage. Cette donnée acquise en continu est précieuse pour tout acteur souhaitant étudier les flux d'usagers dans le réseau de transports. La billettique présente néanmoins un certain nombre d'inconvénients, comme le fait qu'elle ne compte ni les passages sortants du métro (en Ile-de-France) ni les personnes qui fraudent. Il arrive de plus que les

lignes de contrôle soient totalement ouvertes lors de certaines périodes de forte affluence, impliquant un biais dans les données. Dans ces travaux, nous nous intéressons également à une autre source de données numériques, acquises avec un dispositif de comptage installé par la RATP en collaboration avec la Région Ile-de-france, sur le territoire de la Défense. Un ensemble de 14 postes de comptage par capteur stéréoscopique a été placé en différents points d'accès du pôle. Ceux-ci comptent le nombre d'entrants/sortants à chaque minute depuis le début du mois de Mars 2019. Ces deux sources de données de flux piétons peuvent être croisées avec des facteurs impactants, connus comme ayant une influence sur les habitudes de déplacement. Ces facteurs peuvent être calendaires, lorsque l'impact se traduit par l'évolution naturelle de l'activité des usagers au cours du temps. Les facteurs non-calendaires, eux, sont les événements non liés à la temporalité, et ayant un impact court-terme ou long-terme sur les fréquentations. Ces facteurs sont très diversifiés ; on peut parler aussi bien de pandémie, que de grèves ou de concerts.

Nous allons, dans ce chapitre, présenter l'exploration de l'ensemble de ces données. Nous étudierons dans un premier temps les données de fréquentation billettique, et celles issues des capteurs de comptage. Nous passerons notamment par un croisement de ces données, afin de constater les apports de chacune et leur complémentarité. Nous effectuerons ensuite un inventaire des facteurs exogènes pouvant influencer la mobilité des personnes. Cette étude exploratoire est une phase qui nous permettra de mieux comprendre les données en amont de toute mise en place de modèles. Les informations apprises ici pourront en effet nous aider sur les choix et encodages à mettre en place dans les différents travaux de la thèse.

2.2 Les données de mobilité

Nous parlerons régulièrement de « flux piétons » tout au long de ce travail. Il s'agit d'une mesure standard du nombre de passages de piétons, dans un sens ou dans l'autre, qui passent par une porte, un couloir ou une ligne de contrôle dans un intervalle de temps donné.

2.2.1 Données de comptages issues de capteurs stéréoscopiques

La technologie des capteurs stéréoscopiques est un système de collecte de données que l'on retrouve dans de nombreuses applications, et plus particulièrement dans deux domaines : la détection des piétons [Kri+16] et l'aide à la navigation des véhicules [Pel+15]. Situés aux différents points d'accès aux stations Grande Arche et Esplanade, les 14 postes du système de comptage relèvent le nombre d'entrants et sortants agrégés à chaque minute. Les largeurs de sections sur lesquelles les piétons sont comptés peuvent varier d'un lieu à l'autre : il peut s'agir de couloirs d'accès, de portes ou d'escaliers. Ces postes, placés en des lieux variés du pôle, captent le nombre de passages piétons entre le pôle de transport et l'extérieur, avec des contextes divers tels que la proximité d'un centre commercial, un

accès vers les tours de bureau, une gare routière ou une salle de spectacle. Nous présentons dans la figure 2.1 un schéma du pôle de transport de La Défense avec la position des 14 postes et une description des lieux dans lesquels ils se trouvent, ainsi que l'environnement proche représenté par des logos. Les postes P1 à P12 sont situés à la station Grande Arche, et les postes P13 et P14 à Esplanade.

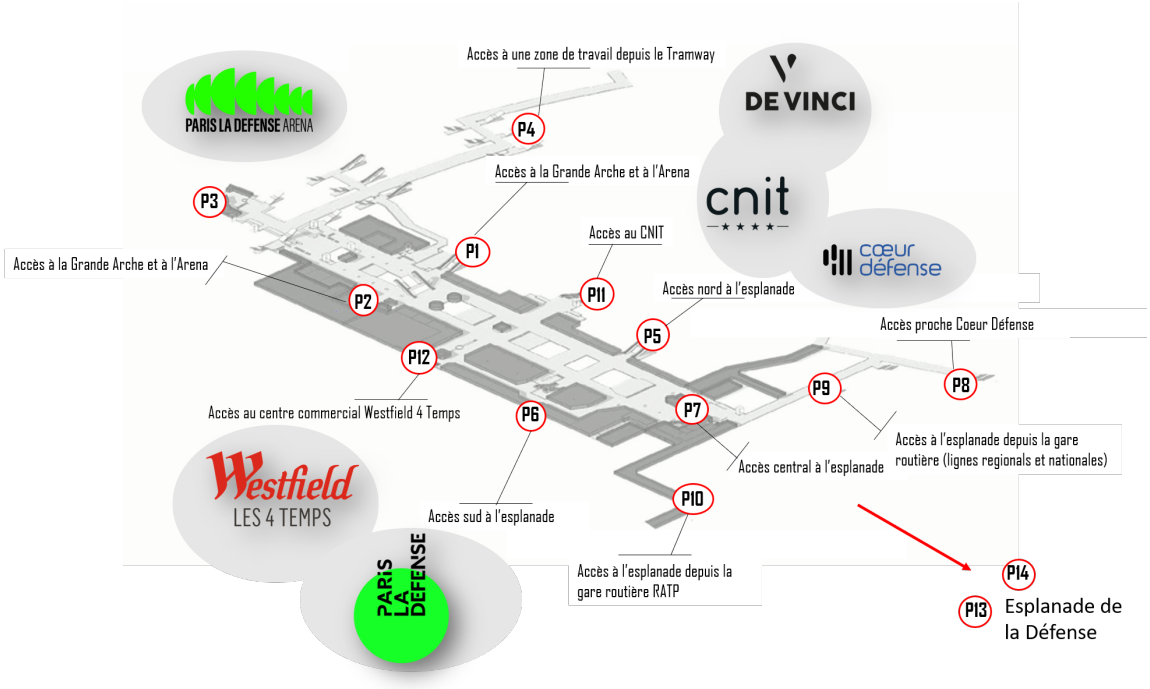


FIGURE 2.1 – Disposition des postes de comptage sur le pôle de La Défense, et environnement proche.

Les environnements proches associés à chaque capteur se reflètent dans les profils de fréquentation captés. Dans les figures 2.2 et 2.3 nous représentons les médianes ainsi que les quartiles 1 et 3 des comptages piétons agrégés à chaque 30 minutes, collectés à chaque capteur pour les entrants et les sortants, pour les jours de travail ou non. Notons d'abord la différence notable entre les profils associés à des capteurs différents, ainsi que la différence entre les jours de travail ou non. Pour les jours ouvrables on constate, comme attendu, la présence de pics de sortants le matin, et des pics d'entrants le soir. Cet effet n'est en revanche pas visible pour les jours non ouvrables. Ce type de profil est courant dans les grandes zones d'emploi, où les employés arrivent en transports en commun le matin et repartent avec ces mêmes transports le soir après le travail. Pour les jours non ouvrables, cette dynamique ne s'applique plus, d'où une si grande différence entre les jours, régie par l'activité du travail. Si l'utilisation du pôle de transport n'est pas homogène dans le temps,

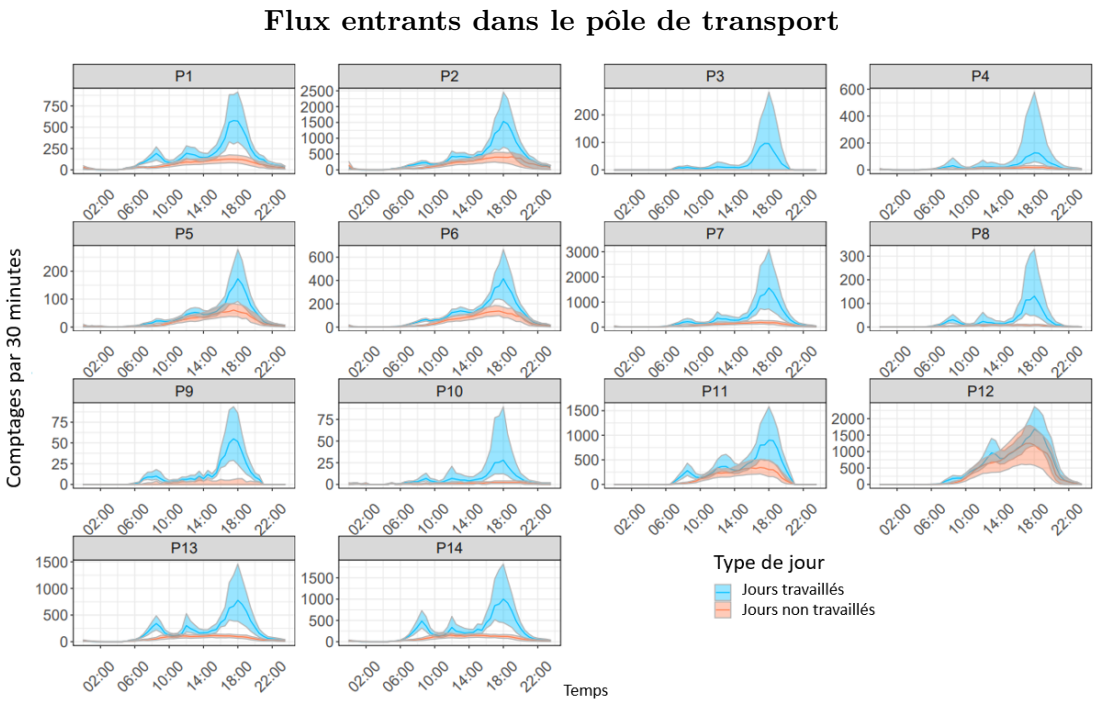


FIGURE 2.2 – Médianes, 1ers et 3èmes quartiles des comptages de flux entrants, par pas de 30 minutes pendant les jours ouvrables (en bleu) et les jours non ouvrables (en rouge) à chacun des 14 postes de comptages.

elle ne l'est pas non plus dans l'espace. Tout d'abord, on constate une grande différence en termes d'affluence selon les lieux, qui dépend des lieux d'activité et des lignes de transports environnantes. Par exemple le poste P7 est situé en un lieu stratégique, proche d'un accès au RER A et d'une zone de travail centrale, ce qui explique qu'il soit très emprunté. Le poste P4 est lui aussi proche d'une zone de travail, mais plus distant des accès aux lignes de transports, excepté au tramway T2, ce qui explique qu'il soit bien moins emprunté que le P7. Un deuxième constat que l'on peut faire est que si l'activité de travail, prédominante, explique en bonne partie les profils de fréquentation, elle n'est pas la seule activité qui attire dans le quartier, et cela se reflète aussi sur les profils. L'activité des jours non ouvrables est moindre, mais pas de manière aussi forte pour tous les capteurs. Les postes P1, P2, P5, P6, P11 et P12 gardent une activité lors de ces jours, ce qui témoigne d'une activité de loisir et commerciale plus ou moins prononcée dans le pôle. Pour les postes P1, P2, P5 et P6, on constate, en étudiant le quartier, que certaines zones d'activité telles que des magasins, des grands centres commerciaux et une salle de concerts (Arena), engendrent cette activité continue (pour les magasins) ou ponctuelle (pour les événements), comme on le verra dans la section 2.3. Pour les postes P11 et P12, il s'agit clairement de l'activité

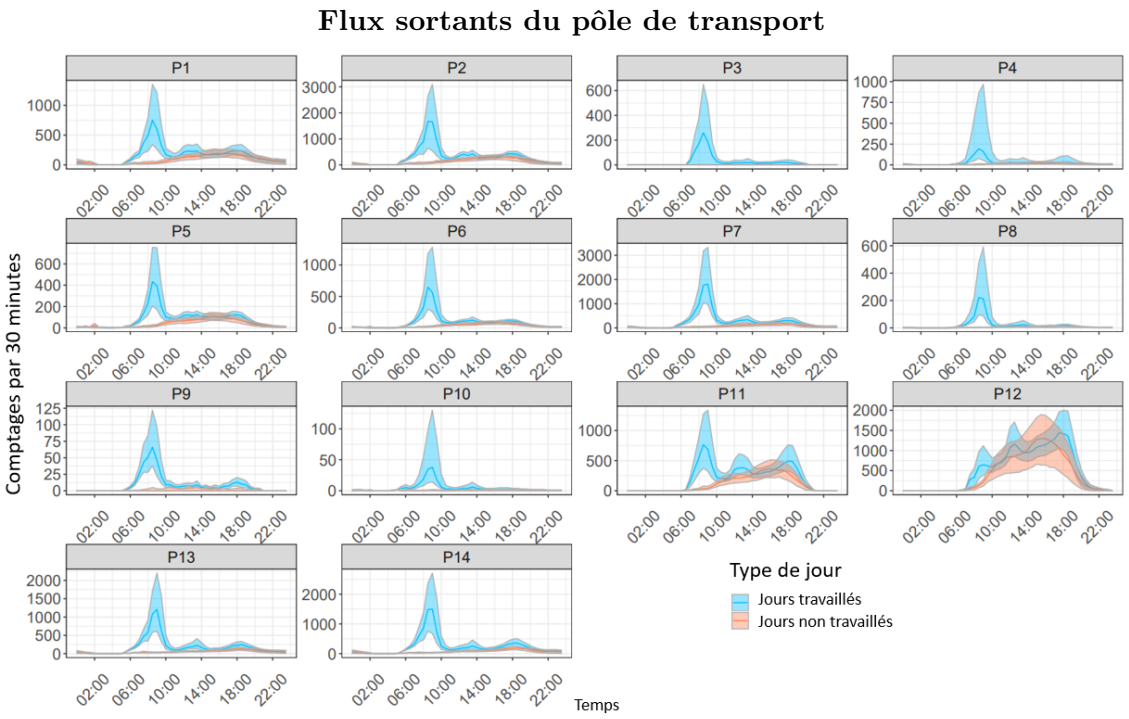


FIGURE 2.3 – Médianes, 1ers et 3èmes quartiles des comptages de flux sortants, par pas de 30 minutes pendant les jours ouvrables (en bleu) et les jours non ouvrables (en rouge) à chacun des 14 postes de comptages.

commerciale très forte provoquée par les deux centres commerciaux Westfield 4 Temps et CNIT. Ces deux postes comptabilisent en effet les flux entre le pôle de transport et ces centres commerciaux majeurs. La table 2.1 résume les types d'activité dans le quartier de La Défense et les capteurs associés.

Activité	**Capteurs**	**Description profil**
Travail	Tous	En semaine : profils pendulaires avec un pic le matin et un pic le soir.
Commercial	P11 et P12	En semaine : forte activité toute la journée. En week-end : l'activité reste importante. Le centre commercial Westfield 4 Temps semble absorber le plus d'activité sur le pôle le week-end.
Loisirs divers	P1, P2, P5, P6, P11, P12	En week-end : activité plus forte que dans les zones dédiées au travail.

TABLE 2.1 – Table des différents types d'activités relevés par les capteurs, avec la description des types de profils engendrés.

2.2.2 Données de comptages issues des validations de billettique

Les cartes à puce et systèmes de billettique électronique (AFC) produisent de grandes quantités de données, qui sont fréquemment utilisées pour analyser la mobilité urbaine. Les travaux de [Bri+17], [PSY20] ou [Wan+21] impliquent par exemple l'utilisation de ce type de données. Dans notre cas, les données de billettique se présentent sous la forme suivante : on a, pour chaque ligne de contrôle, le nombre de passages agrégé par pas de 10 minutes. Chaque ligne de contrôle est par ailleurs associée à une « fonction » catégorisée de F1 à F6, qui permet de connaître la nature du flux (entrée, sortie, correspondance...). Par exemple, si l'on veut comptabiliser tous les nouveaux entrants à la station, on comptabilisera l'ensemble des flux associés à la fonction F1. De ce fait, les comptages issus des données de billettique complètent la vision apportée par les capteurs, en informant sur les mouvements internes au pôle de transport de La Défense. Nous regroupons sous un même identifiant les bornes ayant la même fonction, par exemple l'identifiant « M » correspond aux bornes donnant accès à la ligne de métro 1. Nous représentons dans la figure 2.4 l'ensemble des identifiants de flux captés par la billettique, que nous utiliserons tout au long de cette thèse. Les logos des lignes ferrées proches sont également placés sur cette figure. De la même manière que nous l'avons fait avec les comptages par capteurs, nous pouvons visualiser les profils de fréquentation sous l'angle des données de billettique, dans la figure 2.5 pour la station Grande Arche.

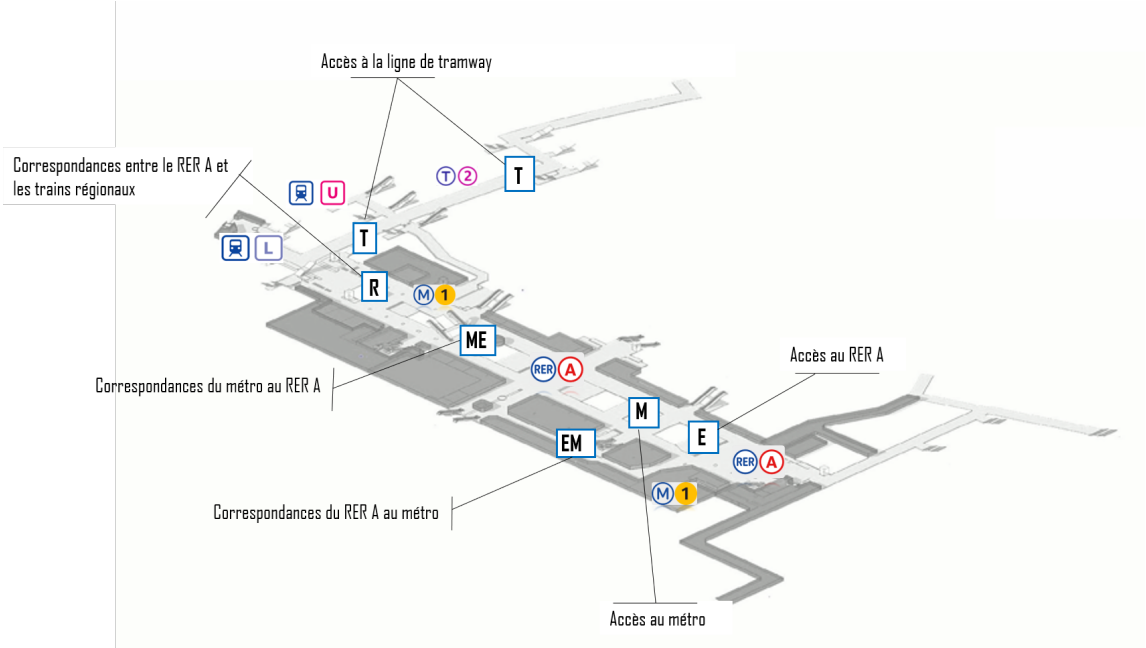


FIGURE 2.4 – Disposition des groupes de lignes de contrôle sur le pôle de La Défense, et placement des lignes ferrées.

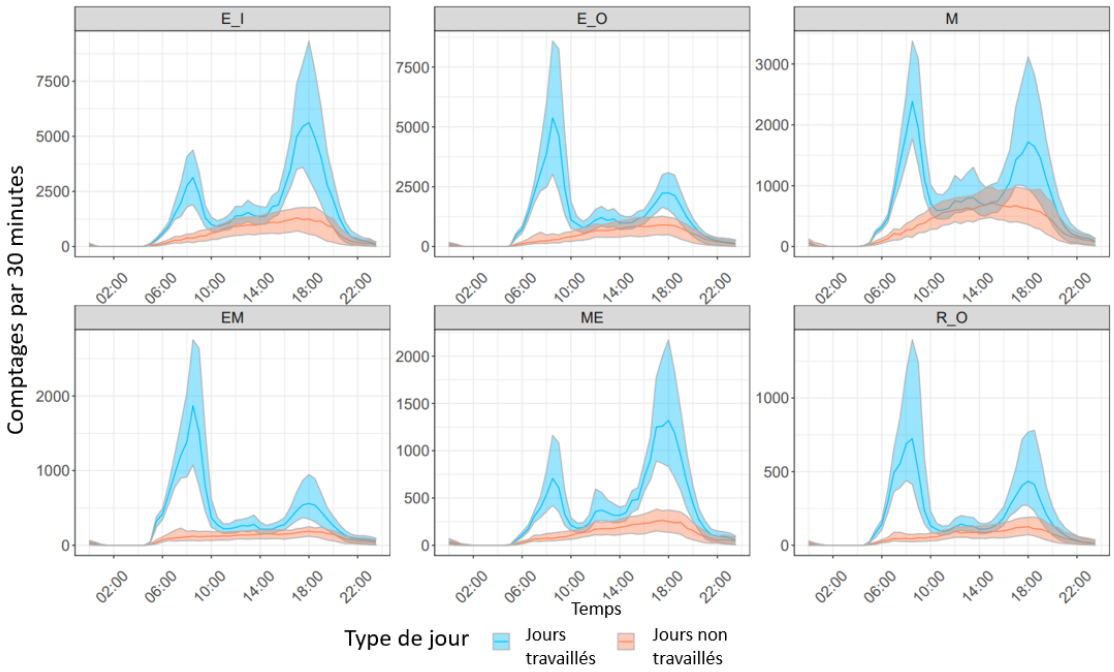


FIGURE 2.5 – Médianes, 1ers et 3èmes quartiles des comptages de flux par pas de 30 minutes pendant les jours ouvrables (en bleu) et les jours non ouvrables (en rouge), pour chaque type de flux capté par la billettique. La chaîne « _O » (pour « Out ») correspond à un flux sortant, et « _I » (pour « In ») à un flux entrant.

Dans le cas des jours ouvrables, on observe sur la figure 2.5 le rôle du RER A comme moyen de transport privilégié pour venir à La Défense le matin (E_O, flux sortants du RER A), et pour en partir le soir (E_I, flux entrants dans le RER A), en raison des flux importants lors des pics du matin et du soir. La ligne de métro 1 est également très empruntée ; même si l'on ne peut pas quantifier les sortants de cette ligne, on peut néanmoins constater qu'il y a beaucoup d'entrants le matin et le soir. Les raisons d'utilisation sont différentes : le matin, il s'agit de personnes arrivant en transports à La Défense, puis empruntant la ligne de métro 1 pour compléter leur trajet (pour aller à la station Esplanade par exemple) ; le soir, il s'agit de personnes quittant le pôle après la journée de travail. De nombreux transferts se font entre le RER A et le métro 1. Dans le sens « RER A vers métro 1 » (EM), il s'agit de personnes qui, en grande majorité le matin, correspondent vers le métro 1 pour finaliser le trajet vers Esplanade. Dans le sens « métro 1 vers RER A » (ME), c'est l'inverse, les usagers arrivent d'Esplanade en métro et viennent à Grande Arche emprunter le RER A. Enfin, les correspondances du RER A vers les Transilien L et U (R_O) se font aussi bien le matin que le soir, dans des proportions moindres. Les profils de fréquentation des jours non ouvrables sont, comme dans le cas des données issues de

capteurs, dénués de pics de fréquentations, comme on peut s'y attendre.

2.2.3 Caractéristiques majeures de la dynamique des données de billettique et de comptage

Les comptages de personnes collectés dans le pôle de La Défense présentent un ensemble de caractéristiques, dont nous allons ici présenter les principales. Par ordre décroissant de durée de chaque caractéristique, nous pouvons commencer avec la tendance long-terme. La tendance dans les fréquentations est un effet qui se voit sur le long terme, car étant la conséquence d'évolutions socio-démographiques, de modifications de l'environnement du pôle de transport (comme l'implantation de nouvelles entreprises ou de nouvelles lignes), des lignes de transport ou encore des changements d'habitudes durables (télétravail). Cet effet ne se révèle qu'avec un grand historique de données. Dans la figure 2.6, nous présentons l'évolution de la fréquentation journalière entrante dans les stations RATP du pôle de La Défense (donc la somme des flux entrants vers RER A, Métro 1 et tramway T2) entre début 2011 et Mai 2022. Si nous omettons ici la période de la pandémie de Covid19 qui modifie brutalement les mobilités depuis 2020, une tendance à l'augmentation des fréquentations semble se dessiner sur les années précédentes. Cette conclusion s'applique d'un point de vue général, mais pourrait différer d'un lieu à l'autre du pôle, selon l'environnement proche (ligne de transport ou zone d'activités).

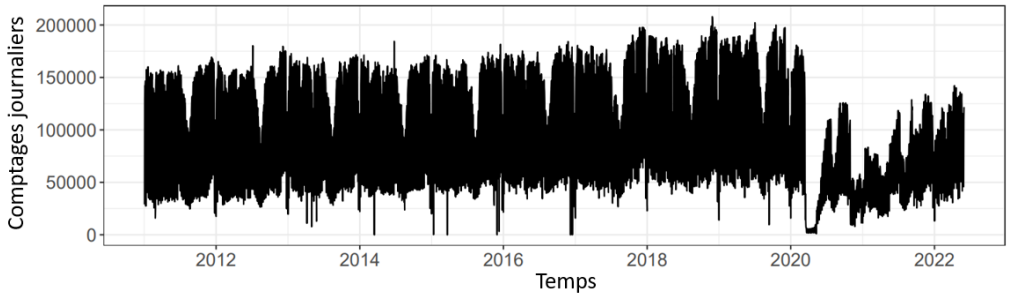


FIGURE 2.6 – Flux journaliers entrants entre 2011 et 2022 à la station Grande Arche. Une tendance à la hausse se dessine lorsque l'on regarde l'évolution des fréquentations sur le long terme, puis un changement brutal se produit lors de la pandémie de Covid19 (mars 2020).

Nous examinons ensuite les évolutions à des échelles de temps plus courtes : tout d'abord, à l'échelle annuelle (voir figure 2.7), où les périodes creuses (vacances) et les périodes non creuses de l'année sont visibles, puis à l'échelle hebdomadaire (voir figure 2.8), où les jours de semaine présentent des flux plus importants que les jours de week-end. En effet, un grand nombre de travailleurs se rendent quotidiennement au pôle de La Défense, avec une part significative d'entre eux utilisant les transports publics. Il est également à

noter que les écarts-types sont élevés, soulignant ainsi une grande diversité dans l'utilisation temporelle du pôle, notamment lors des jours de travail. Les comptages journaliers très bas, proches de zéro, sont à mettre en relation avec les périodes de confinements les plus fortes au cours desquelles le quartier était vidé de ses employés.

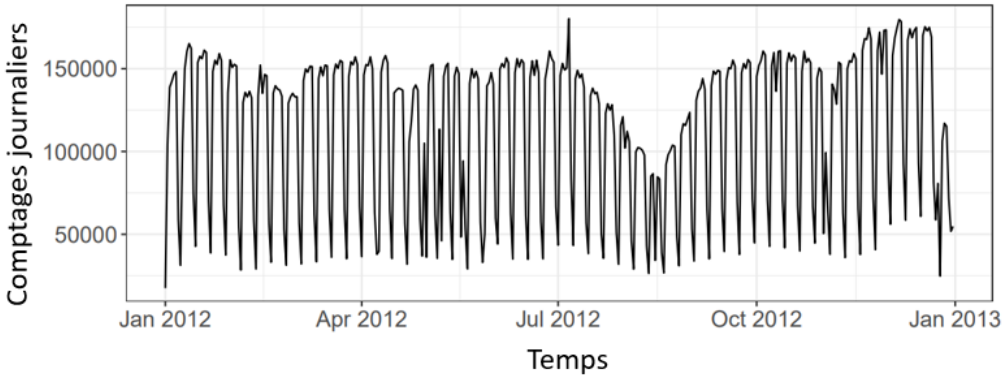


FIGURE 2.7 – Flux journaliers entrants en 2012 à la station Grande Arche.

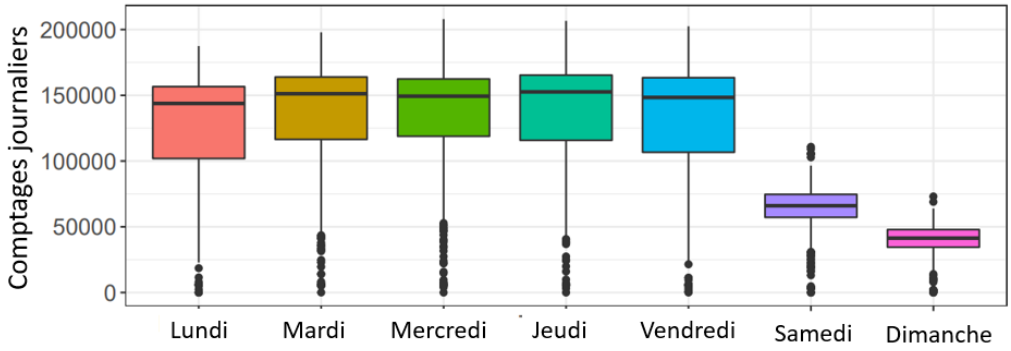


FIGURE 2.8 – Boxplots de fréquentation pour chaque jour de la semaine sur la période 2011-2022 pour les entrants à la station Grande Arche.

Comme on peut le voir dans la figure 2.6 sur l'évolution des fréquentations du pôle de transport, de nouvelles évolutions s'opèrent après 2020, notamment en raison des mesures prises contre la pandémie de Covid19. Ces évolutions sont le fruit de facteurs qui impacteront les fréquentations à plus ou moins long terme. En plus de la pandémie de Covid19, on pense aussi aux périodes de grèves, de travaux ou encore de pollution qui amènent à des changements dans les mobilités. Dans la section suivante (2.3), on établira une liste (non exhaustive) de ces facteurs pouvant impacter à plus ou moins long terme la mobilité des personnes dans le pôle.

Les travaux présentés dans cette thèse sont aussi basés sur des données de comptage à une échelle plus petite que la journée (heures). Plusieurs points caractérisent les données de comptage prises à cette échelle. On constate de la surdispersion et des corrélations entre les séries de comptages captées en différents endroits du pôle de transport, même lorsque l'effet des types de jours (ouvrés ou non ouvrés) et des heures est absent. Nous prenons l'exemple d'une courte période « normale », à savoir avril 2019, pour laquelle nous avons supprimé toutes les tranches horaires comportant des événements spéciaux tels que des perturbations des transports ou des concerts et où nous sélectionnons les jours ouvrés seulement. Nous visualisons d'abord l'effet de surdispersion en calculant les moyennes et les variances empiriques de chaque lieu de comptage à chaque heure sur cette période. Les résultats présentés dans la figure 2.9 montrent que les variances (axe des y) sont beaucoup plus élevées que les moyennes (axe des x), ce qui suggère une surdispersion. La modélisation des données avec une distribution de Poisson, qui fait l'hypothèse d'égalité entre la variance et la moyenne, pourrait ainsi s'avérer difficile (la variance théorique estimée sous une loi de Poisson serait inférieure à la variance observée dans les données).

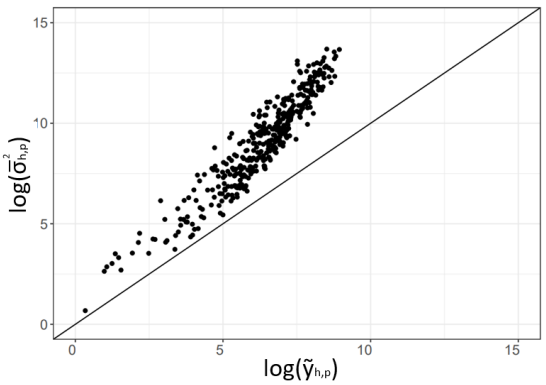


FIGURE 2.9 – Variances empiriques en fonction des moyennes empiriques calculées par lieux de comptages p et tranches horaires h, sur l'ensemble des jours J de la période d'avril 2019 (échelle logarithmique).

Les corrélations entre les séries de comptage sont abondantes dans les données. Pour mettre en évidence ce point, nous avons calculé les corrélations entre les séries de comptages pour trois heures de la journée dans la figure 2.10 : une heure du pic du matin (8h00), l'heure de midi (12h00), et une heure du pic du soir (18h00). Le nom de code du lieu est associé à « O » lorsqu'il s'agit d'un flux sortant, et à « I » lorsqu'il s'agit d'un flux entrant. Notons que les corrélations peuvent être positives ou négatives, mais nous constatons ici une prédominance de corrélations positives.

Les données de billettique et de comptage présentent des caractéristiques que l'on rencontre couramment lorsque l'on étudie des données de mobilité. Ces données sont fortement structurées par de multiples effets saisonniers, qu'ils soient à l'échelle du jour, de la semaine

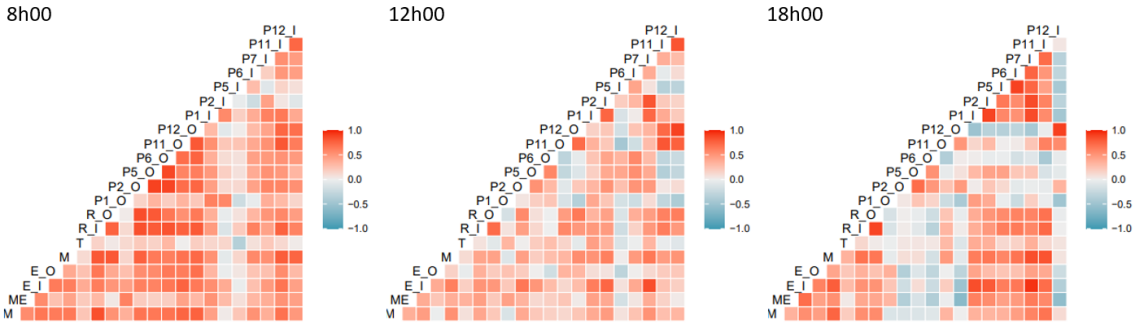


FIGURE 2.10 – Matrices de corrélation entre les différents lieux de comptage pour trois heures particulières de la journée. Notons que les corrélations semblent être davantage influencées par la direction des flux (« O » ou « I ») que par la proximité géographique des lieux. On observe un effet de *vague* avec des flux « O » bien corrélés à l'heure de pointe du matin (8h), et des flux « I » bien corrélés à l'heure de pointe du soir (18h). Les usages du pôle étant moins soumis à ces effets de vagues en dehors des heures de pointe, les corrélations semblent moins fortes à 12h.

ou de l'année. De plus, nos données présentent un aspect très dispersé, et une structure de corrélation qui évolue au cours de la journée. Dans la section suivante, nous croisons les deux sources de données afin de constater les éventuelles ressemblances et différences.

2.2.4 Croiser les données de mobilité

Comme nous l'avons illustré dans les sections précédentes, les deux systèmes de collecte de données capturent les flux de personnes au sein d'un même pôle de transport. Pour autant, ces deux sources n'apportent pas nécessairement la même information, comme nous allons le voir ici. Une ligne de contrôle n'est pas nécessairement associée à un point d'accès (de et vers l'extérieur) en particulier, mais il est possible de comparer l'ensemble des entrées/sorties comptées par les capteurs avec l'ensemble des flux entrants ou sortants de et vers les lignes de transport comptés par la billettique. Une grande partie des sorties sont effectivement quantifiables par la billettique, car la station dispose de sorties RER (équipées de lignes de contrôle, contrairement au métro).

Nous représentons avec la figure 2.11 les relations entre les comptages à 10 minutes capturés par les deux sources de données sur les jours ouvrés avant la pandémie de Covid19, pour les flux entrants puis sortants. En vert, il s'agit des flux observés lors des périodes de pointes du matin (7h-10h30), en bleu, des flux lors des périodes de pointes du soir (16h-20h30) des jours ouvrés, et en rouge, des autres. Les deux droites noires permettent d'identifier les valeurs de comptage issues de capteurs, qui sont distinctes à +/-5% de celles issues de la billettique.

Deux éléments émergent à l'étude de cette comparaison. Le premier est que les deux sources de données se complètent : lorsqu'une source est défaillante, l'autre permet de récupérer l'information manquante. On le voit lorsque les données sont à zéro pour une source et non nulles pour l'autre source. Le deuxième est que les deux sources de données s'enrichissent, c'est-à-dire qu'un ensemble d'informations peut être révélé par une source, mais pas par l'autre. On le voit à travers trois tendances :

1. Pour le cas des sorties, les capteurs comptent l'ensemble des flux, là où la billettique ne comptabilise que les sortants du RER. Le profil résultant permet ainsi de discriminer les flux sortants du RER, des flux sortants des autres lignes de transport.
2. Les valeurs de comptages issues de capteurs sont supérieures à celles issues de la billettique, dans le cas des flux entrants. Cela peut être dû à la fraude d'une part (fraudeurs comptabilisés par les capteurs, mais pas par la billettique), et à l'accès de certains usagers aux magasins internes à la station (sans entrer dans les espaces de transports), d'autre part.
3. Une tendance de valeurs de comptages élevées pour la billettique et faibles pour les capteurs, tous captés lors de la période de pointe du matin, apparaît pour les flux entrants. Il s'agit d'une partie des dynamiques internes au pôle de transport, invisible aux capteurs. Plus spécifiquement, les usagers, lorsqu'ils arrivent le matin, n'accèdent pas nécessairement directement à la sortie la plus proche, mais transitent dans le hall de transport, en (re)validant au passage.

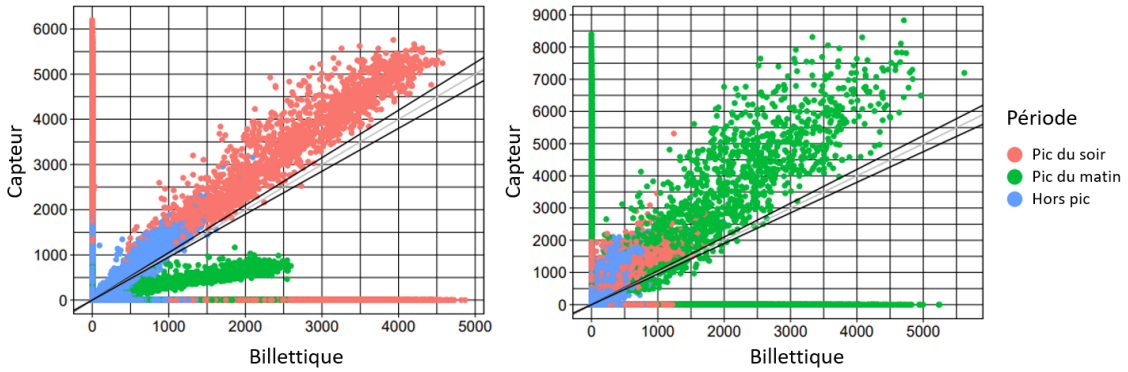


FIGURE 2.11 – Comparaisons entre les comptages issus de capteurs et de la billettique, colorés selon la période de la journée pour l'ensemble des jours ouvrés avant la pandémie de Covid19 à la station Grande Arche. Le graphe de gauche correspond aux flux entrants, celui de droite correspond aux flux sortants.

En conclusion de cette analyse, il y a un fort intérêt à considérer les deux sources de données dans l'analyse des fréquentations du pôle de La Défense. Les deux sources de données se complètent et s'enrichissent dans beaucoup de situations.

2.3 Les données contextuelles

Nous avons, dans la section 2.2, mentionné un ensemble de facteurs calendaires qui expliquent en grande partie les dynamiques de déplacement dans le pôle de transport. Il s'agit notamment des variations naturelles de fréquentations au cours de la journée (figures 2.2, 2.3 et 2.5), de la semaine (figure 2.8) ou de l'année (figure 2.7). Notons dans cette idée que certains jours particuliers ont un impact fort sur les fréquentations, on pense notamment aux jours fériés, de vacances scolaires ou de pont. Dans cette section, nous étudions une autre catégorie de facteurs, non calendaires, ayant un impact important sur les fréquentations du pôle de transport. Les variables étudiées dans cette section englobent des facteurs à l'impact plus ou moins long (quelques heures, plusieurs jours, plusieurs semaines). Nous distinguons également les événements programmés (que les usagers peuvent anticiper) des événements non programmés.

2.3.1 Les événements programmés

Nous illustrons l'effet des événements programmés en visualisant la sur-utilisation ou sous-utilisation de lieux de passages donnés lors de ces événements, par rapport à une situation moyenne. Cette valeur est visualisée à chaque tranche horaire h et lieu de comptage étudié p avec le score suivant :

$$s_{h,p} = \log(\frac{\hat{y}_{h,p}}{\widetilde{y}_{h,p}}) \tag{2.1}$$

avec $\widetilde{y}_{h,p} = M(\{y_{j,h,p}\}_{j\in\{1,\dots,J\}})$ la médiane ($M(.)$) des comptages y calculée sur l'ensemble des jours J et $\hat{y}_{h,p} = M(\{y_{j,h,p}\}_{j\in J_{event},h\in H_{event}})$ la médiane des comptages calculée pour les jours et tranches horaires avec présence de l'événement d'intérêt (respectivement J_{event} et H_{event}).

Les travaux de maintenance du RER A et du métro 1

Durant les vacances d'été, depuis 2015, la ligne de RER A est régulièrement coupée à des fins de maintenance, sur différents tronçons et à différentes périodes (semaine ou week-end, toute la journée ou le soir seulement). Afin d'illustrer l'effet de ces travaux, nous visualisons dans la figure 2.12 l'impact de trois types de coupures, en semaine ou le week-end, sur la sur-utilisation ou sous-utilisation de quatre accès aux lignes de transport : les sorties du RER A (E_O), les entrées vers le métro (M), et les correspondances entre les deux lignes (EM et ME). Les mobilités au sein du pôle de transport de La Défense sont impactées de différentes manières selon la localisation des travaux de maintenance. Dans le cas de travaux où la ligne de RER A est totalement coupée à La Défense (Auber $\leftrightarrow$ Nanterre U), il n'y a aucun flux de et vers le RER A ($s_{h,p}$ à -3 pour les flux E_O, ME et EM). Cet effet est visible toute la journée les jours de weekend, et uniquement le soir en

semaine, comme les travaux n'avaient pas lieu en journée. On constate un flux conséquent plus important vers le métro 1. Dans les deux autres cas, le RER A n'est pas coupé à La Défense, mais les trains n'arrivent pas depuis l'est dans le premier cas (Auber ↔ La Défense) ou n'arrivent que depuis une station située à proximité (Auber) dans le deuxième cas (Auber ↔ Vincennes). La diminution des flux sortants du RER A est donc moins forte ici. On note un transfert important de et vers la ligne de métro 1 (flux M, EM et ME). La ligne de métro 1 semble ainsi servir d'alternative au RER A lorsque celui-ci ne fonctionne pas.

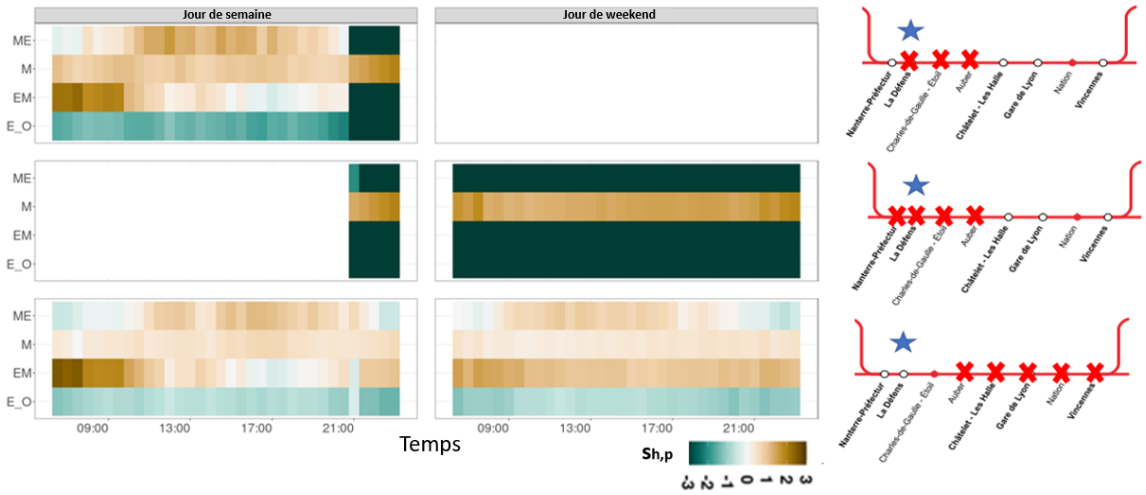


FIGURE 2.12 – Heatmap des scores de sur- ou sous-utilisation $s_{h,p}$ de quatre accès, liés à différentes périodes de travaux du RER A. Les quatre accès sont E_O (sorties du RER A), M (entrées du métro 1), EM (correspondances du RER A vers le métro 1) et ME (correspondances du métro 1 vers le RER A). A droite, sont représentées les stations de la ligne du RER A touchées par les différents travaux (une croix pour désigner une station fermée, l'étoile désignant La Défense). On sépare également les impacts du week-end de ceux de la semaine. Les zones blanches pour tous les accès indiquent des périodes sans travaux.

Les concerts et événements sportifs

Les concerts et événements sportifs organisés au stade (Arena La Défense) situé à proximité de La Défense sont des événements rares dans nos données (en raison de la pandémie de Covid19), mais qui n'en restent pas moins majeurs, car impliquant en l'espace de quelques heures un afflux massif de personnes transitant par le pôle de La Défense. Le cas des concerts est plus simple à traiter, car ces derniers se déroulent toujours le soir, aux mêmes heures (entre 20h et 23h). Dans la figure 2.13 nous représentons l'effet d'un concert

sur la fréquentation d'un lieu proche du stade : le capteur P2 (voir la carte de la figure 2.1). Comme on peut s'y attendre, l'événement implique un accroissement important du nombre de sortants (« P2 O ») avant le concert (usagers quittant la station pour aller vers le stade) et un pic très fort d'entrants après le concert (« P2 I »). Ce type d'événement est rare, mais d'une intensité très forte. Cet effet devrait donc être pris en compte dans nos travaux.

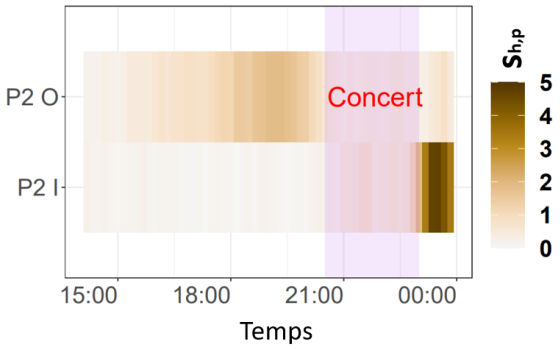


FIGURE 2.13 – Heatmap des scores de sur- ou sous-utilisation $s_{h,p}$ de l'accès P2 en réponse à un concert. La période de déroulement du concert est la zone rosée.

De la même manière, des événements sportifs organisés au stade attirent un public important. Il peut s'agir de grands matchs ou de compétitions ayant lieu plutôt le weekend. Comme ces événements ne sont pas à des heures fixes, nous regardons dans la figure 2.14 leur impact sur la fréquentation des quelques tranches de temps précédant le début et suivant la fin. L'impact de ces événements est moins fort que celui des concerts.

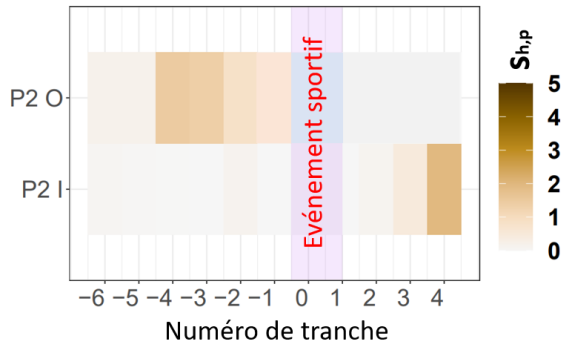


FIGURE 2.14 – Heatmap des scores de sur- ou sous-utilisation $s_{h,p}$ de l'accès P2 en réponse à un événement sportif. L'événement sportif se déroule dans la période couverte par la zone rosée.

Autres

D'autres événements, anticipables à court terme, peuvent influer sur les fréquentations. On pense ici aux phénomènes climatiques particuliers qui pourraient modifier certains comportements. Nous évaluerons l'impact de la pluie et des vagues de chaleur sur les fréquentations du pôle de transport. La pluie pourrait modifier certains comportements comme l'utilisation plus forte des transports couverts à la place de la marche ou du vélo. Les vagues de chaleur pourraient quant à elles influer sur les heures de sorties et d'arrivées. Dans ce type d'événements, on peut également mentionner les pics de pollution qui entraînaient certaines années la gratuité du réseau RATP et donc un impact sur les comptages.

2.3.2 Les événements non programmés

Des événements, non programmés, peuvent également se produire et impacter les flux de passagers. Dans notre cas, il s'agira essentiellement des perturbations du RER A. Cette ligne de transport peut, à elle seule, avoir un impact significatif sur les mobilités lorsque celle-ci est à l'arrêt. L'impact sur les mobilités peut varier fortement selon l'heure, le lieu et le jour de perturbation, ce qui en fait un phénomène difficile à appréhender. Afin de mieux comprendre comment ce type de phénomène impacte les flux de passagers, nous mettons en place une méthode heuristique de catégorisation des impacts possibles sur les fréquentations, que nous mettons ensuite en lien avec des scénarios de perturbations (ex. lieu, heure). La méthodologie, décrite ci-dessous, se décline en trois étapes : récupération des périodes potentiellement perturbées, création d'une base de données à partir des flux de passagers, et application d'une méthode de clustering sur cette base. Les étapes sont détaillées ci-dessous, pour des données agrégées par intervalle de trente minutes :

1. **Détection de fenêtres (2 tranches) potentiellement perturbées**
   Nous utilisons une base de l'historique de l'offre réelle des RER A, c'est-à-dire le nombre de passages de trains par tronçon du RER A à chaque tranche de temps. L'objectif est de sélectionner des tranches de temps pour lesquelles l'offre de transport sur toute la ligne était inférieure à un nombre que nous fixons à 15 trains par rapport à l'offre médiane. Nous créons ensuite des fenêtres correspondantes à la tranche courante détectée comme en situation potentiellement perturbée et la suivante. Nous obtenons ainsi une base de taille $(2H \times 1)$ contenant H fenêtres potentielles perturbations.
2. **Création d'une base des mobilités dans le pôle de La Défense sur des fenêtres potentiellement perturbées**
   Nous créons une base de données des flux voyageurs $(2H \times 3)$, filtrée sur les fenêtres avec une potentielle situation de perturbation, pour trois accès liés au RER A : E_O (sortants du RER A), ME (correspondants du métro 1 vers le RER A) et EM (correspondants du RER A vers le métro 1). De la même manière qu'avec les événements programmés, nous travaillons avec les scores $s_{h,p}$ (formule 2.1) à la différence qu'ils

ne sont ici pas calculés sur les comptages $y_{j,h,p}$ mais sur les rapports $y_{j,h,p}/y_{j,h-1,p}$. L'idée de travailler avec des rapports vient du fait que ces derniers sont moins sujets à être modifiés en profondeur au cours des années, contrairement aux comptages brutes. Il est donc plus aisé d'y révéler des impacts de perturbations, contrairement au cas des événements programmés où les comptages suffisaient.

3. **Classification des perturbations : catégorisation des fenêtres en 5 catégories** Nous appliquons un clustering ascendant hiérarchique (HAC) pour détecter cinq catégories dans les H fenêtres. La méthode est appliquée sur la base de dimension $(H \times 6)$ qui contient pour chaque fenêtre les scores $s_{h,p}$ pour chaque tranche horaire h et chaque type de flux p. Parmi les cinq catégories détectées, deux catégories (dont les scores sont très proches de zéro) représentent des potentielles perturbations n'ayant pas eu d'impact notable sur les flux du pôle de La Défense. Les trois autres catégories présentent en revanche des scores qui s'éloignent de zéro, et peuvent être associées à des impacts-types de perturbations.

Les résultats des trois catégories majeures détectées d'impact de perturbation sont présentés dans la figure 2.15. Les caractéristiques associées sont les suivantes :

— **Catégorie 1** : perturbations ayant eu lieu après le pic d'arrivées du matin, l'après-midi et le soir notamment, et ayant impacté la circulation des trains en direction de Paris ou sur la branche Ouest-Nord vers Cergy. Il s'agit de perturbations dont l'impact reste limité dans le temps (la deuxième tranche est non perturbée). La conséquence de ces perturbations est que les usagers empruntent davantage la ligne de Métro 1 (forte augmentation, puis diminution des entrants vers le métro 1 depuis le RER A).

— **Catégorie 2** : perturbations ayant eu lieu majoritairement lors de la pointe d'arrivée le matin. Il s'agit de perturbations ayant eu lieu dans une grande variété de stations et de branches. Ces perturbations de relativement courte durée expliquent une baisse momentanée des arrivées en RER A le matin, puis d'un effet de rebond par la suite (visible pour les trois types de flux).

— **Catégorie 3** : perturbations longues ayant eu lieu en toute période de la journée, et ayant coupé les deux sens de circulation du RER A. Dans ce cas, comme pour la Catégorie 1, on constate un report important du RER A vers le métro 1. On constate également une augmentation forte des flux du métro 1 vers le RER A avec un retard (tranche 2) : il s'agit des personnes se rendant au pôle de La Défense en métro, qui arrivent donc avec un retard visible, et qui sortent du métro en transitant par le RER A (les accès vers le RER étant alors ouverts pour faciliter l'évacuation des passagers).

2.4 Conclusion

Nous disposons dans cette thèse de deux sources de données principales : les comptages issus de capteurs et la billettique. Ces données possèdent des caractéristiques attendues

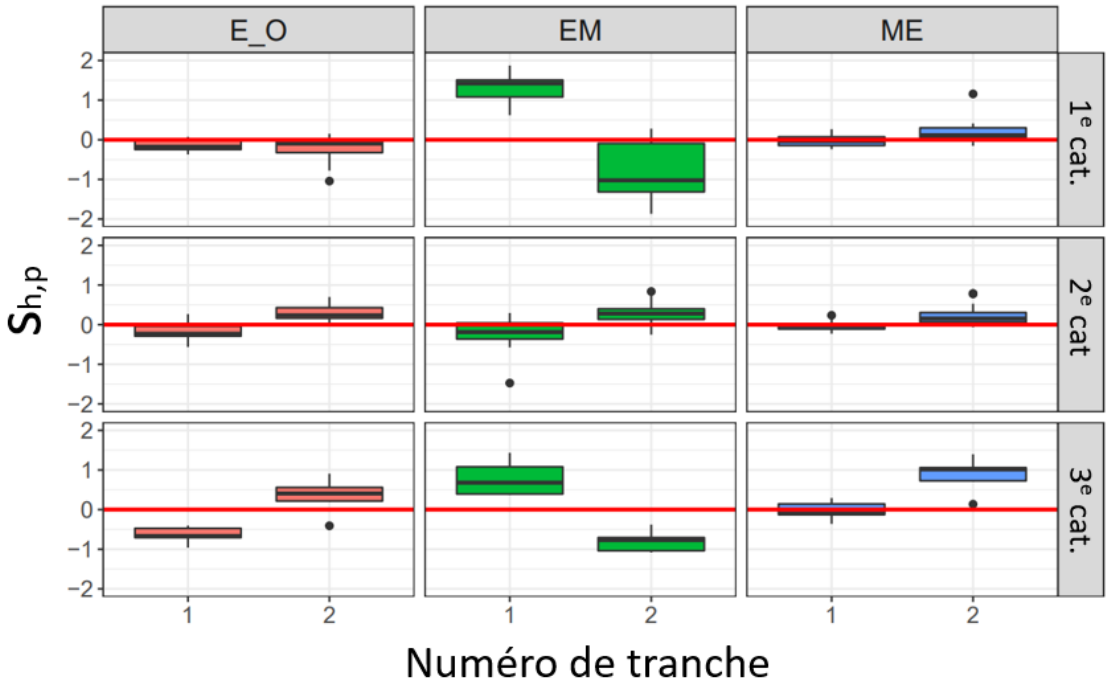


FIGURE 2.15 – Boxplots des scores $s_{h,p}$ pour les trois catégories détectées. Chaque catégorie reflète les effets d'un type de perturbation du RER A sur les flux sortants du RER A, et de correspondance avec le Métro 1.

dans un tel contexte : une structure saisonnière marquée à différentes échelles (semaine, année, ...), une sur-dispersion et une forte corrélation entre les différentes séries. Malgré ces rapprochements, nous avons constaté que les deux sources de données ne captent pas exactement la même information, car étant placées en différents endroits du pôle de transport. Les données s'enrichissent donc mutuellement, et nous permettent d'obtenir une cartographie presque complète du pôle de transport en termes de flux de voyageurs, tant dans les espaces de transport (billettique) que dans les échanges avec le quartier d'affaire environnant (capteurs de comptage). Les deux sources étant complémentaires, lorsqu'une source est défaillante, l'autre source continuera à collecter normalement les données, ce qui permettra de faire une différence entre des problèmes dans les données et de réels impacts d'événements. Les événements sont nombreux à impacter les flux voyageurs, et nous ne pouvons en faire qu'une liste non exhaustive. Nous avons séparé ces événements, entre ceux programmés (concerts, travaux, etc.), et ceux non programmés (perturbations), qui sont plus difficiles à appréhender. Notons qu'il est possible que certains phénomènes, non connus *a priori*, se révèlent au cours des différents travaux. Nous verrons cet aspect particulièrement dans le chapitre 4, où nos modèles auront la possibilité de détecter des périodes

aux mobilités différentes, sans encodage de variable *a priori*.

Connaître les attributs des données de mobilité et des événements impactants est une étape indispensable en amont de tout travail de modélisation. Cette connaissance peut en effet aiguiller les choix et encodages à mettre en place pour la modélisation des données. Par exemple, le fait que les données de comptages soient surdispersées dirige la modélisation vers certaines distributions plutôt que d'autres. Les informations apprises ici seront mises à profit dans les différents travaux de cette thèse.

Chapitre 3

Isoler et quantifier l'impact de facteurs long-terme et journaliers sur les mobilités

3.1 Introduction

L'analyse des mobilités au sein d'un réseau de transport en commun est souvent basée sur les données de billettique, se présentant sous la forme de séries temporelles décrivant le volume des voyageurs au cours du temps. Le profil temporel de ces séries est le résultat d'une multitude de facteurs explicatifs. Une grande partie de ces facteurs est liée aux habitudes de déplacement vers le lieu de travail ou d'études scolaires. Le réseau de transports en commun est ainsi davantage sollicité les jours ouvrés que les jours de week-end en raison de ces comportements. De même, l'affluence fluctue au cours de l'année selon les périodes de travail et de vacances scolaires. D'autres facteurs, plus ponctuels, peuvent également impacter les mobilités, qu'ils soient prévus (travaux sur une ligne) ou non (pandémie). Du fait de la multiplicité des facteurs pouvant impacter les flux voyageurs, la structure des séries temporelles de mobilité est souvent complexe, et les analyses peuvent se révéler difficile. L'une de ces difficultés réside dans l'étude de l'impact de ces facteurs pris isolément les uns des autres. Ce type d'analyse est particulièrement utile lorsque l'on souhaite avoir une idée du comportement collectif des voyageurs résultant d'un événement en particulier. Par exemple, supposons qu'une période de travaux sur une ligne de transport se déroule au moment des vacances d'été. Dans ce cas, l'opérateur de transport ne pourra pas facilement quantifier l'effet des travaux sur les fréquentations, car il devra également prendre en compte l'effet saisonnier des vacances d'été. Le travail présenté ici porte sur la décomposition des séries temporelles de mobilité au travers de modèles stochastiques de décomposition. Ces méthodes bénéficient d'une bonne capacité à modéliser les séries temporelles, tout en étant facilement interprétables. En effet, une fois décomposées avec un modèle dédié, il est pos-

sible de détecter une tendance long-terme, des profils saisonniers répétitifs (jours, semaines, années), des phénomènes calendaires (jours fériés, vacances) ou l'influence d'autres facteurs exogènes. Ces structures sous-jacentes sont appelées « composantes ».

Notre cas d'étude portera ici sur l'analyse comparée de l'utilisation des lignes de RER A et de Métro 1 à la station « La Défense Grande Arche », au travers d'un modèle de décomposition. Ces deux lignes sont connues comme étant en compétition car desservant le même axe Est-Ouest, ainsi que plusieurs gares en commun. Elles sont également complémentaires : le RER A dessert des banlieues plus éloignées que le Métro 1, mais ce dernier dispose d'un réseau de stations plus dense dans Paris. Un travail de décomposition appliqué à la demande d'accès vers ces deux lignes à La Défense est un cas d'étude intéressant. Il nous permet en effet de comparer l'effet de facteurs impactants, comme les travaux de maintenance ou les vacances, sur l'utilisation de ces deux lignes de transport. Des événements ayant fortement impacté les habitudes de déplacement pourront être étudiés en détail : la période de grève contre la réforme des retraites en décembre 2019, et la période du premier confinement mis en place suite à la pandémie de Covid19. Nous utiliserons pour ce travail les données de billettique qui fournissent l'information des volumes entrants vers l'une ou l'autre des lignes de transport à la station « La Défense Grande Arche ». Décomposer les séries temporelles de fréquentation des deux lignes de transport en plusieurs composantes sous-jacentes nous permettra d'étudier leur structure et de répondre à certaines questions :

— Comment les variations des séries se traduisent-elles dans les composantes ?

— Quel est l'impact d'événements particuliers sur la décision des usagers à utiliser l'une des lignes de transport ?

Une approche possible pour décomposer les séries temporelles est de révéler les composantes cachées d'une architecture espace-état. Au sein de cette famille, les modèles structurels sont couramment utilisés en raison de leur structure intuitive et flexible. Cependant, afin de maximiser les capacités d'interprétation et faciliter leur implémentation, ces modèles requièrent des données collectées sur des périodes de temps longues, ainsi que l'incorporation de choix de modélisation faits *a priori*. Pour cela, nous utiliserons des données de flux voyageurs collectées sur une période de 9 années (entre le 1er janvier 2011 et le 31 juillet 2020), notamment afin de bien estimer certaines composantes long terme telles que la tendance ou la saisonnalité annuelle. Ces données seront agrégées à la journée pour cette étude. Nous adresserons les questions critiques liées à l'utilisation et la calibration de ces modèles dans le cadre méthodologique. Pour cela, nous étudierons l'impact de différentes configurations, en nous basant sur les capacités de prédiction de chacune d'elles.

3.2 État de l'art : modèles de décomposition

Les modèles de décomposition subdivisent les séries chronologiques en de multiples composantes sous-jacentes, chacune représentant un aspect de la série chronologique origi-

nale. Les composantes peuvent ensuite être utilisées pour reconstruire la série originale par addition ou multiplication. Chaque composante est caractérisée par un motif, comme une tendance à long terme, une saisonnalité hebdomadaire et/ou annuelle, et un bruit. L'analyse des caractéristiques des composantes résultant de la décomposition permet ainsi une interprétation directe du modèle. Dans le travail de [GTS02] sur la décomposition des séries temporelles de températures mensuelles à l'aide d'un modèle additif généralisé (GAM), les auteurs se posent des questions quant aux évolutions intrinsèques dans les séries :

— Y a-t-il une tendance significative à la hausse ou à la baisse dans les séries ?

— La saisonnalité change-t-elle au fil du temps ?

— Y a-t-il des valeurs extrêmes dans les observations qui ne peuvent être expliquées par les différentes composantes ?

Un type de modèle de décomposition bien connu est la décomposition saisonnière et de tendance à l'aide de LOESS (LOcally Estimated Scatterplot Smoothing). Les auteurs de [ZG17] ont utilisé ce type de modèle pour décomposer les séries temporelles du nombre de courses en taxi dans différents endroits de New York, afin d'isoler la tendance, la saisonnalité et les résidus. Les modèles espace-état sont une autre famille de modèles que nous utiliserons dans la suite de ce travail, et que nous explicitons dans les sections suivantes.

3.2.1 Les modèles espace-état

C'est dans le traitement du signal que sont apparus les modèles espace-état. Ces derniers permettent de mettre en lumière des états internes non observés, qui déterminent les signaux au cours du temps. Comme précisé par [LP+03], ces modèles associent des composantes cachées latentes (ou « états ») avec les séries temporelles observées, à travers un ensemble d'équations qui peuvent être déterministes ou stochastiques :

— Les équations d'observation décrivent de quelle manière les séries observées sont générées par les composantes cachées et les résidus.

— Les équations d'état décrivent l'évolution des composantes cachées, à partir de leur historique et des « innovations » (ou erreurs).

Les modèles espace-état présentent quelques avantages ; ils restent notamment valides en présence de séries non-stationnaires, contrairement à un processus autorégressifs de moyennes mobiles (ARMA, [SSS00, p. 77-90]). De plus, les coefficients du modèle peuvent évoluer au cours de la période d'estimation. Lorsque les opérateurs impliqués dans le système du modèle espace-état sont linéaires et que les erreurs sont distribuées normalement et indépendamment, on parle de modèles linéaires dynamiques (DLM). Considérons $\mathbf{y} = \{y_{t,p'}\}_{t\in 1,\dots,T}$, une série de comptages effectués à chaque tranche de temps t en un lieu de comptage p'. Si l'on simplifie $y_{t,p'}$ par y_t, la formulation des équations d'un modèle linéaire dynamique appliqué à ces données est la suivante :

$$y_t = \mathbf{H}_t \boldsymbol{\varrho}_t + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, r_t) \tag{3.1}$$

$$\boldsymbol{\varrho}_t = \mathbf{M}_t \boldsymbol{\varrho}_{t-1} + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}(0, \mathbf{Q}_t) \tag{3.2}$$

Le terme $\boldsymbol{\varrho}_t$ est le vecteur des composantes cachées de taille m. L'état caché évolue selon la matrice de transition $\mathbf{M}_t$ $(m \times m)$ ainsi que l'innovation $\mathbf{e}_t$. Les composantes cachées sont linéairement combinées avec la matrice de mesure $\mathbf{H}_t$ $(m \times 1)$, et le bruit ϵ_t. Les deux sources d'erreur, ϵ_t et $\mathbf{e}_t$, sont des bruits gaussiens de variances-covariances r_t (scalaire) et $\mathbf{Q}_t$ $(m \times m)$. Les modèles espace-état linéaires possèdent quelques caractéristiques : les équations d'état et d'observation (respectivement $\mathbf{M}_t$ et $\mathbf{H}_t$) sont linéaires, mais peuvent évoluer avec le temps. Dans la version élémentaire des modèles espace-état, les bruits d'observation et d'état sont des bruits blancs. Ces modèles supposent que les états initiaux $\boldsymbol{\varrho}_0$ suivent une loi normale $\boldsymbol{\varrho}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. De plus il y a indépendance entre les bruits d'observation et d'état. On retrouve les DLM sous différentes appellations dans la littérature. Ils sont ainsi appelés modèles à séries temporelles structurelles dans [Har90] et modèles linéaires dynamiques dans [PPC09]. On les trouve dans plusieurs domaines d'application tels que l'économie [KO11], le tourisme [Che+19], la météorologie [NSR19] ou l'énergie [MG19]. Avec de tels modèles, il est possible d'exprimer des composantes déterministes et/ou stochastiques, de les interpréter et de faire des prédictions de manière indépendante.

D'autres modèles dérivent des modèles linéaires dynamiques. Dans les modèles de lissage exponentiel [Hyn+02], aussi appelés *Error, Trend, Seasonal* (ETS), les bruits d'état $\mathbf{e}_t$ et d'observation ϵ_t sont corrélés (ils ont la même valeur à une constante multiplicative près). Cette formulation permet de simplifier l'estimation de la vraisemblance du modèle. Les méthodes de Holt-Winters [Hol04] avec une ou plusieurs saisonnalités sont une extension du modèle ETS. Le modèle *Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components* (TBATS) est lui aussi une extension du modèle ETS [DHS11]. Ce modèle introduit une représentation trigonométrique des composantes saisonnières en se basant sur les séries de Fourier. Il y a dans ce modèle moins de coefficients à estimer que pour Holt-Winters ; de plus, comme il s'agit de fonctions trigonométriques, ce modèle peut prendre en compte des fréquences saisonnières non entières.

La formulation sous forme de modèle linéaire dynamique est nécessaire pour permettre l'inférence des composantes cachées avec l'algorithme du filtre de Kalman [Kal+60]. Cette méthode, dont les formules peuvent être trouvées dans le travail de [SS82], estime le vecteur d'état $\boldsymbol{\varrho}_t$ connaissant l'ensemble des paramètres du modèle. Le principe du filtre de Kalman, ainsi que l'estimation des paramètres, sont détaillés dans les sections suivantes, pour le cas de modèles DLM.

3.2.2 Le filtre de Kalman

Trois étapes d'estimation sont utilisées dans le filtre de Kalman : la prédiction, le filtrage et le lissage. En notant $\boldsymbol{\varrho}_t^s$ l'espérance de $\boldsymbol{\varrho}_t$ conditionnellement aux données observées

jusqu'au temps s, et $\mathbf{P}_t^s$ la matrice de covariance de $\boldsymbol{\varrho}_t$ conditionnellement aux données observées jusqu'au temps s, les étapes de prédiction, de filtrage et de lissage sont définies comme suit :

— **La prédiction** consiste à calculer l'état $\boldsymbol{\varrho}_t^{t-1}$ et la covariance $\mathbf{P}_t^{t-1}$ au temps t sachant toutes les données observées jusqu'au temps $t-1$, en se basant sur les équations du modèle à espace d'état. On part des valeurs connues $\boldsymbol{\varrho}_0^0 = \boldsymbol{\mu}_0$ et $\mathbf{P}_0^0 = \boldsymbol{\Sigma}_0$. Les équations sont les suivantes :

$$\boldsymbol{\varrho}_t^{t-1} = \mathbf{M}_{t-1}\boldsymbol{\varrho}_{t-1}^{t-1}, \tag{3.3}$$

$$\mathbf{P}_t^{t-1} = \mathbf{M}_{t-1}\mathbf{P}_{t-1}^{t-1}\mathbf{M'}_{t-1} + \mathbf{Q}_{t-1} \tag{3.4}$$

— **Le filtrage** se fait lorsque l'on dispose de l'observation à l'instant t (y_t). Il s'agit de corriger les prédictions de l'état et de sa covariance précédemment trouvées, avec cette nouvelle information. Pour cela, est définie une nouvelle quantité : le gain de Kalman $\mathbf{k}_t$ (compris entre 0 et 1), qui représente le poids relatif donné aux observations et à l'estimation de l'état au temps t. Si le gain est proche de 1, l'algorithme accorde plus de poids aux observations les plus récentes. Si le gain est proche de 0, le filtre donnera plus de poids aux prédictions du modèle. L'état filtré est noté $\boldsymbol{\varrho}_t^t$ et la covariance $\mathbf{P}_t^t$. Les équations sont les suivantes :

$$\boldsymbol{\varrho}_t^t = \boldsymbol{\varrho}_t^{t-1} + \mathbf{k}_t(y_t - \mathbf{H}_t\boldsymbol{\varrho}_t^{t-1}), \tag{3.5}$$

$$\mathbf{P}_t^t = [\mathbf{I} - \mathbf{k}_t\mathbf{H}_t]\mathbf{P}_t^{t-1}, \tag{3.6}$$

avec

$$\mathbf{k}_t = \mathbf{P}_t^{t-1}\mathbf{H'}_t[\mathbf{H}_t\mathbf{P}_t^{t-1}\mathbf{H'}_t + r_t]^{-1} \tag{3.7}$$

Le filtre de Kalman permet ainsi de calculer les états prédits $\boldsymbol{\varrho}_t^{t-1}$ et filtrés $\boldsymbol{\varrho}_t^t$ pour $t = 1, ..., T$. Le filtrage permet de calculer deux quantités nécessaires à l'estimation des paramètres du modèle : les erreurs de prédiction ω_t et les variances correspondantes σ_t (ou $\mathbb{V}(\omega_t)$) :

$$\omega_t = y_t - \mathbf{H}_t\boldsymbol{\varrho}_t^{t-1}, \tag{3.8}$$

$$\sigma_t = \mathbb{V}[\mathbf{H}_t(\boldsymbol{\varrho}_t - \boldsymbol{\varrho}_t^{t-1}) + \epsilon_t] = \mathbf{H}_t\mathbf{P}_t^{t-1}\mathbf{H'}_t + r_t \tag{3.9}$$

— **Le lissage** permet de calculer à l'instant t l'espérance de $\boldsymbol{\varrho}_t$ sachant toutes les données $y_1, ..., y_T$. L'estimateur de Kalman lissé $\boldsymbol{\varrho}_t^T = \mathbb{E}(\boldsymbol{\varrho}_t \mid y_1, ..., y_T)$ peut être calculé récursivement, en utilisant les équations définies pour la prédiction et le filtrage,

ainsi qu'un ensemble de récursions $t = T, T-1, ..., 1$. Avec les conditions initiales $\boldsymbol{\varrho}_T^T$ et $\mathbf{P}_T^T$ obtenues *via* l'étape du filtrage, le lissage se calcule comme suit :

$$\boldsymbol{\varrho}_{t-1}^T = \boldsymbol{\varrho}_{t-1}^{t-1} + \mathbf{J}_{t-1}(\boldsymbol{\varrho}_t^T - \boldsymbol{\varrho}_t^{t-1}), \tag{3.10}$$

$$\mathbf{P}_{t-1}^T = \mathbf{P}_{t-1}^{t-1} + \mathbf{J}_{t-1}(\mathbf{P}_t^T - \mathbf{P}_t^{t-1})\mathbf{J'}_{t-1}, \tag{3.11}$$

avec

$$\mathbf{J}_{t-1} = \mathbf{P}_{t-1}^{t-1}\mathbf{M'}_t[\mathbf{P}_t^{t-1}]^{-1} \tag{3.12}$$

Le filtre de Kalman intervient dans l'estimation par maximum de vraisemblance des paramètres du modèle. Nous décrivons cette dernière dans la section suivante.

3.2.3 Estimation des paramètres

Le vecteur de paramètres inconnus $\boldsymbol{\vartheta}$ contient les moyenne et covariance initiales (respectivement $\boldsymbol{\mu}_0$ et $\boldsymbol{\Sigma}_0$), les paramètres associés aux distributions des erreurs ϵ_t et $\mathbf{e}_t$, ainsi que les matrices $\mathbf{M}_t$ et $\mathbf{H}_t$ dans certains cas. La log vraisemblance du modèle s'écrit comme suit :

$$-L_Y(\boldsymbol{\vartheta}) = \frac{1}{2}\sum_{t=1}^{T}\log|\sigma_t(\boldsymbol{\vartheta})| + \frac{1}{2}\sum_{t=1}^{T}\sigma_t(\boldsymbol{\vartheta})^{-1}\omega_t(\boldsymbol{\vartheta})^2 \tag{3.13}$$

avec $\omega_t(\boldsymbol{\vartheta})$ et $\sigma_t(\boldsymbol{\vartheta})$ les erreurs de prédiction et leurs variances, fonctions des paramètres $\boldsymbol{\vartheta}$, et définies dans les équations 3.8 et 3.9.

Plusieurs méthodes peuvent être utilisées pour estimer les paramètres à partir de la fonction 3.13. Une possibilité est d'utiliser l'algorithme espérance-maximisation (EM), développé à l'origine par [DLR77], et ensuite exploité par [SS82] pour l'estimation des paramètres de modèles linéaires espace-état. Dans notre cas, l'algorithme EM a pour objectif de maximiser la log-vraisemblance en lien avec les paramètres $\boldsymbol{\vartheta}$. L'algorithme alterne deux étapes jusqu'à convergence : une étape d'estimation des composantes lissées $\boldsymbol{\varrho}_t^T$, connaissant les paramètres, et une étape de mise à jour des paramètres, connaissant les composantes lissées.

Une autre méthode d'estimation des paramètres utilisée dans le cadre des modèles espace-état est l'algorithme de Newton-Raphson, dont les étapes sont les suivantes :

1. On sélectionne des valeurs initiales $\boldsymbol{\vartheta}^{(0)}$ pour les paramètres.
2. A partir de $\boldsymbol{\vartheta}^{(0)}$, on utilise le filtre de Kalman pour obtenir un ensemble d'erreurs et covariances de prédictions $\omega_t^{(0)}$ et $\sigma_t^{(0)}$ pour $t = 1, ..., T$.
3. On calcule un nouvel ensemble de paramètres $\boldsymbol{\vartheta}^{(1)}$ à travers une itération de la méthode de Newton-Raphson. Le gradient et la hessienne, permettant cette mise à jour, sont calculés à partir de la fonction critère $-L_Y(\boldsymbol{\vartheta})$.

4. A chaque nouvelle itération j, on répète l'étape 2 en utilisant $\boldsymbol{\vartheta}^{(j)}$ (mis à jour) pour obtenir de nouveaux ensembles $\omega_t^{(j)}$ et $\sigma_t^{(j)}$. Ensuite, on répète l'étape 3 pour mettre à jour les paramètres et obtenir $\boldsymbol{\vartheta}^{(j+1)}$. La procédure est arrêtée lorsque la différence entre $-L_Y(\boldsymbol{\vartheta}^{(j+1)})$ et $-L_Y(\boldsymbol{\vartheta}^{(j)})$ est inférieure à un seuil donné.

Une variante de cette procédure passe par la méthode de quasi-Newton implémentée dans l'algorithme Broyden–Fletcher–Goldfarb–Shanno (BFGS), qui est basée sur une projection du gradient et sur l'approximation de la matrice hessienne de log-vraisemblance par une matrice à mémoire limitée [Byr+95]. Cette méthode est populaire dans le cadre des modèles espace-état, car la convergence est généralement plus rapide qu'avec l'algorithme EM, et les restrictions sur les espaces de paramètres y sont prises en compte [PPC09].

Nous allons maintenant nous intéresser à quelques applications courantes des modèles DLM. Trois types d'applications sont régulièrement rencontrées : la décomposition, la prédiction et la détection d'anomalies.

3.2.4 Applications des modèles linéaires dynamiques

Les modèles linéaires dynamiques sont utilisés pour la décomposition dans de nombreux domaines d'application, en raison de leur flexibilité. Ils peuvent servir à isoler un effet jugé intéressant, de ceux d'autres facteurs non pertinents. Les auteurs de [HSA18] ont quantifié l'effet du changement de comportement des clients sur la consommation d'électricité, induit par la mise en place d'une politique d'économie d'énergie au Japon suite à la catastrophe de Fukushima en 2011. Des travaux de prédiction à court terme ont été réalisés par [Dor+08], dans le domaine de la consommation d'électricité. Des prédictions à long terme (quelques années) sont également proposées par [MK21], [RPO20] et [Bia+19], tandis que [Che+19] a proposé un modèle multivarié permettant de prendre en compte la saisonnalité pour prédire la demande touristique à long terme. La détection d'anomalies est une autre application des modèles linéaires dynamiques. Des méthodes intégrant des variables indicatrices dans les modèles ont été développées pour détecter des valeurs atypiques ou des changements de comportement. Il s'agit alors de combiner le potentiel des modèles linéaires dynamiques avec des indicateurs de saturation, pour mettre en évidence les changements de comportement dans les séries. Cette approche a été développée par [MP16] pour détecter des changements de comportement dans la production industrielle de cinq pays européens avec la crise financière de 2008.

Dans le domaine de la mobilité, les modèles linéaires dynamiques ont aussi été utilisés pour des objectifs de prédiction. Dans [AER20], ces modèles servent à prédire le taux d'utilisation d'un système de vélo en libre-service. On peut également citer le travail réalisé par [Doo+14], où ces modèles ont été utilisés pour la prédiction à court terme des flux de vélos. Le filtre de Kalman a permis d'effectuer des prédictions successives sans passer par l'étape de filtrage. Dans leurs travaux sur la prédiction des séries chronologiques de comptages de voitures à certaines intersections de Dublin, [GBO09] ont présenté la décomposition d'une série chronologique en trois composantes : la tendance, la saisonnalité et les résidus. Un

modèle multivarié a été utilisé pour rendre compte d'un ensemble de séries chronologiques de comptages de voitures. Les auteurs de [Bia+19] ont utilisé le même type de modèle pour prévoir le volume de trafic mensuel des douze prochains mois sur un corridor clé du New-Jersey.

Dans la section suivante, nous détaillons la structure d'un modèle linéaire dynamique pour la décomposition de séries temporelles de flux passagers entrants vers deux lignes de transport au sein du pôle de La Défense.

3.3 Modèle de décomposition proposé

Pour nos travaux, nous nous concentrons sur les séries journalières décrivant l'évolution du nombre de passagers qui empruntent une ligne de transport (RER A ou métro 1). Une série temporelle s'écrit ainsi $(y_1, \ldots, y_J)$, où $y_j = \sum_h \{y_{j,h,p'}\}$ est le nombre de personnes entrant sur la ligne de RER A ($p' = E_I$) ou de métro 1 ($p' = M$) au jour j. Un modèle linéaire dynamique additif a été choisi pour représenter la transformation log de la série (Equation 3.14), ce qui revient à modéliser la série brute sous forme multiplicative (Equation 3.15). La transformation logarithmique a l'avantage de forcer les prédictions du nombre de passagers à rester positives, tout en stabilisant la variance des données (voir annexe B.1). Le modèle est défini comme suit :

$$\log(y_j) = l_j + s_j + f_j + \sum_{s=1}^{d} \beta_j^{(s)} \psi_j^{(s)} + \nu_j, \tag{3.14}$$

$$y_j = e^{l_j} \times e^{s_j} \times e^{f_j} \times \prod_{s=1}^{d} e^{\beta_j^{(s)} \psi_j^{(s)}} \times e^{\nu_j}, \tag{3.15}$$

où l_j est la tendance décrivant l'évolution à long terme de la série, s_j est la composante saisonnière hebdomadaire, f_j est la composante saisonnière annuelle, et ν_j est la composante résiduelle, qui est supposée être distribuée suivant une densité normale de moyenne nulle et de variance σ_ν^2. Le modèle décrit par l'équation (3.14) prend également en compte la dépendance des données y_j aux d variables explicatives journalières notées $(\psi_j^{(1)}, \ldots, \psi_j^{(d)})$. Il s'agit de variables indicatrices (0 ou 1) qui soulignent la présence ou non de divers événements (une grève, un jour de pollution, etc.). Les coefficients de régression associés à ces facteurs sont notés $(\beta_j^{(1)}, \ldots, \beta_j^{(d)})$. Les modèles stochastiques décrivant chacune des composantes du modèle sont expliqués ci-dessous.

— La tendance l_j suit un modèle stochastique de niveau local défini par :

$$l_j = l_{j-1} + b + \omega_j^l \tag{3.16}$$

où b est un paramètre de déviation (« drift parameter ») et ω_j^l, un bruit blanc gaussien.

— La composante saisonnière hebdomadaire s_j est modélisée sous la forme stochastique suivante :

$$s_j = -\sum_{i=1}^{6} s_{j-i} + \omega_j^s, \tag{3.17}$$

où ω_j^s est un bruit blanc gaussien. Cette représentation permet aux modèles saisonniers hebdomadaires d'évoluer, tout en garantissant que la somme de 7 termes consécutifs de s_j ait une espérance nulle.

— La composante saisonnière annuelle est modélisée sous la forme trigonométrique suivante, qui a l'avantage de réduire le nombre de ses paramètres :

$$f_j = \sum_{u=1}^{k} f_{u,j} \tag{3.18}$$

$$f_{u,j} = f_{u,j-1} \cos \lambda_u + f^*_{u,j-1} \sin \lambda_u + \omega_j^{f_u} \tag{3.19}$$

$$f^*_{u,j} = -f_{u,j-1} \sin \lambda_u + f^*_{u,j-1} \cos \lambda_u + \omega_j^{f_u^*}, \tag{3.20}$$

où $\omega_j^{f_u}$ et $\omega_j^{f_u *}$ sont des bruits blancs gaussiens de même variance. Il s'agit d'une combinaison de k cycles stochastiques, dont la représentation trigonométrique est définie à partir des fréquences $\lambda_u = 2\pi u/365$ pour $u \in \{1, ..., k\}$.

— Les coefficients de régression associés aux variables exogènes $\psi_j^{(s)}$ sont supposés évoluer suivant une marche aléatoire gaussienne définie par :

$$\beta_j^{(s)} = \beta_{j-1}^{(s)} + \omega_j^{\beta^{(s)}}, \tag{3.21}$$

où $\omega_j^{\beta^{(s)}}$ est un bruit blanc gaussien. Les flux journaliers de personnes sont influencés par de multiples facteurs externes. Dans le cadre de ce travail, nous intégrons les facteurs connus pour avoir un impact sur la fréquentation journalière. La plupart de ces facteurs ont été répartis entre les jours « ouvrés » et les jours « non ouvrés » pour tenir compte de l'effet calendaire. Par jours non ouvrés, nous entendons les jours de week-end ainsi que les jours fériés. Ces jours-là, l'activité de travail est très faible, contrairement aux jours ouvrés (non fériés). Nous prenons également en compte les jours de travaux du RER A et du Métro 1, les jours de grève, et les jours de premier confinement et post-confinement dus à la pandémie de Covid19. Enfin, nous prenons en compte les jours où le réseau de transport urbain était gratuit, pendant les pics de pollution ou les journées sans voiture ; les portiques de validation étant ouverts, très peu de données de validation ont été rapportées pour ces jours spécifiques.

Notons que pour toutes les composantes, tous les termes résiduels sont des bruits blancs gaussiens dont les variances sont résumées dans le tableau 3.1. Les valeurs initiales des composantes suivent également des distributions gaussiennes dont les paramètres sont spécifiés

dans le tableau 3.1. Le vecteur des paramètres inconnus $\boldsymbol{\vartheta}$ contient l'ensemble des paramètres décrits dans le tableau 3.1 : $\boldsymbol{\vartheta} = (\sigma_l^2, \sigma_s^2, ...)$. Le vecteur d'état $\boldsymbol{\varrho}_j$ est le suivant :

$$\boldsymbol{\varrho}_j = (l_j, s_j, s_{j-1}, \ldots, s_{j-6}, f_{1,j}, f^*_{1,j}, \ldots, f_{k,j}, f^*_{k,j}, \beta_j^{(1)}, \ldots, \beta_j^{(d)})^T$$

L'estimation du vecteur d'état $\boldsymbol{\varrho}_j$, connaissant l'ensemble des paramètres $\boldsymbol{\vartheta}$, est basée sur le filtre de Kalman. L'estimation de $\boldsymbol{\vartheta}$ se fait avec la méthode de quasi-Newton.

Composante	**Variance**	**Param. composante initiale**	
		Espérance	**Variance**
Tendance (Equation 3.16)	σ_l^2	m_{l_0}	C_{l_0}
Saisonnalité hebdomadaire (Equation 3.17)	σ_s^2	m_{s_0}	C_{s_0}
Saisonnalité annuelle (Equations 3.19 and 3.20)	$\sigma^2_{f_u} = \sigma^2_{f_u^*}$	$m_{f_{u,0}}$ $m_{f^*_{u,0}}$	$C_{f_{u,0}}$ $C_{f^*_{u,0}}$
Coefficients de régression (Equation 3.21)	$\sigma^2_{\beta^{(j)}}$	$m_{\beta_0^{(j)}}$	$C_{\beta_0^{(j)}}$
Résidus (Equation 3.14)	σ_ν^2		

TABLE 3.1 – Tableau récapitulatif des paramètres des composantes

3.4 Résultats et discussions

Dans le cadre de la calibration du modèle, nous cherchons à réduire sa complexité tout en maintenant une qualité satisfaisante de représentation des données. Nous avons d'abord choisi les composantes à prendre en compte, ainsi que leur mode d'expression, afin d'obtenir une décomposition propre. Certains paramètres du modèle ne pouvant être déduits *a priori*, nous avons utilisé des critères de vraisemblance (critère d'information d'Akaike, AIC) et de mesure d'erreurs de prédiction (racine de l'erreur quadratique moyenne, RMSE) pour les choisir. Les critères AIC et RMSE sont présentés en annexe A. Une fois la configuration fixée, les paramètres du modèle résultant ont été estimés par la méthode du maximum de vraisemblance, en utilisant la méthode quasi-Newton mise en œuvre par l'algorithme BFGS. Nous nous sommes appuyés sur le package R dlm [Pet10]. La calibration détaillée du modèle est présentée dans la section suivante.

3.4.1 Calibration des composantes

Le modèle devrait privilégier l'aspect descriptif plutôt que l'aspect prédictif, pour donner le plus de sens possible à la décomposition ; une préférence pour les composants déterministes matérialisera ce point. Pour déterminer un modèle adapté à nos données, nous disposons de certaines connaissances *a priori* sur la configuration qu'il devrait prendre. Il s'agit tout d'abord du choix des composantes à inclure dans le modèle. Dans la section 2.2.3, nous avons déterminé que les tendances à long terme, et les saisonnalités hebdomadaires et annuelles, semblaient expliquer une grande partie des variations de la fréquentation des lignes de transport. Plus particulièrement, nous recherchons un modèle avec une évolution lente des tendances, des composantes saisonnières déterministes et stables dans lequel il y a peu de variabilité. Certains effets exogènes doivent également être ajoutés, avec la possibilité de varier en intensité pour quantifier leur impact à différentes périodes. Ces choix sont formalisés dans les composantes comme suit :

— **Tendance**
Pour s'assurer que la tendance (équation 3.16) ne reflète que les changements de fréquentation sur le long terme (9 ans), nous avons limité la variance de la tendance à une limite supérieure ($< 3 \times 10^{-7}$).

— **Saisonnalité hebdomadaire**
Pour mieux visualiser l'effet moyen des différents jours de la semaine sur la fréquentation des stations, sans modification au cours du temps, nous avons choisi une composante déterministe qui ne contient pas de modifications stochastiques ($\omega_j^s = 0$). Les modifications de saisonnalité hebdomadaire devraient être captées par les composantes associées aux facteurs exogènes non calendaires.

— **Saisonnalité annuelle**
De même que pour la composante saisonnière hebdomadaire, nous avons opté pour une composante annuelle déterministe. Les variances $\sigma^2_{\omega^{f_u}}$ et $\sigma^2_{\omega^{f_u^*}}$ des termes d'erreur ont été fixées à zéro, afin que les deux erreurs $\omega_j^{f_u}$ et $\omega_j^{f_u^*}$ soient nulles (voir équations 3.18 à 3.20). L'objectif était ici de mettre en évidence les variations de fréquentation, en moyenne sur une année.

— **Coefficients de régression**
L'effet des facteurs exogènes doit pouvoir varier dans le temps pour tenir compte de leur évolution temporelle. Ainsi, aucune contrainte n'a été imposée sur les paramètres $\sigma^2_{\beta^{(s)}}$ de cette composante.

Le nombre d'harmoniques dans la composante saisonnière annuelle n'est pas un paramètre qui peut être calibré *a priori*, donc ce choix sera fait sur la base des métriques AIC et RMSE. Compte tenu des contraintes décrites ci-dessus, plusieurs configurations de modèles ont été comparées pour les deux lignes de transport : une absence de composante saisonnière annuelle ($k = 0$), et un nombre d'harmoniques k prenant les valeurs

Ligne de transport	Harmoniques	AIC	
		Avec covariables	Sans covariables
RER A	0	836	-3,880
	2	822	-4,108
	4	800	-4,280
	6	766	-4,430
	8	762	-4,504
	10	754	-4,514
	12	750	-4,540
Metro 1	0	942	-4,262
	2	930	-4,634
	4	916	-4,846
	6	900	-4,994
	8	874	-5,122
	10	872	-5,158
	12	872	-5,168

TABLE 3.2 – Critère AIC obtenu pour différentes configurations du modèle : nombre d'harmoniques (composante saisonnière annuelle), présence ou absence de covariables (énumérées dans la section 3.3).

$2, 4, 6, 8, 10, 12$. Ces différents modèles ont été testés, avec et sans la présence de variables explicatives. Le choix du nombre d'harmoniques a été basé sur la minimisation du critère AIC évalué sur la période d'apprentissage (années 2011 à 2015), et du RMSE calculé sur la base de tests (années 2016 et 2017). Les critères obtenus pour chacune des configurations sont donnés dans le tableau 3.2 et la figure 3.1. Le tableau 3.2 montre que le critère AIC est meilleur lorsque le nombre d'harmoniques augmente. Ce résultat est cohérent, car le nombre de paramètres à estimer n'augmente pas lorsque le nombre d'harmoniques de la composante annuelle croît, mais la vraisemblance, elle, augmente. Le critère AIC obtenu à partir des modèles avec covariables est meilleur que celui obtenu à partir des modèles sans covariable. Entre les modèles avec 10 et 12 harmoniques, il n'y a pas d'amélioration significative du critère AIC. Comme le critère AIC ne permet pas de sélectionner un bon nombre d'harmoniques, nous nous appuyons sur le critère RMSE pour distinguer les modèles. Comme prévu, l'augmentation de l'horizon de prédiction h conduit à une augmentation de l'erreur de prédiction. Les modèles avec covariables améliorent les performances de prédiction. Plusieurs observations ressortent :

— Pour les modèles sans variables exogènes, le choix de dix harmoniques est meilleur pour le RER A, et le modèle sans composante annuelle est meilleur pour les données métro. Si une composante annuelle capable de capturer les périodes de pointe et les périodes creuses est un bon ajout au modèle du premier cas, elle est pénalisante dans le second cas. Les profils de fréquentation des étés 2017 et 2018 sont totalement modifiés en raison des travaux de maintenance de la ligne RER.

— Pour les modèles avec variables exogènes, malgré des capacités de prédiction proches entre les différents modèles, les versions avec $k = 6$ harmoniques semblent être

légèrement meilleures que les autres, pour les deux lignes de transport (Métro 1 pour h = 4, 5, 6, 7 et RER A pour h = 2, 3, 4, 5, 6, 7).

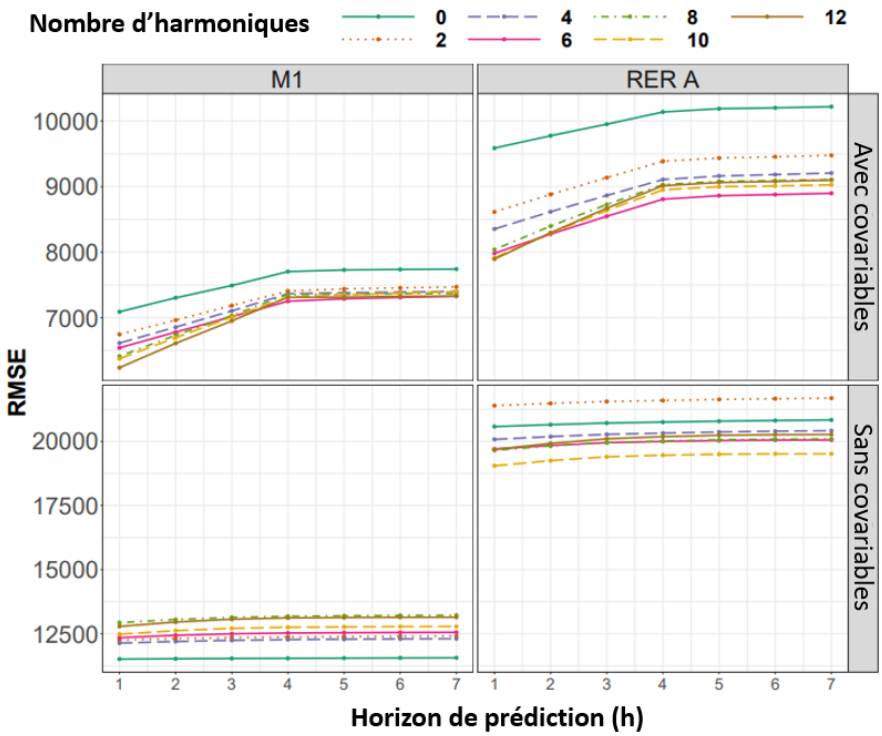


FIGURE 3.1 – Critère RMSE obtenu pour différentes configurations du modèle sur les années 2016 et 2017, en faisant varier l'horizon de prévision h de 1 à 7.

Le modèle retenu a des saisonnalités annuelles et hebdomadaires déterministes. Sa tendance est contrainte pour ne conserver que les évolutions de long terme. Il intègre les covariables avec un coefficient de régression stochastique, et la saisonnalité annuelle est basée sur une décomposition en six harmoniques. Les composantes estimées à partir du modèle sélectionné sont analysées dans la section suivante.

3.4.2 Analyse des composantes du modèle

Dans notre situation, la tendance et les composantes saisonnières ont été utilisées pour analyser les variations naturelles de la fréquentation des stations, tandis que les composantes résultant des variables explicatives ont été utilisées pour analyser l'effet des perturbations anticipées (tels les travaux de maintenance sur les lignes de métro/RER) ou non anticipées (dont les grèves, la crise sanitaire de Covid19). Les résultats de la décomposition sont présentés pour chaque composante : l'estimation de Kalman lissée $\boldsymbol{\varrho}_j^J$, et l'intervalle de

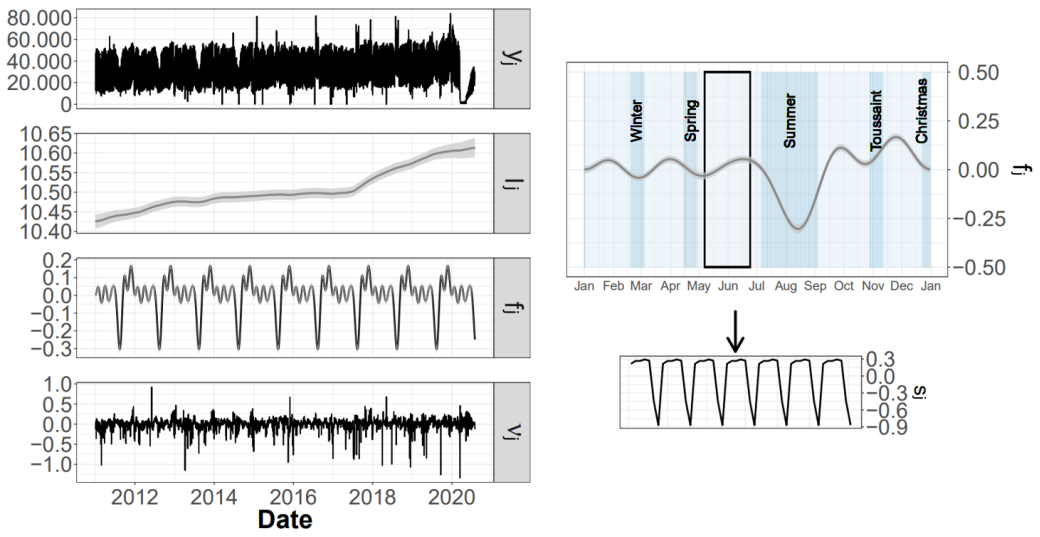
confiance à 95% calculé avec l'estimation lissée de la covariance $\mathbf{P}_j^J$, pour chaque jour entre 2011 et 2020. Ils sont respectivement représentés par des lignes noires et des zones grises remplies sur les différentes figures de résultats. Les composantes ont été mises à l'échelle logarithmique, afin de mieux visualiser les intervalles de confiance. Pour certaines composantes, il sera important de différencier les jours ouvrés des jours non ouvrés, qui seront représentés respectivement par des points rouges et bleus. Nous quantifierons, pour certains points d'intérêt, l'impact des variables exogènes s sur les fréquentations, en étudiant la transformation exponentielle de leurs coefficients de régression $e^{\beta_j^{(s)}}$. Pour un jour donné j, la variable s multiplie la fréquentation par la valeur $e^{\beta_j^{(s)}}$ par rapport à un niveau de fréquentation de référence. Par exemple, prenons l'effet de la variable $s =$ « Jours de travaux de maintenance sur la ligne RER » qui a pour valeur $e^{\beta_j^{(s)}} = 0,7$ un jour donné : cela implique que les jours de travaux de maintenance impactent la fréquentation, et explique qu'il n'y ait que 70% des fréquentations habituelles.

Les variations naturelles de la fréquentation des stations

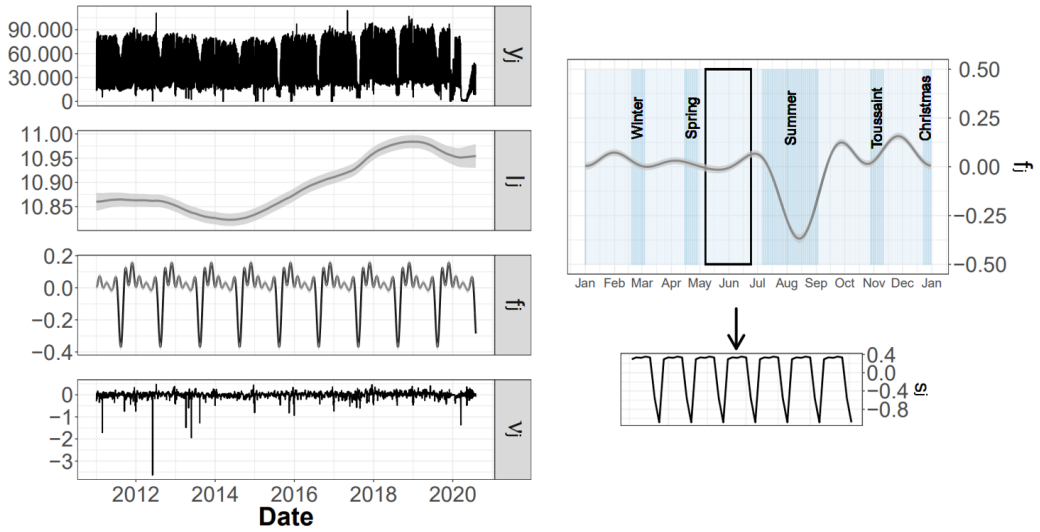
La figure 3.2a (métro) et la figure 3.2b (RER) présentent les séries temporelles de fréquentation des lignes de transport étudiées, ainsi que les composantes de tendance et de saisonnalité.

Les tendances d'évolution de la fréquentation au fil des années montrent des différences entre les deux lignes de transport. Pour le métro (figure 3.2a), on note une augmentation de la fréquentation au fil des années, avec une accélération à partir de 2017. Cela peut s'expliquer par l'augmentation du trafic sur l'axe est-ouest parisien, induit par une concentration croissante des emplois à l'ouest, notamment dans le quartier de La Défense. Pour le RER (figure 3.2b), la tendance l_j présente un profil plus difficile à interpréter, puisqu'elle diminue entre 2011 et 2014, contrairement au métro. Nous associons cette baisse à la création de deux nouvelles sorties depuis le tramway T2 après 2012, ce qui a sans doute modifié la dynamique des déplacements dans toute la partie ouest du pôle d'échange. L'augmentation qui suit s'inscrit dans la continuité de la dynamique observée dans le métro. La diminution de cette composante, ainsi que l'augmentation de son incertitude au cours de la période allant de fin 2019 à l'année 2020, sont imputables aux périodes de grève et de pandémie de Covid19.

Les deux lignes de transport présentent des composantes saisonnières annuelles f_j similaires. Les creux importants entre juillet et octobre correspondent aux périodes de vacances estivales associées à une baisse considérable de la fréquentation des lignes ($e^{f_t} = 0,67$ pour le RER, et $0,74$ pour le métro). Les vacances d'été sont ainsi responsables d'une baisse de près de 30% de la fréquentation sur ces deux lignes. Les autres vacances scolaires (hiver, printemps, automne et Noël) sont également visibles. A noter la présence de périodes de sur-fréquentation juste avant les vacances de Noël. Il existe cependant une différence entre les deux profils durant la première partie de chaque année, avant les vacances d'été : la



(a)



(b)

FIGURE 3.2 – Décomposition de la série temporelle y_j des flux entrants vers la ligne de métro 1 (a) et de RER A (b), en tendance à échelle logarithmique (l_j), en saisonnalité annuelle à échelle logarithmique (f_j), et en résidus à échelle logarithmique (ν_j) (panneaux de gauche). Profils annuels en échelle logarithmique de la composante f_j, avec les différentes périodes de vacances en bleu, et élargissement sur quatre semaines de la saisonnalité hebdomadaire en échelle logarithmique s_j (panneaux de droite).

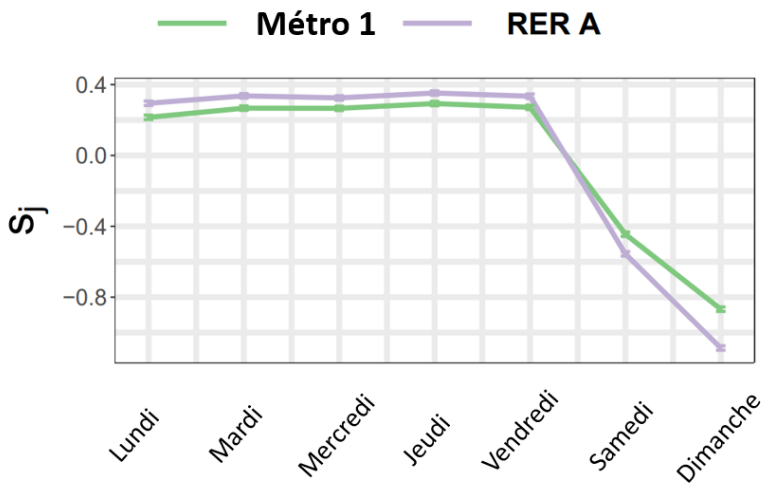


FIGURE 3.3 – Saisonnalité hebdomadaire à l'échelle logarithmique s_j, et intervalles de confiance à 95% pour les flux entrants vers la ligne du métro 1 (vert), et vers la ligne du RER A (violet).

fréquentation du métro fluctue davantage entre les périodes de vacances et de travail que celle du RER. L'incertitude est plus forte pendant les vacances d'été que pendant le reste de l'année. Comme nous le verrons plus loin, les travaux de maintenance interviennent souvent à ces moments-là, modifiant sensiblement la fréquentation d'une année à l'autre.
Les composantes saisonnières hebdomadaires s_j sont difficiles à visualiser à l'échelle des années. Nous présentons donc un agrandissement comparatif des deux profils dans la figure 3.3. Les coefficients de la composante saisonnière hebdomadaire e^{s_j} associent chaque jour de la semaine à un pourcentage de fréquentation par rapport au niveau de référence. Ces profils reflètent ainsi le niveau de fréquentation de chacun de ces jours. Les profils des deux lignes de transport sont très similaires, avec des poids élevés alloués aux jours de la semaine ($e^{\beta_j^{(s)}} > 1,2$; fréquentation supérieure de 20% au niveau de référence), et de faibles poids attribués aux week-ends (fréquentation autour de 60% du niveau de référence le samedi, et 40% le dimanche). Une légère différence, observée entre les jours de semaine, peut être associée aux habitudes connues en matière de déplacements dans les transports publics : les lundis sont légèrement moins fréquentés, et on observe également une légère baisse le mercredi (jour sans école pour certains enfants). Les mardis et jeudis attirent plus de monde. A noter qu'il semble y avoir des différences plus importantes entre les jours de semaine et les week-ends pour le RER que pour le métro.

Les résidus ν_j nous permettent de détecter les jours où le modèle n'a pas, ou mal, pris en compte un effet. Pour les deux lignes de transport, nous constatons que certains jours avec des pannes de transport (non anticipées), ou des événements nécessitant la

fermeture des lignes pendant une grande partie de la journée, sont associés à des résidus importants. Par exemple, le 21 septembre 2019 présente un résidu important, car une manifestation a empêché l'arrivée du métro à la station « La Défense Grande Arche ». Notons qu'une analyse plus approfondie de ces résidus pourrait être mise à profit pour détecter des situations atypiques dans l'affluence vers les deux lignes de transport.

Analyse de l'impact des travaux de maintenance

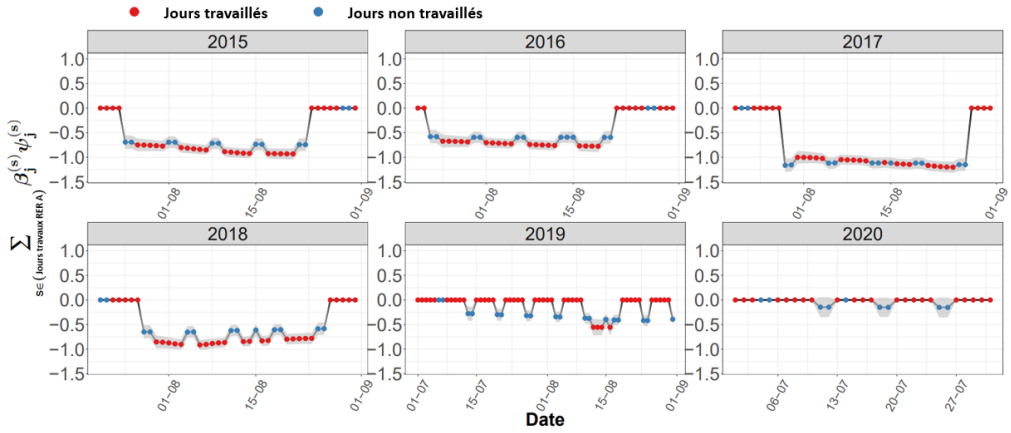
Les travaux de maintenance ont un fort impact sur la dynamique de fréquentation du pôle. Commençons par les périodes de travaux de maintenance de la ligne RER : chaque été depuis 2015, celle-ci fait l'objet de travaux qui nécessitent sa fermeture en semaine et/ou le week-end. Pour visualiser l'effet de ces travaux, nous allons représenter les coefficients de régression multipliés par les indicateurs de classe associés. Les résultats sont présentés dans la Figure 3.4a pour l'impact des travaux sur la fréquentation du RER A, et dans la Figure 3.4b pour l'impact sur la ligne de métro 1.

Ces résultats révèlent des différences notables entre les deux lignes de transport : alors que l'impact est négatif sur la fréquentation de la ligne de RER, c'est l'inverse pour la ligne de métro. Ce résultat souligne l'importance de la ligne 1 du métro comme substitut de la ligne A du RER pour traverser Paris. L'impact des travaux de maintenance sur la fréquentation de la ligne RER n'est pas total, celle-ci atteint un minimum de $e^{\beta_j^{(s)}} = 0,3$ en 2017. « La Défense Grande Arche » étant une gare d'échange, les flux de personnes peuvent continuer à transiter par la zone d'accès à la ligne RER pour rejoindre d'autres lignes ; la baisse souligne néanmoins que la plupart des personnes ne transitent plus par cette zone. Par ailleurs, notons la moindre importance des travaux sur la baisse de fréquentation de la ligne RER durant les étés 2019 et 2020, par rapport aux autres étés (figure 3.4a). Durant ces deux étés, les travaux de maintenance n'ont eu lieu que les week-ends (sauf la semaine du 15 août 2019, avec des travaux de maintenance tous les jours). Nous émettons deux hypothèses pour expliquer ce phénomène :

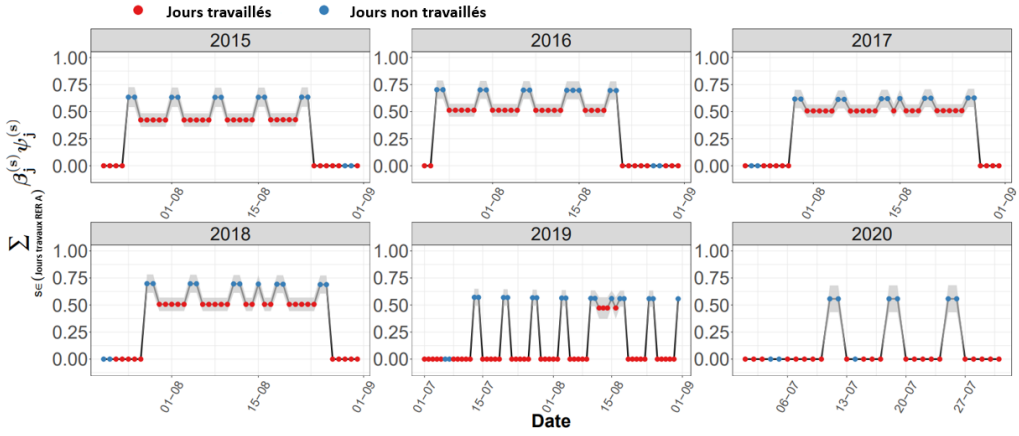
— Comme les travaux ont eu lieu tous les jours pour les années 2015 à 2018, davantage de personnes se sont reportées sur un autre mode de transport, ou ont privilégié l'évitement de leur lieu de travail (ex. congés, recours au télétravail, utilisation d'un tiers lieu).

— Les travaux ont eu lieu entre le pôle de transport et Paris de 2015 à 2018, puis au sein même de Paris en 2019 et 2020 : une alternative à la ligne 1 du métro était possible pour les habitants de l'Est parisien souhaitant se rendre à La Défense en 2019 et 2020. C'est ce que montrent les coefficients de régression pour le métro, qui sont plus faibles en 2019 et 2020 que les autres années.

Comme la ligne de RER, la ligne de métro est parfois arrêtée pour cause de travaux de maintenance. La figure 3.5 montre les coefficients de régression associés. Les travaux de maintenance sur la ligne de métro ont eu un impact considérable sur la fréquentation



(a)



(b)

FIGURE 3.4 – Composante (à l'échelle logarithmique) associée aux travaux de maintenance sur la ligne RER, et intervalle de confiance à 95% pour les flux entrants vers la ligne RER (a) ou métro (b). Les étés entre 2015 et 2020 sont représentés.

de cette ligne. Contrairement à la zone d'accès à la ligne de RER, l'espace de la ligne de métro n'est pas une zone de correspondance, et presque personne n'y entre lorsque la ligne est arrêtée. De même, comme pour les travaux de maintenance sur la ligne de RER, on note un phénomène de transfert de flux, cette fois du métro vers la ligne de RER, dont la fréquentation a légèrement augmenté. De larges intervalles de confiance soulignent pour ce cas une grande incertitude. En raison du peu de jours de travaux de maintenance sur la ligne de métro, l'incertitude est importante.

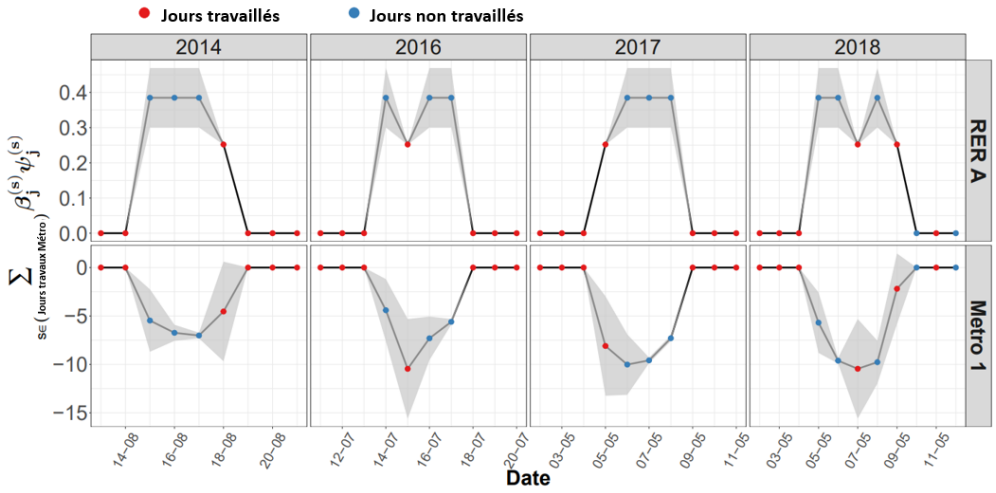


FIGURE 3.5 – Composante (à l'échelle logarithmique) associée aux travaux de maintenance sur la ligne de métro 1, et intervalle de confiance à 95% pour les flux entrants vers les lignes RER et métro.

Notons que la complémentarité entre les deux lignes étudiées permet de les rendre moins vulnérables aux périodes de travaux. Les usagers peuvent continuer à se déplacer lorsque l'une des lignes est à l'arrêt. La conservation de stations de correspondance entre les deux ligne lors de ces périodes semble ainsi indispensable comme on a pu le voir ici.

Analyse de la période de grève de Décembre 2019 - Janvier 2020

La période de décembre 2019 à janvier 2020 a été caractérisée par une mobilisation massive contre la réforme du système de retraite français. Le soutien à la grève a été très fort chez l'opérateur de transport RATP, perturbant l'ensemble de son réseau de transport public. La comparaison de l'impact de cette grève sur les fréquentations des deux lignes est visualisée dans la figure 3.6. L'étude de ces profils montre que, si l'effet de la grève a été néfaste pour la fréquentation de la ligne RER, il a été positif pour la ligne de métro à partir du 7 décembre. Ce phénomène est dû à un transfert de flux de la ligne de RER vers la ligne de métro. Le métro 1, étant automatique, a maintenu un service normal pendant

la grève, alors que le trafic du RER a été fortement perturbé. Ce phénomène a également entraîné de nombreuses situations de congestion et de surcharge du métro. Le trafic du RER s'est amélioré après les vacances de Noël. Dans la figure 3.6, nous avons encadré la période des vacances de Noël, qui sépare deux tendances :

— Avant les vacances : les premières semaines de la grève ont été particulièrement difficiles pour la circulation des RER. La grève a eu un effet très marqué sur le trafic RER, puisqu'elle explique une baisse de plus de la moitié de la fréquentation des lignes du RER ($e^{\beta_j^{(s)}}$ inférieur à 0,5). En revanche, les flux de voyageurs sur la ligne de métro ont augmenté régulièrement : le coefficient passe de $e^{\beta_j^{(s)}} = 1$ au 3ème jour de grève, à plus de 1,5 au début des vacances. Au fur et à mesure de la grève, alors que le trafic du RER est fortement perturbé, l'effet du report sur la ligne de métro est de plus en plus important, au point de surcharger la ligne de plus de 50% par rapport à l'utilisation de référence.

— Après les vacances : au fur et à mesure que la situation s'est améliorée pour le trafic RER, on assiste à un retour progressif à une situation normale de fréquentation, du RER comme du métro.

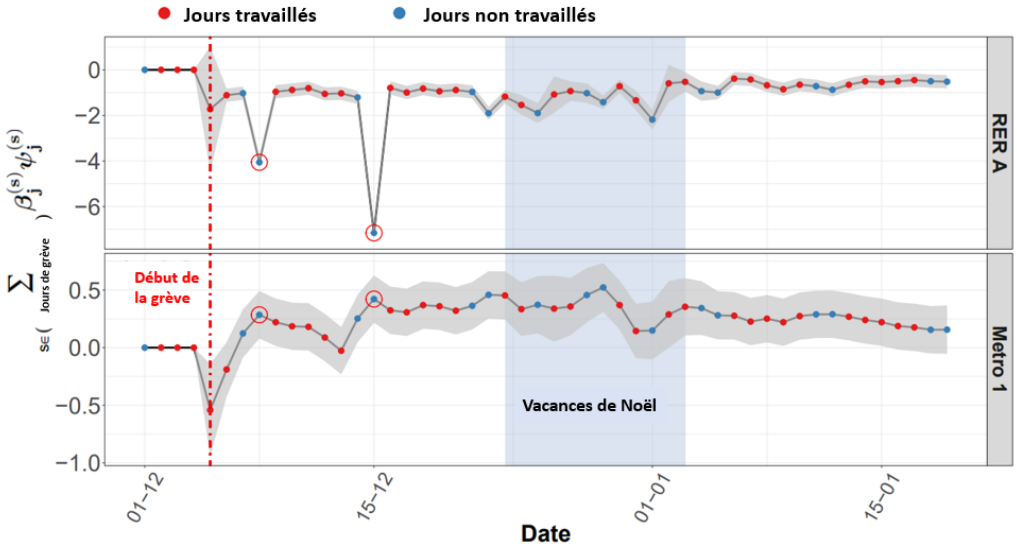


FIGURE 3.6 – Composantes (à l'échelle logarithmique) associées à l'effet des grèves, et intervalles de confiance à 95% pour les flux entrants dans les lignes RER et métro, pour la période de décembre 2019 à janvier 2020.

On note la présence des deux dimanches 8 et 15 décembre (entourés en rouge sur la figure 3.6), durant lesquels l'utilisation de la ligne de RER a été quasi nulle : le trafic du RER a en effet été totalement interrompu ces jours-là. L'incertitude reste pratiquement

constante sur l'ensemble de la période, à l'exception du jour de Noël et du jour de l'an : l'effet cumulé de ces journées particulières et de la grève rend l'estimation de l'impact de la grève plus incertaine.

Périodes du premier confinement et post-confinement liés à la pandémie de Covid19

Le premier confinement dû à la pandémie de Covid19 a, sans surprise, eu un impact considérable sur la fréquentation du pôle de transport. La plupart des employés de bureau qui se rendent quotidiennement à La Défense n'ont alors pas pu se rendre au travail. La période post-confinement a également eu un impact significatif sur l'utilisation du pôle, puisque le télétravail était fortement recommandé. Nous présentons les coefficients de régression associés aux périodes de confinement et de post-confinement dans la Figure 3.7.

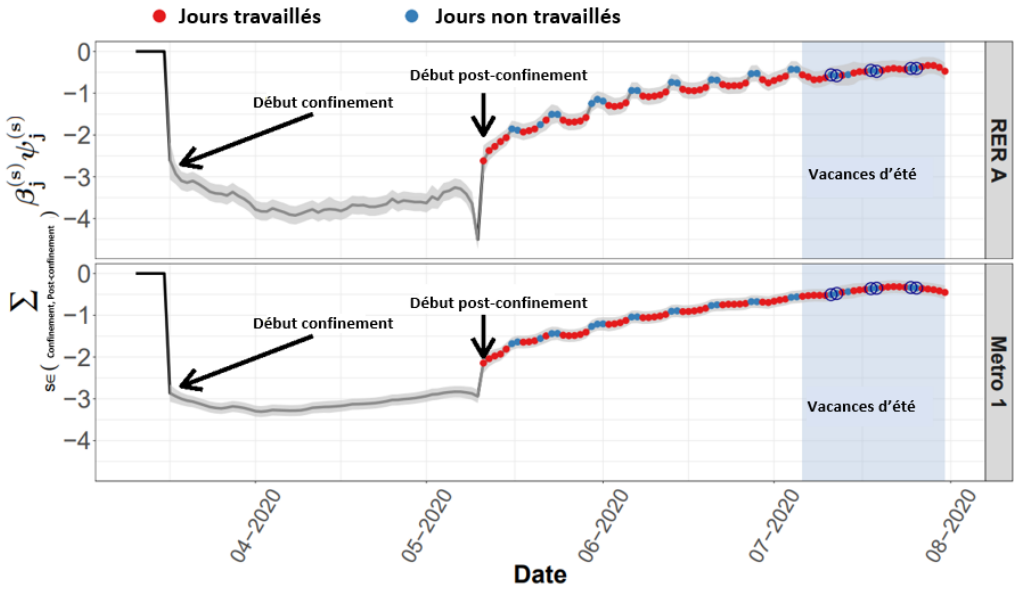


FIGURE 3.7 – Composantes (à l'échelle logarithmique) associées aux périodes de 1er confinement/post-confinement, et intervalles de confiance à 95% pour les flux entrants vers le RER A et vers le métro 1.

Pour les deux lignes de transport (voir la figure 3.7), il y a un effet de perte presque totale de la fréquentation pendant la période de confinement du 17 mars 2020 au 11 mai 2020. Le coefficient $\beta_j^{(s)}$ est autour de -3 pendant cette période, et donc $e^{\beta_j^{(s)}}$, autour de zéro. Il n'y a pas de différence notable entre les jours de semaine et les week-ends : personne n'était présent quel que soit le jour. On note un comportement atypique qui maximise l'effet du confinement lors de son dernier jour pour le RER.

L'impact de la période de post-confinement (à partir du 11 mai), très visible sur la Figure 3.7, montre des similitudes et des différences entre les deux lignes. Plusieurs observations émergent :

— On observe, pour les deux cas, un lent retour à une situation normale, avec un surprenant profil « en vagues » : le post-confinement a eu un effet plus fort sur la baisse de fréquentation les jours de semaine, que les jours de week-end. Les entreprises ayant fortement encouragé le télétravail au-delà de la fin du confinement, nous pouvons faire l'hypothèse qu'il y a eu lors de cette période une acceptabilité plus forte à sortir pour le shopping et les loisirs individuels (en particulier le week-end), que pour le travail.

— Le profil en vagues est plus marqué pour le RER que pour le métro : comme le RER relie le pôle de La Défense à un plus grand nombre de banlieues que la ligne 1 du métro, il touche un plus grand nombre de personnes, ce qui rend notre hypothèse précédente plus visible sur la fréquentation de la ligne de RER, que sur celle du métro.

— La période des vacances d'été semble supprimer complètement le profil du retour à la normale en vagues. Pendant les vacances d'été, la ligne RER était en travaux de maintenance les week-ends (entouré en bleu sur la figure 3.7). Le phénomène de baisse de la fréquentation des gares peut donc s'expliquer par deux variables : la période post-confinement et les travaux de maintenance.

Notons que la situation « normale » post-Covid19 n'est pas la même que celle d'avant Covid19 : le coefficient $\beta_j^{(s)}$ lié au post-confinement reste négatif, et les tendances l_j ont diminué, notamment pour le RER A (figure 3.2b).

3.5 Conclusion

Ce chapitre a présenté l'analyse des séries temporelles de flux de personnes entrants dans un hub de transport (La Défense), sur la base de modèles de décomposition. Nous avons concentré notre analyse sur la fréquentation quotidienne collectée pour neuf années sur les lignes de RER A et de métro 1. La force de tels modèles réside dans leur capacité explicative. Ils sont capables de mettre en évidence les dynamiques à long terme des fréquentations, telles que la tendance et la saisonnalité (annuelle, hebdomadaire), et l'impact de facteurs exogènes, qu'ils soient anticipés, comme les travaux de maintenance, ou non, comme les grèves ou la crise sanitaire Covid19. Grâce aux coefficients de régression du modèle de décomposition, nous pouvons quantifier les impacts séparés de chaque facteur, ce qui est impossible avec d'autres modèles basés sur l'apprentissage. Ces modèles nous permettent également de prédire la fréquentation quotidienne sur plusieurs horizons temporels. Comme les deux lignes de transport desservent les mêmes stations, la décomposition a permis

d'identifier et de quantifier certains transferts modaux, notamment pendant les périodes de travaux de maintenance.

Dans ce travail, nous avons abordé toutes les questions relatives à l'estimation des paramètres de ces modèles, afin d'obtenir une décomposition appropriée aux données considérées. Cependant, malgré leur fort aspect descriptif, les modèles linéaires dynamiques nécessitent une calibration fine, et des connaissances statistiques approfondies. De plus, il est nécessaire de réussir à combiner la connaissance métier avec les choix de modélisation.

Un transfert de ce travail à d'autres études de cas, où il y a un besoin de quantifier l'impact d'événements typiques par exemple, est possible grâce à l'adaptabilité de ces modèles. C'est pourquoi nous mettons à disposition le code source des analyses à l'adresse https://github.com/pdenailly/TransportHub_TimeSeriesDecomposition. Les décompositions y sont appliquées à des données de fréquentation des stations de transport du réseau parisien, disponibles en ligne mais différentes de celles que nous avons utilisé dans le cadre de ce travail. Ces données proviennent de : https://data.sncf.com/explore/dataset/validations-reseau-ferre-nombre-validations-par-jour-1er-semestre/table/.

Dans le cadre de nos travaux sur les fréquentations du pôle de La Défense, ce chapitre nous a apporté une vision plus claire sur leur évolution, ainsi que sur l'impact d'événements long-terme. Nous nous sommes ici concentré sur l'aspect temporel des séries, et non sur leur aspect spatial. Pourtant, de nombreuses informations pourraient être retirées de l'étude de la répartition spatiale des flux voyageurs au sein du pôle de transport, en réponse à certains événements. D'autres limites se posent également dans le chapitre abordé :

— L'étude d'événements long-terme demande à ce qu'ils soient connus *a priori*. D'autres événements inconnus pourraient aussi impacter la mobilité des personnes au sein du pôle de transport.

— L'agrégation à l'échelle journalière empêche d'étudier finement des événements de courte durée, comme des concerts ou des perturbations dans les transports.

— Une analyse multivariée semble nécessaire. Nous nous sommes concentrés ici sur deux zones particulières du pôle de La Défense. D'autres lieux, comme les accès aux centres commerciaux ou aux gares routières, pourraient aussi être impactés par de multiples événements. La prise en compte de la structure de dépendance (corrélations) entre ces séries pourrait également avoir un intérêt.

Ces questions sont abordées dans le chapitre suivant, qui a pour objectif de mettre en place une méthode de synthétisation, des dynamiques de déplacements piétons au sein du pôle de La Défense, en profils de mobilité temporels facilement interprétables.

Chapitre 4

Identifier des groupes similaires de profils de mobilité

4.1 Introduction

La modélisation statistique des flux piétons est un sujet de recherche actif qui peut fournir des informations précieuses pour la planification des services de transport, la gestion des flux aux sein des centres de transport, ou bien encore l'aide à la négociation de baux immobiliers ou publicitaires. Néanmoins, l'extraction d'informations à partir d'un grand ensemble de ces données, souvent bruitées, est une tâche difficile.

Un pôle de transport multimodal est un lieu d'étude complexe pour les données de comptages, car il s'y croise quotidiennement un grand nombre de flux de passagers. Ces flux peuvent aller et venir de lieux d'intérêt présents dans l'environnement du pôle (zones de travail, centres commerciaux, etc.) ou transiter entre les lignes de transport desservant le pôle. Cependant, en fonction de l'heure de la journée, de la période de l'année ou de divers événements locaux, ces flux ne transitent pas nécessairement par les mêmes endroits. Trois éléments devraient être pris en compte lors de la modélisation des données de comptage de personnes dans un pôle de transport :

— La mobilité des personnes est influencée par divers facteurs qui peuvent être calendaires (par exemple, l'heure et le type de jour, les jours fériés) ou non calendaires, comme les concerts ou les perturbations de transports.

— Les mobilités sont soumises à des effets de longue durée tels que des tendances, des effets saisonniers, ou l'impact de phénomènes de grande ampleur comme la pandémie de Covid19 [Nai+21]. Ces séries présentent donc un aspect non stationnaire.

— Des dépendances potentielles peuvent exister entre les différents lieux de captage des données, ce qui est difficile à modéliser pour des données de comptage.

En plus de partager une dynamique commune en réponse à certains événements, ces séries de comptages peuvent aussi avoir leur propre dynamique pour des événements plus loca-

lisés. En considérant ces différents aspects, ainsi que la nature souvent bruitée des flux de personnes, une approche de clustering s'avère être indispensable lorsque l'on souhaite analyser de manière synthétique les dynamiques de déplacement. Le clustering est un moyen de partitionner les séries temporelles de comptages, en un ensemble réduit de catégories partageant des dynamiques communes, que l'on appellera des « profils de mobilité ». Comme mentionné dans le travail de [Gha+17], les trois catégories suivantes de profils de mobilité peuvent être analysées :

1. Les profils spatiaux, qui fournissent des informations sur des stations ou des lieux utilisés de la même manière ;
2. Les profils temporels, qui rassemblent des périodes qui se ressemblent en termes de manière dont le réseau de transport est utilisé au fil du temps [Bri+17 ; Gha+17] ;
3. Les profils spatio-temporels qui combinent les deux aspects. On peut penser à des profils qui se ressemblent en termes d'évolution temporelle de l'utilisation spatiale d'un espace de transport. Par exemple, le travail de [PSY20] détermine les dates présentant des schémas similaires de mobilité quotidienne (à savoir, des profils origine-destination similaires).

Comme vu dans le chapitre 3, les séries temporelles de flux de passagers dans le pôle de La Défense sont susceptibles d'évoluer en fonction de diverses périodes caractérisées par des phénomènes calendaires ou non (des grèves, des travaux de maintenance ou des mesures sanitaires contre la pandémie de Covid19). De plus, des facteurs ponctuels exogènes, tels que des concerts et des perturbations des transports, peuvent également avoir un impact sur la mobilité.

Dans un souci de compréhension du fonctionnement du pôle de transport de La Défense en termes de flux de passagers, ce chapitre sera dédié à la mise en place d'une approche de clustering basée sur un modèle de mélange, pour analyser les séries temporelles de comptage et d'en extraire des profils de mobilité temporels interprétables. Nous travaillerons sur l'ensemble des séries (multivarié) et chercherons à détecter de manière flexible des segments de temps au sein desquels les séries forment des profils de mobilité homogènes, en lien avec du contexte. Ce travail, le coeur et la contribution de ce chapitre, s'appuiera sur les données de capteurs stéréoscopiques et de billettique, collectées à la station Grande Arche, entre le 1er Avril 2019 et le 1er Septembre 2020. Dans la section suivante, nous présentons quelques notions et méthodes de clustering, avant de nous intéresser à la modélisation de données de comptage multivariées.

4.2 Etat de l'art

Le clustering est un ensemble d'approches non-supervisées de partitionnement des données en S catégories. Ce type d'approche aide notamment à donner du sens aux données, en y révélant des catégories permettant de mieux les appréhender. Plusieurs approches ont

été proposées pour le clustering des données de mobilité. Il s'agit notamment d'approches à base de distance, comme la classification ascendante hiérarchique (HAC) ou les algorithmes K-means [Lat+13]. L'algorithme DBSCAN est une autre méthode de clustering basée sur la densité des points de données. Cette méthode a été utilisée par [MZB18] pour détecter quelles stations de métro ou lignes de bus sont régulièrement utilisées, quelles lignes sont reliées, ou dans quelles tranches horaires de la journée les passagers utilisent le plus fréquemment le réseau de transport londonien.

Nous nous appuierons dans ce travail sur d'autres méthodes basées sur des modèles probabilistes qui, contrairement aux méthodes basées sur la distance, regroupent les données en utilisant un mélange de distributions de probabilité [Bou+19 ; MLR19]. En définissant $\mathbf{z}_t \in \{0,1\}^S$ comme étant le vecteur d'association de l'observation $\mathbf{y}_t \in \mathbb{N}^P$ à l'un des S clusters, chaque observation est modélisée selon un processus en deux étapes :

1. La variable latente $\mathbf{z}_t$ est distribuée suivant une loi multinomiale $\mathcal{M}$ avec les proportions $\pi = (\pi_1, ..., \pi_S)$
2. Conditionnellement à $\mathbf{z}_{t,s} = 1$, l'observation $\mathbf{y}_t$ suit une distribution de paramètres $\boldsymbol{\theta}_s$ associés au cluster d'appartenances.

   Le schéma d'échantillonnage est donc le suivant :

$$\mathbf{z}_t|\pi \sim \mathcal{M}(1, \pi) \tag{4.1}$$

$$\mathbf{y}_t|z_{t,s} = 1, \boldsymbol{\theta}_s \sim p(\mathbf{y}_t|\boldsymbol{\theta}_s). \tag{4.2}$$

Ces modèles sont plus flexibles que les approches à base de distance, et permettent l'incorporation de facteurs exogènes dans la modélisation. Par conséquent, ils semblent adaptés à notre cas d'étude des mobilités dans le pôle de La Défense, en raison de la grande variété de distributions au sein de chaque segment, qui intègrent des effets des facteurs exogènes et des interdépendances entre les lieux de comptage. De plus, ces modèles offrent une interprétabilité appréciable, permettant de mieux comprendre les dynamiques de mobilité au sein du pôle de transport.

Les données observées étant des comptages surdispersés (les variances sont supérieures aux moyennes, voir section 2.2.3), les travaux s'orienteraient vers des modèles prenant en compte cette spécificité. Les mélanges de régression de Poisson peuvent être ajustés avec succès à ces comptages surdispersés, en présence de facteurs exogènes [EL14 ; Moh+16]. Néanmoins, l'utilité de ces méthodes peut être limitée lorsque les données sont fortement surdispersées, car elles font l'hypothèse que l'espérance et la variance sont égales conditionnellement aux segments et aux covariables. L'utilisation de régressions binomiales négatives peut atténuer ce problème, car les moyennes et les variances diffèrent [Hil11]. Des mélanges de ces modèles sont utilisés dans l'analyse transcriptomique comme dans [Li+21], mais nous n'avons pas trouvé d'application de ces modèles dans le domaine de la mobilité.

Dans le cas de données de comptage multivariées, chaque observation $\mathbf{y}_t$ est un vecteur de taille P $\{y_{t,p}\}_{p\in\{1,...,P\}}$, où chaque $y_{t,p}$ est un comptage. Un problème qui se pose,

lorsqu'on travaille avec des données de comptage multivariées, est celui des relations de dépendance entre les séries de comptage. Un défi consiste alors à quantifier ces relations. Les régressions multinomiales et Dirichlet-multinomiales sont les modèles généralement appliqués aux cas multivariés [Zha+17a]. Cependant, ces modèles ne peuvent pas traiter les comptages totaux $\sum_p y_{t,p}$. Comme indiqué par [PFD21], cela induit une contrainte en termes de dépendances entre les séries de comptage car, dans le cas où les totaux ne sont pas distribués de manière aléatoire, toute série est déterministe lorsque toutes les autres séries sont connues. Une première solution consiste à écrire les distributions de comptage comme des distributions de type « sommes et partages », ce qui permet de prendre en compte une dépendance simple entre les séries, par le biais d'une contrainte sur la somme. Ce type de modèle est décrit dans le travail de [JM19]. Ces modèles supposent que les comptages totaux suivent une distribution univariée (par exemple Poisson ou binomiale négative), et que la distribution jointe des comptages est multivariée (par exemple Multinomiale ou Dirichlet-multinomiale). Il est par exemple possible d'obtenir une distribution en combinant une binomiale négative avec une distribution multinomiale comme dans [SYS64]. Une autre façon de ne pas supposer l'indépendance entre les séries de comptages consiste à utiliser la distribution de Poisson log-normale multivariée proposée pour la première fois par [AH89], également bien expliquée par [CMR21], et utilisée de façon intéressante dans un travail de clustering par [Sil+19]. Dans ce modèle, la structure de dépendance entre les séries est prise en compte par une matrice de covariance (qui pourra être pleine ou non). Les différents modèles évoqués ici se distinguent en particulier par les hypothèses faites sur la structure de corrélation du bruit.

Les méthodes présentées ici semblent convenir à la modélisation de données de comptage multivariées, surdispersées et corrélées (caractéristiques vues dans le chapitre 2). Dans les travaux qui suivent, nous mettons en place un modèle adapté à la recherche de profils de mobilité dans un cas multivarié, avec prise en compte du contexte. La section suivante précise notre positionnement et les contributions de ce chapitre. Une étude exploratoire qui avait été menée en amont afin de détecter des profils de mobilité temporels dans des cas univariés, sans prise en compte de données contextuelles est quant à elle proposée dans l'annexe C.1.

4.3 Positionnement et contributions

Notre objectif dans ce travail est de modéliser une série temporelle multivariée de comptages par un ensemble de covariables, afin d'obtenir une vision synthétique, utile pour comprendre le fonctionnement du pôle de transport, ou comme base de travail pour la prédiction. En raison de leur caractère non stationnaire, et de la variation temporelle de l'effet des covariables, les séries temporelles doivent être subdivisées en plusieurs segments [TOV20]. On peut envisager ce travail de segmentation temporelle des données multivariées de comptage, comme un moyen de capturer la régularité locale par des profils de mobilité,

qui peuvent être facilement interprétés, et où les covariables ajoutent des caractéristiques distinctives à ces profils [Zho+15]. La modélisation régressive par les covariables permettra de prendre en compte les phénomènes explicables impactant les mobilités à une échelle horaire. La segmentation quant à elle permettra de déterminer les périodes aux dynamiques communes. La surdispersion et les corrélations entre les séries sont des caractéristiques fréquemment rencontrées avec les données de comptage [08]. Afin de prendre en compte ces spécificités dans la segmentation, nous nous sommes appuyés sur deux stratégies, trouvées dans la littérature, permettant de modéliser des données de comptage volumineuses, bruitées et éventuellement corrélées. Ces deux stratégies, à savoir les modèles de type « sommes et partages » [JM19] et les modèles de Poisson log-normaux [CMR21], abordent la modélisation des données de comptage selon des philosophies distinctes. Dans la première approche, les périodes sont considérées comme homogènes si, conditionnellement aux covariables et aux segments, les totaux (i.e. « sommes ») des personnes observées dans le pôle de transport et leur distribution (i.e. « partages ») entre plusieurs lieux sont similaires. Dans la seconde, les périodes sont considérées comme homogènes si, conditionnellement aux covariables et aux segments, les séries ont des effectifs moyens similaires et des structures de corrélation semblables. Les deux stratégies sont discutées dans la section 4.4. Notre travail se positionne donc comme la recherche d'un modèle qui puisse à la fois prendre en compte des données multiples, dispersées et corrélées, et regrouper les périodes d'observation en des segments les plus continus possibles. Les contributions apportées par ce travail sont les suivantes :

1. Nous synthétisons les données de comptage en profils de mobilité temporels réguliers. À cette fin, nous proposons des méthodes pour détecter de manière flexible les segments temporels dans les séries de comptage multivariées. Nous construisons des modèles de mélange temporellement « lisses », dans lesquels nous modélisons la transition entre les segments en utilisant des fonctions logistiques (de manière semblable à ce qui est proposé par les auteurs de [SAO21]). Ces fonctions logistiques intègrent des fonctions splines pour assurer une régularité temporelle à la segmentation, en particulier entre les jours proches. Au niveau supérieur, les probabilités d'occurrence des segments évoluent ; elles sont conçues pour faciliter la détection des segments. Au niveau inférieur, et pour chaque segment, des modèles *spécifiques* de régression des données de comptage gèrent la dynamique des flux de passagers.
2. Au sein des segments, nous utilisons des modèles de régression de Poisson log-normaux, et de « sommes et partages », bien adaptés au traitement de données de comptage multivariées, corrélées et surdispersées. À notre connaissance, aucune étude n'a été menée pour comparer les deux modèles présentés.
3. La segmentation temporelle est réalisée à l'échelle de la journée, mais le modèle gère également l'échelle des heures. Cette représentation tire des informations utiles pour les opérateurs et les gestionnaires d'espaces urbains, en valorisant les données collectées. La segmentation, sur une base journalière, donne une idée de la dynamique

à long terme des flux de passagers, tandis que la modélisation sur une base horaire permet d'étudier en détail l'impact des facteurs calendaires et non calendaires.

4.4 Structures et estimation des modèles

Les modèles de régression, « sommes et partages » et Poisson log-normal, sont d'abord introduits. En particulier, nous expliquerons comment ces différentes stratégies gèrent (ou non) les phénomènes de corrélation et de surdispersion. Nous expliquerons ensuite comment nous avons transformé ces modèles en modèles de mélange, capables de détecter les changements de régime dans les séries temporelles.

4.4.1 Modèles de régression pour des données de comptage multivariées, corrélées et surdispersées

Ci-après, Y est considéré comme une variable aléatoire sous la forme d'un P-vecteur de comptages parmi P emplacements : $Y = (Y_p)_{p\in\{1,..,P\}}$. Nous proposons de modéliser ces données de comptage par deux modèles de régression distincts, à savoir : « somme et partages » et Poisson log-normal. La notation générale de ces modèles de régression est la suivante :

$$Y|\mathbf{x}, \boldsymbol{\theta} \sim \mathcal{D}(\mathbf{x}, \boldsymbol{\theta}), \tag{4.3}$$

avec $\boldsymbol{\theta}$ un ensemble de paramètres contrôlant les distributions conditionnelles $\mathcal{D}$, que nous recherchons. Nous notons $\mathbf{x}$ un vecteur $D \times 1$ de D facteurs exogènes. Dans ce qui suit, nous présentons les différents modèles de régression que nous souhaitons comparer.

Les modèles de régression « sommes et partages »

La première stratégie s'inspire des travaux présentés par [JM19]. Soit $V = \sum_p Y_P$ la somme des comptages ; la stratégie proposée modélise les comptages multivariés comme suit :

1. La somme V suit une distribution $\mathcal{G}(\mathbf{x}, \boldsymbol{\theta})$;
2. Conditionnellement à $V = v$, Y suit une distribution $\mathcal{H}(v, \mathbf{x}, \boldsymbol{\theta})$ sur le simplexe défini par $\{0, ..., V\}^P$.

Par conséquent, la distribution de probabilité conditionnelle des observations est donnée par :

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = h(\mathbf{y}|v, \mathbf{x}, \boldsymbol{\theta})g(v|\mathbf{x}, \boldsymbol{\theta}),$$

avec g la distribution de la somme et h la distribution du partage. Dans ce qui suit, nous nous concentrerons sur deux modèles « sommes et partages » : Poisson-multinomiale et Binomiale négative-Dirichlet multinomiale.

1. Modèle de régression avec une loi de Poisson sur la somme et une loi multinomiale sur le partage :
   Ce modèle applique une régression de Poisson pour la distribution de la somme $\mathcal{G}(\mathbf{x}, \boldsymbol{\theta})$, et une régression multinomiale pour la distribution conjointe des comptages $\mathcal{H}(v, \mathbf{x}, \boldsymbol{\theta})$. Comme l'indique le Lemme 4.1 de [Zho+12], cette distribution conjointe est la même que celle produite par P variables de Poisson indépendantes avec des paramètres $r_1 = \lambda u_1, ..., r_P = \lambda u_P$, où λ est le paramètre de la distribution de Poisson, et $u_1, ..., u_P$ sont les paramètres de la distribution multinomiale. Ce modèle peut donc être considéré comme une base de comparaison dans notre étude, puisque les composantes $Y_1, ..., Y_P$ sont indépendantes. La moyenne et la variance de Y_p sont $\mathbb{E}(Y_p|\mathbf{x}) = \mathbb{V}(Y_p|\mathbf{x}) = \lambda u_p$; la covariance entre Y_p et $Y_{p'}$ est $\text{Cov}(Y_p, Y_{p'}|\mathbf{x}) = 0$. Dans ce modèle, ni la surdispersion ni la corrélation ne sont traitées.
2. Modèle de régression avec une loi binomiale négative sur la somme et une loi de Dirichlet-multinomiale (ou Pólya) sur le partage :
   Ce modèle introduit des corrélations entre les séries de comptage et modélise mieux les comptages surdispersés. En effet, il est possible de mélanger λ et $u_1, ..., u_P$ sur des distributions pour $\Lambda > 0$ et $U_1, ..., U_P \in (0, 1)$ telles que $U_1 + ... + U_P = 1$. Ici, Λ suit une distribution Gamma, et $U_1, ..., U_P$ une distribution de Dirichlet, de sorte que $\mathcal{G}(\mathbf{x}, \boldsymbol{\theta})$ est une distribution binomiale négative ($\mathcal{NB}$), et $\mathcal{H}(v, \mathbf{x}, \boldsymbol{\theta})$ une distribution Dirichlet-multinomiale (ou de Pólya) ($\mathcal{DM}$). Avec cette spécification, le modèle peut être écrit comme suit :

$$V|\mathbf{x} \sim \mathcal{NB}(\exp(\mathbf{x}^T\boldsymbol{\gamma}), r) \tag{4.4}$$

$$Y|\mathbf{x}, V \sim \mathcal{DM}(V, (\exp(\mathbf{x}^T\boldsymbol{\xi}_p))_{p\in 1,...P}), \tag{4.5}$$

avec r le paramètre de forme et $\boldsymbol{\gamma}$ le vecteur $(D \times 1)$ des paramètres de régression de $\mathcal{NB}$. $\boldsymbol{\xi_p}$ est le vecteur $(D \times 1)$ des coefficients de régression $\mathcal{DM}$ liés aux effets exogènes $\mathbf{x}$. Notons que $\boldsymbol{\theta} = (r, \boldsymbol{\gamma}, \boldsymbol{\xi})$ ici.

Propriétés 1 *Les moments de Y, du modèle binomial négatif-Dirichlet multinomiale décrit dans [JM19], s'écrivent comme suit :*

- $\mathbb{E}(Y_p|\mathbf{x}) = \frac{rq_p}{\frac{r}{k}q_.} = k\frac{q_p}{q_.}$
- $\mathbb{V}(Y_p|\mathbf{x}) = \frac{rq_p}{(\frac{r}{k})^2 q_.^2(1+q_.)}\left[q_.\{r + 1 + (1 + q_.)\frac{r}{k}\} + (q_. - r)q_p\right]$
- $\text{Cov}(Y_p, Y_{p'}|\mathbf{x}) = \frac{r(q_.-r)q_pq_{p'}}{(\frac{r}{k})^2(q_.)^2(1+q_.)}$,

avec $k = \exp(\mathbf{x}^T\boldsymbol{\gamma})$, $q_p = \exp(\mathbf{x}^T\boldsymbol{\xi}_p)$ et $q_. = \sum_{p=1}^P q_p$.

D'après l'expression de la variance, nous pouvons voir que la surdispersion est prise en compte. De plus, les signes des covariances sont les mêmes pour tous les p, p', et dépendent du signe de $q. - r$.

Les modèles de régression Poisson log-normaux

La seconde stratégie consiste à utiliser un modèle hiérarchique à deux couches, avec une couche d'observation modélisant les données de comptage, et une couche cachée qui estime les dépendances entre les comptages en différents lieux Y_p. La distribution de Poisson log-normale multivariée aborde ce problème, en modélisant les données de comptage avec une couche latente gaussienne multivariée, qui est mise à l'échelle exponentielle, avant d'être utilisée pour paramétrer des distributions de Poisson indépendantes. Ce modèle est expliqué dans le travail de [CMR21], et appliqué dans un cadre de clustering dans [Sil+19]. Les avantages de ce modèle résident dans le fait qu'il n'est pas nécessaire de supposer une indépendance entre les séries, et que toute surdispersion peut être prise en compte. Les équations définissant ce modèle sont :

$$\boldsymbol{\varphi}|\mathbf{x} \sim \mathcal{N}(\mathbf{x}^T\boldsymbol{\rho}, \boldsymbol{\Sigma}) \tag{4.6}$$

$$Y|\boldsymbol{\varphi} \sim \mathcal{P}(\exp(\boldsymbol{\varphi})), \tag{4.7}$$

où $\mathcal{N}$ est une distribution gaussienne. Chaque Y est modélisé *via* un vecteur latent gaussien $\boldsymbol{\varphi}$. La moyenne du vecteur latent est une combinaison des covariables $\mathbf{x}$ et de la matrice $(D \times P)$ de paramètres de régression $\boldsymbol{\rho}$. La matrice $(P \times P)$ de covariance $\boldsymbol{\Sigma}$ décrit la structure sous-jacente des dépendances entre les éléments P. Dans cette étude, nous considérons deux cas concernant $\boldsymbol{\Sigma}$. Dans le premier cas, $\boldsymbol{\Sigma}$ est diagonale, ce qui signifie que seules les variances sont estimées, et les covariances sont supposées nulles. Dans le second cas, $\boldsymbol{\Sigma}$ est estimée sans restriction, c'est-à-dire que toutes les covariances sont estimées. Notons que $\boldsymbol{\theta} = (\boldsymbol{\rho}, \boldsymbol{\Sigma})$ ici.

Propriétés 2 *Les moments de Y, tirés d'un modèle Poisson log-normal tel que décrit dans [CMR21], s'écrivent comme suit :*

- $\mathbb{E}(Y_p|\mathbf{x}) = exp\big(\mu_p + \frac{1}{2}\Sigma_{p,p}\big)$
- $\mathbb{V}(Y_p|\mathbf{x}) = \mathbb{E}(Y_p|\mathbf{x}) + exp\big(\mu_p + \frac{1}{2}\Sigma_{p,p}\big)^2\big(exp(\Sigma_{p,p}) - 1\big)$
- $\mathrm{Cov}(Y_p, Y_{p'}|\mathbf{x}) = \mathbb{E}(Y_p|\mathbf{x})\,\mathbb{E}(Y_{p'}|\mathbf{x})\big(exp(\Sigma_{p,p'}) - 1\big)$,

où $\boldsymbol{\mu} = \mathbf{x}^T\boldsymbol{\rho}$.

À partir de ces équations, nous voyons que ce modèle tient compte de la surdispersion. Il prend également en compte les corrélations négatives et positives, lorsqu'il n'y a pas de restriction sur la matrice de covariance.

Résumé des modèles

Le tableau 4.1 répertorie tous les modèles comparés dans notre étude. Chaque modèle est associé à un acronyme utilisé dans le reste du travail. Des éléments sur le nombre de paramètres estimés et la gestion de la surdispersion/corrélation sont également affichés. Les corrélations entre les séries sont traitées de manière distincte. Pour le modèle « NegPol », les corrélations sont capturées par le vecteur de paramètres de la distribution Dirichlet-multinomiale, et leur signe est régi par les paramètres de la binomiale négative. Pour le modèle « PLNfull », c'est la matrice de covariance estimée directement qui capture les corrélations. Nous pensons qu'il n'est pas nécessaire de modéliser l'autocorrélation, en raison de l'accent mis sur une bonne segmentation des données, plutôt que sur la prédiction. Nous ne prenons pas ici en compte les phénomènes d'autocorrélation qui pourraient expliquer certaines variations journalières. Ce type de phénomène pourrait être considéré dans une autre version des modèles proposés.

Acronyme	**Modèle**	**Nombre de paramètres**	**Surdispersion prise en compte**	**Corrélation prise en compte**
PoiMult	Poisson-multinomial	$b1 = D + (D \times (P-1))$	✗	✗
NegPol	Binomiale négative-Pólya	$b2 = D + 1 + (D \times P)$	✓	✓[a]
PLNdiag	Poisson log-normal Σ diagonale	$b3 = D \times P + P$	✓	✗
PLNfull	Poisson log-normal Σ non contrainte	$b4 = D \times P + P^2/2 + P$	✓	✓

TABLE 4.1 – Modèles de mélanges étudiés avec les acronymes, les nombres de paramètres, et la gestion de la corrélation et de la surdispersion. Notons que $b1 < b2 < b3 < b4$ si $D < P + 1$, et $b1 < b3 < b2 < b4$ si $D > P + 1$, avec D le nombre de facteurs exogènes, et P le nombre d'emplacements de comptage.
[a] Toutes les corrélations sont positives ou négatives

4.4.2 Modèles de mélange

Nous considérons maintenant $Y_{j,h}$, une variable sous la forme d'un P-vecteur des comptages pour la tranche h du jour j. Nous supposons que chaque jour j peut être associé à la dynamique d'un segment s, parmi S segments possibles, avec une certaine probabilité. Associer des dynamiques à l'échelle du jour, plutôt qu'à l'échelle de la plage horaire est, à notre avis, un bon moyen de synthétiser l'information sur des périodes longues, telles que

les années. Si nous adaptons les modèles de régression présentés précédemment au cadre des modèles de mélange, nous nous retrouvons avec des modèles génératifs qui incluent les variables indicatrices Z_j ($Z_j \in \{1, \ldots, S\}$) encodant l'appartenance des jours au segment. Le nombre de segments S est choisi *a priori*. Le modèle génératif suivant est adopté pour les données observées :

$$Y_{j,h}|Z_j = s, \mathbf{x}_{j,h}, \boldsymbol{\theta}_s \sim \mathcal{D}(\mathbf{x}_{j,h}, \boldsymbol{\theta}_s), \tag{4.8}$$

avec $\boldsymbol{\theta}_s$ l'ensemble des paramètres contrôlant les distributions conditionnelles au sein du segment s. Ce modèle génératif suppose que, connaissant le segment d'appartenance de la journée et les valeurs des covariables, les comptages à chaque tranche h suivent une distribution de paramètres spécifiques à chaque segment. La variable Z_j suit une distribution multinomiale ($\mathcal{M}$) de paramètre $\boldsymbol{\pi}$ (qui est le vecteur des poids d'association). Nous allons comparer deux façons de modéliser les appartenances aux segments Z_j, comme détaillé ci-dessous :

— Dans le premier schéma, les poids d'association π_s sont fixes, et ne changeront pas au cours des jours.

$$Z_j \sim \mathcal{M}(1, (\pi_s)_{s\in 1,\ldots,S}) \tag{4.9}$$

— Dans le second schéma, désigné comme « lissé », les poids d'association évoluent avec les jours j en suivant une transformation logistique de fonctions splines cubiques avec M nœuds. L'idée derrière ce schéma est d'aider le modèle à détecter les changements de régime, qui sont difficiles à détecter à partir des données, en prenant en compte une relation entre les jours proximaux à travers des fonctions splines. De plus, les fonctions splines cubiques peuvent aider à détecter des jours similaires, mais très séparés dans le temps.

$$Z_j|j \sim \mathcal{M}(1, (\pi_s(j;\alpha))_{s\in 1,\ldots,S}) \tag{4.10}$$

et

$$\pi_s(j;\alpha) = \frac{exp(\sum_{m=1}^{M+4} \alpha_{s,m} a_m(j))}{\sum_h exp(\sum_{m=1}^{M+4} \alpha_{h,m} a_m(j))}, \tag{4.11}$$

avec $\alpha_{s,m}$ un poids à estimer, et

$$a_m(j) = j^{m-1}, m \in \{1, ..., 4\} \tag{4.12}$$

$$a_{m+4}(j) = (j - \kappa_m)^3, m \in \{1, ..., M\}. \tag{4.13}$$

Chaque spline cubique est un polynôme cubique par morceaux, avec les nœuds κ_m, $m \in \{1, ..., M\}$. Dans les notations mathématiques, nous utiliserons les poids d'association « lissés » $\pi_s(j; \alpha)$, afin d'inclure le formalisme de l'estimation des paramètres $\boldsymbol{\alpha}$.

Dans la suite de l'étude, nous utiliserons l'acronyme des modèles, tel que défini dans le tableau 4.1 et précédé d'un « s » pour désigner les versions lissées (par exemple, « sNegPol » est le « NegPol » lissé). Ces modèles nécessitent l'estimation supplémentaire de $(S-1) \times M$ paramètres $\boldsymbol{\alpha}$. Les modèles graphiques, pour les modèles de mélange « sommes et partages » et Poisson log-normaux, sont présentés dans la figure 4.1.

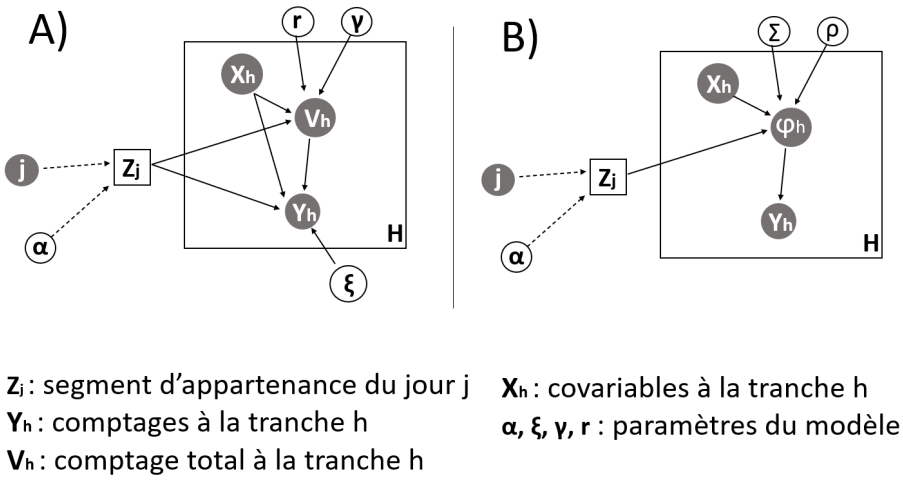


FIGURE 4.1 – Représentations graphiques du modèle de mélange « sommes et partages » (figure A), et du modèle de mélange Poisson log-normal (figure B) . Les cercles gris représentent les données observées. Les cercles blancs sont les paramètres à estimer. Z_j est la couche cachée du modèle génératif. Les liens en pointillés représentent l'intervention potentielle des poids d'association lissés $\pi_s(j; \alpha)$. H est le nombre total de tranches de temps pour le jour j.

4.4.3 Estimation des paramètres

Les paramètres du modèle sont estimés par la méthode du maximum de vraisemblance, résolue par l'algorithme Espérance-Maximisation (EM) ([DLR77]), comme expliqué dans l'algorithme 1 (le texte en bleu y représente les estimations du modèle « lissé »). Les détails des différentes étapes de l'estimation des paramètres sont développés dans l'annexe C.3, pour les modèles de mélange « sommes et partages », et dans l'annexe C.4, pour le modèle

de mélange Poisson log-normal. Pour chaque estimation des modèles de S segments, nous avons effectué cinq essais, chacun avec une initialisation de segment comme expliqué dans l'algorithme 2. La façon dont les paramètres sont initialisés est cruciale pour permettre à l'algorithme EM de converger plus rapidement et de fournir des solutions acceptables. La même procédure d'initialisation a été mise en œuvre pour les deux modèles, en utilisant un clustering ascendant hiérarchique (HAC) (voir Algorithme 2). Cette initialisation non aléatoire a été choisie en particulier pour le modèle de mélange de Poisson log-normal, car une initialisation aléatoire aurait impliqué de très grands calculs de covariance Σ_s sur des données hétérogènes, ce qui aurait rendu difficile l'identification de segments homogènes. Pour chaque expérience, nous utilisons un ensemble de cinq jours par segment, choisis au hasard, pour l'initialisation des paramètres, car nous considérons que chaque expérience ne doit pas avoir exactement le même point de départ, afin de permettre aux modèles de mélange de trouver potentiellement des solutions différentes [LG07]. Les itérations successives de l'algorithme EM peuvent mener à la disparition d'un segment, surtout lorsque S est grand. C'est pourquoi il est utile de lancer plusieurs fois l'algorithme EM pour chaque expérience. De plus, nous avons ajouté une étape de "hot restart" dans la procédure EM, afin de gérer la disparition des segments. Cette étape consiste à réinitialiser le segment disparu, en utilisant les jours (d'autres segments) ayant les plus petites valeurs de $\tau_{j,s}$ (voir équations C.8 et C.19), c'est-à-dire les jours qui sont les moins susceptibles d'appartenir à ces segments. Chaque expérience est arrêtée lorsque la différence de décroissance entre deux log-vraisemblances successives est inférieure à un seuil donné, que nous avons fixé à 10^{-6}. Le modèle présentant la meilleure log-vraisemblance est finalement choisi. Tous les modèles ont été construits dans l'environnement R, en utilisant la fonction glm du paquet stats, la fonction glm.nb du paquet MASS [Rip+13], la fonction multinom du paquet nnet [RVR16], ainsi que les paquets MGLM ([Kim+18]) et PLNmodels [CMR21]. La segmentation temporelle est obtenue en actualisant, à chaque étape de l'algorithme EM, l'espérance conditionnelle d'appartenance des jours aux segments $(\tau_{j,s})_{s=1,\dots,S}$.

Dans les sections suivantes, nous comparons les différents modèles, sur données simulées puis réelles (du cas d'étude de La Défense). Avec les données simulées, l'objectif est d'évaluer la capacité de chaque modèle à bien classer les jours dans des contextes contrôlés. Nous mettrons également en évidence les bonnes performances de ces modèles dans leurs versions lissées (avec évolution des poids de mélange) ou non lissées. Nous appliquerons ensuite les modèles aux comptages réels des flux de personnes à La Défense. Tout d'abord, nous comparerons les différents modèles en fonction de leur capacité à bien modéliser les données, et à détecter des segments continus, avec une variation du nombre de segments. Ensuite, pour le modèle choisi, trois résultats seront détaillés : la segmentation de la période totale en S segments temporels, l'analyse des profils de mobilité typiques au sein de ces segments, et l'impact des variables exogènes.

Algorithme 1 Algorithme EM pour estimer les paramètres $\boldsymbol{\theta}$

Entrées : Y tenseur (J jours $\times$ H tranches de temps $\times$ P lieux), X tenseur (J jours $\times$ H tranches de temps$\times$ D covariables), nombre de segments S.
Sorties : Paramètres estimés ($\boldsymbol{\theta}$), probabilités *a posteriori* $\tau_{j,s}$

Initialisation
Initialiser $\boldsymbol{\theta}^{(0)}$, $\tau^{(0)}$ et $\alpha^{(0)}$ ▷ voir Algorithme 2
c $\longleftarrow$ 0
répéter
    Etape E : calculer les probabilités *a posteriori*
    pour chaque segment $s \in \{1, \ldots, S\}$ **faire**
        Calculer $\tau_{j,s}^{(c)}$ à chaque jour $j \in \{1, \ldots, J\}$ ▷ voir équations C.8 et C.19
    pour chaque segment $s \in \{1, \ldots, S\}$ **faire**
        si le nombre de jours dans le segment $s = 0$ **alors**
            Lancer un "hot restart"
            break
    Etape M : Mise à jour des paramètres
    Calculer $\alpha^{(c+1)}$ ▷ voir équations C.9 et C.18
    pour chaque segment $s \in \{1, \ldots, S\}$ **faire**
        Calculer $\boldsymbol{\theta}_s^{(c+1)}$ ▷ voir équations C.9 et C.18
jusqu'à *convergence*

Algorithme 2 Initialisation pour les paramètres $\boldsymbol{\theta}^{(0)}$ et les probabilités *a posteriori* $\tau^{(0)}$

Entrées : Y tenseur (J jours $\times$ H tranches de temps $\times$ P lieux), X tenseur (J jours $\times$ H tranches de temps $\times$ D covariables), nombre de segments S.
Sorties : Paramètres initialisés ($\boldsymbol{\theta}^{(0)}$), probabilités *a posteriori* $\tau^{(0)}$

Appliquer un clustering ascendant hiérarchique (CAH) sur Y, les individus étant les jours, afin d'associer chaque jour j à l'un des S clusters (segments) basé sur Y_j, le vecteur ($H \times P$) de comptages.
pour chaque segment $s \in \{1, \ldots, S\}$ **faire**
    Échantillonner aléatoirement 5 jours j du segment s
    Calculer $\boldsymbol{\theta}_s^{(0)}$ en se basant sur ces jours, à travers une passe de l'étape M
    Calculer $\tau_{j,s}^{(0)}$ avec un *a priori* uniforme pour $\pi_s(j; \alpha)$ (ou π_s) pour chaque jour $j \in (1, ..., J)$ ▷ voir équations C.8 et C.19
Calculer $\alpha^{(0)}$

4.5 Résultats

4.5.1 Expérimentations sur données simulées

L'objectif de travailler avec des données simulées est double : évaluer les capacités des modèles à classer correctement les jours provenant de séries temporelles soumises à des changements de régime (globaux ou locaux) contrôlés ; étudier l'impact du nombre de nœuds M pour les modèles lissés. Nous testons ainsi des modèles avec $M = 5, 20, 50$ et 80 nœuds. Le protocole de génération des données est le suivant : nous créons $P = 5$ séries de comptages, soumises à $S = 3$ changements de régime pendant une période de deux cents jours. Ces séries sont générées à partir d'un modèle PLNdiag ou PoiMult (voir tableau 4.1), dont les paramètres sont spécifiés *a priori* (c'est-à-dire $\boldsymbol{\mu}$ pour PLNdiag, λ et $\mathbf{u}$ pour PoiMult). Pour le modèle de simulation PLNdiag, nous avons limité les valeurs de variance $\boldsymbol{\Sigma}_{\boldsymbol{p},\boldsymbol{p'}}$ à 1×10^{-3}. Les changements de régime sont générés en augmentant ou en diminuant les valeurs de comptages avec un taux moyen de d%. Notons que ce taux est légèrement différent d'une série de comptages à l'autre. Comme le montre la figure 4.2, nous générons trois segments selon le protocole suivant :

1. Le segment 1 comprend les jours 1 à 60, et se caractérise par une augmentation moyenne de d% des comptages.
2. Le segment 2 comprend les jours 61 à 100 et 181 à 200, et n'est affecté par aucun changement.
3. Le segment 3 comprend les jours 101 à 180, et se caractérise par une diminution moyenne de d% des comptages.

Un changement de régime peut avoir un impact sur toutes les séries temporelles (c'est-à-dire un impact global) ou sur une seule série (c'est-à-dire un impact local). Un exemple de données simulées de comptage, générées selon un modèle PoiMult, est présenté dans la figure 4.2.

La série d'expérimentations suivantes vise à étudier l'impact du taux d de changement des comptages sur les capacités de segmentation de quatre modèles : PoiMult, sPoiMult, PLNdiag et sPLNdiag (voir tableau 4.1). Dans un premier temps, l'impact des changements de régime sera global (sur toutes les séries), puis nous testerons un impact local (sur une seule série). Chaque expérience implique qu'un modèle soit calibré sur un ensemble de données simulées avec un taux spécifique d. Dans chaque expérience, le modèle est testé avec $S = 2$, 3 et 4 segments, en espérant que le modèle avec le nombre de segments $S = 3$ utilisé pour générer les données simulées soit le plus performant. En raison de la nature de l'initialisation du modèle (voir Algorithme 2), chaque test est répété trois fois. Ce processus de génération de données est effectué 20 fois, chaque fois avec un nouvel ensemble de données simulées. Deux critères sont estimés pour comparer les capacités des modèles : le taux d'erreur de segmentation des jours, et le pourcentage de fois où le modèle à 3 segments était le meilleur selon le *critère d'information bayésien* (BIC) [Sch78] (voir

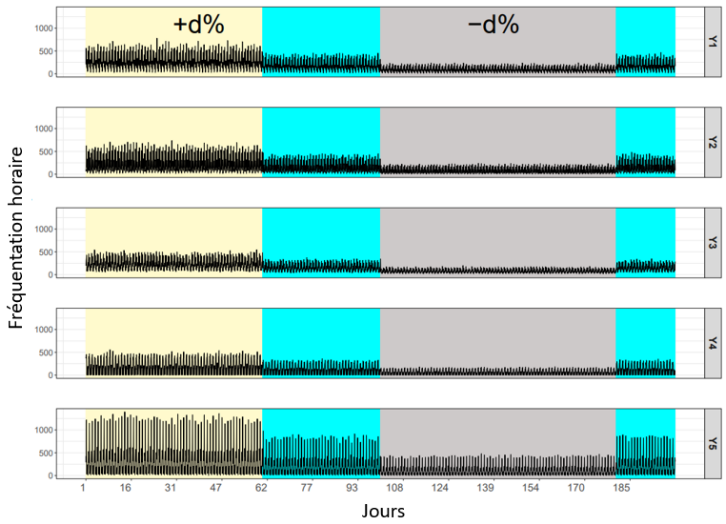


FIGURE 4.2 – Données simulées sur 3 segments avec des taux de changement des données $d = 50$ et le modèle PoiMult. Une différence est spécifiée *a priori* pour les ensembles λ et u à chaque h. Les couleurs jaune, cyan et gris correspondent respectivement aux segments 1, 2 et 3.

annexe A). Pour les expériences ayant un impact global, les résultats sont présentés dans les figures 4.3 et 4.4.

Comme prévu, plus les segments sont différents, en d'autres termes plus les taux d sont élevés, plus la segmentation est aisée pour tous les modèles. Cela se voit dans les taux plus faibles de mauvaise classification des jours (voir la figure 4.3), et dans la sélection plus fréquente de modèles à 3 segments (voir la figure 4.4). Il semble ensuite que les versions lissées avec un faible nombre de nœuds (M=5, 20) donnent de meilleurs résultats pour les deux critères. Ce résultat a été obtenu dans le contexte de données simulées, faiblement bruitées et à faible dimension, mais indique néanmoins un avantage potentiel des modèles lissés avec peu de nœuds par rapport aux versions non lissées.

Pour les expériences avec un impact local, nous avons choisi un taux $d = 10$. Les résultats sont présentés dans le tableau 4.2. Dans ce cas difficile à détecter, seuls les modèles sPoiMult avec moins de nœuds (c'est-à-dire 5 ou 20) semblent réussir à considérer la version à trois segments comme la meilleure. Les modèles « sommes et partages » semblent avoir un avantage sur les modèles de Poisson log-normaux dans cette situation, lorsqu'on considère les deux critères. De plus, les modèles lissés avec moins de nœuds semblent mieux classer les jours.

Tous ces résultats, obtenus sur données simulées, démontrent l'intérêt de considérer des modèles de mélange « sommes et partages » ainsi que des versions lissées, pour catégoriser correctement les données de comptage soumises à des changements de régime. L'étude des

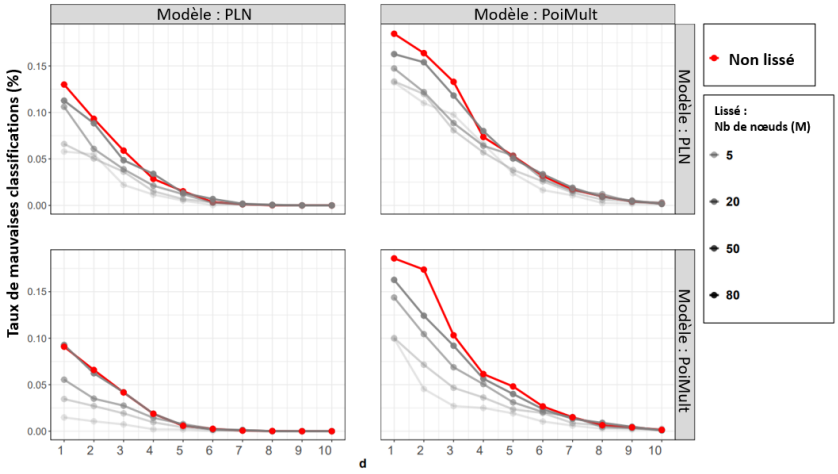


FIGURE 4.3 – Taux d'erreur de segmentation pour le modèle PoiMult et les modèles PLN, avec les modèles à 3 segments. Les graphiques montrent l'évolution de l'impact de d : une fois que d tombe en dessous de 10%, le taux d'erreur de segmentation augmente (différemment selon le modèle).

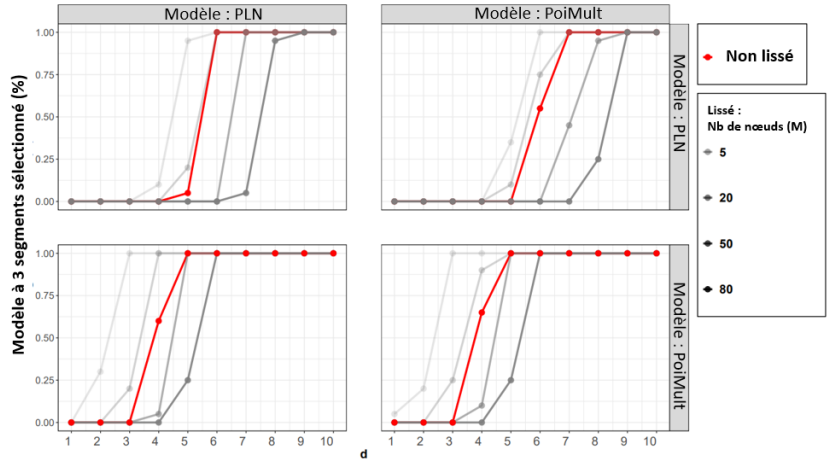


FIGURE 4.4 – Pourcentage de fois où le modèle à 3 segments était le meilleur selon le BIC, en fonction de d.

modèles sur des données réelles permettra d'approfondir ces conclusions, valables dans un cas simple à faible bruit. Le code source du script R, et de l'application sur des données simulées, est disponible à l'adresse https://github.com/pdenailly/segmentation_models.

TABLE 4.2 – Expérimentations avec un impact local d = 10. Les modèles sPoiMult avec peu de noeuds sont meilleurs que les autres modèles, selon les deux critères.

Modèle	**Génération : PoiMult**		**Génération : PLNdiag**	
	Taux err. (%)	Sel. 3 segments (%)	Taux err. (%)	Sel. 3 segments (%)
PoiMult	0.05	0	0.03	0.15
sPoiMult (5)	**0.01**	**1**	**0**	**1**
sPoiMult (20)	**0.01**	**1**	**0.01**	**1**
sPoiMult (50)	0.03	0.05	0.02	0
sPoiMult (80)	0.04	0	0.03	0
PLN	0.08	0	0.05	0
sPLN (5)	0.06	0	0.02	0
sPLN (20)	0.04	0	0.02	0
sPLN (50)	0.08	0	0.04	0
sPLN (80)	0.08	0	0.04	0

4.5.2 Expérimentations sur les données de mobilité

Cette section compare les différents modèles de mélange de régressions (voir tableau 4.1), en utilisant les données de comptage de personnes captées au pôle de transport de La Défense. L'objectif est d'explorer la capacité des modèles sur deux bases : (i) leur capacité à s'ajuster aux données de comptage grâce au critère BIC, et (ii) leur capacité à détecter les segments propres grâce au critère d'entropie [CS96]. Les critères sont calculés pour chaque modèle, pour un nombre donné de segments S. Les mêmes facteurs exogènes sont utilisés dans tous les modèles. Les facteurs exogènes intègrent le vecteur $\mathbf{x}_{j,h}$ tel que présenté dans le tableau 4.3. Comme indiqué dans la section sur les données simulées, des versions « lissées » (avec évolution des poids de mélange) de chaque modèle seront appliquées. Ces versions lissées considéreront $M = 25$ nœuds, ce qui est faible par rapport au nombre de jours de l'étude de cas (> 1 an de données).

TABLE 4.3 – Variables exogènes

Position	**Nom**	**Description**
$\mathbf{x}_{j,h}^{1,\dots,8}$	$\mathsf{heure}_{j,h}$	splines cubiques avec 8 degrés de liberté sur les tranches de 1 heure de h = 7h à h = 0h
$\mathbf{x}_{j,h}^{9}$	$\mathsf{concOut}_{j,h}$	Encodage pour les tranches après un concert : 1 s'il y a un concert au jour j et h = 23h
$\mathbf{x}_{j,h}^{10}$	$\mathsf{concIn}_{j,h}$	Encodage pour les tranches avant un concert : 1 s'il y a un concert au jour j et h se situe entre 16h et 22h
$\mathbf{x}_{j,h}^{13}$	$\mathsf{disturbanceRERmorn}_{j,h}$	Transformation log de la durée totale de la perturbation du RER, au cours du pic du matin (7h à 9h)
$\mathbf{x}_{j,h}^{14}$	$\mathsf{disturbanceRER}_{j,h}$	Transformation log de la durée totale de la perturbation du RER, après le pic du matin (9h à 0h)

Comparaison des capacités de modélisation

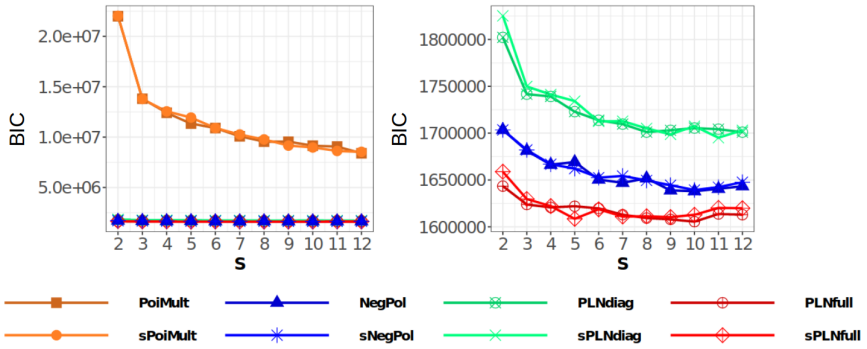


FIGURE 4.5 – Critère BIC calculé pour tous les modèles de mélange, sur l'ensemble de données de comptage de La Défense, et pour $S \in \{2, ..., 12\}$. La figure de gauche inclut tous les modèles. La figure de droite n'inclut pas les modèles de base (PoiMult et sPoiMult), afin de mieux visualiser le critère BIC pour les autres modèles.

Le critère BIC (voir annexe A), calculé pour les différents modèles, est affiché dans la figure 4.5. Comme prévu, les modèles PLNfull et NegPol, dans leurs versions lissées et non lissées, sont plus performants que les autres modèles en termes de critère BIC, car ils gèrent la surdispersion ainsi que la covariance (voir tableau 4.1). Les modèles PLNfull et sPLNfull semblent toutefois meilleurs que NegPol et sNegPol, selon ce critère calculé sur toute la période (figure 4.5). Une deuxième façon d'analyser les résultats de ces quatre modèles est de calculer les critères sur des périodes connues comme homogènes a priori. Les résultats concernant la log vraisemblance, le critère BIC et le nombre de segments détectés, calculés avec des modèles à $S = 10$ segments, sont présentés dans le tableau 4.4. Les modèles NegPol et sNegPol sont meilleurs ici, sauf pour la période bruitée de la grève de décembre 2019. Les vraisemblances calculées par les différents modèles sont proches, mais NegPol et sNegPol sont plus performants en termes de critère BIC (sauf pour la période de grèves), car ils nécessitent moins de paramètres, comme le montre le tableau 4.1. En outre, moins de segments sont nécessaires avec ces modèles.

Ce dernier point nous renvoie à notre objectif d'identifier une segmentation plus « propre », c'est-à-dire avec des segments aussi continus que possible. Pour mesurer cette continuité, nous introduisons le critère d'entropie ([CS96]) sur les poids d'association $\pi_s(j, \boldsymbol{\alpha})$, qui se calcule comme suit :

$$-\sum_s \sum_j \pi_s(j, \boldsymbol{\alpha}) log(\pi_s(j, \boldsymbol{\alpha})).$$

Nous ne calculons ce critère que sur des modèles lissés, en raison de l'évolution dyna-

TABLE 4.4 – Log-vraisemblance ($L_Y()$), critère BIC et nombre de segments détectés, sur quatre périodes bien connues. Ces périodes sont : une période dite « normale » (début 2019), la période du premier confinement contre la pandémie de Covid19, une période de grèves, et une période de travaux sur la ligne de tramway T2.

Période	**PLNfull**		
	#seg	LL	BIC
Normale	5	-141565	316611
Grève	1	-21598	**49893**
Confinement	2	-21256	55905
Travaux tram	2	-24432	62257

Période	**sPLNfull**		
	#seg	LL	BIC
Normale	5	-143043	321339
Grève	2	-20424	54948
Confinement	2	-21618	57336
Travaux tram	2	-25407	64915

Période	**NegPol**		
	#seg	LL	BIC
Normale	4	-143829	304286
Grève	2	-22031	52376
Confinement	2	-22163	**52641**
Travaux tram	1	-28007	60171

Période	**sNegPol**		
	#seg	LL	BIC
Normale	2	-144822	**298430**
Grève	2	-21732	52251
Confinement	2	-22152	53091
Travaux tram	1	-27733	**59860**

mique de $\pi_s(j, \boldsymbol{\alpha})$, et de la meilleure performance des modèles lissés. Les valeurs d'entropie sont affichées dans la figure 4.6, pour les modèles sNegPol et sPLNfull. Les entropies sont plus petites pour le modèle sNegPol, ce qui souligne la capacité de ce modèle à détecter plus de segments continus que le modèle sPLNfull.

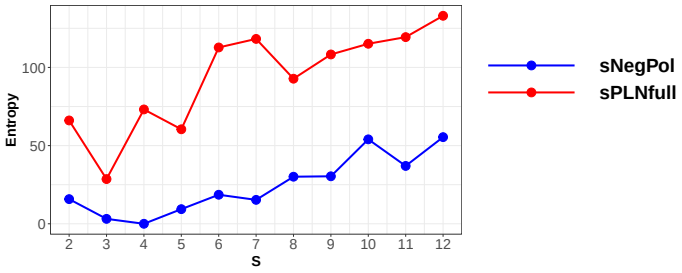


FIGURE 4.6 – Critère d'entropie calculé pour les modèles sNegPol et sPLNfull, sur le jeu de données de comptage de La Défense, et pour $S \in \{2, ..., 12\}$.

Les modèles de Poisson log-normaux, en estimant les variances au sein des segments, permettent l'affectation d'un plus grand nombre de jours distincts aux mêmes segments, pas nécessairement continus. Par rapport aux modèles de Poisson log-normaux, les modèles « sommes et partages » semblent avoir une meilleure capacité à résumer les données en segments continus. NegPol et sNegPol offrent un compromis raisonnable entre la capacité de détecter des segments continus, et celle de traiter des données de comptage surdispersées

et corrélées.

Pour ces raisons, et les nombreux avantages trouvés dans les versions lissées avec les données simulées, nous nous concentrerons dans la section suivante sur les résultats associés au modèle de mélange (sNegPol) lissé, binomial négatif sur la somme, et Pólya pour le partage, avec $S = 10$ segments (d'après le critère BIC).

Résultats de segmentation obtenus avec le modèle choisi

Segmentation temporelle

La segmentation temporelle obtenue avec le modèle est présentée dans la figure 4.7. Nous pouvons observer une richesse des segments induits par divers changements de contexte, tels que des travaux de maintenance, des grèves, ou des mesures sanitaires contre la pandémie de Covid19. Une grande diversité de segments, avec peu de « retours »[1], souligne la nécessité pour les opérateurs urbains de s'adapter à une situation qui change régulièrement. Nous détaillons chaque segment dans le tableau 4.5 (voir la colonne *Caractéristiques de la période*)[2]. Dans la figure 4.8, nous associons ces segments à des flux totaux typiques dans le pôle, obtenus avec l'espérance de la binomiale négative pour chaque segment ($\mathbb{E}(V|\mathbf{x}, Z_{j=s}) = \exp(\mathbf{x}^T\gamma_s)$). On comprend que les flux totaux ont largement diminué depuis le début de la pandémie de Covid19, ce qui est visible dans tous les segments au-delà du segment *Premier confinement*. Ce résultat souligne, qu'au moment de la rédaction de ce travail, l'utilisation du hub n'est pas revenue à la normale (c'est-à-dire « Normal 2019 » et « Fermeture d'une sortie » dans la figure 4.7), depuis le début de la pandémie de Covid19.

Répartitions caractéristiques des flux de personnes

D'un point de vue « répartition », on peut étudier la distribution caractéristique des flux de personnes dans le pôle de transport, au sein de chaque segment. En effet, chaque segment s est associé à un ensemble de distributions typiques parmi les P lieux $\mathbf{u}^{(s)}_{j,h} = ((u^{(s)}_{j,h,p})_{p\in 1,\dots P})$, affichées dans la figure 4.9[2]. Selon les segments obtenus, on constate une sur-utilisation (en rouge) ou une sous-utilisation (en bleu) à certains endroits, par rapport à « Normal 2019 ». Par exemple, dans le segment « Grève / travaux RER A », on observe une sur-utilisation des accès vers le métro (M, EM), et une sous-utilisation des accès vers et depuis le RER (E_O, E_I) et les lignes régionales (R_O, R_I), mettant en évidence un transfert attendu des usagers vers la ligne de métro lorsque les autres transports ferrés sont arrêtés.

Si l'on considère à la fois les flux totaux (figure 4.8) et les distributions spatiales (figure

1. On parle de « retour » lorsque le segment est détecté en plusieurs périodes distinctes.

2. Pour ces sections, seuls les $\mathbf{x}^{1,\dots,8}_{j,h}$ (voir tableau 4.3) sont utilisés pour le calcul des résultats, afin d'exclure les effets non calendaires. Ainsi, les profils totaux et les répartitions spatiales sont invariants par jour.

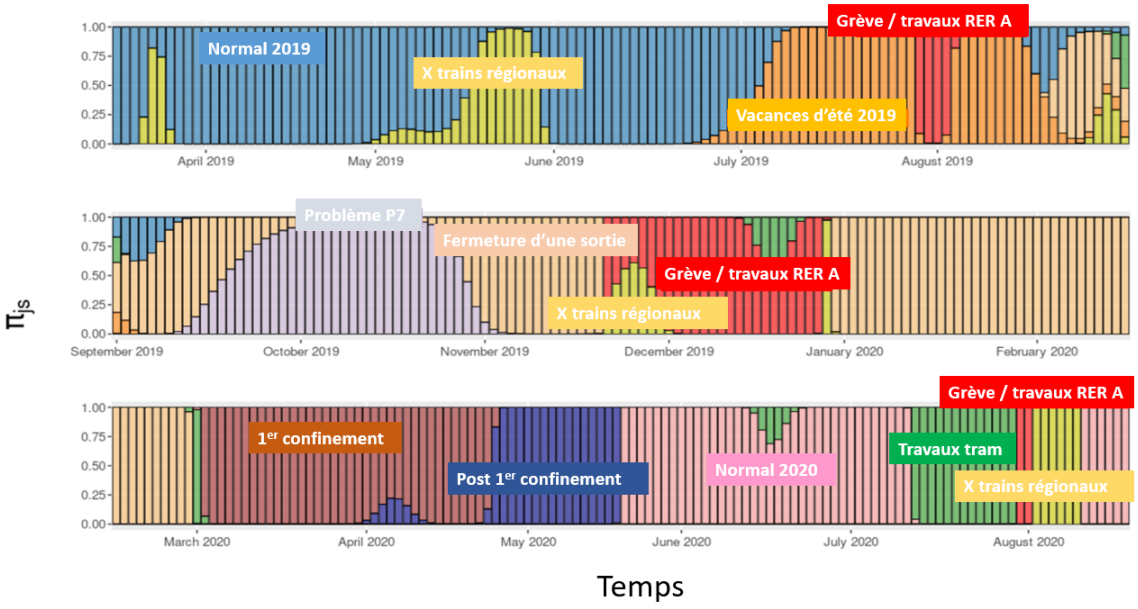


FIGURE 4.7 – Histogramme de la segmentation temporelle. Chaque jour est associé aux probabilités d'appartenir à chaque segment. Chacun des S (=10) segments a sa propre couleur et son propre label.

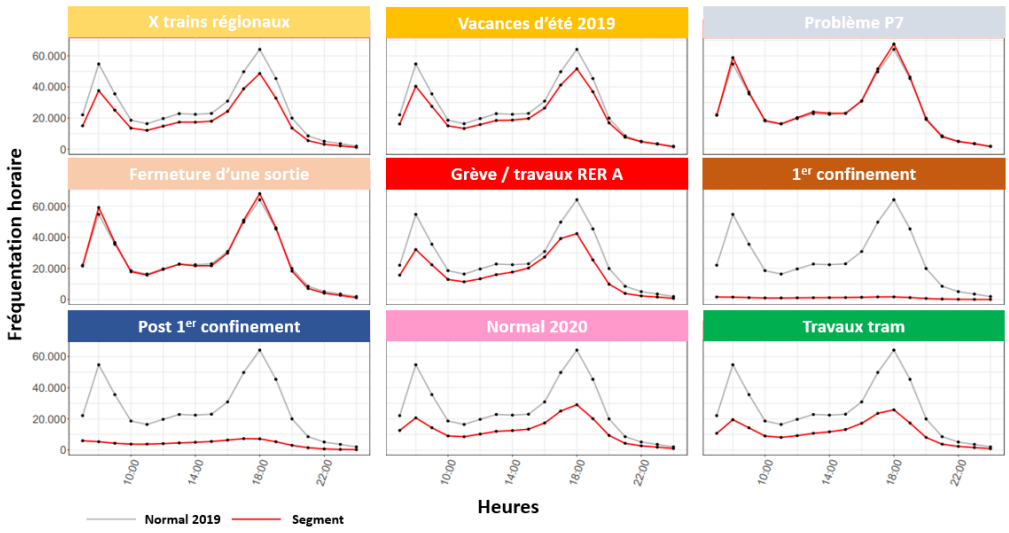


FIGURE 4.8 – Profils typiques des flux totaux obtenus dans chaque segment s. Chaque profil est comparé à celui de *Normal 2019*, c'est-à-dire le segment de référence (en gris).

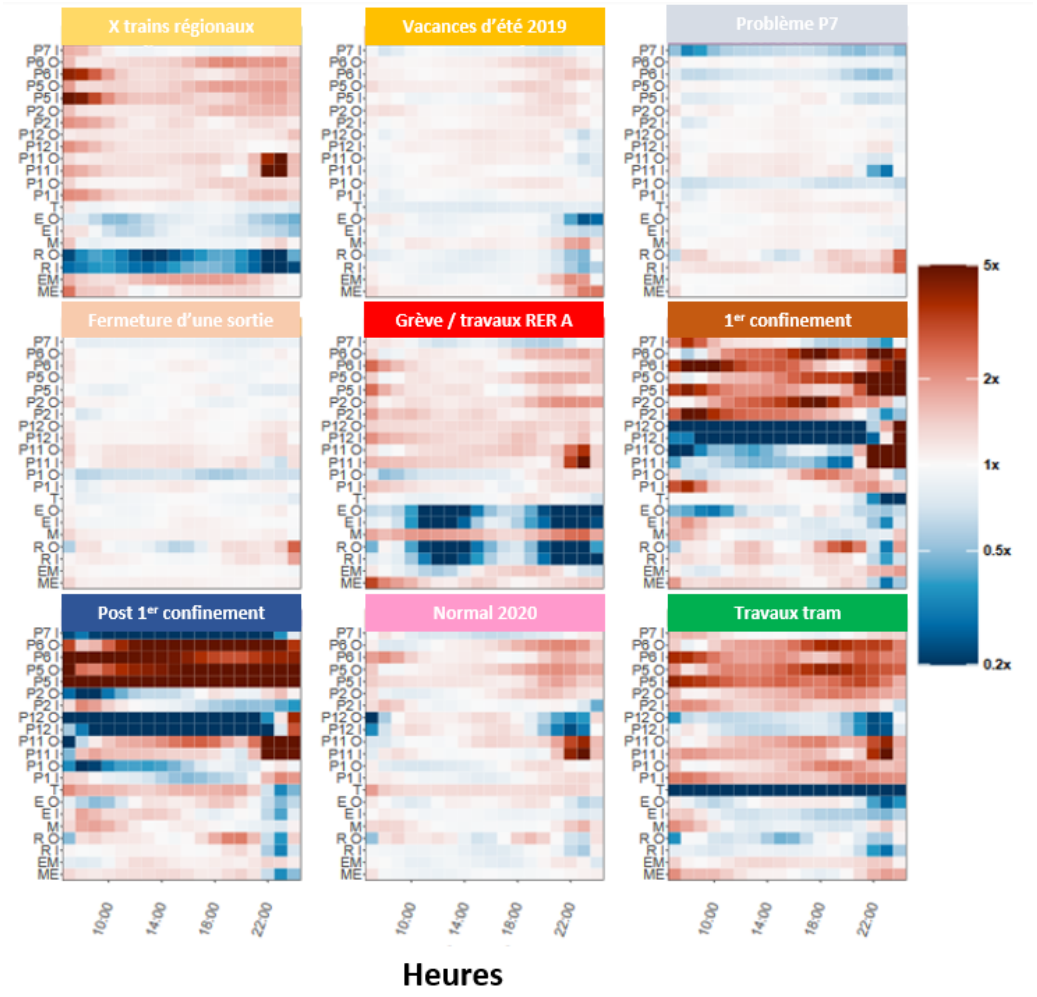


FIGURE 4.9 – Répartitions spatiales typiques parmi les P (=21) emplacements trouvés dans chaque segment s. Pour rappel, "O" correspond à un flux sortant, et "I" à un flux entrant. Chaque cellule des cartes de chaleur correspond à une tranche horaire et à un endroit donné. Pour une cellule donnée, la couleur reflète le logarithme du rapport entre la proportion de flux dans le segment courant, et celle dans le segment de référence *Normal 2019*. Les couleurs reflètent donc les différences entre les proportions de flux de chaque segment et celles du segment de référence, en ce qui concerne la distribution spatiale.

4.9), on peut voir que les segments temporels sont généralement reconnaissables à travers les profils totaux et les distributions spatiales. Deux d'entre eux (*Problème P7* et *Fermeture d'une sortie*), cependant, semblent se distinguer principalement par les distributions spatiales uniquement. Une description des distributions spatiales est présentée dans le tableau

4.5.

TABLE 4.5 – Segmentation temporelle. Description de profils types. Les descriptions sont faites en comparaison avec le profil de référence *Normal 2019*.

Nom	**Caractéristiques de la période**	**Flux total et distribution spatiale**
Normal 2019	Début 2019 sans les vacances d'été. Période de référence.	
X trains régionaux	Regroupe les périodes pendant lesquelles les échanges entre les lignes de Transilien et de RER diminuent fortement. Un problème de voie en juin 2019 interrompant la ligne U, la grève de décembre 2019, et d'importants travaux de maintenance sur la ligne RER en août 2020.	Forte baisse des flux entrants et sortants vers et depuis l'accès aux lignes Transilien (R)
Vacances été 2019	Juillet et Août 2019	Diminution des flux totaux. Pas de différences majeures dans les distributions spatiales.
Fermeture d'une sortie	Fermeture d'une sortie des lignes Transilien vers l'esplanade	Aucun changement dans les flux totaux. Pour les distributions spatiales, les différences sont à peine visibles, à l'exception d'une diminution du nombre de passages par la sortie P1 et de quelques modifications de l'accès R.
Problème P7	Les conditions ici sont les mêmes que celles du segment *Closure of an exit*	Pas de différences pour les flux totaux. Diminution inexpliquée de l'utilisation de la sortie P7 ; il peut s'agir d'une période pendant laquelle le capteur de comptage a eu un problème technique. Les autres lieux subissent les mêmes conséquences que dans le segment *Closure of an exit*.
Grèves / travaux RER A	Période caractérisée par une mobilisation massive contre la réforme du système de retraite français, et quelques périodes pendant lesquelles des travaux de maintenance ont eu lieu sur la ligne RER à Paris.	Forte diminution de l'utilisation du RER (entrants et sortants) et de l'utilisation des lignes Transilien, en raison de leur exploitation partielle ou du manque de transits.
1er confinement	Première période de confinement due à la pandémie de Covid19.	Perte presque totale des flux totaux. L'accès Westfield Les 4 Temps (P12) a été fermé, et n'a donc pas été utilisé. En raison du très petit nombre de personnes qui ont visité le hub pendant cette période, il est difficile de relier les changements importants dans l'utilisation des autres accès avec les changements réels dans le comportement de choix d'itinéraire.
Post 1er confinement	Première étape de la levée du confinement, avec des restrictions.	Perte presque totale des flux totaux. Perte totale aux points d'accès au centre commercial Westfield Les 4 Temps (P12) et à l'esplanade piétonne centrale (P7). Augmentation conséquente de l'utilisation des accès nord (P5) et sud (P6) de l'esplanade.
Normal 2020	Période sans couvre-feu ni confinement en 2020, une période de "retour à la normale".	Forte baisse des flux totaux. Diminution de l'utilisation des accès du centre commercial Westfield Les 4 Temps (P12) en soirée.
Travaux tram	Périodes de travaux de maintenance sur la ligne de tramway.	Perte totale des flux entrants à l'accès au tramway (T).

Impact de facteurs non calendaires

L'impact des facteurs non liés au calendrier est analysé par une comparaison, avec et sans facteur, des profils types et des distributions spatiales. Nous avons donné la priorité à la compréhension de l'impact de ces facteurs dans des conditions normales, c'est-à-dire dans le segment *Normal 2019*. Pour chaque variable exogène, nous étudions le profil standard des flux totaux, avec et sans l'impact du facteur atypique. Pour toutes les variables, une carte de chaleur est produite pour comparer, dans un même segment, les distributions $\mathbf{u}_{j,h}^{(s)}$ du modèle construit avec le facteur atypique, et les $\mathbf{u}_{j,h}^{(s)}$ du modèle sans ce facteur, en utilisant les log-ratios. Dans ces cartes de chaleur, la hauteur de chaque cellule est proportionnelle au comptage moyen du lieu et de la tranche correspondants, sur l'ensemble de la période d'étude. L'impact des concerts et des perturbations des transports est détaillé ci-dessous.

1. L'impact d'un concert sur l'utilisation de la station « La Défense Grande Arche » est représenté sur la figure 4.10. Comme prévu, les profils-types montrent une augmentation de la fréquentation totale au cours de l'après-midi, car les usagers arrivent au pôle de transport en prévision du concert. Lorsqu'un concert a lieu, il y a également un pic d'entrées à 23h, qui représente les personnes quittant le concert et prenant les transports en commun. D'un point de vue spatial, les entrées vers le hub qui sont proches de la salle de concert ($P1_I$ et $P2_I$) sont privilégiées. Le point d'entrée de la ligne de métro (M) est également plus utilisé, par rapport à la référence, ce qui n'est pas le cas de l'accès à la ligne de RER (E_I). Nous supposons que les personnes quittant le concert et souhaitant emprunter le RER préfèrent une autre station (Nanterre-Préfecture), plus proche de la salle de concert. Les personnes souhaitant emprunter la ligne de métro n'ont pas d'autre choix que la station « La Défense Grande Arche », puisqu'il s'agit du terminus, et sont donc plus susceptibles d'être détectées dans notre étude.
2. Les perturbations du RER ont un effet substantiel sur l'utilisation des espaces du pôle de transport, comme le montre la figure 4.11. Le modèle sélectionné met en évidence le phénomène de retard pour les arrivées du matin (c'est-à-dire une baisse de fréquentation sur les tranches horaires du matin). Comme prévu, ce phénomène n'est pas visible sur la pointe du soir, car les personnes sont déjà présentes sur le pôle de transport. Il y a un impact sur les transferts entre les lignes de RER et de métro (voir ME et EM). La perturbation pendant la période de pointe du matin tend à augmenter les flux piétons entre le métro et le RER (c'est-à-dire que les gens arrivent en métro, puis sortent du pôle de transport par la station RER). Elle augmente également l'accès du RER au métro, les personnes se reportant sur la ligne de métro pour quitter le pôle et se rendre à Paris. La ligne de RER est fortement impactée, surtout le matin (E_O), en raison de l'absence de tous les usagers qui n'ont pas réussi à atteindre le pôle de transport.

Ces deux résultats d'impacts de facteurs non calendaires sur les profils types d'affluences sont cohérents avec ce qui est observé dans la réalité. Interpréter ces résultats permet de

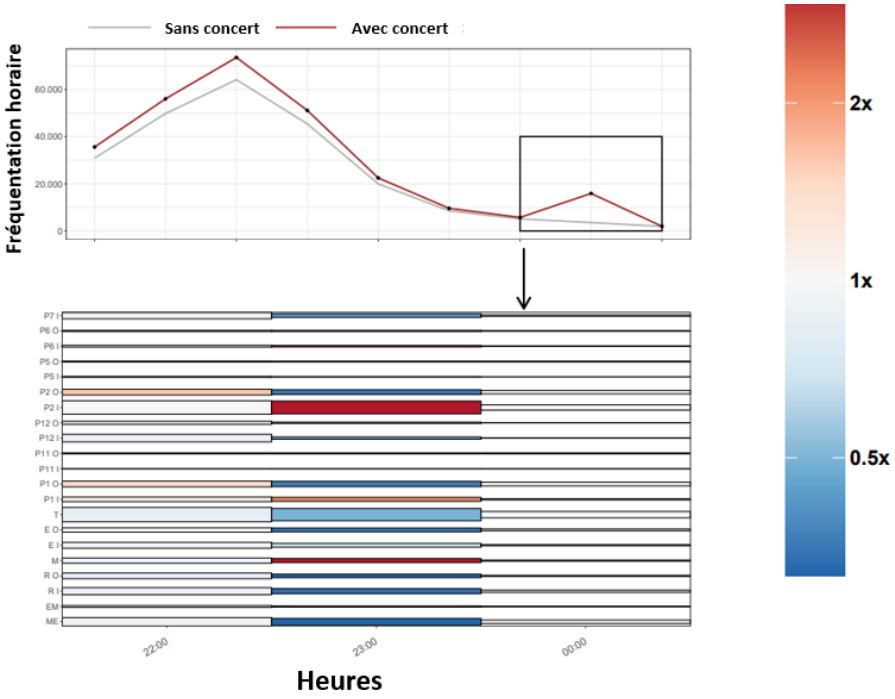


FIGURE 4.10 – Profils typiques et distribution spatiale. Comparaison, avec et sans concert, dans le segment *Normal 2019*.

quantifier et de détailler l'impact des différents événements sur les affluences dans l'espace de transport. Pour un gestionnaire de ces espaces, cela peut permettre d'identifier des situations "types" qui mènent à la saturation de l'un ou l'autre des points d'accès et ainsi lui donner la possibilité de savoir où et quand gérer efficacement les flux piétons (par exemple via l'ouverture de bornes d'accès, la proposition de sorties alternatives, etc.).

4.6 Conclusion

Ce travail nous a permis de mettre en place des modèles statistiques, pour segmenter des séries temporelles multidimensionnelles de mobilité, dont la dynamique est caractérisée par des changements de régime. Deux stratégies inspirées de la littérature, à savoir les modèles « sommes et partages » et les modèles Poisson log-normaux, sont comparées pour cette tâche, tant en termes de vraisemblance que de cohérence des segments. Chaque stratégie présente des avantages et des inconvénients. Par exemple, le modèle Poisson-multinomial ne peut pas prendre en compte les surdispersions, ni les corrélations, contrairement aux autres modèles. Le modèle binomial négatif - Dirichlet multinomial peut prendre en compte les corrélations, mais elles seront toujours du même signe (soit positives, soit négatives).

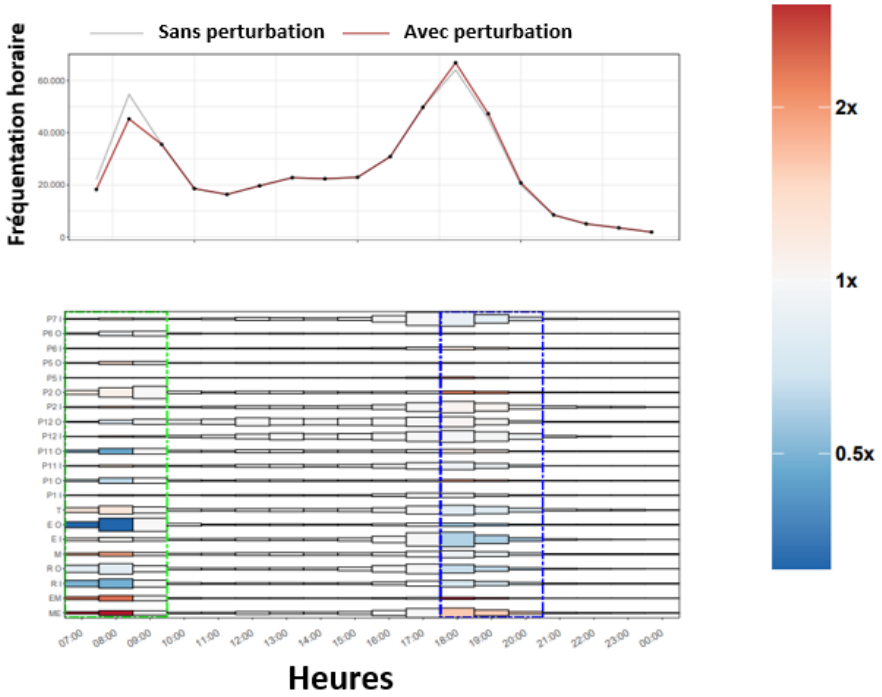


FIGURE 4.11 – Profils temporels et distributions spatiales typiques. Comparaison, avec et sans perturbations du RER A, dans le segment *Normal 2019*. Deux perturbations distinctes sont modélisées ici : une pendant la pointe du matin (rectangle vert), et une pendant la pointe du soir (rectangle bleu).

Les modèles de Poisson log-normaux semblent être plus flexibles, et s'adaptent mieux aux données observées. Les modèles « sommes et partages » semblent mieux détecter les segments continus, ce qui correspond davantage à la réalité de notre étude de cas.

De plus, il y a des avantages à utiliser des régressions logistiques de fonctions splines, pour exprimer la probabilité que chaque jour appartienne à un segment donné. Ce formalisme semble fournir au modèle une meilleure capacité de détection des événements localisés et/ou à faible impact. Nous avons choisi d'appliquer un modèle de mélange lissé, avec une binomiale négative pour la somme, et une Dirichlet-multinomiale pour le partage, pour analyser les données de mobilité collectées dans le pôle de transport de La Défense. Les coefficients de régression de ces modèles dépendent des segments auxquels ils appartiennent. De plus, un ensemble d'événements atypiques a été incorporé au modèle pour que leurs impacts soient étudiés : nous avons ainsi considéré les concerts, et les perturbations du RER A.

Sur le plan opérationnel, ce travail révèle comment les diverses restrictions visant à lutter contre la pandémie de Covid19 ont considérablement affecté la dynamique des flux

piétons dans le pôle de transport. Ces restrictions ne sont pas les seuls événements qui ont impacté l'utilisation du pôle sur le long terme. L'étude de l'impact de facteurs atypiques révèle comment les flux piétons réagissent en conséquence. Par exemple, un concert va augmenter la fréquentation du pôle, et modifier les entrées et sorties privilégiées. Nous avons constaté que des situations données, qu'il s'agisse d'un segment de temps ou d'un facteur exogène, peuvent entraîner une sur- ou sous-utilisation spécifique de certains lieux. Ce type d'étude est reproductible à toute situation où l'on dispose d'un large ensemble de données de comptage, et où l'on cherche à les synthétiser au sein de profils caractéristiques associés à des périodes distinctes. Nous pensons au domaine de la mobilité, étendu à l'étude d'un réseau de transport public ou d'une ville, et pour lequel la caractérisation des modes de déplacement humain revêt une grande importance. Ce type de problème peut également émerger en écologie ou en transcriptomique, entres autres. Ces modèles peuvent être utiles pour étudier l'impact des covariables, isolées du reste, et potentiellement dans différentes périodes.

Des recherches supplémentaires sont nécessaires pour surmonter certaines des limites de ce type de modélisation, notamment :

— Un facteur exogène peut être au choix codé explicitement dans le modèle, ou bien son impact peut être détecté par le modèle comme constituant un segment à part. Cette décision induit une variabilité lors de la construction des segments.

— Il est nécessaire de disposer d'une quantité suffisante de données exogènes. Augmenter le nombre de segments signifie que les facteurs exogènes seront modélisés dans des contextes de plus en plus spécifiques, pour lesquels moins de données sont disponibles.

Cette étude ouvre la voie à des travaux plus avancés de clustering ou de modélisation prédictive. Elle permet notamment de distinguer les périodes à dynamique de flux variable, ce qui peut être utile pour prédire la fréquentation dans des contextes spécifiques.

Chapitre 5

Prédire les mobilités

5.1 Introduction

La prédiction des séries temporelles est un problème majeur pouvant aider à la planification et à la prise de décision. Des méthodes bien connues d'apprentissage statistique supervisé sont régulièrement appliquées pour résoudre ce problème, dans des domaines variés tels que la finance, la logistique, la mobilité ou encore la météorologie. Dans le domaine de la mobilité dans les transports urbains, des travaux de recherche portent régulièrement sur la prévision des flux piétons en se basant sur la richesse des données numériques (capteurs, téléphones mobiles, billettique, etc.). Les flux piétons, dans les espaces de transport, sont amenés à varier en fonction de l'heure de la journée, du jour de la semaine, de la période de l'année, ou encore de divers facteurs non calendaires. La gestion des flux de personnes peut, à certaines périodes, devenir difficile en raison d'une grande affluence périodique ou ponctuelle (due à un concert ou à une perturbation, par exemple). Tout opérateur de transport souhaite anticiper le mieux possible les flux de passagers à venir, notamment pour éviter des situations de forte affluence mal gérées. Comme précisé par [Toq+18], les objectifs visés peuvent changer en fonction de l'horizon temporel de la prédiction. Dans le cas de prévisions à long terme, les données disponibles (calendaires, événements programmés à l'avance, etc.) aident à planifier l'offre de transport très en amont. Pour les prévisions à court terme, il est possible de prendre en compte, dans le modèle, l'historique récent de la demande. Ces prévisions peuvent aider à mieux gérer des situations atypiques survenues dans le passé proche, et à améliorer l'information voyageurs par exemple. Les objectifs visés peuvent également changer, selon que les prédictions faites soient déterministes ou probabilistes. Les prédictions déterministes, proposées le plus souvent, permettent de quantifier l'effet que les événements auront en moyenne sur l'affluence dans les transports. Dans certains cas en revanche, l'incorporation de l'incertitude est de première importance dans les processus de prise de décision. Cette vision mène au champ de la prédiction probabiliste, qui prédit les quantiles conditionnels pour les prochains pas de temps, sachant le passé.

Le travail que nous menons dans ce chapitre part de la constatation que la majorité des études sur la prédiction de séries multivariées, avec prise en compte des corrélations du bruit, se concentrent sur la prédiction sans modélisation de l'incertitude. Ce type de prévision semble pourtant particulièrement adapté dans le domaine des transports, où le risque (même faible) d'une forte affluence mal gérée est à éviter. La flexibilité et la capacité d'abstraction de l'apprentissage profond, adapté à la prévision probabiliste, en font un bon candidat pour ce travail. Le succès de la modélisation de données de comptage, multivariées, corrélées et surdispersées, permise par la vision « somme et partages » que nous avons développée dans le chapitre 4, s'adapte également à notre problématique de prédiction. Ainsi, nous nous inspirons dans ce travail de ces deux aspects, pour apprendre une représentation latente des données en entrée avec un réseau de neurone, puis la traduire en prévisions de flux de passagers en plusieurs points *via* la modélisation « somme et partages ». Nous montrons dans ce travail que notre modèle obtient des résultats de prédiction comparables aux autres méthodes de l'état de l'art, et les prédictions qu'il fournit restent bonnes, lorsque les données sont de très grande dimension et présentent des régularités temporelles.

5.2 Etat de l'art

Avec la multiplication de la quantité de données disponibles, différents types de modèles ont été développés pour exploiter la précision et la richesse des données spatiales et temporelles de mobilité urbaine. Nous proposons dans cette section un état de l'art de certaines de ces méthodes de prédiction, ainsi que leurs éventuelles applications dans le cadre de l'étude multivariée des mobilités. Ces méthodes peuvent être rangées en deux grandes catégories : les méthodes « classiques », et les méthodes basées sur l'apprentissage profond. Dans les sections suivantes, nous détaillons quelques méthodes présentes dans chacune de ces catégories. Nous introduirons ensuite des méthodes permettant d'effectuer une prédiction probabiliste des données, avant de nous positionner dans cette riche littérature.

5.2.1 Les méthodes classiques

Parmi les méthodes traditionnellement utilisées pour la prédiction des séries temporelles de mobilité, on retrouve les méthodes statistiques, qui offrent une bonne explicabilité. Ces dernières englobent les modèles de régression linéaire et les modèles autorégressifs. Pour les séries temporelles stationnaires, contenant des phénomènes d'autocorrélation entre des événements survenant à des moments différents, les méthodes généralement utilisées sont basées sur les processus autorégressifs (AR), les processus à moyennes mobiles (MA), et les processus autorégressifs de moyennes mobiles (ARMA), qui sont une combinaison des deux modèles précédents [SSS00, p. 77-90]. Pour les séries temporelles non stationnaires, une opération de différenciation est nécessaire pour obtenir une série stationnaire. Cette différenciation est prise en compte dans les modèles ARIMA [SSS00, p. 133-137], et ARIMA

saisonniers (SARIMA), qui conviennent aux séries chronologiques comportant une composante saisonnière [SSS00, p. 148-156]. Les modèles vecteur autorégressif (VAR) permettent de traiter simultanément plusieurs séries temporelles, afin de prévoir la prochaine valeur de chacune d'elles [Lüt13]. Chaque série est prédite avec une équation incluant ses valeurs passées, les valeurs passées des autres séries, ainsi qu'un terme d'erreur. Dans le cadre des travaux sur les mobilités, les auteurs de [Sri+] utilisent des modèles ARIMA et VAR pour la prédiction de flux piétons en différents points de la ville de Melbourne. Les modèles *generalized autoregressive conditional heteroskedastic* (GARCH) proposent une modélisation dynamique, des moyennes et covariances conditionnelles de systèmes multivariés [BLR06]. Les auteurs de [XSC15] ont par exemple appliqué un modèle hybride intégrant des modèles ARIMA, SARIMA et GARCH, pour prédire à court terme les flux de passagers sur une ligne de bus de Shenzhen.

Les méthodes basées sur le *machine learning* ou l'apprentissage automatique sont également utilisées pour la prédiction des données de mobilité, notamment dans le cas de problèmes de régressions non linéaires. Il est possible de reformaliser un problème non linéaire, en problème linéaire dans un nouvel espace de représentation des données, avec les méthodes à noyau. Parmi ces méthodes, les processus gaussiens (GP) prédisent les nouvelles données à partir d'une distribution gaussienne [Ras03]. Les auteurs de [GV22] utilisent un modèle de régression à processus gaussien pour prédire le trafic passager dans un réseau de bus. Les modèles de type espace-état, qui ont été abordés dans le chapitre 3, peuvent également servir à la prédiction, et ont l'avantage d'estimer une structure latente facilement interprétable des séries temporelles. On trouve ces modèles dans de nombreux travaux de prédictions liés à la mobilité [Doo+14 ; GBO09 ; Bia+19]. Enfin, d'autres modèles, comme les méthodes basées sur les forêts aléatoires (*Random Forest*) [Pra85] et le *Gradient boosting*, sont elles aussi largement exploitées pour des problèmes de classification ou de régression [Din+16]. Dans le travail de [EB21], des méthodes de *random forest* sur séries « détendancées » sont appliquées pour prédire à long terme la fréquentation des différentes lignes de transport à Lyon.

5.2.2 Les méthodes basées sur l'apprentissage profond

Récemment, les méthodes basées sur l'apprentissage profond (*deep learning*) ont apporté une approche alternative pour répondre à des enjeux de prédiction. Ces méthodes sont facilement modulables, et ont l'avantage de permettre l'intégration d'un grand nombre de covariables en entrée. En plus, ces modèles peuvent être utilisés pour de l'apprentissage non supervisé, comme le clustering. Les modèles basés sur l'apprentissage profond se basent sur une diversité notable d'architectures, de manière non exhaustive : les réseaux neuronaux convolutifs (CNN), les réseaux de neurones récurrents (RNN), les transformers et les réseaux basés sur des graphes (GNN), où différentes fonctions de coût sont optimisées pour calculer les prédictions. Les CNNs sont régulièrement utilisés pour prendre en compte des dépendances spatiales, mais aussi temporelles, entre les données. Le tra-

vail de [Bap+21] utilise par exemple un réseau de neurones convolutionnel de type U-net, pour prédire la charge des trains à différentes stations d'une ligne, en se basant sur une représentation en image des charges et informations associées. Les transformers, développés plus récemment, sont basés sur le principe de l'attention [Vas+17] et sont de bons candidats pour la modélisation des séries temporelles. Le mécanisme sous-jacent à l'attention permet en outre de capturer des dépendances à court ou à long terme dans ces dernières. Les GNNs sont une autre catégorie de réseaux permettant la prise en compte de dépendances spatiales *via* l'utilisation de graphes. Par exemple pour des sujets de prédiction du trafic, les GNNs présentent de bonnes performances comme ils permettent de modéliser la structure en graphe des systèmes de transports [JL22]. Les RNNs sont largement utilisés dans la prédiction de séries temporelles, du fait de leur capacité à intégrer les prévisions passées dans la recherche de prédictions actuelles (récurrence). Récemment, les RNNs ont montré une très bonne capacité à prédire la demande des passagers, de par leur prise en compte des dépendances temporelles [Zha+17b ; Pas+19]. Les RNNs présentent néanmoins des difficultés à mémoriser des informations sur de longues périodes. D'autres architectures découlant des RNNs ont ainsi été proposées pour pallier ce problème, notamment le Long Short-Term Memory (LSTM, [HS97]). Dans beaucoup de situations, des modèles combinant des CNNs et des RNNs, pour capturer les dépendances spatiales et les corrélations temporelles, sont appliqués à des données de mobilité [Ke+17 ; WT16]. Enfin, de plus en plus de travaux combinent des méthodes basées sur l'apprentissage profond avec des méthodes classiques, en vue d'améliorer les résultats de prédiction. Par exemple, les auteurs de [Zha+20] utilisent un réseau de neurones récurrent (LSTM) sur des séries décomposées, afin de prédire les entrants/sortants des stations de métro à Shanghai.

5.2.3 Les prévisions probabilistes

Dans certains domaines, les prédictions probabilistes peuvent être davantage recherchées que les prédictions déterministes. Une prédiction probabiliste permet par exemple de prendre des décisions sous un risque maximal (ou minimal). Certains modèles comme le modèle VAR permettent de prendre en compte du bruit dans les prédictions mais ce bruit n'est pas modélisé et est indépendant d'une série à l'autre. D'autres modèles ont la capacité de modéliser la structure du bruit en lien avec du contexte, ce sont ces modèles qui nous intéressent davantage dans ce travail. Afin d'intégrer l'incertitude dans les prédictions de temps de trajets en bus, les auteurs de [Che+22] ont développé un modèle bayésien probabiliste avec des mélanges de lois normales multivariées pour la prise en compte des interactions entre les bus consécutifs. Les paramètres du modèle proposé contiennent des informations de valeur pour améliorer les services de bus. Dans ce travail, nous allons nous intéresser aux méthodes de prédiction probabilistes basées sur des approches d'apprentissage profond. Le modèle DeepAR proposé par [Sal+20] est l'un des premiers à avoir proposé une approche de *deep learning* pour la prédiction probabiliste des séries temporelles. Les sorties de ce modèle sont les paramètres d'une distribution. La distribution choisie dépendra

des données, il pourra ainsi s'agir d'une distribution gaussienne pour des données continues, ou d'une distribution binomiale négative pour des données de comptage. Néanmoins, le modèle DeepAR ne fonctionne que dans le cas univarié, ce qui empêche la prise en compte d'éventuelles dépendances entre séries, qui pourraient améliorer la précision. Dans notre cas, un événement impactant une zone du pôle de transport pourrait impacter les flux piétons (donc les comptages issus de la billettique et des capteurs) en divers points proches de cet événement. Le modèle DeepAR ne serait pas ici en mesure de prendre en compte cet effet de dépendance. Les modèles DeepVAR et GPVAR proposés par [Sal+19] appliquent des distributions gaussiennes multivariées, afin de prendre en compte et prédire des dépendances entre les séries temporelles. Pour des données non distribuées selon une loi normale, ces méthodes appliquent une transformation inversible de manière à ce que, marginalement, elles suivent une distribution normale. Avec le modèle DeepVAR, un seul LSTM prédit l'ensemble des P séries en une fois. Le modèle GPVAR est une alternative, notamment lorsque P devient grand, car ici un réseau LSTM est appliqué sur chaque série séparément, avant de reconstruire la distribution jointe. Dans les deux cas, il est possible d'utiliser, comme loi de distribution, une gaussienne de faible rang, pour prendre en compte des cas où P est grand. [Ras+20] proposent d'utiliser un modèle à normalisation de flux en sortie, comme le *Masked Autoregressive Flows* (MAF, [PPM17]). La normalisation des flux permet de transformer une distribution d'un espace en entrée en une distribution plus simple dans un autre espace. La séquence de transformations est obtenue avec des fonctions inversibles. Les auteurs obtiennent des résultats de prédiction meilleurs que ceux de [Sal+19], sur un ensemble de bases de données disponibles en ligne.

5.2.4 Notre positionnement

La modélisation de données de comptage est soumise à certaines limitations dans les différentes stratégies mentionnées précédemment. DeepAR peut prendre en compte des comptages *via* des distributions binomiales négatives, mais ne pourra en revanche pas considérer les dépendances entre séries. DeepVAR et GPVAR répondent à ce problème, mais le font à travers des lois normales multivariées nécessitant une transformation des données en entrée. Le présent travail s'applique à mettre en place un nouveau modèle de prédiction probabiliste de données de comptage multivariées, corrélées et éventuellement sur-dispersées. Nous nous référons pour cela à la stratégie des modèles « sommes et partages » que nous avons déjà utilisés dans le chapitre 4. Pour rappel, ces modèles subdivisent la modélisation avec deux lois de distribution : la somme des comptages suit une distribution univariée (Poisson, binomiale négative) et la distribution de cette somme entre P séries suit une loi multivariée (Multinomiale, Dirichlet multinomiale). Nous avons observé lors de ce travail un intérêt à utiliser un modèle prenant en compte une binomiale négative pour la somme, puis une Dirichlet multinomiale pour le partage, en raison de sa capacité à prendre en compte des phénomènes de sur-dispersion et de corrélation, tout en restant parcimonieux (modélisation simple des dépendances par la somme). L'avantage, par rap-

port à un modèle nécessitant le calcul d'une matrice de covariance, est le nombre réduit de paramètres à calculer, notamment lorsque le nombre P de séries devient grand. Un avantage attendu de ce modèle est qu'il puisse bien prendre en compte des phénomènes globaux de diminution ou d'augmentation des fréquentations, à travers la distribution sur la somme, puis l'impact de phénomènes localisés, à travers la distribution sur le partage. La force d'abstraction des réseaux de neurones, combinée à la modélisation des comptages *via* un modèle « sommes et partages », le tout intégré dans une méthode de prédiction probabiliste des flux de passagers dans les espaces de transport, est un travail prometteur que nous développons dans les sections suivantes.

5.3 Méthodologie

5.3.1 La prévision probabiliste de séries temporelles à l'aide de réseaux de neurones récurrents

Considérons les éléments d'une série temporelle multivariée $\mathbf{y}_t$ (un vecteur $\{y_{t,1}, ..., y_{t,P}\}$ des comptages faits en P points) où t désigne le temps. La prédiction des futures séries peut se faire de manière déterministe ou probabiliste. De manière déterministe, il s'agit d'estimer les futures valeurs $\hat{\mathbf{y}}_t$ pour $t \in \{t_0, ..., T\}$, connaissant les observations passées $\mathbf{y}_t$ entre 1 et $t_0 - 1$, et les données exogènes $\boldsymbol{\chi}_t$ sur tout l'historique $t \in \{1, ..., T\}$. Nous allons nous intéresser à la prédiction probabiliste dans ce travail. Cela revient à estimer $p(\mathbf{y}_{t_0:T}|\mathbf{y}_{1:t_0-1}, \boldsymbol{\chi}_{1:T})$, les futures distributions de probabilité des séries temporelles de t_0 à T.

Structure des modèles

Les réseaux de neurones récurrents (RNNs) sont des modèles issus de l'apprentissage profond, efficaces pour prévoir des données issues de séries temporelles. Ces modèles contiennent une couche cachée $\mathbf{h}_t$ (au temps t) qui leur permet de garder en mémoire les observations passées. A chaque pas de temps t, $\mathbf{h}_t$ est ainsi calculée avec une fonction f de l'entrée courante $\mathbf{z}_t = \{\mathbf{y}_{t-1}, \boldsymbol{\chi}_t\}$, et de la couche cachée du pas de temps précédent $\mathbf{h}_{t-1}$. Les sorties $\mathbf{y}_t$ du modèle sont ensuite calculées à partir d'un tirage d'une distribution g dont les paramètres sont obtenus avec une transformation Φ de la couche cachée. Les équations sont les suivantes :

$$\mathbf{h}_t = f(\mathbf{U}\mathbf{z}_t + \mathbf{W}\mathbf{h}_{t-1}), \tag{5.1}$$

$$\mathbf{y}_t|\mathbf{h}_t \sim g(\mathbf{y}_t; \Phi(\mathbf{V}\mathbf{h}_t)) \tag{5.2}$$

$$\tag{5.3}$$

où $\mathbf{U}$, $\mathbf{W}$ et $\mathbf{V}$ sont des matrices de poids à estimer, $\mathbf{z}_t$ les entrées et $\mathbf{y}_t$ les prédictions désirées. Avec l'augmentation de la taille des séries temporelles, les réseaux de neurones

récurrents sont victimes d'un phénomène de disparition du gradient (*vanishing gradient*), qui les empêche de mémoriser l'information sur de longues périodes de temps. Le LSTM intègre une cellule mémoire contenant un système de « gates » pour répondre à ce problème. La structure d'un LSTM est présentée dans l'annexe D.1.

Une architecture couramment utilisée pour la prédiction est le « Sequence to Sequence » [SVL14], que nous représentons dans le schéma de la figure 5.1. Dans le cadre de la prédiction, un modèle *Sequence to Sequence* a pour objectif d'associer une entrée $\mathbf{y}_{1:t_0-1}$, de longueur fixe $t_0 - 1$, avec une sortie $\mathbf{y}_{t_0:T}$, elle aussi de longueur fixe $T - t_0 + 1$.

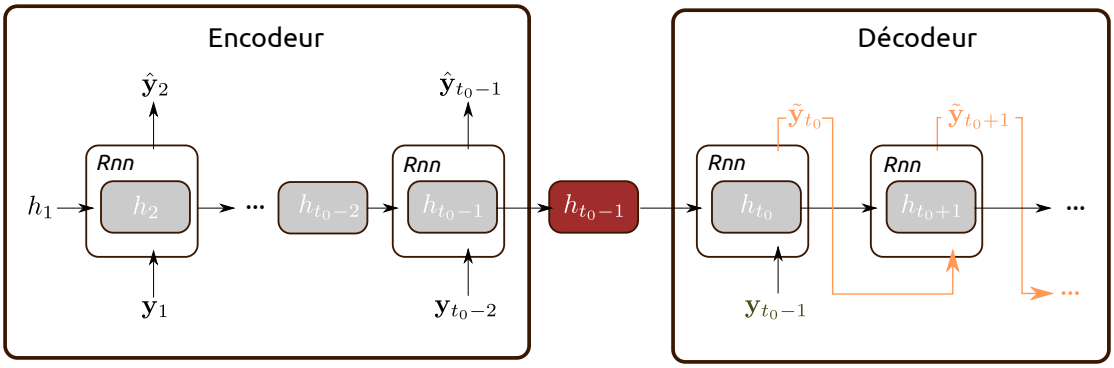


FIGURE 5.1 – Illustration d'une architecture de type Encodeur-Décodeur (illustration inspirée de [Kos]).

Deux composantes intègrent ce type de modèle :

— L'**encodeur** intègre en entrée une séquence d'observations, et la transforme en « contexte » (la dernière couche cachée).

— Le **décodeur** récupère ensuite la sortie émise par l'encodeur, pour l'associer à un ensemble d'observations.

Deux RNNs sont utilisés dans l'encodeur et le décodeur, et sont entraînés ensemble pour maximiser la fonction de coût (*loss*) de l'ensemble (ou minimiser, selon la nature de la *loss*).

Dans le cadre d'un modèle probabiliste, la fonction Φ (équation 5.2) consiste à transformer la couche cachée $\mathbf{h}_t$ en paramètres $\boldsymbol{\theta}_t$ d'une loi de probabilité (par exemple les paramètres de moyenne et de matrice de covariances d'une loi normale multivariée). Il s'agira typiquement d'une fonction linéaire, suivie d'une fonction d'activation (ReLU, softmax, etc.). L'objectif est de ramener la couche cachée aux paramètres de distribution, tout en prenant en compte leurs spécificités (toujours positifs, entre 0 et 1, etc.). La prédiction $\hat{\mathbf{y}}_t$ peut ensuite être tirée de la loi de distribution appropriée $p(\mathbf{y}_t|\boldsymbol{\theta}_t)$. D'après la figure 5.1,

il est possible d'écrire

$$p(\mathbf{y}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{h}_t) \tag{5.4}$$

Connaissant les paramètres ($\mathbf{U}$, $\mathbf{W}$ et $\mathbf{V}$ dans les équations 5.1 et 5.2) et $\mathbf{h}_{t_0}$, les prédictions $\hat{\mathbf{y}}_{t_0:T}$ se calculent en répétant un certain nombre de fois le processus d'estimation de $p(\mathbf{y}_{t_0:T}|\mathbf{y}_{1:t_0-1}, \boldsymbol{\chi}_{1:T})$. Cela revient à produire des échantillons de Monte Carlo à partir de :

$$p(\mathbf{y}_{t_0:T}|\mathbf{y}_{1:t_0-1}, \boldsymbol{\chi}_{1:T}) = p(\mathbf{y}_{t_0:T}|\mathbf{h}_{t_0}, \boldsymbol{\chi}_{1:T}) = \prod_{t=t_0}^{T} p(\mathbf{y}_t|\mathbf{h}_t(\mathbf{h}_{t-1}, \mathbf{y}_{t-1}, \boldsymbol{\chi}_t)), \tag{5.5}$$

via le calcul itératif de $p(\mathbf{y}_t|\mathbf{h}_t)$ et de mises à jour récursives de $\mathbf{h}_t$ avec la fonction f (équation 5.1). Nous présentons dans les sections suivantes les deux étapes nécessaires à la mise en place d'une méthode d'apprentissage profond : l'apprentissage et l'évaluation sur une base de test.

Processus d'entraînement et de calage des hyperparamètres

Dans tout travail d'apprentissage supervisé, il y a la nécessité de minimiser (ou maximiser) une fonction de coût sur une base d'entraînement. Dans le cas de l'apprentissage profond, cet entraînement passe par un algorithme d'optimisation de type « descente de gradient stochastique » (SGD). L'objectif est de trouver l'ensemble de paramètres optimal ($\mathbf{U}$, $\mathbf{W}$ et $\mathbf{V}$ dans les équations 5.1 et 5.2), de manière à maximiser (pour la vraisemblance, par exemple) ou minimiser (pour l'erreur quadratique moyenne, par exemple) la fonction de coût. Dans le cas d'un modèle de prédiction probabiliste, il s'agira de maximiser la log vraisemblance.

La mise en place d'un modèle de prédiction passe par la recherche des paramètres optimaux, pour lesquels la fonction de coût est la meilleur possible. Des hyperparamètres intègrent également les modèles mais, à la différence des paramètres, ces derniers ne sont pas estimés lors de l'apprentissage, et doivent être spécifiés *a priori*. En général, il faut éliciter des modèles avec différentes valeurs d'hyperparamètres, et tester quels modèles performent le mieux en termes de fonction de coût. Les combinaisons possibles de valeurs d'hyperparamètres peuvent former une grille ; on parle alors de procédure « grid-search ». Plusieurs étapes sont nécessaires pour procéder à la recherche des meilleurs hyperparamètres :

1. Subdiviser la base de données, en une base d'apprentissage et une base de validation
2. Lancer l'apprentissage du modèle avec une certaine combinaison d'hyperparamètres
3. Calculer des métriques de qualité de la prédiction, sur la base de validation (moyenne sur plusieurs sous-ensembles de la base de validation)

TABLE 5.1 – Hyperparamètres d'un réseau de neurone récurrent

Hyperparamètre	**Description**
Pas d'apprentissage	Contrôle à quel point les valeurs de paramètres sont changées en réponse à la fonction de coût, à chaque mise à jour de l'apprentissage.
Taille de couche cachée	Nombre d'unités constitutives de chaque couche cachée du RNN.
Nombre de couches	Nombre de couches cachées successives dans le RNN.
Dropout	Probabilité à laquelle les sorties de chaque couche sont retenues ou non. Il s'agit d'une manière d'éviter le sur-apprentissage des exemples de la base d'entraînement.
Taille de lot (*batch*)	Nombre d'échantillons qui seront propagés dans le réseau d'apprentissage.

4. Réitérer le processus depuis l'étape 2, plusieurs fois pour le même modèle, puis pour les autres combinaisons d'hyperparamètres de la grille
5. Moyenner, pour chaque modèle, les métriques de qualité de prédiction, et choisir le modèle pour lequel ce score est le meilleur

Dans le cas d'un modèle RNN, quelques hyperparamètres sont souvent amenés à être recherchés. Nous résumons dans le tableau 5.1 quelques-uns de ces hyperparamètres, ainsi que leur signification.

Durant la phase d'entraînement, nous échantillonnons des fenêtres depuis plusieurs points de départ dans le jeu d'entraînement. La tranche $t = 1$ peut ainsi se situer en plusieurs positions dans la base d'entraînement. Ces fenêtres font toutes la même taille T (un intervalle de comptages « contexte » intégrant l'encodeur, et un intervalle de comptages « prédictions » intégrant le décodeur). Notons que le modèle utilisé pour l'encodeur est le même que celui du décodeur. Au cours de l'entraînement, il n'est pas fait de différence entre les deux pour le calcul de la vraisemblance. Nous maximisons la log vraisemblance $L_Y(\mathbf{U}, \mathbf{V}, \mathbf{W})$ qui correspond à la moyenne des vraisemblances individuelles de chaque fenêtre (*batch*) :

$$-L_Y(\mathbf{U}, \mathbf{V}, \mathbf{W}) = -\frac{1}{|Ba|(T)} \sum_{Ba} \sum_{t=1}^{T} \log p(\mathbf{y}_t | \mathbf{h}_t, \boldsymbol{\chi}_t) \tag{5.6}$$

avec Ba l'ensemble des fenêtres (*batches*) sélectionnées. Pour l'ensemble du processus d'entraînement, nous utilisons pour cela l'optimisateur ADAM ([KB14]), qui intègre un certain pas d'apprentissage (*learning rate*), et une taille de lot $|Ba|$ (voir table 5.1).

Les données en entrée

Les données en entrée $\mathbf{z}_t$ intègrent les covariables $\boldsymbol{\chi}$, et les comptages $\mathbf{y}$ collectés aux pas de temps passés. Les covariables $\boldsymbol{\chi}$ intègrent un ensemble de caractéristiques temporelles liées aux cas d'études rencontrés. En fonction de la fréquence à laquelle sont collectées les données (jour, heure, etc.), les caractéristiques saisonnières peuvent intégrer des informations comme l'heure de la journée, le jour de la semaine et sa position dans l'année. D'autres événements saisonniers peuvent être intégrés dans les données exogènes, comme les vacances, les jours fériés ou les ponts. Des données non calendaires peuvent également être intégrées, et utilisées en tant que variables continues ou facteurs. Les modèles étudiés dans ce travail sont autorégressifs ; les comptages $\mathbf{y}$ collectés aux pas de temps passés peuvent ainsi être intégrés en entrée. Par exemple, si les données sont des comptages horaires, il pourra être intéressant de prendre en compte les comptages des périodes passées -1 pour l'heure précédente, -24 pour la même heure à la journée précédente, et -168 pour la même heure à la semaine précédente.

Évaluation du modèle sur une base de test

En plus de la base d'entraînement nécessaire à l'apprentissage, et de la base de validation permettant le calage des hyperparamètres du modèle, une base de test est mise en place pour juger de la qualité prédictive en généralisation des modèles. Des métriques d'évaluation sont calculées, pour comparer les sorties théoriques des modèles avec les données de la base de test. Ces métriques sont calculées avec une fenêtre glissante, afin de couvrir l'ensemble de la base de test (comme décrit dans [YRD16]). Le modèle est ainsi entraîné une fois sur les données d'entraînement, donc sur toutes les données situées avant la première fenêtre de prédiction. Le schéma de la figure 5.2 illustre les processus d'entraînement et de test.

Nous analysons la capacité d'un modèle à prédire les données correctement sous un angle déterministe et probabiliste. La performance de prédiction déterministe des différents modèles peut être mesurée avec l'erreur quadratique moyenne (MSE), dont la valeur est proportionnelle à l'importance de l'erreur quadratique (les erreurs importantes ont un effet d'autant plus fort sur cette métrique). Le calcul est le suivant :

$$MSE = \frac{1}{P(T - t_0 + 1)} \sum_{t,p} (y_{t,p} - \hat{y}_{t,p})^2 \tag{5.7}$$

où $\hat{y}$ est une prédiction. La performance de prédiction probabiliste des différents modèles peut être estimée à travers le *continuously ranked probability score* (CRPS) [MW76]. Cette métrique mesure à quelle distance la prédiction se trouve de l'observation, dans un contexte probabiliste. La meilleur prédiction probabiliste est ainsi celle pour laquelle toute la fonction de masse se trouve au niveau de l'observation. La métrique CRPS se calcule comme suit :

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(\alpha) - 1(y \leq \alpha))^2 d\alpha, \tag{5.8}$$

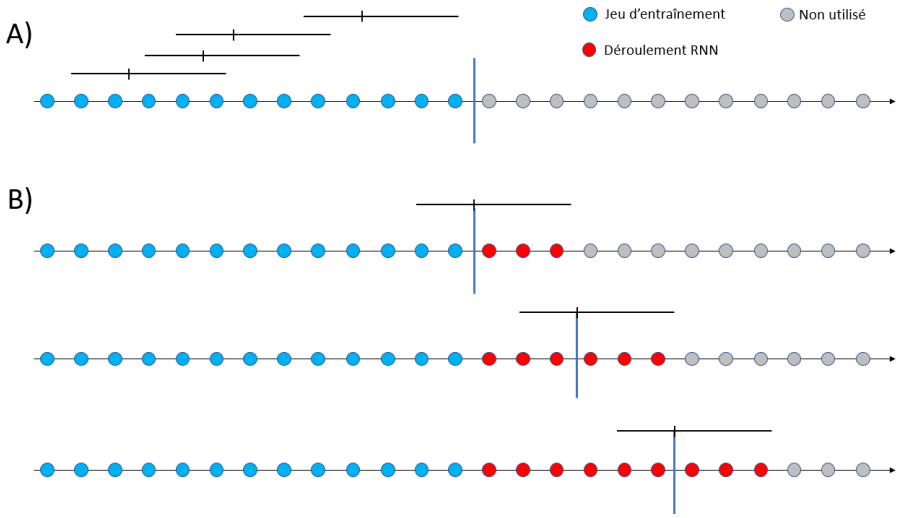


FIGURE 5.2 – Processus d'entraînement (partie A) et de test (partie B), pour un modèle basé sur un réseau de neurones récurrent. La ligne verticale sépare les données connues, des données à prédire (il peut s'agir ici de la base de validation aussi). A) Phase d'entraînement. Les différents intervalles noirs représentent les fenêtres utilisées par le modèle pour l'entraînement, avec à gauche, l'intervalle de l'encodeur, et à droite, celui du décodeur. La vraisemblance est calculée sur l'ensemble de ces intervalles. B) Phase de prédiction. L'encodeur seulement se trouve dans l'intervalle des données connues, alors que le décodeur se situe dans la période à prédire. Une fenêtre glissante est ainsi mise en place pour couvrir l'ensemble de la base de test.

avec F la distribution cumulée prédite. $1(y \leq \alpha)$ est une fonction indicatrice, qui vaut 1 si $y \leq \alpha$ et 0 sinon. Ici, F est calculé de manière empirique, avec S échantillons comme $\hat{F}(\alpha) = \frac{1}{S}\sum_s 1(y^{(s)} \leq \alpha)$. Sa version calculée sur la somme des séries (CRPS-sum) est aussi considérée.

5.3.2 Les modèles « sommes et partages »

Les modèles de distribution « sommes et partages » sont inspirés des travaux présentés par [JM19]. Nous détaillons ces modèles dans le chapitre 4.

Dans ce qui suit, nous nous concentrerons sur le modèle avec une distribution binomiale négative sur la somme, et Dirichlet-multinomiale (ou Pólya) sur le partage. Ce modèle introduit des corrélations entre les séries de comptage, et peut modéliser des comptages surdispersés. Pour rappel, ce modèle s'écrit comme suit :

$$V \sim \mathcal{G}(\mu, \sigma) \tag{5.9}$$

$$\mathbf{Y}|V \sim \mathcal{H}(V, (\alpha_p)_{p\in 1,\dots P}), \tag{5.10}$$

avec $V = \sum_p Y_p$ la somme des comptages, μ et σ, les paramètres de moyenne et de somme de la binomiale négative sur la somme, et α_p, le paramètre de poids pour la p-ème série de la Dirichlet-Multinomiale.

5.3.3 Méthodologie proposée

Nous considérons dans ce travail une distribution de type « sommes et partages » pour la fonction g (voir équation 5.2). Une distribution binomiale négative est utilisée pour modéliser la somme des comptages v_t. Une distribution Dirichlet-Multinomiale est utilisée pour modéliser la répartition du total en P comptages $y_{t,1}, ..., y_{t,P}$. Dans notre approche, un premier réseau récurrent de type LSTM prédit les paramètres, de moyenne $\mu \in \mathbb{R}^+$ et de forme $\sigma \in \mathbb{R}^+$, de la distribution binomiale négative, et un deuxième réseau LSTM est appliqué aux paramètres $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_P) \in \mathbb{R}^+$ de la distribution Dirichlet-Multinomiale. Tous les paramètres sont obtenus en appliquant une transformation linéaire, puis une activation *softplus*, afin de garantir qu'ils soient tous positifs. La formulation du modèle est la suivante :

$$g(v|\mu,\sigma) = \frac{\Gamma(v+\frac{1}{\sigma})}{\Gamma(v+1)\Gamma(\frac{1}{\sigma})}(\frac{1}{1+\sigma\mu})^{\frac{1}{\sigma}}(\frac{\sigma\mu}{1+\sigma\mu})^{v} \tag{5.11}$$

$$h(\mathbf{y}|v,\boldsymbol{\alpha}) = \frac{\Gamma(\sum \alpha_p)\Gamma(v+1)}{\Gamma(n+\sum \alpha_p)} \prod_{p=1}^{P} \frac{\Gamma(y_p+\alpha_p)}{\Gamma(\alpha_p)\Gamma(y_p+1)} \tag{5.12}$$

$$\alpha(\mathbf{h}_t^1) = \log(1+\exp(\mathbf{w}_\alpha^T \mathbf{h}_t^1 + b_\alpha)) \tag{5.13}$$

$$\mu(\mathbf{h}_t^2) = \log(1+\exp(\mathbf{w}_\mu^T \mathbf{h}_t^2 + b_\mu)) \tag{5.14}$$

$$\sigma(\mathbf{h}_t^2) = \log(1+\exp(\mathbf{w}_\sigma^T \mathbf{h}_t^2 + b_\sigma)), \tag{5.15}$$

Nous appelons **DeepNegPol** notre modèle de prédiction probabiliste de séries de comptages, basé sur un modèle de type « sommes et partages » prenant en entrée les sorties de réseaux de neurones récurrents. Le modèle intègre les covariables $\boldsymbol{\chi}$, les comptages du (ou des) pas de temps précédent(s), ainsi que la sortie du modèle $\mathbf{h}$ au pas de temps précédent. Un premier LSTM permet de calculer les paramètres associés à une distribution Dirichlet-Multinomiale. Un deuxième LSTM est, lui, adapté au calcul des paramètres d'une distribution binomiale négative qui traite la somme des comptages. Au cours de l'entraînement du modèle, les paramètres de la distribution « sommes et partages » sont calculés et utilisés pour obtenir la vraisemblance à maximiser. Cette vraisemblance intègre les sorties des deux LSTMs. Les fenêtres de contexte (encodeur) et de prédiction (décodeur) sont toutes deux utilisées pour le calcul de la vraisemblance lors de cette phase. Lors de la prédiction, l'historique des comptages pour $t < t_0$ est utilisé pour dérouler les LSTM, puis pour $t \geq t_0$, des valeurs $\hat{\mathbf{y}}_t$ sont échantillonnées depuis la distribution, puis réintégrées

comme entrées des LSTM pour prédire les pas de temps suivants. La taille de la prédiction correspond au nombre de pas de temps spécifiés dans la fenêtre de prédiction. Ce processus est répété plusieurs fois pour obtenir des quantiles de prédictions. Une représentation graphique du modèle **DeepNegPol** est dévoilée dans la figure 5.3.

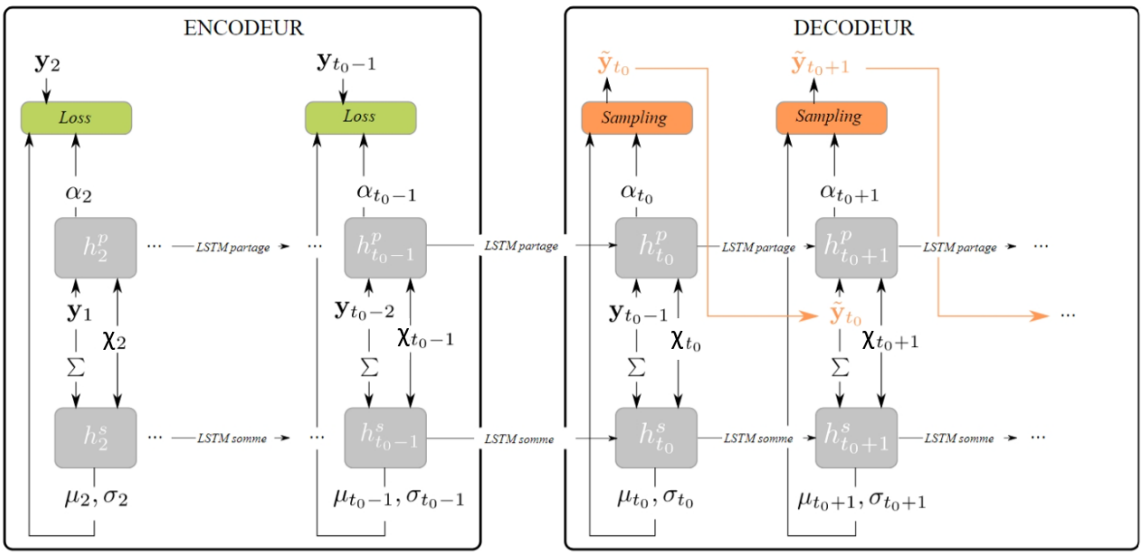


FIGURE 5.3 – Représentation schématique de **DeepNegPol**, notre modèle de prédiction probabiliste basé sur un réseau de neurone récurrent et une distribution « somme et partages » en sortie.

5.4 Résultats

Dans cette section, nous comparons les performances du modèle proposé (**DeepNegPol**) à celles d'autres modèles issus de l'état de l'art, sur quelques données issues de l'open data dans un premier temps, puis sur le cas d'étude des comptages à La Défense dans un deuxième temps. Les modèles de prédiction probabiliste comparés sont basés sur des réseaux de neurones. La plupart des modèles proviennent de l'article de [Sal+19].

— **VAR** est un modèle autoregressif multivarié dont les paramètres sont estimés avec une pénalité Lasso.

— **Vec-LSTM-ind-scaling** intègre une distribution normale indépendante avec une matrice de covariance diagonale, et où les données en entrée sont mises à l'échelle par la moyenne, comme proposé par [Sal+19].

— **Vec-LSTM-lowrank-Copula** intègre une distribution normale de faible rang, et transforme les données en entrée avec des copules gaussiennes (représentation des

données sous forme gaussienne à partir de leur fonction de répartition).

— **GP-scaling** applique un LSTM sur chaque série mise à l'échelle, avant de reconstruire la distribution jointe avec une gaussienne de faible rang.

— **GP-Copula** opère de la même manière que le modèle GP-scaling, à la différence qu'il applique une transformation basée sur des copules en entrée.

— **LSTM-MAF**, issu du travail de [Ras+20], utilise en sortie un modèle par normalisation de flux MAF.

— **DeepNegPol**, notre modèle, applique un modèle de type « somme et partages » en sortie. Un premier LSTM est utilisé pour modéliser la distribution de la somme des séries et un deuxième est dédié à la modélisation de la répartition entre les P points de comptage.

Nous construisons le modèle **DeepNegPol** avec la librairie PytorchTS, qui encapsule la construction de modèles sous Pytorch. La plupart des autres modèles testés sont ceux proposés par la librairie GluonTS, qui implémente les modèles sous MXNET principalement. Le modèle **VAR** est encodé de la même manière que dans [Sal+19] avec la librairie GLMNET de Python. Dans nos expériences, nous générons 200 échantillons avec le décodeur pour la prédiction.

5.4.1 Expériences sur données *open* de comptage

En raison de l'application préférentielle de notre modèle **DeepNegPol** à des données de comptage, nous recherchons des données en *open source* adaptées. Les jeux de données suivants sont utilisés :

— Bike : comptages horaires de passages de vélos en 80 lieux de la ville de Paris (nous avons retiré les séries avec des valeurs manquantes). La période d'entraînement s'étend sur les mois de janvier à fin mai 2022 (soit une matrice 3575×80). 30 fenêtres glissantes sont utilisées pour le test, afin de couvrir le mois de juin 2022, avec 24 tranches horaires de prédites par fenêtre. Source : https://parisdata.opendatasoft.com/explore/dataset/comptage-velo-donnees-compteurs.

— Railway : comptages journaliers de flux piétons entrants en 502 stations du réseau ferré d'Ile-de-France (nous avons retiré les séries avec des valeurs manquantes). La période d'entraînement s'étend sur les mois de janvier à fin septembre 2021 (soit une matrice 273×502). 3 fenêtres glissantes sont utilisées pour le test, afin de couvrir les mois d'octobre, novembre et décembre 2021, avec 30 jours de prédits par fenêtre. Source : https://data.sncf.com/explore/.

— Wikipedia : nombre journalier de vues de 2000 pages Wikipédia. 792 jours sont utilisés pour l'entraînement (soit une matrice 792×2000), et 5 fenêtres glissantes sont utilisées pour le test, avec 30 jours de prédits dans chaque fenêtre. Données disponibles en open data : https://github.com/mbohlkeschneider/gluon-ts/tree/mv_release/datasets.

— Taxi : passages de taxis en 1214 lieux de la ville de New York toutes les 30 minutes. Le mois de janvier 2015 sert de base d'entraînement (soit une matrice 1488×1214). Le mois de janvier 2016 sert de test ; il est constitué de 57 fenêtres glissantes, avec 24 tranches de 30 minutes de prédites pour chacune. Données disponible en open data : https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Chaque expérience est répétée trois fois, et les moyennes et écarts-types des métriques sont ensuite calculés. De la même manière que dans [Sal+19] nous utilisons des tailles de lot (*batch size*) de 16, avec 100 lots par itération et 100 itérations avec un pas d'apprentissage à 1e-3. Les hyperparamètres utilisés pour les LSTMs sont ceux de [Sal+19], notamment le nombre de cellules utilisés y est de 40. Nous utilisons par ailleurs la même configuration que [Ras+20] pour le modèle **LSTM-MAF**. L'architecture de notre modèle **DeepNegPol** est très proche de celle des autres modèles, nous reprenons donc les mêmes valeurs d'hyperparamètres : une taille de 40 cellules est utilisée pour le LSTM appliqué à la Dirichlet-Multinomiale, 20 cellules sont utilisées pour le LSTM appliqué à la binomiale négative, un nombre plus bas en raison de la moindre quantité d'informations qui lui est présentée. Pour l'encodage des covariables, nous reprenons les options proposées par GluonTS : une covariable d'âge (temps passé depuis la première observation), le jour de la semaine, et l'heure (ou la demi-heure) de la journée. Les résultats sont présentés dans la table 5.2, pour la métrique CRPS-Sum, 5.3 pour la métrique CRPS, et 5.4 pour la métrique MSE. Notons que les résultats du modèle **VAR** pour les données Taxi et Wikipedia sont ceux de [Sal+19], nous avons utilisé le même plan d'expérience pour les données Bike et Railway.

TABLE 5.2 – Métrique CRPS-Sum calculée pour les données de comptage proposées. Les deux meilleurs modèles sont mis en gras.

Modèle	Bike	Railway	Taxi	Wikipedia
VAR	0.391+/-0.001	/	0.292+/-0.000	3.400+/-0.003
Vec-LSTM-ind-scaling	0.495+/-0.041	0.367+/-0.024	0.451+/-0.011	0.160+/-0.009
Vec-LSTM-lowrank-Copula	0.460+/-0.015	0.207+/-0.005	0.361+/-0.005	0.175+/-0.009
GP-scaling	0.159+/-0.003	0.143+/-0.014	**0.097 +/-0.009**	0.666+/-0.209
GP-Copula	0.146+/-0.003	**0.142+/-0.001**	0.099 +/-0.005	**0.047+/-0.005**
LSTM-MAF	**0.138+/-0.012**	0.202+/-0.014	0.126 +/- 0.003	**0.061+/-0.002**
DeepNegPol	**0.131+/-0.006**	**0.125+/-0.021**	**0.074 +/- 0.009**	0.065+/-0.014

Quelques observations peuvent être émises à la lumière de ces résultats. Tout d'abord, nous pouvons souligner une meilleure performance des modèles basés sur un processus gaussien (GP) (**GP-scaling** et **GPCOP**), **LSTM-MAF** et notre modèle **DeepNegPol**. Notre modèle est ainsi le meilleur modèle, tous critères confondus, pour les jeux de données Bike et Taxi. Ces deux jeux de données présentent des périodes d'apprentissage conséquentes, permettant aux modèles d'apprendre sur un nombre diversifié d'exemples de fenêtres. Taxi

TABLE 5.3 – Métrique CRPS calculée pour les données de comptage proposées. Les deux meilleurs modèles sont mis en gras.

Modèle	Bike	Railway	Taxi	Wikipedia
VAR	0.418+/-0.001	/	0.410+/-0.000	4.101+/-0.002
Vec-LSTM-ind-scaling	0.525 +/- 0.051	0.405+/-0.026	0.552+/-0.019	0.564+/-0.146
Vec-LSTM-lowrank-Copula	0.480+/-0.014	0.233+/-0.005	0.550+/-0.004	0.313+/-0.005
GP-scaling	0.187+/-0.003	**0.163+/-0.012**	0.282 +/-0.004	0.880+/-0.192
GP-Copula	0.179+/-0.002	**0.154+/-0.001**	**0.282 +/-0.001**	**0.223+/-0.007**
LSTM-MAF	**0.168+/-0.011**	0.219+/-0.015	0.301 +/- 0.002	**0.278+/-0.008**
DeepNegPol	**0.169+/-0.007**	0.172+/-0.012	**0.270 +/- 0.005**	0.381+/-0.024

TABLE 5.4 – Métrique MSE calculée pour les données de comptage proposées. Les deux meilleurs modèles sont mis en gras.

Modèle	Bike	Railway	Taxi	Wikipedia
VAR	6.62×10^3	/	/	/
Vec-LSTM-ind-scaling	1.08×10^4	3.65×10^7	6.76×10^1	7.33×10^7
Vec-LSTM-lowrank-Copula	8.71×10^3	2.37×10^7	5.94×10^1	7.72×10^7
GP-scaling	1.89×10^3	$\mathbf{1.15 \times 10^7}$	1.99×10^1	5.76×10^7
GP-Copula	1.60×10^3	$\mathbf{1.13 \times 10^7}$	$\mathbf{1.94 \times 10^1}$	$\mathbf{4.59 \times 10^7}$
LSTM-MAF	$\mathbf{1.43 \times 10^3}$	/	2.34×10^1	$\mathbf{3.77 \times 10^7}$
DeepNegPol	$\mathbf{1.42 \times 10^3}$	1.32×10^7	$\mathbf{1.88 \times 10^1}$	5.02×10^7

est un cas difficile, de grande dimension, et présentant des profils très bruités, mais très bien pris en compte par notre modèle, par rapport aux modèles de l'état de l'art. La vision « sommes et partages » semble ici avoir un apport notable sur la bonne prédiction des données. **DeepNegPol** présente également de bonnes performances pour le jeu de données Railway, mais ne se détache des autres qu'à travers la métrique « CRPS-Sum ». Notre modèle modélisant la somme des séries, cela peut avoir un impact dans ce sens. Railway présente une période d'apprentissage courte de 273 jours, dans un contexte qui évolue rapidement (l'année 2021 est caractérisée par de nombreux changements de dynamique de déplacement, en raison de la pandémie de Covid19). Ce sont les modèles GP qui performent le mieux ici. En revanche, le modèle **LSTM-MAF** présente une instabilité ici. Certaines simulations mènent à des erreurs, et l'erreur MSE est impossible à calculer en raison de certaines prédictions bien trop grandes (notées « / » dans les tableaux). Comme précisé dans [BMM21], les métriques CRPS et CRPS-Sum pénalisent moins les valeurs extrêmes et peuvent donc être utilisées plus souvent ici. Le modèle **DeepNegPol** est enfin moins per-

formant pour le jeu de données Wikipedia. Il s'agit d'un exemple en grande dimension qui, contrairement aux cas précédents, contient une majorité de séries sans profils répétitifs en particulier, et donc plus chaotiques. Sur ce type de cas, notre modèle semble moins performant, bien que l'écart avec les meilleurs modèles se réduise en termes de MSE et CRPS-Sum. Le modèle **DeepNegPol** fait l'hypothèse d'une certaine distribution en sortie qu'il apprend à associer à un contexte durant la phase d'apprentissage. Pour des séries régulières dans le temps cette stratégie fonctionne bien mais lorsque ce n'est pas le cas, notre modèle présente une moins bonne capacité d'adaptation que le modèle **LSTM-MAF**. Les codes et exemples sur données en *open source* sont disponibles sur le dépôt github à l'adresse suivante : https://github.com/pdenailly/Probabilistic_forecasting.

5.4.2 Etude du cas de La Défense

Mise en forme des données

L'objectif de cette section est d'appliquer les différents modèles basés sur l'apprentissage profond sur le cas d'étude du pôle de La Défense. Il s'agit d'un travail où la structure des données en entrée prend une plus grande importance, car une grande diversité d'événements impactants devraient être pris en compte.

Nous travaillons avec les comptages issus des capteurs et des lignes de contrôle de billettique, au sein de la station Grande Arche. Nous travaillons à terme avec 25 séries de comptage de flux entrants, sortants et de transfert. La période d'entraînement recouvre la durée d'Avril 2019 à fin Janvier 2022. Cette période d'entraînement, longue, a permis d'englober un maximum d'événements, raréfiés par la période de Covid19. Le mois de février 2022 nous sert de base de validation, sur laquelle une sélection d'hyperparamètres peut être menée. Les mois de Mars et Avril 2022 seront la base de test. Les modèles utilisent tous un ensemble d'informations contextuelles, calendaires et non-calendaires. Pour les données calendaires, nous nous inspirons de l'encodage proposé dans le travail de [Pas+19] :

— Position du jour dans l'année (8 dimensions), encodée par cosinus et sinus, avec des fréquences (2×4).

— Type de jour (7 dimensions) : position du jour dans la semaine, encodage *one-hot*.

— Tranche de 30 minutes dans la journée (8 dimensions), encodée par cosinus et sinus, avec des fréquences (2×4).

Cet encodage permet de limiter à quelques dimensions la prise en compte de nombreux facteurs, pour la position du jour (365 possibilités) et de la tranche (48 possibilités). En plus des facteurs calendaires, des facteurs non-calendaires sont utilisés dans ce travail :

— Les travaux sur la ligne de RER A : il y a 7 catégories de travaux, selon la zone coupée sur la ligne (Auber $\leftrightarrow$ La Défense, Auber $\leftrightarrow$ Vincennes, etc.). Ces catégories sont mises en encodage *one-hot* (passage à 7 variables binaires).

— Les événements climatiques particuliers comme la pluie (en mm tombés les 3 dernières heures) et les températures anormales. Pour la température, nous retirons l'effet saisonnier, en appliquant une décomposition par moyennes mobiles, puis en ne gardant que la composante irrégulière (donc sans tendance, ni saisonnalité).

— Les événements au stade de l'Arena de La Défense (concerts ou événements autres) : nous choisissons un encodage *one-hot* des 4 tranches précédant un événement (6 pour un concert), et des 3 tranches suivant l'événement (ou le concert).

— Les perturbations du RER A pour chaque branche du RER A, ainsi que pour le tronçon central (voir plan de la ligne sur la figure 1.3), nous disposons de l'offre réelle du nombre de trains, par intervalle de 30 minutes. Nous considérons 10 catégories représentant chacune une branche et une direction. Nous créons pour chaque tranche t, et chaque catégorie, un score de perturbation calculé comme le log du ratio du nombre de trains courant, et du nombre médian de trains qui circulent normalement, pour une catégorie et une tranche donnée. Nous ne considérons pas ici les périodes de confinement, les grèves et les périodes de travaux. Chaque score de perturbation est associé à une catégorie basée sur la « sévérité » de la perturbation (5 catégories). Finalement, nous mettons en place un encodage *one-hot* avec, pour chaque tranche t, les ($10\times$ 5) vecteurs de perturbation. Nous utilisons ici les données de la situation de la ligne à une tranche donnée, pour anticiper la fréquentation à la tranche suivante.

Nous intégrons également les comptages passés, avec des retards de 1, 2, 4, 12, 24, 48 et 336 demi-heures en entrée des modèles (48 pour la même demi-heure de la journée précédente, et 336 pour la semaine précédente).

Calibration des hyperparamètres

Pour l'ensemble des modèles, nous choisissons un jeu d'hyperparamètres, dont certains sont issus d'une sélection faite sur la base de validation, qui correspond au mois de février 2022 dans nos données. Nous considérons un *dropout* à 0.01 pour tous les modèles, comme dans [Sal+19]. La longueur de la fenêtre des comptages passés équivaut à trois fois celle de la période à prédire ; dans les faits, nous prédisons les prochaines 48 tranches de temps (à savoir, la prochaine journée). Lors de l'entraînement, nous effectuons 80 itérations avec, pour chacune, un entraînement sur 100 *batches* de tailles 128.

Pour chaque entraînement, nous échantillonnons les fenêtres de comptages $\mathbf{y}$ (et les covariables x correspondantes), de manière pondérée. La pondération se fait par la présence d'événements rares pour la sélection des fenêtres. Certains événements, comme les concerts ou les perturbations, sont en effet rares (voire très rares) à l'échelle de l'historique des données, et devraient donc être « sur-représentés » dans l'entraînement, afin de bien pouvoir les intégrer. Supposons que B soit la taille totale de la base d'entraînement, C un ensemble de covariables non-calendaires, et b_c le nombre d'occurrences de la covariable c

dans la base d'entraînement. Plus une covariable est rare, plus la valeur associée $\frac{B}{b_c}$ sera grande. On peut nommer cette valeur la **rareté** (r(.)). Pour chaque position t de la base d'entraînement, il est possible de calculer une valeur représentant la somme des raretés des covariables non-calendaires pour les T pas de temps suivants, soit la taille d'une fenêtre :

$$r(t) = \sum_{T} \sum_{C} \mathbb{1}_{t,c} \frac{B}{b_c}$$

Sur la période de validation, nous testons différents modèles, où nous faisons varier les tailles de couches cachées des LSTMs entre 80, 160 et 320 (pour DeepNegPol, nous testons 80 ; 160 pour le LSTM des totaux ; et 80, 160, 320 pour le LSTM des répartitions). Nous testons également les modèles avec 2, 3 et 4 couches de LSTM. Nous testons enfin des taux d'apprentissage (*learning rate*) à 1e-2 et 1e-3. Pour chaque modèle, nous calculons le CRPS sur la période de validation, et comparons les différentes combinaisons pour faire un choix sur des valeurs judicieuses d'hyperparamètres. Chaque expérience est répétée trois fois, et les valeurs de métriques sont moyennées. Pour les modèles Vec-LSTM et GP, nous utilisons les jeux d'hyperparamètres que nous trouvons pour le modèle Vec-LSTM-lowrank-Copula, comme il s'agit de modèles similaires (le même type de calage a été mené dans le papier de [Sal+19]). La table 5.5 résume l'ensemble des hyperparamètres, ainsi que leurs valeurs, communes aux modèles pour certaines, et différentes pour d'autres.

TABLE 5.5 – Hyperparamètres pour le cas des données de La Défense.

Hyperparamètres	**Valeur**
Dim. couche cachée LSTM	80/160/320
Learning rate	1e-2/1e-3
Taille *batch*	128
Nb *batchs* par *epoch*	100
Epochs	80
Nombre de fenêtres de test	30
Nb échantillons évaluation	200
Nb couches	2/3/4
Dropout	0.01
Longueur contexte et prédiction	(144,48)

Les valeurs optimales de taille de couche / nombre de couches / *learning rate* trouvées pour les différents modèles sont respectivement de 320 / 3 / 1e-3 pour Vec-LSTM-lowrank-Copula, 160-320 / 4 / 1e-3 pour DeepNegPol, et 160 / 4 / 1e-3 pour LSTM-MAF. Notons que LSTM-MAF présente des métriques très légèrement meilleures avec un *learning rate* à 1e-2, mais dans un souci de comparaison des modèles, nous préférons lui conserver un *learning rate* à 1e-3.

Résultats de la prédiction

La comparaison des modèles est effectuée avec le calcul de métriques d'évaluations calculées entre les prédictions et les observations de la base de test (mois de mars-avril 2022, comprenant une période de vacances à partir du 23 avril). Chaque expérience est ici répétée six fois, et les métriques de comparaison sont moyennées. La comparaison des métriques CRPS, CRPS-sum et MSE, pour les données du cas pratique de La Défense, est présentée dans la table 5.6.

TABLE 5.6 – Métriques CRPS, CRPS-sum et MSE calculées pour les données de mobilité à La Défense.

Modèle	**CRPS**	**CRPS-sum**	**MSE**
VAR	0.624+/-0.015	0.567+/-0.020	413501
Vec-LSTM-ind-scaling	0.424+/-0.557	0.218+/-0.286	186244
Vec-LSTM-lowrank-Copula	0.224+/-0.013	0.185+/-0.016	184615
GP-scaling	0.138+/-0.002	0.122+/-0.011	75677
GP-Copula	0.177+/-0.001	0.143+/-0.003	151354
LSTM-MAF	**0.053+/-0.006**	**0.024+/-0.002**	**6646**
DeepNegPol	**0.051+/-0.023**	**0.008+/-0.003**	**5207**

Les deux modèles **DeepNegPol** et **LSTM-MAF** sont les plus performants, quelle que soit la métrique considérée dans la table 5.5.

Le modèle **LSTM-MAF** ne fait pas l'hypothèse *a priori* d'une distribution définie pour les données, et peut ainsi s'adapter à un large panel de distributions de données. De plus, ce modèle n'a pas besoin de passer par l'estimation, approximée, d'une matrice de covariance entre les séries de comptages. Le modèle **DeepNegPol**, à travers la vision « somme et partage » a bien intégré la structure des séries de comptages et de quelle manière elles évoluent en conséquence du contexte exogène. Il se détache bien du modèle **LSTM-MAF** à travers la métrique CRPS-sum en raison de sa spécificité à prédire la somme des séries.

Le modèle **GP-scaling** présente lui aussi de bonnes performances sur notre cas d'étude. Cela pourrait venir du fait qu'à chaque itération de l'entraînement, les modèles GP prédisent différents groupes de séries, ce qui les rend moins sujets à l'*overfitting*. Cette particularité rend par ailleurs les modèles avec un processus gaussien (GP) plus robustes que les modèles **Vec-LSTM-ind-scaling** et **Vec-LSTM-lowrank-Copula** dans cette étude.

Ci-après nous présentons quelques résultats de prédiction, faits avec le modèle **DeepNegPol** sur les données de La Défense, pour quelques événements particuliers. Dans chaque situation, nous représentons, en bleu, les flux observés en quelques lieux sélectionnés. Les intervalles de confiance des prédictions sont représentés en rouge ; il s'agit des intervalles entre les quantiles 5% et 95% des échantillons prédits. Nous représentons dans la figure 5.4 les résultats, pour des périodes avec ou sans événements particuliers au stade Arena, pour le poste P2 (accès privilégié entre les transports et le stade). Dans le cas du concert, le modèle

est en capacité de bien prédire un surplus de flux sortant (P2 sortants) du pôle de transport en amont du concert, et un retour massif, après le concert (P2 entrants). La capacité de prédiction a été également correcte pour la prise en compte de l'influence d'un meeting politique dans ce même stade (colonne « Événement »). Notons que les résultats pour ce type d'événement sont variables d'un entraînement à l'autre, contrairement aux concerts. Les événements sportifs (ou autres) ont des effets plus variables sur les fréquentations, car pouvant arriver à différentes périodes de la journée ; contrairement aux concerts, ils sont donc plus difficiles à bien prédire.

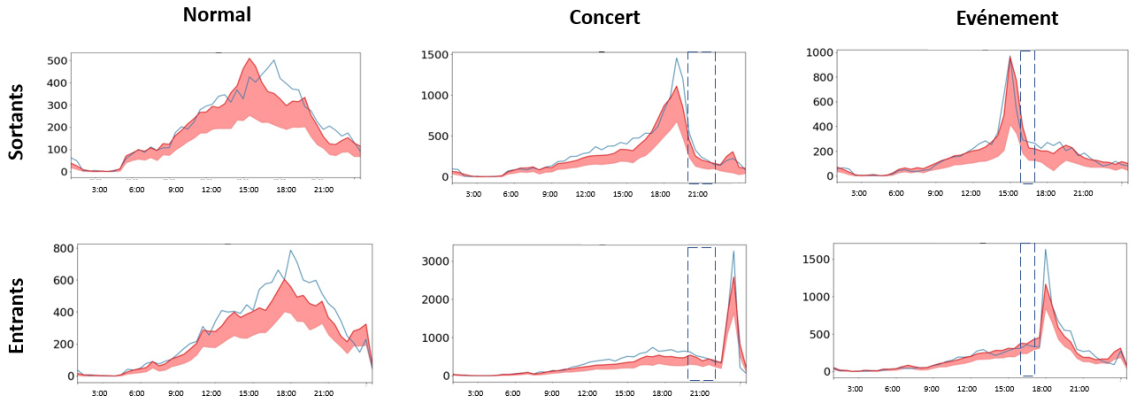


FIGURE 5.4 – Flux observés (en bleu) et prédictions (en rouge), pour la zone de comptage P2, lors d'une période sans événement, un concert et un meeting politique à l'Arena. Le concert et le meeting politique se sont déroulés au cours des périodes encadrées.

L'impact des perturbations est bien plus difficile à prévoir. Pour l'ensemble des modèles, la prédiction de l'impact des petites perturbations était souvent mal prise en compte, tant les effets sont variables. Pour les perturbations plus importantes en revanche, l'encodage que nous avons utilisé semble apporter des résultats satisfaisants. Nous présentons dans la figure 5.5 les résultats de prédiction pour deux périodes de perturbations, ayant coupé la circulation sur le tronçon central du RER A. La première perturbation a eu lieu lors de la pointe du soir, et a notamment poussé un grand nombre d'usagers à transiter du RER A vers le métro 1 pour effectuer leur trajet. Cet impact se voit sur le flux EM avec une augmentation forte le soir, bien prise en compte par le modèle. La deuxième perturbation a eu lieu lors de la pointe du matin. Une diminution momentanée des flux sortants du RER A (ES) a eu lieu lors de cette perturbation, mais a mal été prise en compte. Ici, les usagers ont plus tendance à arriver en métro qu'en RER A, ce qui se voit au niveau du flux ME, qui traduit les usagers issus du métro et transitant par les espaces du RER A pour sortir. L'augmentation de ce flux est observée, et partiellement prise en compte par le modèle. Notons que, comme pour les événements à l'Arena, les perturbations sont des événements dont l'impact est difficile à prédire. Certains entraînements arrivent à bien

prendre en compte ces événements, mais pas tous.

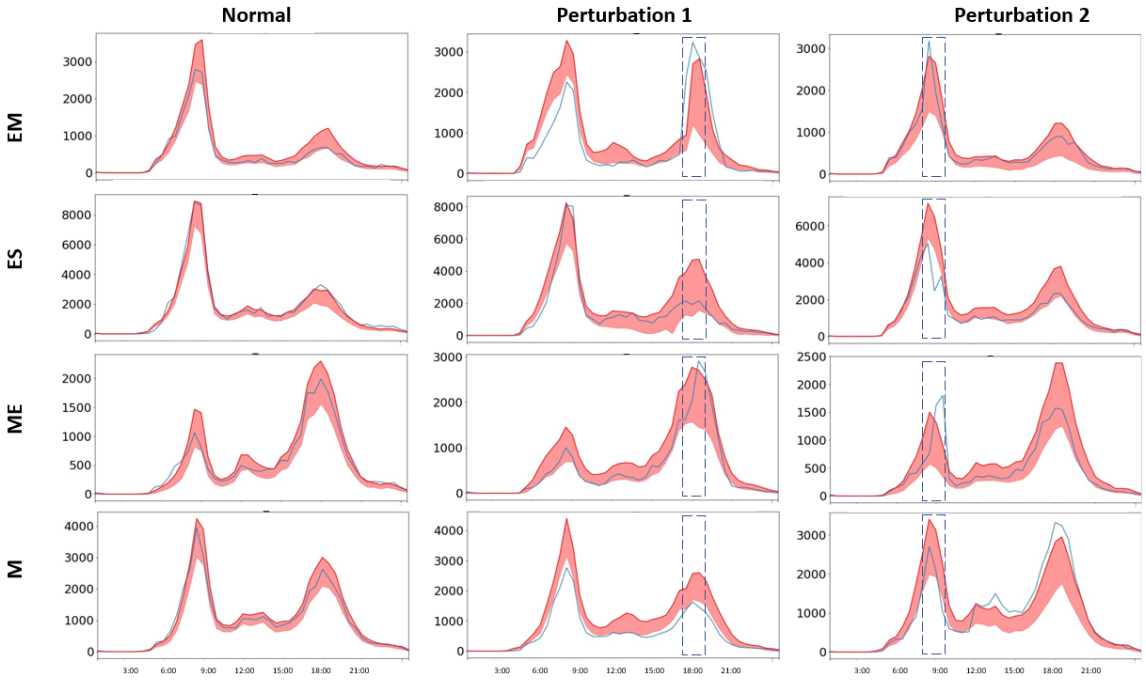


FIGURE 5.5 – Flux observés (en bleu) et prédictions (en rouge), pour quatre zones de comptage (EM, ME, M et ES), lors de deux périodes de coupure du tronçon central du RER A pendant la pointe du soir et du matin.

5.5 Conclusion

Dans ce travail, nous avons proposé et implémenté un modèle de prédiction probabiliste, qui s'inspire du modèle « sommes et partages » et des modèles autorégressifs de prédictions probabilistes (comme DeepAR). L'idée d'un tel modèle est de prédire les séries temporelles *via* une décomposition du problème en deux parties. Le modèle prend en compte et prédit la totalité de l'affluence dans le système étudié d'une part, puis prédit la répartition de cette totalité en différents points de comptages d'autre part. Cette modélisation présente un double intérêt :

— elle permet d'éviter la modélisation explicite des dépendances entre les séries étudiées,

— elle permet de différencier la modélisation du total et de la répartition, si les effets influençant l'un et l'autre sont différents par exemple.

Nous avons appliqué ce modèle à plusieurs bases de données en *open access*, puis à notre cas d'étude du pôle de La Défense. L'architecture présente un premier modèle qui

prédit le total des flux voyageurs entrants, sortants et de transit à La Défense. Ensuite, un deuxième modèle utilise la prédiction du premier modèle, et estime la répartition de la somme en de multiples points de comptages. Cette méthode est comparée à d'autres méthodes de l'état de l'art de prédictions probabilistes, basées sur le *deep learning*. Notre méthode est originale par rapport aux autres, dans le sens où elle se passe de l'estimation d'une matrice de covariance entre les différentes séries temporelles. Les résultats nous ont montré que notre méthode pourrait trouver son avantage dans les cas où il y a un besoin de prédire un des séries temporelles régulières dans le temps, tout en évitant d'estimer les corrélations entre chaque série.

Nous avons constaté, dans le cas d'étude des données de La Défense, que notre modèle et le modèle LSTM-MAF étaient meilleurs que les autres. La vision « somme et partage » de notre modèle et le fait que LSTM-MAF ne fait pas l'hypothèse d'une distribution en particulier (mais passe plutôt par l'assemblage de plusieurs couches de flux conditionnel, ce qui lui confère une meilleure adaptabilité aux données) expliquent leurs supériorité.

Les trois modèles de prédiction DeepNegPol, LSTM-MAF et GP-Copula sont de bons modèles à utiliser pour de la prédiction de séries multivariées de comptages. DeepNegPol semble ainsi particulièrement adapté pour des séries régulières, en grande dimension ou non et devrait être appliqué dans ce type de cas d'études. Les modèles GP-Copula et LSTM-MAF semblent être de meilleurs alternatives pour des séries moins régulières ou avec une évolution rapide du contexte (pour GP-Copula). Le modèle DeepNegPol vient ainsi enrichir la littérature des réseaux de neurone récurrents pour la prédiction probabiliste de données multivariées.

Chapitre 6

Conclusion et perspectives

La collecte de grandes quantités de données d'affluence dans les espaces publics permet de connaître finement les comportements collectifs, dans un contexte qui évolue. Dans le cadre de politiques de lissage des heures de pointe, des analyses peuvent ainsi permettre de quantifier à quel point la politique a fonctionné ou non. Des événements majeurs comme la crise Covid19 ont nécessité, dans un contexte plus urgent, une gestion fine des déplacements collectifs, pour éviter la reprise épidémique. Dans cette thèse, nous avons d'abord mis en place des méthodes permettant de mieux comprendre la dynamique de l'affluence piétonne, dans l'espace très dense du pôle de transport du quartier de La Défense. Nous nous sommes concentrés pour cela, dans un premier temps, sur l'aspect temporel de l'affluence *via* un travail de décomposition des séries temporelles, qui nous a permis d'isoler et de quantifier l'impact de facteurs exogènes sur l'affluence et ce sur une longue période (9 ans). Nous avons ensuite entrepris un travail sur la caractérisation de profils-types d'affluences dans le pôle de transport. Ce travail, central dans la thèse, a permis la mise en place d'une méthode efficace de détection de ces profils, permettant ainsi une synthèse de séries multivariées de comptages bruitées. Nous nous sommes ensuite intéressés à l'anticipation des affluences futures à court terme. Nous avons pour cela proposé un modèle de prédiction probabiliste des affluences. Ce modèle a des performances de prédiction semblables aux modèles issus de l'état de l'art.

Cette thèse apporte, à travers les différents chapitres abordés, certaines contributions d'un point de vue méthodologique et opérationnel. Elle valorise les données issues des capteurs du pôle de La Défense à travers plusieurs applications qui étendent leur utilisation au-delà de l'expérimentation de lissage des heures de pointe. De plus, à travers nos travaux, nous introduisons dans le domaine de la mobilité dans les transports en commun des modèles majoritairement utilisés dans d'autres domaines tels que l'énergie, l'écologie, la finance ou la santé. Nous enrichissons ainsi la diversité des modèles utilisés dans l'analyse des mobilités dans les transports en commun. Enfin nous mettons en place de nouveaux modèles sur la base de distributions « sommes et partages ». Cette distribution, présentée

de manière théorique dans la littérature, a ici été utilisée dans des modélisations où elle n'avait pas été appliquée jusqu'à présent : un cadre de modèles de régression, un modèle de mélange, utile pour le clustering de séries multivariées et un modèle de prédiction probabiliste basé sur l'apprentissage profond. La vision « sommes et partages » a prouvé une pertinence forte à être utilisée dans ces différents cadres lorsqu'elle a été comparée aux autres méthodes issues de l'état de l'art.

Certaines limites ont été rencontrées dans cette thèse. Pour les travaux de décomposition, nous ne nous sommes pas penché sur une vision multivariée des modèles. Pourtant ce type de décomposition aurait pu être pertinente car elle aurait permis d'introduire des termes de dépendances entre les erreurs sur les composantes entre les différentes séries (par exemple entre les tendances des séries d'affluences vers le métro et le RER A). Une autre limite se pose avec la représentation « sommes et partages » abondamment utilisée dans cette thèse. Cette distribution semble très adaptée à des données sur les mobilités où les habitudes de déplacements permettent la formation de séries de comptage régulières et périodiques dans le temps. La distribution pourrait ainsi être plus limitée dans son utilisation pour des séries multivariées présentant moins de régularités temporelles.

Les présents travaux ont vocation à être étendus à d'autres cas applicatifs ou servir de base pour la recherche d'autres modèles. D'un point de vue opérationnel, l'opérateur des transports en commun RATP répond régulièrement à de nouveaux cas d'étude. Les travaux présentés dans cette thèse pourraient être appliqués à d'autres pôles ou lignes de transport en commun, à condition de disposer de suffisamment de données en entrée. Une autre application envisagée est de tester la robustesse du modèle de prédiction DeepNegPol face à une dégradation des sources de données en entrée, comme la disparition de certains capteurs de comptages. Une première application a pu émerger lors des derniers mois de la thèse, avec l'adaptation des modèles de catégorisation des dynamiques d'affluence, dans le cadre d'un stage de master 2. Les modèles ont ainsi pu servir de base pour l'analyse des flux passagers vers la ligne 13 du métro parisien. Toujours d'un point de vue applicatif, les travaux présentés dans cette thèse pourraient être étendus et adaptés à d'autres domaines. Nous nous sommes concentrés ici à l'étude spécifique d'un pôle de transport, soit une zone d'activités dense en terme de trafic mais restreinte dans l'espace. Les travaux pourraient par exemple être appliqués au cas du trafic routier, un problème en grande dimension où des portions du réseau fonctionnent plus ou moins de la même manière en fonction des événements. Cette notion d'adaptation des travaux nous mène aux perspectives d'un point de vue modélisation. Un aspect que nous n'avons pas traité au cours de cette thèse est la notion de modèles hybrides. La combinaison entre un modèle de segmentation et un modèle de prédiction peut être utile pour aider à la prédiction selon deux points de vue :

— Segmenter temporellement les données afin d'extraire des périodes homogènes puis appliquer des modèles de prédiction spécifiques à chacun de ces segments. Si l'on suppose que les prédictions proches doivent se faire dans le contexte d'un segment similaire au segment courant, la prédiction est plus aisée. Ce type de travail est

proposé dans [Ni+20].

— Segmenter spatialement les données pour appliquer la prédiction sur des groupes de données proches en terme de comportements. Dans le cadre du trafic routier on retrouve cette idée dans [LKG22].

Il est également possible de combiner les aspects décomposition et prédiction des séries temporelles. Une étape de décomposition en amont de la prédiction permet en effet d'isoler les résidus d'une série des composantes telles que la tendance ou les saisonnalités. Il suffit ensuite d'appliquer la prédiction sur les différentes composantes avec un choix judicieux de modèles pour chaque composante. Ce type de travail est proposé dans [Zha+20] ou [Che+20]. Au delà de l'adaptation des présents travaux à d'autres cas d'études ou modèles, nous n'avons pas eu la possibilité de travailler sur la détection d'anomalies et de tendances non expliquées dans nos données. Cet aspect était pourtant recherché pour constater ou non l'impact de la politique de lissage des heures de pointe à La Défense. Notamment, le modèle de décomposition aurait pu aider à isoler ce phénomène du reste dans les données. La pandémie de Covid19 et ses multiples impacts sur les mobilités a en revanche empêché le bon déroulement de cette étude. Cette dernière pourrait être reprise en cas de remise en place d'une expérimentation de lissage.

Il nous a été possible de montrer à travers cette thèse le potentiel que ces études peuvent apporter, pour mieux gérer l'affluence dans un espace de transport en particulier. Les outils mis en place ici ont vocation à être utilisés comme bases pour d'autres recherches, pour toute personne travaillant sur l'analyse des flux piétons, dans tout espace équipé d'un système de collecte de données. Nous espérons que nos travaux pourront, de cette manière, contribuer à d'autres recherches.

Annexe A

Métriques de comparaison des modèles

Pour évaluer différents modèles, afin de sélectionner le meilleur, nous avons basé nos différents travaux sur les critères suivants :

— Le critère d'information d'Akaike ou AIC ([Aka74]), défini par

$$AIC = -2L + 2\kappa, \tag{A.1}$$

où L est la log-vraisemblance du modèle, et κ le nombre de paramètres estimés. Le meilleur modèle est celui qui minimise le critère AIC.

— Le critère bayésien d'information ou BIC, défini par

$$BIC = -2L + \log(n_{obs})\kappa, \tag{A.2}$$

où n_{obs} est le nombre d'observations.

— Le critère de vraisemblance intégré complet ou ICL ([BCG00]) est défini par

$$ICL = -2L_c + \log(n_{obs})\kappa, \tag{A.3}$$

où L_c est la log-vraisemblance complétée du modèle. Le critère ICL est étroitement lié au critère BIC, mais est aussi pénalisé par l'entropie moyenne estimée.

— La racine de l'erreur quadratique moyenne entre les observations et les prédictions (RMSE), évaluée sur un jeu de test pour différents horizons de prédiction $h \geq 1$:

$$RMSE(h) = \sqrt{\sum_{i=1}^{n-h} \frac{(\hat{y}_{i+h}(i) - y_{i+h})^2}{n-h}}, \tag{A.4}$$

où $\hat{y}_{i+h}(i)$ est la prédiction de y_{i+h} obtenue depuis les observations $(y_1, \ldots, y_i)$.

— Le score de probabilité continu rangé (*Continuous Ranked Probability Score*, CRPS) ([MW76]) est bien adapté aux cas de modèles probabilistes. Cette métrique mesure à quelle distance la prédiction se trouve de l'observation, dans un contexte probabiliste. La meilleure prédiction probabiliste est ainsi celle pour laquelle toute la fonction de masse se trouve au niveau de l'observation. CRPS se calcule comme suit :

$$CRPS(F, x) = \int_{-\infty}^{\infty} (F(y) - 1(x \leq y))^2 dy, \tag{A.5}$$

avec F la distribution cumulée prédite, et x une observation. $1(x \leq y)$ est une fonction indicatrice qui vaut 1 si $x \leq y$ et 0 sinon. Ici, F est calculée de manière empirique avec S échantillons, comme $\hat{F}(y) = \frac{1}{S} \sum_s 1(x^{(s)} \leq y)$.

— L'erreur quadratique moyenne (MSE) est calculée comme :

$$MSE = \frac{1}{NT} \sum_{i,t} (y_{i,t} - \hat{y}_{i,t})^2 \tag{A.6}$$

pour N séries, T pas de temps et $\hat{y}$ une prédiction

Annexe B

Isoler et quantifier l'impact de facteurs long-terme et journaliers sur les mobilités

B.1 Choix d'un type de modèle DLM : additif ou multiplicatif

L'objectif de cette annexe est de comparer les versions, multiplicative (équation B.1) et additive (équation B.2), du modèle de décomposition, afin de sélectionner le modèle le plus adapté pour expliquer nos données de flux.

$$\log(y_j) = l_j + s_j + f_j + \sum_{s=1}^{d} \beta_j^{(s)} \psi_j^{(s)} + \nu_j, \tag{B.1}$$

$$y_j = l_j + s_j + f_j + \sum_{s=1}^{d} \beta_j^{(s)} \psi_j^{(s)} + \nu_j, \tag{B.2}$$

Il existe deux façons de choisir entre ces deux types de modèles :

— **Observations empiriques**
Contrairement au modèle additif, le modèle multiplicatif peut aborder l'effet cumulatif de plusieurs composantes qui affectent négativement les flux, sans pour autant considérer des valeurs négatives de flux. C'est ce que nous avons observé en visualisant les prédictions à un jour, réalisées avec les deux formes de modèles. Par exemple, prenons le mois d'août 2017, où le 15 août était férié, pendant les vacances d'été. La figure B.1 montre le trafic entrant dans la gare du RER A, et les prédictions faites par les deux modèles. On constate que le modèle additif prend en compte un effet

très fort qui, combiné à l'effet des vacances d'été, conduit à une prédiction faussée pendant le jour férié.

— **Comparaisons sur les capacités de prédiction des deux modèles**
Les deux modèles sont comparés sur la base de leurs capacités de prédiction, sur plusieurs horizons temporels. Pour ce faire, nous avons utilisé la notion de RMSE par horizon de prévision (voir annexe A). Les périodes d'apprentissage couvraient les années 2011 à 2015, et la base de test était l'année 2016. Les résultats sont présentés dans la figure B.2. Le modèle multiplicatif fournit de meilleures prédictions que le modèle additif, pour les flux entrants journaliers du RER A, quel que soit l'horizon de prévision (h). Ce résultat est moins marqué pour le métro 1, où le modèle multiplicatif n'est meilleur que pour les prévisions à h = 4 à h = 7. Dans ce cas, des travaux de maintenance ont été effectués en été. Aux petits horizons de prévision, la phase de filtrage de l'algorithme de Kalman a plus de mal à corriger les composantes, dans le cas des données logarithmiques du modèle multiplicatif, que dans le cas des données non logarithmiques du modèle additif, lorsque la période de maintenance est rencontrée. A des horizons de prévision plus larges, la différence disparaît, car les composantes ont eu plus de temps pour s'adapter au changement.

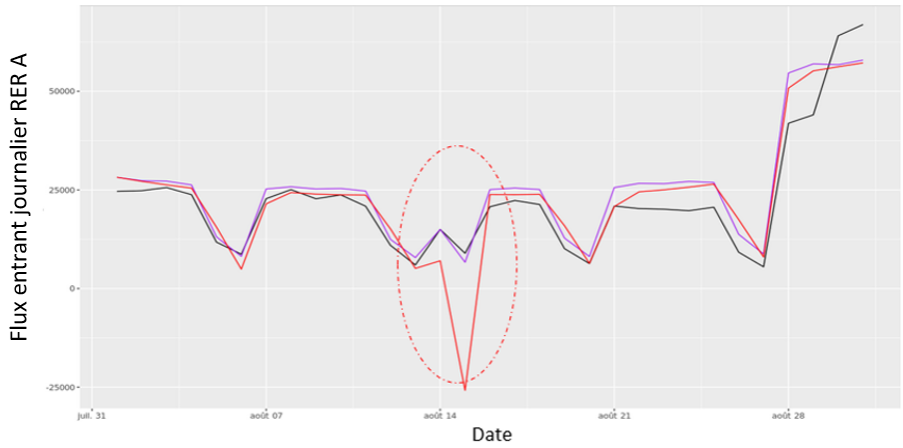


FIGURE B.1 – Flux journaliers entrants, observés et prédits à un jour, sur la ligne du RER A, à la station La Défense Grande Arche. Les comptages observés sont en noir, les prédictions faites avec le modèle additif en rouge, et les prédictions faites avec le modèle multiplicatif en violet.

Nous avons choisi le modèle multiplicatif pour la décomposition des séries temporelles. Ce modèle a l'avantage de forcer les valeurs observées à rester positives, et ses capacités de prédiction sont proches de celles du modèle additif.

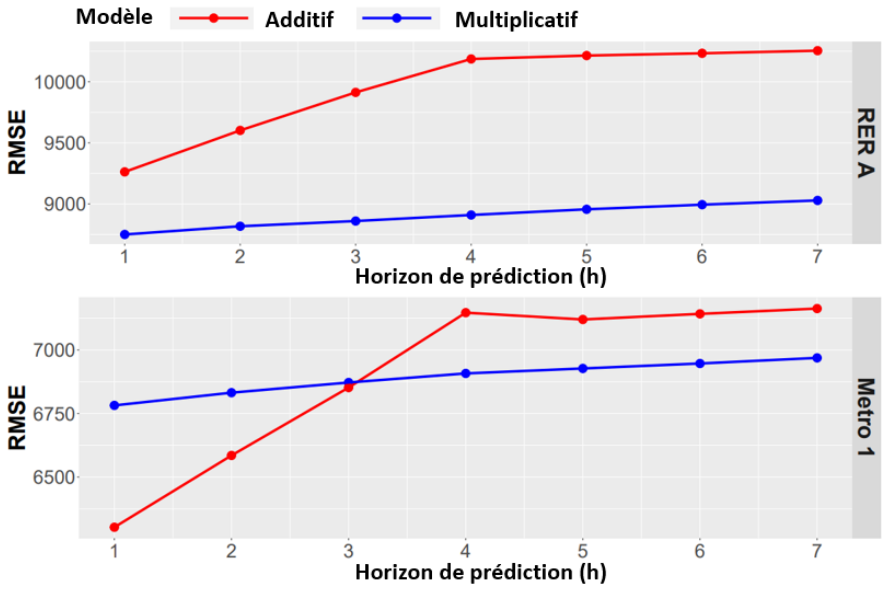


FIGURE B.2 – RMSE(h) calculées pour un horizon de prévision (h) allant de 1 à 7 sur l'année 2016, pour les modèles additif et multiplicatif appliqués aux flux journaliers entrant dans le RER A (en haut), et le métro 1 (en bas).

Annexe C

Identifier des groupes similaires de profils de mobilité

C.1 Recherche de profils de mobilité temporels, en univarié et sans contexte, dans le pôle de La Défense

C.1.1 Introduction

Nous visons dans ce travail exploratoire à mettre en évidence des profils de mobilité temporels, en regroupant en classes homogènes les jours ayant des profils de mobilité similaires, au sein du pôle de transport de La Défense. Les variabilités temporelles détectées traduisent des différences d'utilisation du pôle de transport selon la période. L'étude peut se faire à échelle locale, sur les fréquentations captées par quelques postes de comptage en particulier, ou à échelle globale, celle du pôle de transport dans son ensemble. Classifier les jours de fréquentation à une échelle locale permet de constater comment un certain type d'activité (travail, commerce, loisir) évolue selon les jours. A l'échelle globale, l'intérêt est de constater comment évolue la fréquentation du pôle dans son ensemble. Le travail présenté ici est basé sur l'utilisation de modèles de mélange de Dirichlet-multinomiales, pour la recherche de similarités temporelles des dynamiques de fréquentation, en se basant sur les séries en univarié uniquement et sans la prise en compte de contexte. Les variabilités spatiales de déplacements traduisent les différences de comportement dans l'usage des différentes zones du pôle de la Défense. Selon que l'environnement proche d'un point de comptage soit composé de zones de travail, de commerces, de congrès ou de loisirs, les flux captés vont fortement différer. Afin d'illustrer l'évolution de l'utilisation spatiale du pôle de transport dans les différentes catégories temporelles détectées à l'échelle globale, nous utilisons une méthode heuristique pour calculer des similarités spatiales entre zones du pôle. Ce travail permet de constater le degré d'homogénéité avec lequel le pôle est utilisé, selon les périodes. Dans la section suivante, le modèle de mélange de Dirichlet-multinomiales est

détaillé, ainsi que son application dans la recherche de profils de mobilité temporels.

C.1.2 Le modèle de mélange Dirichlet-Multinomial

Considérons un vecteur $\mathbf{y} = \{\mathbf{y}_i\}_{i\in\{1,..,I\}}$ avec chaque $\mathbf{y}_i$ de taille L. Le modèle de mélange de Dirichlet-multinomiales catégorise les I observations en S clusters, selon le modèle génératif suivant :

$$(\pi_s)_{s\in 1,...,S} \sim \mathcal{D}(\alpha) \tag{C.1}$$
$$Z_i \sim \mathcal{M}(1, (\pi_s)_{s\in 1,...,S}) \tag{C.2}$$
$$\boldsymbol{\theta}_s \sim \mathcal{D}(\beta) \tag{C.3}$$
$$\mathbf{y}_i | Z_{i,s} = 1 \sim \mathcal{M}(v_i, \boldsymbol{\theta}_s), \tag{C.4}$$

où $v_i = \sum_l y_{i,l}$, α et β sont les paramètres des loi *a priori* et appartiennent respectivement à $\mathbb{R}_+^S$ et $\mathbb{R}_+^L$. Le vecteur $\boldsymbol{\theta}_s$ est un vecteur de L probabilités, défini comme $\boldsymbol{\theta}_s \in]0,1[^L$.

Des modèles de mélange de Dirichlet-multinomiales sont proposés dans le package R greed, associé aux travaux réalisés par [Côm+21], qui implémente une méthode hiérarchique basée sur la maximisation d'une modification du critère *Integrated Completed Likelihood* (ou ICL, voir annexe A). Le package propose de trouver une partition optimale pour S clusters, à travers un algorithme génétique hybride. Une fois qu'une solution de partitionnement est obtenue pour S^* clusters, une méthode agglomérative hiérarchique est proposée pour chercher des solutions de partitionnement pour toutes valeurs $S \in 1,..,S^*$. Considérons Q comme la différence d'ICL entre les partitions à un cluster et à S^* : $ICL(S = 1) - ICL(S^*) = Q$. Dans notre méthodologie, on sélectionnera le nombre optimal de clusters S^*, comme le premier pour lequel $S^* < ICL(S^*) + 0.01Q$.

Les profils de mobilité temporels journaliers à détecter sont soit globaux (échelle du pôle), soit locaux (une zone en particulier). Il s'agit ici de détecter des catégories de profils journaliers où les flux de personnes sont répartis de la même manière temporellement. Pour ce travail, on veut catégoriser les J jours en S catégories. Avec nos données de comptage, le vecteur d'observations $\mathbf{y}$ peut s'écrire sous la forme $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_J\}$, avec chaque $\mathbf{y}_j = \{y_{h,p}\}_{h\in\{1,...,H\},p\in\mathbb{P}*}$ un vecteur de taille $H|\mathbb{P}*|$ de tous les comptages relevés à chaque tranche horaire h et chaque poste p d'un sous-ensemble $\mathbb{P}*$ au jour j. Dans le cas où nous étudions une zone du pôle en particulier (échelle locale), l'ensemble $\mathbb{P}*$ se réduit à un poste en particulier ($|\mathbb{P}*| = 2$ car nous considérons pour un même poste les flux entrants et sortants), tandis que $|\mathbb{P}*| = P$ lorsque l'on travaille pour tout le pôle (échelle globale). Dans le cadre de la recherche de zones du pôle fonctionnant de la même manière, nous mettons en place une méthode heuristique présentée dans l'annexe C.2.

C.1.3 Résultats

Classification des jours à une échelle locale ($|\mathbb{P}*| = 2$)

Nous explorons ici trois postes connus pour capter trois types d'activité : le poste 7, qui capte les flux essentiellement liés aux activités de travail ; le poste 2, qui capte occasionnellement des flux liés à des événements festifs (concerts à l'Arena) ; et le poste 12, qui capte les flux liés à une activité commerciale (Les 4 Temps).

Poste 7 : Accès vers la dalle et les tours de bureau

Le poste 7 capte des flux de personnes qui transitent entre la salle d'échanges de la station Grande Arche et la dalle, à partir de laquelle elles peuvent se rendre aux différents gratte-ciel du pôle, qui hébergent les activités de travail. Nous représentons dans la figure C.1 les catégories de profils de fréquentation détectées, ainsi que leur positionnement dans le temps (calendrier).

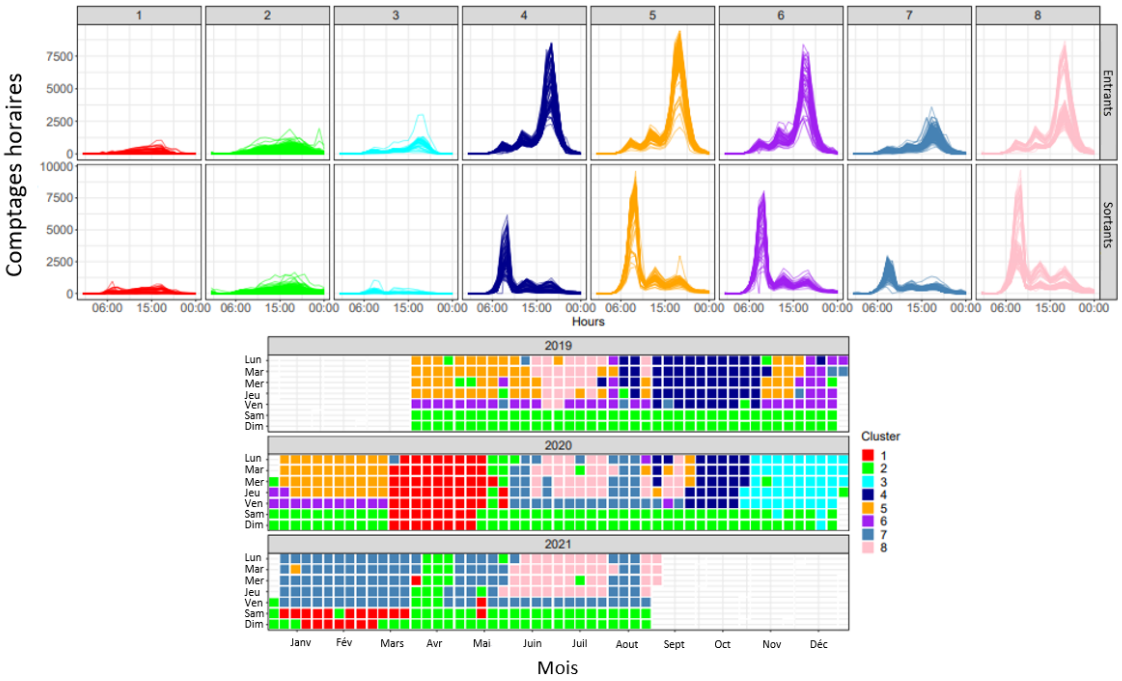


FIGURE C.1 – Nombre de validations en entrée (haut) ou sortie (bas) par heure, durant la période du 01 Avril 2019 Mars au 31 Août 2022, pour le poste 7. Les jours sont colorés selon le cluster auquel ils sont affiliés. En dessous, le calendrier de la classification des jours, obtenu à l'aide du clustering.

Les catégories de profils rencontrées sont fortement associées à l'alternance « jours de semaine/jours de weekend ». Avant la période de Covid19 (débute en mars 2020), les jours

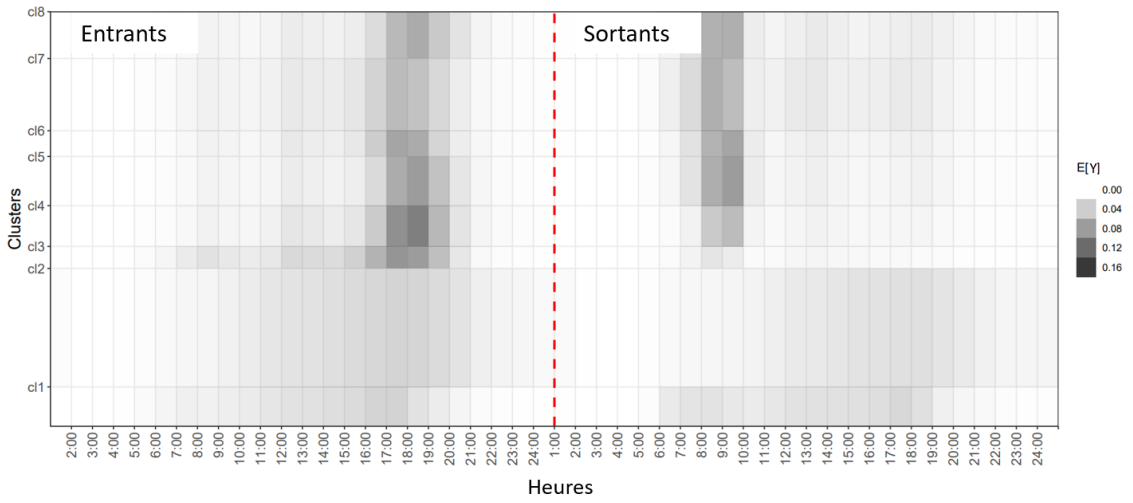


FIGURE C.2 – Représentation des 8 catégories de profils pour le poste P7

de semaine sont associés à différents types de profils. Le cluster 5 rassemble en grande partie les jours travaillés. Les vendredis sont souvent associés au cluster 6. D'après la figure C.2, cette différence réside dans le fait que les profils du cluster 6 présentent des pointes du soir moins chargées entre 18h et 19h, ce qui souligne des départs plus tôt le vendredi. Le cluster 8 est bien représenté pendant les périodes de beaux jours (juin-juillet) avant travaux du RER A en 2019, 2020 et 2021. Le cluster 8 a peu de différences avec le cluster 5, excepté que la pointe du matin est moins prononcée, et qu'il y a davantage de passages l'après-midi, en raison des beaux jours et d'un afflux conséquent plus important vers la dalle. Le cluster 4 est associé à une période où le poste 7 était défaillant pour comptabiliser tous les comptages sortants. Pendant la période de Covid19, le premier confinement est totalement contenu dans le cluster 1. Sur l'ensemble de la période, les weekends sont associés au cluster 2. Les clusters 1 et 2 sont totalement dissociés des jours de travail, à la différence que les profils du cluster 2 s'étalent davantage sur le soir, car il s'agit de périodes le plus souvent hors restrictions. Le deuxième confinement est contenu dans le cluster 3. On peut mentionner les périodes de couvre-feux contenus dans le cluster 7, avec, comme on peut le voir dans la figure C.2, des flux entrants concentrés plus tôt que pour les autres profils.

Poste 2 : Accès vers l'Arena

Le poste 2 se situe entre la dalle et la salle d'échange, et se trouve directement sur le point de passage entre la station Grande Arche et la salle de concert Défense Arena. Cette particularité est visible au niveau des courbes de fréquentations, où l'on voit pour certains jours une affluence très importante le soir, notamment un nombre important de personnes entrant dans la station après avoir quitté l'événement. La figure C.3 rassemble les résultats.

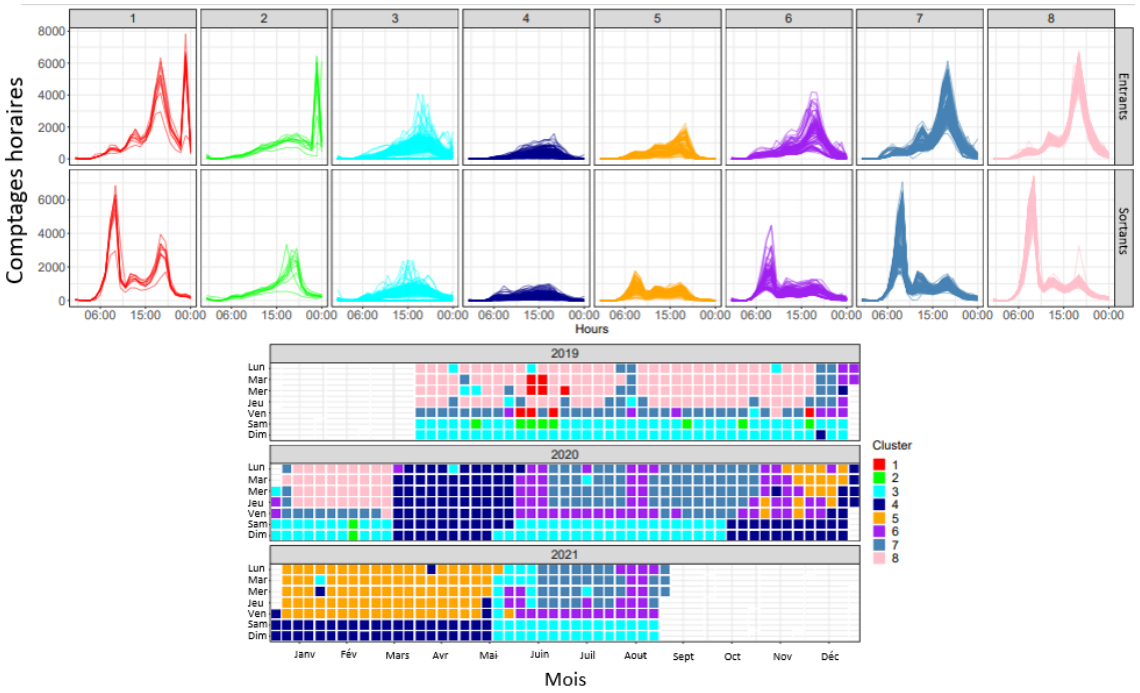


FIGURE C.3 – Nombre de validations en entrée (haut) ou sortie (bas) par heure, durant la période du 01 Avril 2019 Mars au 31 Août 2022, pour le poste 2. Les jours sont colorés selon le cluster auquel ils sont affiliés. En dessous, le calendrier de la classification des jours, obtenu à l'aide du clustering.

Il apparaît un ensemble de catégories de profils journaliers reflétant l'historique des dynamiques de fréquentation à ce poste. On peut mentionner les périodes de travail « normales », non impactées par les confinements liés à la crise du Covid19, qui sont associées aux clusters 1, 7 et 8 pour les jours de semaine, et 3 pour les jours de week-end. Les vendredis sont dissociés des autres jours de semaine (cluster 7 pour les vendredis, et 8 pour les lundis aux jeudis). Les clusters 6 et 7 peuvent être considérés comme des catégories de transition entre périodes de confinement et périodes normales. Les clusters 4 et 5 sont associés à des périodes de confinement et de couvre-feux, où la fréquentation était très limitée. Les clusters 1 et 2 témoignent de la spécificité de ce lieu : ils sont associés à des jours de concert à l'Arena, en semaine et le week-end respectivement. Ces catégories sont caractérisées par un flux important de personnes sortant du pôle avant le concert, et un pic important d'entrées après le concert.

Poste 12 : accès aux 4 Temps

Le poste 12 compte les passages entre la salle d'échange de la station Grande Arche et Les

4 Temps. Les résultats de clustering temporel sont présentés dans la figure C.4.

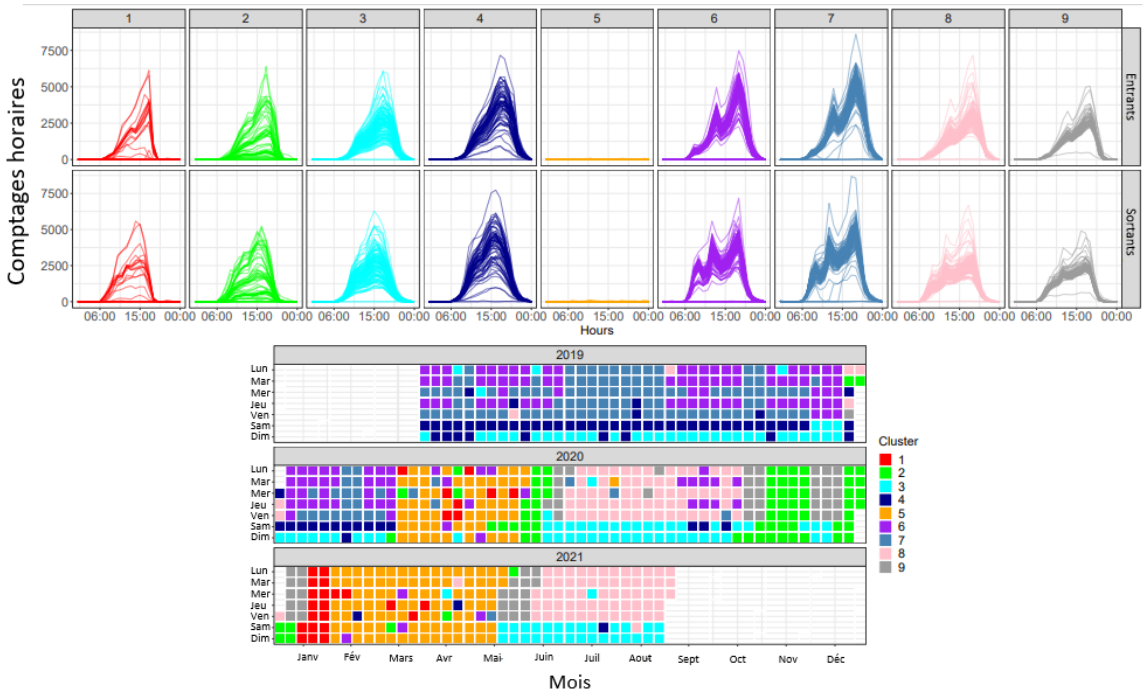


FIGURE C.4 – Nombre de validations en entrée (haut) ou sortie (bas) par heure, durant la période du 01 Avril 2019 Mars au 31 Août 2022, pour le poste 12. Les jours sont colorés selon le cluster auquel ils sont affiliés. En dessous, le calendrier de la classification des jours, obtenu à l'aide du clustering.

Les catégories de jours classifiées semblent refléter des jours-types de consommation. Avant la pandémie, et hors vacances, on note ainsi que les lundis, mardis et jeudis rentrent dans le cluster 6, tandis que les mercredis et vendredis rentrent dans le cluster 7. Les vacances sont elles aussi associées au cluster 7. La différence entre ces deux profils est très légère, on peut l'observer dans la figure C.5 : les profils du cluster 7 ont une pointe du matin (9h00-10h00) légèrement moins chargée, et des transits du centre commercial vers le pôle de transport qui s'étalent un peu plus tard le soir. Les mercredis, vendredis et jours de vacances semblent être favorisés pour la consommation. Les samedis et dimanches sont respectivement associés aux clusters 4 et 3. Dans la figure C.5, on constate que la différence entre ces deux profils réside principalement dans une fréquentation le soir plus tardive pour les profils du samedi, que du dimanche. Les périodes associées à la pandémie de Covid19 sont bien visibles, avec des périodes de fermeture totale du centre commercial, captées par le cluster 5. Le deuxième confinement intègre le cluster 2, tandis que les clusters 8 et 9 sont associés à des périodes de transition. Il apparaît un net impact du couvre-feu dans le

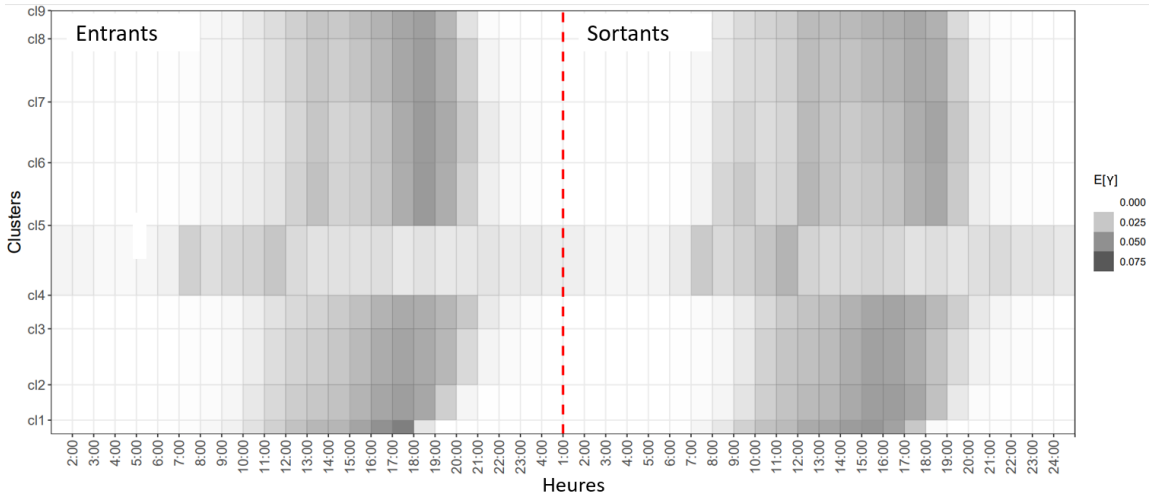


FIGURE C.5 – Représentation des 9 catégories de profils pour le poste P12

cluster 1, visible dans la figure C.5.

L'étude à échelle locale illustre ainsi comment l'environnement proche des postes influence les dynamiques de fréquentation différemment selon les zones. Nous explorerons plus en détail les similarités spatiales qui en découlent avec l'aide de la méthode heuristique que nous aborderons dans la suite.

Classification des jours à une échelle globale ($|\mathbb{P}*| = P$)

A l'échelle de tous les flux entrants et sortants du pôle de transport, l'analyse des profils de mobilité journaliers peut s'avérer utile pour détecter des périodes de fonctionnement du pôle. Les résultats sont rassemblés dans la figure C.6. De la même manière que pour les catégorisations par poste, on remarque une rupture entre les catégories détectées avant le début de la pandémie de Covid19 et après. Avant la pandémie, on trouve les clusters 7 et 5 associés aux jours travaillés, avec le premier présent hors vacances du lundi au jeudi, tandis que le deuxième est représenté en périodes de vacances et les vendredis. Lors de cette période pré-Covid19, les jours de weekend sont rassemblés dans le cluster 1. Les clusters 3 et 4 sont associés à la période atypique de la grève de décembre 2019. La période de la pandémie de Covid19 est caractérisée par d'autres catégories de profils temporels. Le premier confinement est contenu dans le cluster 4, tandis que le deuxième confinement est associé aux profils du clusters 3 en semaine. Le cluster 3 représente également la période où des mesures de couvre-feu ont été prises en semaine. Le cluster 2 rassemble les jours de weekend, pendant les périodes de couvre-feux entre octobre 2020 et juin 2021. Enfin, on peut mentionner le cluster 5, qui représente des périodes de reprise vers un trafic normal après les vagues de pandémie.

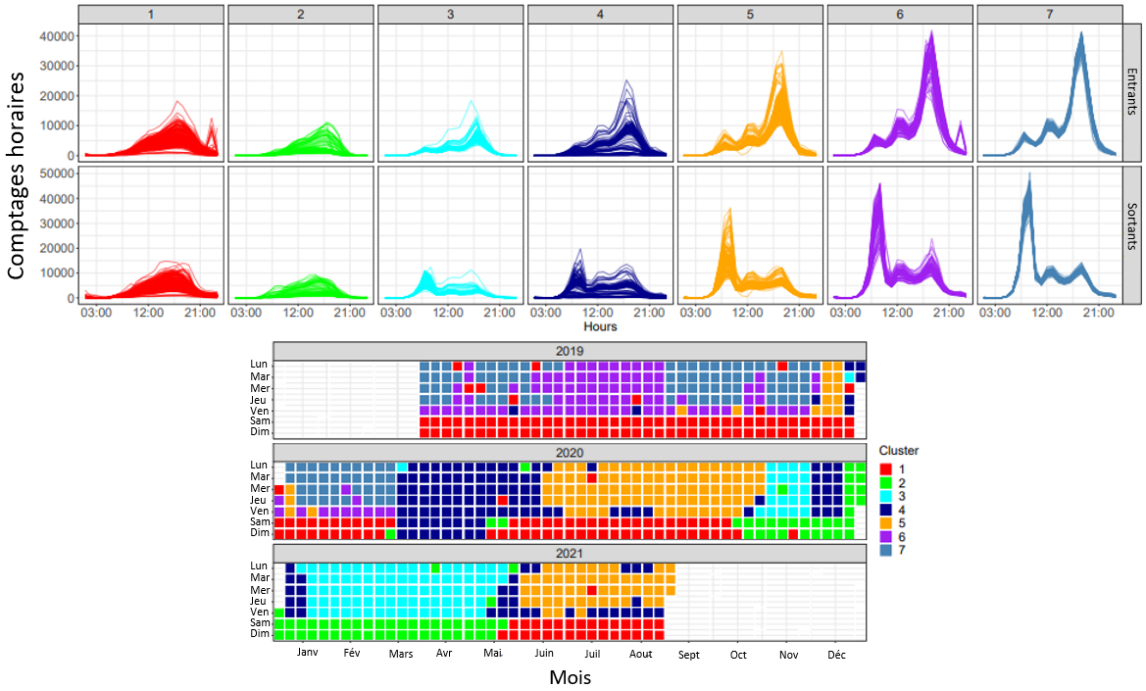


FIGURE C.6 – Nombre de validations en entrée (haut) ou sortie (bas) par heure, durant la période du 01 Avril 2019 Mars au 31 Août 2022, pour le total des flux entrants et sortants dans le pôle de transport. Les jours sont colorés selon le cluster auquel ils sont affiliés. En dessous, le calendrier de la classification des jours, obtenu à l'aide du clustering.

Depuis le début de nos travaux, le pôle de transport est passé par plusieurs périodes qui ont modifié en profondeur les dynamiques journalières de fréquentation. Si un impact temporel de ces périodes est visible, nous pouvons également nous demander si l'impact se joue aussi d'un point de vue spatial. C'est avec le travail exploratoire de la section suivante que nous allons tenter de répondre à cette question. Nous allons notamment calculer, pour différentes périodes détectées par le clustering temporel précédemment mené, un score de similarité entre les différentes zones du pôle de transport.

Recherche de similarités spatiales avec une méthode heuristique

Nous avons calculé les scores de similarité entre les lieux du pôle avec la méthode détaillée dans l'annexe C.2, pour différentes périodes de fonctionnement détectées avec l'étude sur la classification des jours à une échelle globale (voir figure C.6). Nous avons choisi d'analyser les similarités pour trois périodes : les jours de semaine en période normale (cluster 7), les jours de week-end (cluster 1) et des jours de semaine sous restriction, en période de pandémie (cluster 3). Pour les jours de semaine, nous obtenons les similarités

représentées sous forme d'un arbre de distances dans la figure C.7. Un premier élément que l'on peut constater est que ce sont les centres commerciaux (P11 et P12) qui se différencient le plus du reste du pôle dans le cas des flux sortants, tandis que ce sont les accès aux lignes (E et M) qui sont les plus séparés du reste dans le cas des flux entrants. L'activité de travail contribue, comme attendu, à un rapprochement des dynamiques de fréquentation des zones qui desservent les lieux de travail (P4, P6, P7, P8, P9, P10). Les postes P1 et P2, proches d'une université, se distinguent des autres catégories.

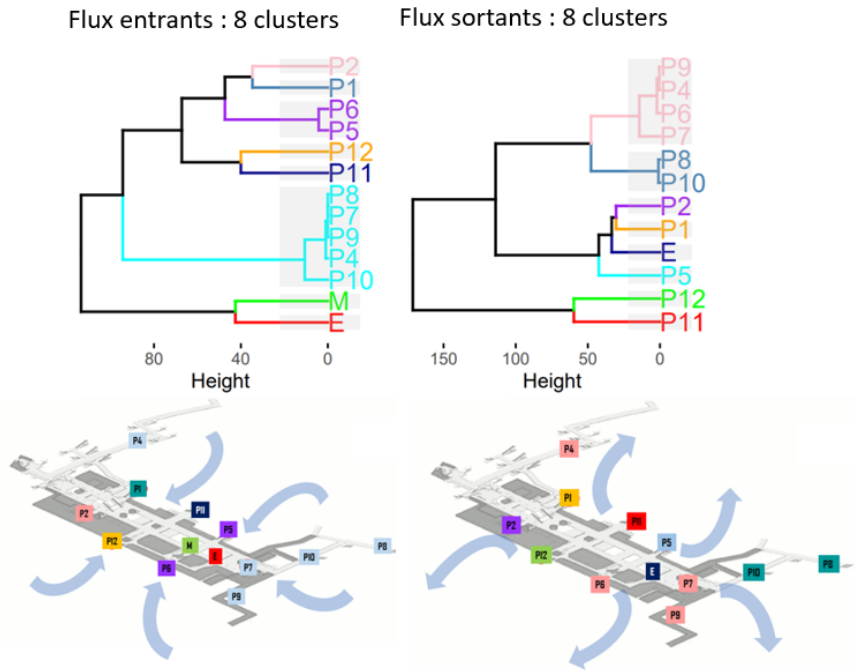


FIGURE C.7 – Arbre des similarités entre lieux, pour les jours de semaine, hors vacances et avant Covid19

Les scores de similarité pour les jours de week-end sont présentés dans la figure C.8. Un premier élément que révèle cet arbre est que les différentes zones du pôle sont davantage dissociées le week-end que les jours de semaine (davantage de clusters, hauteurs plus importantes). Un rapprochement se fait néanmoins entre les centres commerciaux dans le sens des flux sortants. Cela souligne l'importance de ces centres comme attracteurs vers le pôle lors du week-end. En revanche, le travail ne régit plus les flux entrants et sortants, ce qui élimine le rapprochement qui pouvait être fait entre les postes P4, P6, P7, P8, P9 et P10. Notons également que, dans le sens des flux entrants, les accès vers le métro et le RER sont utilisés de manière plus similaire que dans le cas des jours de semaine.

Pour le cas des jours de semaine en période de restriction contre la pandémie de Covid19,

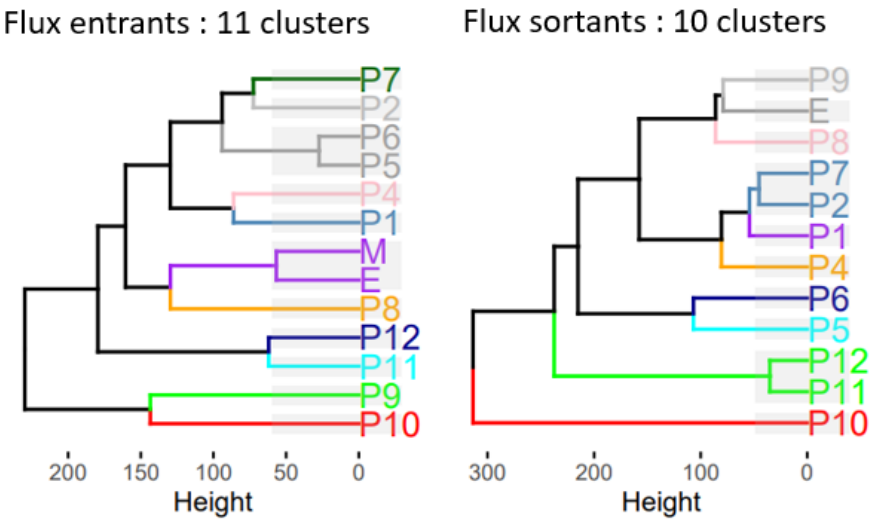


FIGURE C.8 – Arbre des similarités entre lieux, pour les jours de week-end avant Covid19

nous pouvons tirer en partie les mêmes conclusions qu'avec les jours de week-end : comme le travail est très peu présent, les postes traditionnellement proches ne le sont plus ici. De plus, les centres commerciaux ne sont eux-même plus utilisés lors de cette période, les postes associés sont donc encore plus différents du reste du pôle.

C.1.4 Conclusion

Ce travail exploratoire a permis de se faire une première idée de la dynamique des mobilités de personnes au sein du pôle de transport de La Défense au cours du temps. Il a notamment été mis en lumière qu'il existait des tendances globales ayant impacté l'ensemble du pôle au cours de l'historique, on pense aux périodes de restrictions contre la pandémie de Covid19, aux vacances ou encore aux grèves (voir figure C.6). A un niveau plus local, certaines zones du pôle présentent également des profils de mobilité spécifiques, liés à leur environnement proche. On pense par exemple aux périodes de concerts pour le poste P2 (figure C.3), ou à la différence entre samedi et dimanche pour la consommation au poste P12 (figure C.4). Les facteurs impactants à long terme, que sont les périodes de Covid19 ou de grève, sont par ailleurs à différencier des facteurs plus ponctuels comme les concerts. Nous avons ainsi constaté que les mobilités de personnes dans le pôle ont été amenées à changer en termes d'affluence totale (figure C.6), mais aussi en termes d'utilisation des différentes zones du pôle. Ce dernier point est visible dans les spécificités locales que nous avons constatées dans les figures C.1, C.3 et C.4, mais aussi dans le travail exploratoire des similarités entre zones que nous avons mené après. Un aspect important

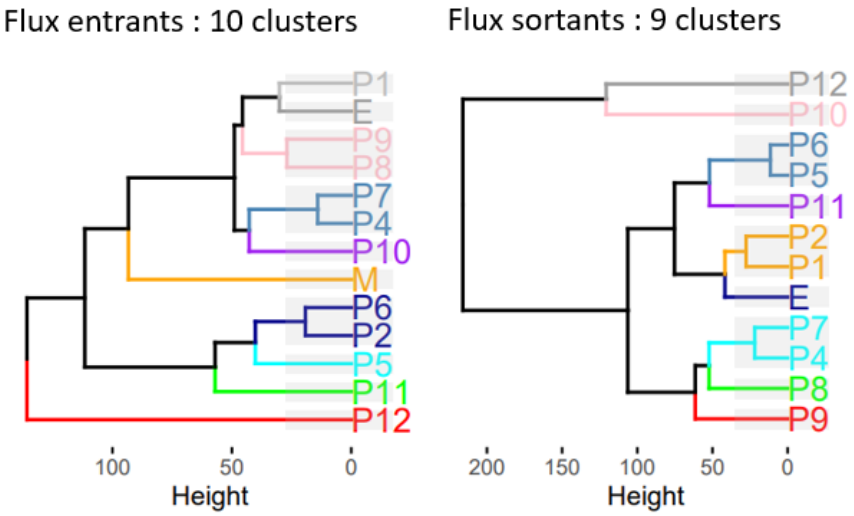


FIGURE C.9 – Arbre des similarités entre lieux, pour les jours de couvre-feu pendant la pandémie de Covid19

que nous avons constaté dans cette étude est que l'activité de travail régit en grande partie le rapprochement entre un grand nombre de zones du pôle (figure C.7), tandis que les activités de consommation prennent plus d'importance lors des jours de week-end (figure C.8).

Dans ce travail, on a détecté des profils de mobilité temporels, dans un cas univarié et sans contexte, ce qui pose un certain nombre de problèmes. Dans le cas d'une étude à échelle locale, seules les spécificités liées à des zones restreintes sont prises en compte ce qui rend difficile d'identifier des phénomènes n'agissant pas aux mêmes endroits. Par exemple une période de travaux, pendant laquelle il y aurait vraisemblablement autant de monde à venir travailler, mais par des moyens de transport différents, serait visible localement au niveau des accès vers les transports mais pas nécessairement au niveau des accès vers la dalle. Dans le cas d'une étude à échelle globale, où toute l'information est prise en compte, des phénomènes locaux pourraient être « noyés » dans une trop grande quantité d'informations et donc non détectés.

L'idée serait donc de formaliser le problème différemment de manière à pouvoir détecter des effets ayant des conséquences sur l'ensemble du pôle mais aussi à des échelles plus locales. Une vision multivariée pourrait ainsi répondre à ce problème. Afin d'appréhender cette dernière, il faut prendre en compte des phénomènes nouveaux comme la corrélation entre les séries temporelles. Dans le chapitre 4, nous proposons des modèles de mélange capables de prendre en compte l'ensemble des séries temporelles, surdispersées et correlées ainsi que l'impact de facteurs exogènes pour détecter des profils de mobilité temporels

facilement interprétables.

C.2 Méthode heuristique pour le calcul de similarités spatiales

La méthode proposée ici a pour objectif de quantifier le degré de similarité entre différentes zones du pôle de La Défense en terme de dynamiques de déplacement puis comment ces similarités évoluent selon différentes périodes considérées. Les périodes sont ici celles ayant été détectées par la catégorisation temporelle à une échelle globale. La méthodologie utilisée comprend deux étapes : d'abord, nous catégorisons l'ensemble des profils journaliers de mobilité du pôle de La Défense, puis nous utilisons ce résultat pour calculer un score de similarité entre différentes zones du pôle, en nous basant sur la ressemblance entre les catégories de profils journaliers détectées.

Nous explorons les similarités spatiales dans le cadre des flux entrants dans le pôle d'une part, puis sortants d'autre part. Pour cela, nous subdivisons les données en deux ensembles : $\mathbf{y}^{(I)}$ dont les comptages sont sélectionnés sur les flux entrants ($P^{(I)}$), et $\mathbf{y}^{(O)}$ dont les comptages sont des flux sortants uniquement. Dans le cadre des flux entrants, cette méthode exige dans un premier temps une catégorisation des $J|P^{(I)}|$ profils journaliers de tous les postes en $S^{(I)}$ catégories. Les observations $\mathbf{y}^{(I)}$ peuvent s'écrire sous la forme d'un vecteur $\mathbf{y}^{(I)} = \{\mathbf{y}_{j,p}\}_{j\in\{1,...,J\},p\in P^{(I)}}$ avec chaque $\mathbf{y}_{j,p}$ un vecteur de taille H de tous les comptages entrants relevés à chaque tranche horaire h au jour j et poste p. Le même processus est appliqué aux flux sortants $\mathbf{y}^{(O)}$ pour obtenir $S^{(O)}$ catégories. Un score de similarité est ensuite calculé entre chaque poste $p \in P^{(I)}$ ou $p \in P^{(O)}$, basé sur une comparaison pour chaque couple de clusters journaliers rencontrés pour chaque jour $j \in 1, ..., J$. La méthode de calcul du score de similarité est présentée dans le schéma de la figure C.10, et peut être décrite comme suit :

1. Calcul de courbes médianes pour chaque ensemble de vecteurs $\mathbf{y}_{j,p}$ affectés au même cluster :

$$\mathbf{y}_c^{(O)} = \left(\mathbf{y}_O^{c_1}, ..., \mathbf{y}_O^{c_s}, ..., \mathbf{y}_O^{c_{S^{(O)}}}\right)$$
$$\mathbf{y}_c^{(I)} = \left(\mathbf{y}_I^{c_1}, ..., \mathbf{y}_I^{c_s}, ..., \mathbf{y}_I^{c_{S^{(I)}}}\right)$$

   Les vecteurs $\mathbf{y}_c^{(O)}$ et $\mathbf{y}_c^{(I)}$ sont de tailles respectives $S^{(O)}$ et $S^{(I)}$. Leurs éléments $\mathbf{y}_O^{c_s}$ et $\mathbf{y}_I^{c_s}$ sont de taille H et sont calculés comme $\mathbf{y}_O^{c_s} = M(\{\mathbf{y}_{j,p}\}_{j\cap p\in C_s^{(O)}})$ et $\mathbf{y}_I^{c_s} = M(\{\mathbf{y}_{j,p}\}_{j\cap p\in C_s^{(I)}})$ avec $M(.)$ la médiane.

2. Une distance du chi2 est calculée entre chaque profil médian de $\mathbf{y}_c^{(O)}$ et de $\mathbf{y}_c^{(I)}$. Pour deux vecteurs médians issus de deux clusters c_1 et c_2, la distance se calcule comme :

$$D(y^{c_1}, y^{c_2}) = \sum_H \frac{1}{y_h^{c_1} + y_h^{c_2}} \left(\frac{y_h^{c_1}}{y_\cdot^{c_1}} - \frac{y_h^{c_2}}{y_\cdot^{c_2}}\right)^2,$$

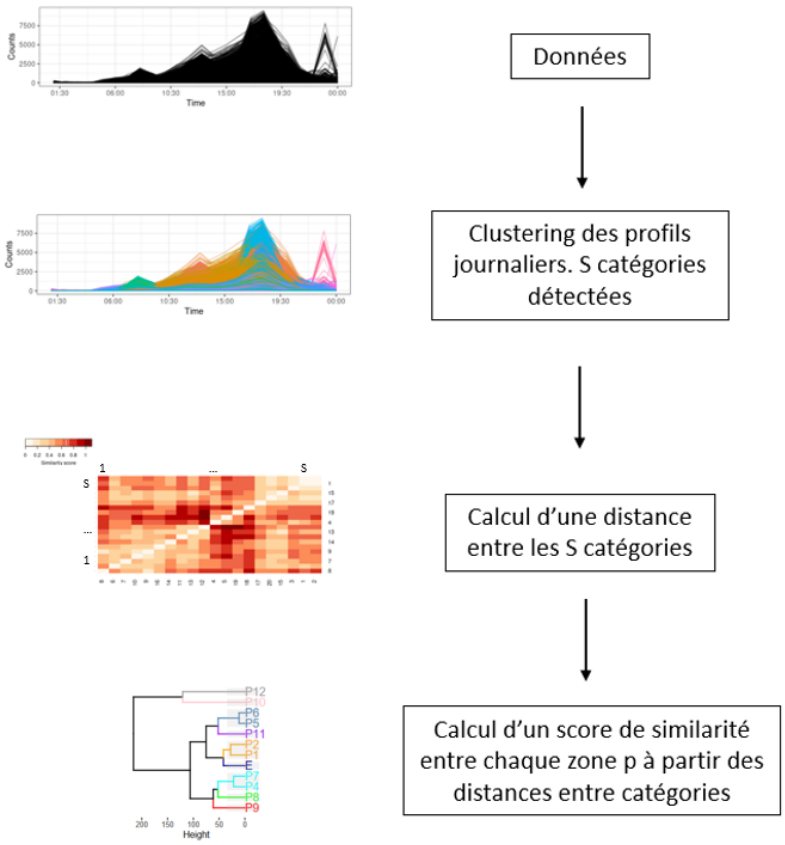


FIGURE C.10 – Méthodologie pour la recherche de similarités spatiales

avec $y_{.}^{c1} = \sum_H y_h^{c1}$ et $y_{.}^{c2} = \sum_H y_h^{c2}$.

3. Un score de dissimilarité est ensuite calculé pour chaque paire de zones $(p1, p2)$, en sommant les distances du chi2 calculées entre les médianes des clusters rencontrés à chaque jour $j \in 1, ..., J$.

Cette méthodologie peut être appliquée à des données issues de périodes homogènes. Par exemple, nous pourrions explorer les similarités spatiales lors des jours ouvrés seulement, ou lors des vacances. Les périodes homogènes sont celles ayant été détectées par la catégorisation temporelle appliquée à une échelle globale.

C.3 Estimation des modèles de mélanges de modèles « sommes et partages » avec l'algorithme d'espérance-maximisation (EM)

Étant donné $z_{j,s} = 1$ et $\mathbf{x}_{j,h}$, la série $\mathbf{y}_{j,h}$ est distribuée selon le modèle de mélange suivant :

$$p(\mathbf{y}_{j,h}; \boldsymbol{\alpha}, \boldsymbol{\gamma}, r, \boldsymbol{\xi}) = \sum_{s=1}^{S} \pi_s(j; \boldsymbol{\alpha}) g(v_{j,h}|\mathbf{x}_{j,h}, \boldsymbol{\gamma}_s, r_s) h(\mathbf{y}_{j,h}|v_{j,h}, \mathbf{x}_{j,h}, \boldsymbol{\xi}_s), \tag{C.5}$$

avec $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_s)_{s=1,\dots,S}$, $r = (r_s)_{s=1,\dots,S}$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_s)_{s=1,\dots,S}$. Les paramètres du modèle sont estimés à l'aide de l'algorithme de maximisation de l'espérance (EM) [DLR77], qui nécessite une maximisation de la log-vraisemblance complète des données. La log-vraisemblance complète des données peut être écrite :

$$\mathcal{CL}_Y(\boldsymbol{\alpha}, \boldsymbol{\gamma}, r, \boldsymbol{\xi}) = \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{h=1}^{H} z_{j,s} \log \left(\pi_s(j; \boldsymbol{\alpha}) g(v_{j,h}|\mathbf{x}_{j,h}, \boldsymbol{\gamma}_s, r_s) h(\mathbf{y}_{j,h}|v_{j,h}, \mathbf{x}_{j,h}, \boldsymbol{\xi}_s) \right). \tag{C.6}$$

Étant donné la valeur initiale des paramètres $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, $r^{(0)}$ et $\boldsymbol{\alpha}^{(0)}$, les deux étapes suivantes sont répétées jusqu'à convergence.

— *Etape d'espérance (E)*
L'espérance de la log-vraisemblance complétée est évaluée, connaissant les données observées Y et l'ensemble des paramètres actuels : $\boldsymbol{\xi}^{(c)}$, $\boldsymbol{\gamma}^{(c)}$, $r^{(c)}$ et $\boldsymbol{\alpha}^{(c)}$.

$$\begin{aligned} Q(\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\xi}^{(c)}) = & \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{h=1}^{H} E_{\boldsymbol{\xi}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\alpha}^{(c)}}[z_{j,s}|Y] \\ & log \left(\pi_s(j; \boldsymbol{\alpha}^{(c)}) g(v_{j,h}|\mathbf{x}_{j,h}, \boldsymbol{\gamma}_s^{(c)}, r_s^{(c)}) h(\mathbf{y}_{j,h}|v_{j,h}, \mathbf{x}_{j,h}, \boldsymbol{\xi}_s^{(c)}) \right), \end{aligned} \tag{C.7}$$

où

$$\begin{aligned} E_{\boldsymbol{\xi}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\alpha}^{(c)}}[z_{j,s}|Y] &= \tau_{j,s}^{(c)} \\ &= \frac{\pi_s(j; \boldsymbol{\alpha}^{(c)}) \prod_H g(v_{j,h}|\boldsymbol{x}_{j,h}, \boldsymbol{\gamma}_s^{(c)}, r_s^{(c)}) h(\mathbf{y}_{j,h}|\mathbf{x}_{j,h}, v_{j,h}, \boldsymbol{\xi}_s^{(c)})}{\sum_{s'} \pi_{s'}(j; \boldsymbol{\alpha}^{(c)}) \prod_H g(v_{j,h}|\mathbf{x}_{j,h}, \boldsymbol{\gamma}_{s'}^{(c)}, r_{s'}^{(c)}) h(\mathbf{y}_{j,h}|\mathbf{x}_{j,h}, v_{j,h}, \boldsymbol{\xi}_{s'}^{(c)})}. \end{aligned} \tag{C.8}$$

Les poids $\tau_{j,s}^{(c)}$ sont mis à jour à chaque itération de l'étape E.

— *Etape de maximisation (M)*
Les paramètres $\boldsymbol{\xi}^{(c+1)}$, $\boldsymbol{\gamma}^{(c+1)}$, $r^{(c+1)}$ et $\boldsymbol{\alpha}^{(c+1)}$ qui maximisent $Q(\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\xi}^{(c)})$ sont calculés. Il est possible de réécrire cette quantité comme suit :

$$Q(\boldsymbol{\alpha}^{(c)}, \boldsymbol{\gamma}^{(c)}, r^{(c)}, \boldsymbol{\xi}^{(c)}) = Q_1(\boldsymbol{\alpha}^{(c)}) + Q_2(\boldsymbol{\gamma}^{(c)}, r^{(c)}) + Q_3(\boldsymbol{\xi}^{(c)}) \tag{C.9}$$

où

$$Q_1(\boldsymbol{\alpha}^{(c)}) = \sum_{s=1}^{S} \sum_{j=1}^{J} \tau_{j,s}^{(c)} log(\pi_s(j; \boldsymbol{\alpha}^{(c)})) \tag{C.10}$$

$$Q_2(\boldsymbol{\gamma}^{(c)}, r^{(c)}) = \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{h=1}^{H} \tau_{j,s}^{(c)} log(g(v_{j,h}|\mathbf{x}_{j,h}, \boldsymbol{\gamma}_s^{(c)}, r_s^{(c)})) \tag{C.11}$$

$$Q_3(\boldsymbol{\xi}^{(c)}) = \sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{h=1}^{H} \tau_{j,s}^{(c)} log(h(\mathbf{y}_{j,h}|\mathbf{x}_{j,h}, v_{j,h}, \boldsymbol{\xi}_s^{(c)})). \tag{C.12}$$

La maximisation de $Q1$ consiste à résoudre une régression logistique multinomiale pondérée. De nouvelles valeurs de $\boldsymbol{\alpha}$ peuvent être trouvées en utilisant des procédures itératives, telles que les moindres carrés itérativement repondérés (IRLS) [HW77]. Ce problème est résolu avec la fonction multinom du paquet nnet [RVR16]. Q2 est la log-vraisemblance correspondant à un modèle linéaire généralisé binomial négatif. Sa maximisation est résolue par un processus d'itération alternée, fourni par la fonction glm.nb du paquet MASS [Rip+13]. Dans chaque segment s, pour une valeur donnée de $r_s^{(c)}$, le modèle linéaire est ajusté en utilisant une méthode IRLS. Ensuite, pour les paramètres $\boldsymbol{\gamma}_s^{(c)}$ trouvés fixes, le paramètre $r_s^{(c)}$ est estimé avec des itérations de score et d'information. Les deux étapes sont alternées jusqu'à convergence, et obtention de $\boldsymbol{\gamma}_s^{(c+1)}$ et $r_s^{(c+1)}$. Notons que $\tau_{j,s}^{(c)}$ sont ici utilisés comme poids *a priori* dans le processus d'ajustement. Le critère Q3, qui est associé à un modèle de régression Dirichlet-Multinomiale pondéré, est résolu avec le paquet MGLM [Kim+18]. La distribution Dirichlet-Multinomiale n'appartenant pas à la famille exponentielle, la méthode IRLS n'est pas utilisée, car la matrice d'information attendue est difficile à calculer. La méthode utilisée ici combine l'algorithme de minorisation-maximisation (MM) [LHY00], et la méthode de Newton. Les mises à jour MM et Newton sont calculées à chaque itération, et celle dont la log-vraisemblance est la plus élevée est choisie.

C.4 Estimation des modèles de mélange Poisson log-normal avec l'algorithme d'espérance-maximisation variationnel (VEM)

Les séries $\mathbf{y}_{j,h}$ sont distribuées selon le modèle de mélange suivant :

$$p(\mathbf{y}_{j,h};\boldsymbol{\alpha},\boldsymbol{\rho},\boldsymbol{\Sigma}) = \sum_{s=1}^{S} \pi_s(j;\boldsymbol{\alpha}) \int_{R^P} [\prod_{p=1}^{P} g(\mathbf{y}_{j,h,p}|\boldsymbol{\varphi}_{j,h,p})] m(\boldsymbol{\varphi}_{j,h}|\boldsymbol{x}_{j,h},\boldsymbol{\rho}_s,\boldsymbol{\Sigma}_s) d\boldsymbol{\varphi}_{j,h}, \tag{C.13}$$

avec $\boldsymbol{\rho} = (\boldsymbol{\rho}_s)_{s=1,\dots,S}$ et $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_s)_{s=1,\dots,S}$. g est une distribution de Poisson, et m une fonction de distribution gaussienne. L'algorithme EM peut être utilisé pour l'estimation des paramètres, mais pour trouver la valeur attendue de la log-vraisemblance complète des données, il faut estimer les espérances conditionnelles $\mathbb{E}(Z_{js}\boldsymbol{\varphi}_{j,h}|\mathbf{y}_{j,h},\boldsymbol{\rho}_s,\boldsymbol{\Sigma}_s)$ et $\mathbb{E}(Z_{js}\boldsymbol{\varphi}_{j,h}\boldsymbol{\varphi}'_{j,h}|\mathbf{y}_{j,h},\boldsymbol{\rho}_s,\boldsymbol{\Sigma}_s)$, qui ne sont pas calculées exactement. Ces espérances conditionnelles peuvent être calculées à l'aide d'un algorithme EM, couplé à un algorithme de chaîne de Markov Monte Carlo (MCMC-EM), comme présenté par [Sil+19], qui s'accompagne cependant d'une lourde charge de calcul. Nous nous référons plutôt au travail présenté par [CRM19], qui utilise l'approximation variationnelle, à savoir une technique d'inférence approximative. L'idée derrière l'inférence variationnelle est d'utiliser des densités gaussiennes, et d'approximer des distributions postérieures complexes, en minimisant la divergence de Kullback-Leibler entre les densités réelles et approximatives $q(\boldsymbol{\varphi})$. La log-vraisemblance marginale pour $\mathbf{y}_{j,h}$ peut s'écrire comme suit

$$\log p(\mathbf{y}_{j,h}) = F(q(\boldsymbol{\varphi}_{j,h}),\mathbf{y}_{j,h}) + D_{KL}(q(\boldsymbol{\varphi}_{j,h})|p(\boldsymbol{\varphi}_{j,h})), \tag{C.14}$$

avec $D_{KL}(q(\boldsymbol{\varphi}_{j,h})|p(\boldsymbol{\varphi}_{j,h}))$ la divergence de Kullback-Leibler entre $p(\boldsymbol{\varphi}_{j,h})$ et $q(\boldsymbol{\varphi}_{j,h})$. $F(q(\boldsymbol{\varphi}_{j,h}),\mathbf{y}_{j,h})$ est l'expression de la borne inférieure variationnelle de la log-vraisemblance. Il s'agit du critère que nous cherchons à maximiser dans le processus d'estimation des paramètres. Dans le cas du modèle de Poisson log-normal, on suppose que q est une distribution gaussienne :

$$q(\boldsymbol{\varphi}_{j,h};\mathbf{m}_{j,h},\mathbf{S}_{j,h}) = \mathcal{N}(\boldsymbol{\varphi}_{j,h};\mathbf{m}_{j,h},\mathbf{S}_{j,h}), \tag{C.15}$$

avec $\mathbf{m}_{j,h}$ et $\mathbf{S}_{j,h} = diag(\mathbf{S}_{j,h})$, les paramètres variationnels associés à l'échantillon $\mathbf{y}_{j,h}$ au jour j et à la tranche h. Pour minimiser la divergence de Kullback-Leibler, la limite inférieure variationnelle doit être maximisée. La log-vraisemblance complète

des données peut être écrite comme suit :

$$\mathcal{CL}_Y(\boldsymbol{\alpha}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, \boldsymbol{m}, \boldsymbol{S}) = \sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{h=1}^{H} z_{j,s} \log(\pi_s(j;\boldsymbol{\alpha}))+ \\ \sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{h=1}^{H} z_{j,s}[F(q^{(s)}(\boldsymbol{\varphi}_{j,h}), \mathbf{y}_{j,h}) + D_{KL}(q^{(s)}(\boldsymbol{\varphi}_{j,h})|p^{(s)}(\boldsymbol{\varphi}_{j,h}))], \tag{C.16}$$

où $D_{KL}(q^{(s)}(\boldsymbol{\varphi}_{j,h})|p^{(s)}(\boldsymbol{\varphi}_{j,h}))$ est la divergence de Kullback-Leibler entre $p(\boldsymbol{\varphi}_{j,h}|\mathbf{y}_{j,h}, z_j = s)$ et $q^{(s)}(\boldsymbol{\varphi}_{j,h})$, avec $q^{(s)}(\boldsymbol{\varphi}_{j,h}) = \mathcal{N}(\mathbf{m}_{j,h}^{(s)}, \mathbf{S}_{j,h}^{(s)})$. La borne inférieure variationnelle de la log-vraisemblance pour chaque observation $\mathbf{y}_{j,h}$ est

$$F(q^{(s)}(\boldsymbol{\varphi}_{j,h}), \mathbf{y}_{j,h}) = \frac{1}{2}\log|\mathbf{S}_{j,h}^{(s)}| - \frac{1}{2}(\mathbf{m}_{j,h}^{(s)} - \mathbf{x}_{j,h}^T\boldsymbol{\rho}_s)'\boldsymbol{\Sigma}_s^{-1}(\mathbf{m}_{j,h}^{(s)} - \mathbf{x}_{j,h}^T\boldsymbol{\rho}_s) - tr(\boldsymbol{\Sigma}_s^{-1}\mathbf{S}_{j,h}^{(s)})- \\ \frac{1}{2}\log|\boldsymbol{\Sigma}_s| - \frac{P}{2} + \mathbf{m}^{(s)\prime}_{j,h}\mathbf{y}_{j,h} - \sum_{p=1}^{P}(\exp(m_{j,h,p}^{(s)} + \frac{1}{2}s_{j,h,p}^{(s)}) + \log(y_{j,h,p}!)). \tag{C.17}$$

avec $tr(.)$ la trace. L'algorithme EM est utilisé pour estimer les paramètres, et les deux étapes suivantes sont répétées jusqu'à convergence.

— *Etape d'espérance (E)*
L'espérance de la log-vraisemblance complétée est évaluée, connaissant les données observées Y, l'ensemble des paramètres $\boldsymbol{\rho}^{(c)}$, $\boldsymbol{\Sigma}^{(c)}$ et $\boldsymbol{\alpha}^{(c)}$, et les paramètres variationnels $\mathbf{m}_{j,h}^{(c)}$ et $\mathbf{S}_{j,h}^{(c)}$.

$$Q(\boldsymbol{\rho}^{(c)}, \boldsymbol{\Sigma}^{(c)}, \boldsymbol{\alpha}^{(c)}, \mathbf{m}^{(c)}, \mathbf{S}^{(c)}) = \sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{h=1}^{H} \tau_{j,s}^{(c)} \log(\pi_s(j;\boldsymbol{\alpha}^{(c)}))+ \\ \sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{h=1}^{H} \tau_{j,s}^{(c)} E_{\boldsymbol{\rho}^{(c)},\boldsymbol{\Sigma}^{(c)},\boldsymbol{\alpha}^{(c)},m_{j,h}^{(c)},S_{j,h}^{(c)}}[F(q^{(s)}(\boldsymbol{\varphi}_{j,t}), \mathbf{y}_{j,h})+ \\ D_{KL}(q^{(s)}(\theta_{j,h})|p^{(s)}(\theta_{j,h}))], \tag{C.18}$$

avec $\tau_{j,s}^{(c)} = E_{\boldsymbol{\rho}^{(c)},\boldsymbol{\Sigma}^{(c)},\boldsymbol{\alpha}^{(c)},m_{j,h}^{(c)},S_{j,h}^{(c)}}[z_{j,s}|Y]$. La borne inférieure variationnelle de la log-vraisemblance est utilisée pour approximer $\tau_{j,s}^{(c)}$:

$$\tau_{j,s}^{(c)} = \frac{\pi_s(j;\boldsymbol{\alpha}^{(c)}) \prod_{h=1}^{H} \exp(F(q^{(s)}(\boldsymbol{\varphi}_{j,h}), \mathbf{y}_{j,h}))}{\sum_{s'=1}^{S} \pi_{s'}(j;\boldsymbol{\alpha}^{(c)}) \prod_{h=1}^{H} \exp(F(q^{(s')}(\theta_{j,h}), \mathbf{y}_{j,h}))}. \tag{C.19}$$

Cette approximation est utilisée dans le package R PLNmodels.

— *Etape de maximisation (M)*
L'étape de maximisation est divisée en deux parties :

— Conditionnellement à $\boldsymbol{\rho}_s$, $\boldsymbol{\Sigma}_s$ et $\tau_{j,s}$ donnés, les paramètres variationnels $\mathbf{m}_{j,h}^{(c)}$ et $\mathbf{S}_{j,h}^{(c)}$ sont mis à jour. $F(q^{(s)}(\boldsymbol{\varphi}_{j,h}), \mathbf{y}_{j,h})$ étant strictement concave avec $\mathbf{m}_{j,h}^{(c)}$ et $\mathbf{S}_{j,h}^{(c)}$, il est possible d'obtenir $\mathbf{S}_{j,h}^{(c+1)}$ avec la méthode des points-fixes, et $\mathbf{m}_{j,h}^{(c+1)}$ avec la méthode de Newton.

— Connaissant les paramètres $\tau_{j,s}^{(c)}$, $\mathbf{m}_{j,h}^{(c+1)}$ et $\mathbf{S}_{j,h}^{(c+1)}$; $\boldsymbol{\rho}^{(c+1)}$, $\boldsymbol{\Sigma}^{(c+1)}$ et $\boldsymbol{\alpha}^{(c+1)}$ sont obtenus.

Annexe D

Prédire les futures mobilités

D.1 Le *Long Short Term Memory* (LSTM)

Les *Long Short Term Memory* (LSTM) sont une sous-branche des réseaux de neurones récurrents (RNN), plus adaptée pour la conservation d'informations sur de longues périodes de temps [HS97]. En plus de la couche cachée $\mathbf{h}$ d'un RNN classique, le LSTM contient une cellule mémoire $\mathbf{c}$ incluant un système de portes (*gates*), efficace pour contrer le phénomène de disparition du gradient (*vanishing gradient*). Considérons x_t comme une entrée, et y_t une sortie au temps t ; les équations d'une unité de LSTM sont les suivantes :

$$
\begin{aligned}
F_t &= \sigma(b^F + x_t \mathbf{U}^F + \mathbf{h}_{t-1}\mathbf{W}^F) \\
I_t &= \sigma(b^I + x_t \mathbf{U}^I + \mathbf{h}_{t-1}\mathbf{W}^I) \\
O_t &= \sigma(b^O + x_t \mathbf{U}^O + \mathbf{h}_{t-1}\mathbf{W}^O) \\
c_t &= F_t ct - 1 + I_t tanh(b + x_t \mathbf{U} + \mathbf{h}_{t-1}\mathbf{W}) \\
\mathbf{h}_t &= tanh(c_t) O_t \\
y_t &\sim g(y_t; \Phi(b^y + \mathbf{h}_t W^y))
\end{aligned}
$$

Ici, F est la porte d'oubli (*forget gate*), utile à la mise à jour de la cellule mémoire c. I est la porte d'entrée des données d'entrée x et $\mathbf{h}_{t-1}$ (état caché du pas de temps précédent). O est l'équivalent de la porte d'entrée, mais pour les sorties vers l'état caché suivant. Les données en entrée x (associées à des poids $\mathbf{U}$) et l'état caché du pas de temps précédent (associé à des poids $\mathbf{W}$) intègrent le calcul de la cellule mémoire du LSTM. Enfin, g est une fonction d'activation, permettant de passer de la couche cachée courante $\mathbf{h}_t$, à la sortie $\hat{y}$.

Notons qu'il existe une autre sous-branche des RNN, appelée *Gated Recurrent Unit* (GRU), introduite par [Cho+14], que nous ne traitons pas dans cette thèse. Une unité GRU n'a pas de portes d'oubli (F) et d'entrée (I) séparées comme le LSTM, ce qui rend

les LSTMs à la fois plus détaillés et complexes. Le choix entre les deux modèles n'est pas simple, car l'un est plus sophistiqué, au prix d'un plus grand temps de calcul. Nous nous référons au modèle LSTM dans cette thèse, car ne disposant pas de temps de calcul trop conséquents, et cherchant à estimer des détails complexes dans les séries temporelles (ex. impact d'une perturbation).

Bibliographie

[AH89] John Aitchison et CH Ho. « The multivariate Poisson-log normal distribution ». In : *Biometrika* 76.4 (1989), p. 643-653.

[Aka74] Hirotugu Akaike. « A new look at the statistical model identification ». In : *IEEE transactions on automatic control* 19.6 (1974), p. 716-723.

[AER20] Mohammed H Almannaa, Mohammed Elhenawy et Hesham A Rakha. « Dynamic linear models to predict bike availability in a bike sharing system ». In : *International journal of sustainable transportation* 14.3 (2020), p. 232-242.

[Bap+21] Thomas Bapaume et al. « Image inpainting and deep learning to forecast short-term train loads ». In : *IEEE Access* 9 (2021), p. 98506-98522.

[BLR06] Luc Bauwens, Sébastien Laurent et Jeroen VK Rombouts. « Multivariate GARCH models : a survey ». In : *Journal of applied econometrics* 21.1 (2006), p. 79-109.

[Bia+19] Zheyong Bian et al. « Unobserved component model for predicting monthly traffic volume ». In : *Journal of Transportation Engineering, Part A : Systems* 145.12 (2019), p. 04019052.

[BCG00] Christophe Biernacki, Gilles Celeux et Gérard Govaert. « Assessing a mixture model for clustering with the integrated completed likelihood ». In : *IEEE transactions on pattern analysis and machine intelligence* 22.7 (2000), p. 719-725.

[BMM21] Mathias Blicher Bjerregård, Jan Kloppenborg Møller et Henrik Madsen. « An introduction to multivariate probabilistic forecast evaluation ». In : *Energy and AI* 4 (2021), p. 100058.

[BM83] Alfred Blumstein et Harold D Miller. « Making do : The effects of a mass transit strike on travel behavior ». In : *Transportation* 11.4 (1983), p. 361-382.

[Bou+19] Charles Bouveyron et al. *Model-based clustering and classification for data science : with applications in R*. T. 50. Cambridge University Press, 2019.

[Bri+17] Anne-Sarah Briand et al. « Analyzing year-to-year changes in public transport passenger behaviour using smart card data ». In : *Transportation Research Part C : Emerging Technologies* 79 (2017), p. 274-289.

[Byr+95] Richard H Byrd et al. « A limited memory algorithm for bound constrained optimization ». In : *SIAM Journal on scientific computing* 16.5 (1995), p. 1190-1208.

[CS96] Gilles Celeux et Gilda Soromenho. « An entropy criterion for assessing the number of clusters in a mixture model ». In : *Journal of classification* 13.2 (1996), p. 195-212.

[Che+19] Jason Li Chen et al. « Forecasting seasonal tourism demand using a multiseries structural time series method ». In : *Journal of Travel Research* 58.1 (2019), p. 92-103.

[Che+20] Li Chen et al. « Short-term traffic flow prediction : From the perspective of traffic flow decomposition ». In : *Neurocomputing* 413 (2020), p. 444-456.

[Che+22] Xiaoxu Chen et al. « Probabilistic forecasting of bus travel time with a Bayesian Gaussian mixture model ». In : *arXiv preprint arXiv :2206.06915* (2022).

[CMR21] Julien Chiquet, Mahendra Mariadassou et Stéphane Robin. « The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances ». In : *Frontiers in Ecology and Evolution* 9 (2021), p. 188.

[CRM19] Julien Chiquet, Stephane Robin et Mahendra Mariadassou. « Variational inference for sparse network reconstruction from count data ». In : *International Conference on Machine Learning*. PMLR. 2019, p. 1162-1171.

[Cho+14] Kyunghyun Cho et al. « On the properties of neural machine translation : Encoder-decoder approaches ». In : *arXiv preprint arXiv :1409.1259* (2014).

[Côm+21] Etienne Côme et al. « Hierarchical clustering with discrete latent variable models and the integrated classification likelihood ». In : *Advances in Data Analysis and Classification* 15.4 (2021), p. 957-986.

[08] « Correlated Count Data ». In : *Econometric Analysis of Count Data*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 203-239.

[Cui+13] Jianqiang Cui et al. « Underground pedestrian systems development in cities : Influencing factors and implications ». In : *Tunnelling and underground space technology* 35 (2013), p. 152-160.

[Cur10] Graham Currie. « Quick and effective solution to rail overcrowding : free early bird ticket experience in Melbourne, Australia ». In : *Transportation research record* 2146.1 (2010), p. 35-42.

[DHS11] Alysha M De Livera, Rob J Hyndman et Ralph D Snyder. « Forecasting time series with complex seasonal patterns using exponential smoothing ». In : *Journal of the American statistical association* 106.496 (2011), p. 1513-1527.

[DLR77] Arthur P Dempster, Nan M Laird et Donald B Rubin. « Maximum likelihood from incomplete data via the EM algorithm ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 39.1 (1977), p. 1-22.

[Din+16] Chuan Ding et al. « Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees ». In : *Sustainability* 8.11 (2016), p. 1100.

[Doo+14] Ronan Doorley et al. « Short-term forecasting of bicycle traffic using structural time series models ». In : *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2014, p. 1764-1769.

[Dor+08] V Dordonnat et al. « An hourly periodic state space model for modelling French national electricity load ». In : *International Journal of Forecasting* 24.4 (2008), p. 566-587.

[DA11] Martine Drozdz et Manuel Appert. « Re-understanding CBD : a landscape perspective ». In : (2011).

[EB21] Oscar Egu et Patrick Bonnel. « Medium-term public transit route ridership forecasting : What, how and why ? A case study in Lyon ». In : *Transport Policy* 105 (2021), p. 124-133.

[EL14] Côme Etienne et 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. 1-21.

[Gha+17] Mohammad Sajjad Ghaemi et al. « A visual segmentation method for temporal smart card data ». In : *Transportmetrica A : Transport Science* 13.5 (2017), p. 381-404.

[GBO09] Bidisha Ghosh, Biswajit Basu et Margaret O'Mahony. « Multivariate short-term traffic flow forecasting using time-series analysis ». In : *IEEE transactions on intelligent transportation systems* 10.2 (2009), p. 246-254.

[GC21] Konstantinos Gkiotsalitis et Oded Cats. « Public transport planning adaption under the COVID-19 pandemic crisis : literature review of research needs and directions ». In : *Transport Reviews* 41.3 (2021), p. 374-392.

[GTS02] Jürgen Grieser, Silke Trömel et C-D Schönwiese. « Statistical time series decomposition into significant components and application to European temperature ». In : *Theoretical and applied climatology* 71.3 (2002), p. 171-183.

[GV22] Vidya GS et Hari VS. « Prediction of Bus Passenger Traffic using Gaussian Process Regression ». In : *Journal of Signal Processing Systems* (2022), p. 1-12.

[Gui22] Faure Guillemette. « Tours à moitié vides, parvis déserté, restaurants sans clients... La crise existentielle du quartier de La Défense ». In : *Le Monde* (11 juin 2022). URL : https://www.lemonde.fr/m-perso/article/2022/06/11/qui-veut-encore-travailler-a-la-defense_6129788_4497916.html.

[Har90] Andrew C Harvey. « Forecasting, structural time series models and the Kalman filter ». In : (1990).

[Hil11] Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

[HS97] Sepp Hochreiter et Jürgen Schmidhuber. « Long short-term memory ». In : *Neural computation* 9.8 (1997), p. 1735-1780.

[HW77] Paul W Holland et Roy E Welsch. « Robust regression using iteratively reweighted least-squares ». In : *Communications in Statistics-theory and Methods* 6.9 (1977), p. 813-827.

[Hol04] Charles C Holt. « Forecasting seasonals and trends by exponentially weighted moving averages ». In : *International journal of forecasting* 20.1 (2004), p. 5-10.

[HSA18] Keita Honjo, Hiroto Shiraki et Shuichi Ashina. « Dynamic linear modeling of monthly electricity demand in Japan : Time variation of electricity conservation effect ». In : *PloS one* 13.4 (2018), e0196331.

[Hyn+02] Rob J Hyndman et al. « A state space framework for automatic forecasting using exponential smoothing methods ». In : *International Journal of forecasting* 18.3 (2002), p. 439-454.

[JL22] Weiwei Jiang et Jiayun Luo. « Graph neural network for traffic forecasting : A survey ». In : *Expert Systems with Applications* (2022), p. 117921.

[JM19] MC Jones et Éric Marchand. « Multivariate discrete distributions via sums and shares ». In : *Journal of Multivariate Analysis* 171 (2019), p. 83-93.

[Kal+60] Rudolf Emil Kalman et al. « Contributions to the theory of optimal control ». In : *Bol. soc. mat. mexicana* 5.2 (1960), p. 102-119.

[Ke+17] Jintao Ke et al. « 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), p. 591-608.

[Kim+18] Juhyun Kim et al. « MGLM : an R package for multivariate categorical data analysis ». In : *The R journal* 10.1 (2018), p. 73.

[KB14] Diederik P Kingma et Jimmy Ba. « Adam : A method for stochastic optimization ». In : *arXiv preprint arXiv :1412.6980* (2014).

[KO11] S Koopman et Marius Ooms. *Forecasting economic time series using unobserved components time series models*. 2011.

[Kos] Simeon Kostadinov. *Understanding Encoder-Decoder Sequence to Sequence Model*. url : https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346.

[Kri+16] Miklas S Kristoffersen et al. « Pedestrian counting with occlusion handling using stereo thermal cameras ». In : *Sensors* 16.1 (2016), p. 62.

[Kro+14] Eric Kroes et al. « Value of crowding on public transport in ıle-de-France, France ». In : *Transportation Research Record* 2417.1 (2014), p. 37-45.

[LWL10] Lawrence W Lan, Chieh-Hua Wen et Hsiang-Yi Lee. « Effects of Temporally Differential Fares on Taipei Metro Riders' Mode and Time-of-Day Choiches ». In : *Effects of Temporally Differential Fares on Taipei Metro Riders' Mode and Time-of-Day Choiches* (2010), p. 1000-1022.

[LHY00] Kenneth Lange, David R Hunter et Ilsoon Yang. « Optimization transfer using surrogate objective functions ». In : *Journal of computational and graphical statistics* 9.1 (2000), p. 1-20.

[LG07] Danial Lashkari et Polina Golland. « Convex clustering with exemplar-based models ». In : *Advances in neural information processing systems* 20 (2007).

[Lat+13] Neal Lathia et al. « Individuals among commuters : Building personalised transport information services from fare collection systems ». In : *Pervasive and Mobile Computing* 9.5 (2013), p. 643-664.

[LP+03] Matthieu Lemoine, Florian Pelgrin et al. « Introduction aux modèles espace-état et au filtre de Kalman ». In : *Revue de l'OFCE* 86.3 (2003), p. 203-229.

[LKG22] Danya Li, Semin Kwak et Nikolas Geroliminis. « TwoResNet : Two-level resolution neural network for traffic forecasting on freeway networks ». In : *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2022, p. 3963-3969.

[LT03] Li-mei Li et Wei Tao. « Spatial structure evolution of system of recreation business district ». In : *Chinese geographical science* 13.4 (2003), p. 370-377.

[Li+21] Yujia Li et al. « A sparse negative binomial mixture model for clustering RNA-seq count data ». In : *Biostatistics* (2021).

[Lüt13] Helmut Lütkepohl. « Vector autoregressive models ». In : *Handbook of Research Methods and Applications in Empirical Macroeconomics*. Edward Elgar Publishing, 2013, p. 139-164.

[MZB18] Ed Manley, Chen Zhong et Michael Batty. « Spatiotemporal variation in travel regularity through transit user profiling ». In : *Transportation* 45.3 (2018), p. 703-732.

[MP16] Martyna Marczak et Tommaso Proietti. « Outlier detection in structural time series models : The indicator saturation approach ». In : *International Journal of Forecasting* 32.1 (2016), p. 180-202.

[MW76] James E Matheson et Robert L Winkler. « Scoring rules for continuous probability distributions ». In : *Management science* 22.10 (1976), p. 1087-1096.

[MLR19] Geoffrey J McLachlan, Sharon X Lee et Suren I Rathnayake. « Finite mixture models ». In : *Annual review of statistics and its application* 6 (2019), p. 355-378.

[Moh+16] K Mohamed et al. « Clustering smart card data for urban mobility analysis ». In : *IEEE Transactions on intelligent transportation systems* 18.3 (2016), p. 712-728.

[MG19] Mir Hossein Mousavi et Saleh Ghavidel. « Structural time series model for energy demand in Iran's transportation sector ». In : *Case Studies on Transport Policy* 7.2 (2019), p. 423-432.

[MK21] KV Murthy et G Kishore Kumar. « Structural time-series modelling for seasonal surface air temperature patterns in India 1951–2016 ». In : *Meteorology and Atmospheric Physics* 133.1 (2021), p. 27-39.

[Nai+21] Paul de Nailly et al. « What can we learn from 9 years of ticketing data at a major transport hub ? A structural time series decomposition ». In : *Transportmetrica A : Transport Science* (2021), p. 1-25.

[NSR19] KV Narasimha Murthy, R Saravana et P Rajendra. « Unobserved component modeling for seasonal rainfall patterns in Rayalaseema region, India 1951–2015 ». In : *Meteorology and Atmospheric Physics* 131.5 (2019), p. 1387-1399.

[Ni+20] Lingling Ni et al. « Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model ». In : *Journal of Hydrology* 586 (2020), p. 124901.

[PPM17] George Papamakarios, Theo Pavlakou et Iain Murray. « Masked autoregressive flow for density estimation ». In : *Advances in neural information processing systems* 30 (2017).

[Pas+19] Kevin Pasini et al. « LSTM encoder-predictor for short-term train load forecasting ». In : *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, p. 535-551.

[PSY20] Dmitry Pavlyuk, Nadežda Spiridovska et Irina Yatskiv. « Spatiotemporal dynamics of public transport demand : a case study of Riga ». In : *Transport* 35.6 (2020), p. 576-587.

[Pel+15] GA Peláez et al. « Road detection with thermal cameras through 3D information ». In : *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2015, p. 255-260.

[Pen+20] Fang-Le Peng et al. « Planning and implementation of underground space in Chinese central business district (CBD) : A case of Shanghai Hongqiao CBD ». In : *Tunnelling and Underground Space Technology* 95 (2020), p. 103176.

[Pet10] Giovanni Petris. « An R package for dynamic linear models ». In : *Journal of statistical software* 36 (2010), p. 1-16.

[PPC09] Giovanni Petris, Sonia Petrone et Patrizia Campagnoli. « Dynamic linear models ». In : *Dynamic Linear Models with R*. Springer, 2009, p. 31-84.

[PFD21] Jean Peyhardi, Pierre Fernique et Jean-Baptiste Durand. « Splitting models for multivariate count data ». In : *Journal of Multivariate Analysis* 181 (2021), p. 104677.

[Pra85] J Praagman. *Classification and regression trees : Leo BREIMAN, Jerome H. FRIEDMAN, Richard A. OLSHEN and Charles J. STONE The Wadsworth Statistics/Probability Series, Wadsworth, Belmont, 1984, x+ 358 pages*. 1985.

[QPW17] Yong-Kang Qiao, Fang-Le Peng et Yang Wang. « Monetary valuation of urban underground space : A critical issue for the decision-making of urban underground space development ». In : *Land Use Policy* 69 (2017), p. 12-24.

[Ras03] Carl Edward Rasmussen. « Gaussian processes in machine learning ». In : *Summer school on machine learning*. Springer. 2003, p. 63-71.

[Ras+20] Kashif Rasul et al. « Multivariate probabilistic time series forecasting via conditioned normalizing flows ». In : *arXiv preprint arXiv :2002.06103* (2020).

[RVR16] Brian Ripley, William Venables et Maintainer Brian Ripley. « Package 'nnet' ». In : *R package version* 7.3-12 (2016), p. 700.

[Rip+13] Brian Ripley et al. « Package 'mass' ». In : *Cran r* 538 (2013), p. 113-120.

[RPO20] Yesid Rodriguez, Wilmer Pineda et Oscar Diaz Olariaga. « Air traffic forecast in post-liberalization context : a Dynamic Linear Models approach ». In : *Aviation* 24.1 (2020), p. 10-19.

[Sal+19] David Salinas et al. « High-dimensional multivariate forecasting with low-rank gaussian copula processes ». In : *Advances in neural information processing systems* 32 (2019).

[Sal+20] David Salinas et al. « DeepAR : Probabilistic forecasting with autoregressive recurrent networks ». In : *International Journal of Forecasting* 36.3 (2020), p. 1181-1191.

[SAO21] A Samé, ML Abadi et L Oukhellou. « Change Detection in Smart Grids Using Dynamic Mixtures of t-Distributions ». In : *Advances in Condition Monitoring and Structural Health Monitoring*. Springer, 2021, p. 53-68.

[Sch78] Gideon Schwarz. « Estimating the dimension of a model ». In : *The annals of statistics* (1978), p. 461-464.

[SS82] Robert H Shumway et David S Stoffer. « An approach to time series smoothing and forecasting using the EM algorithm ». In : *Journal of time series analysis* 3.4 (1982), p. 253-264.

[SSS00] Robert H Shumway, David S Stoffer et David S Stoffer. *Time series analysis and its applications*. T. 3. Springer, 2000.

[SYS64] Masaaki Sibuya, Isao Yoshimura et Ryoichi Shimizu. « Negative multinomial distribution ». In : *Annals of the Institute of Statistical Mathematics* 16.1 (1964), p. 409-426.

[Sil+19] Anjali Silva et al. « A multivariate Poisson-log normal mixture model for clustering transcriptome sequencing data ». In : *BMC bioinformatics* 20.1 (2019), p. 1-11.

[Sri+] BABU Sridhar et al. « Forecasting of Pedestrian Counts at City Location Points ». In : ().

[SVL14] Ilya Sutskever, Oriol Vinyals et Quoc V Le. « Sequence to sequence learning with neural networks ». In : *Advances in neural information processing systems* 27 (2014).

[Tho+21] Francene MF Thomas et al. « Commuting before and after COVID-19 ». In : *Transportation Research Interdisciplinary Perspectives* 11 (2021), p. 100423.

[Toq+18] Florian Toqué et al. « 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*. 2018, 15p.

[TOV20] Charles Truong, Laurent Oudre et Nicolas Vayatis. « Selective review of offline change point detection methods ». In : *Signal Processing* 167 (2020), p. 107299.

[Uni19] United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects : The 2018 Revision*. Rapp. tech. 2019.

[Vas+17] Ashish Vaswani et al. « Attention is all you need ». In : *Advances in neural information processing systems* 30 (2017).

[Wan+21] Zijia Wang et al. « Identifying Urban Functional Areas and Their Dynamic Changes in Beijing : Using Multiyear Transit Smart Card Data ». In : *Journal of Urban Planning and Development* 147.2 (2021), p. 04021002.

[WR09] Glen Weisbrod et Arlee Reno. *Economic impact of public transportation investment*. Citeseer, 2009.

[WT16] Yuankai Wu et Huachun Tan. « Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework ». In : *arXiv preprint arXiv :1612.01022* (2016).

[XSC15] Rui Xue, Daniel Jian Sun et Shukai Chen. « Short-term bus passenger demand prediction based on time series model and interactive multiple model approach ». In : *Discrete Dynamics in Nature and Society* 2015 (2015).

[YT18] Hai Yang et Yili Tang. « Managing rail transit peak-hour congestion with a fare-reward scheme ». In : *Transportation Research Part B : Methodological* 110 (2018), p. 122-136.

[YRD16] Hsiang-Fu Yu, Nikhil Rao et Inderjit S Dhillon. « Temporal regularized matrix factorization for high-dimensional time series prediction ». In : *Advances in neural information processing systems* 29 (2016).

[Zha+17a] Yiwen Zhang et al. « Regression models for multivariate count data ». In : *Journal of Computational and Graphical Statistics* 26.1 (2017), p. 1-13.

[Zha+20] Yangyang Zhao et al. « Short-term passenger flow prediction with decomposition in urban railway systems ». In : *IEEE Access* 8 (2020), p. 107876-107886.

[Zha+17b] Zheng Zhao et al. « LSTM network : a deep learning approach for short-term traffic forecast ». In : *IET Intelligent Transport Systems* 11.2 (2017), p. 68-75.

[Zho+15] Chen Zhong et al. « Measuring variability of mobility patterns from multiday smart-card data ». In : *Journal of Computational Science* 9 (2015), p. 125-130.

[Zho+12] Mingyuan Zhou et al. « Beta-negative binomial process and Poisson factor analysis ». In : *Artificial Intelligence and Statistics*. PMLR. 2012, p. 1462-1471.

[Zhu+15] He Zhu et al. « A spatial-temporal analysis of urban recreational business districts : A case study in Beijing, China ». In : *Journal of Geographical Sciences* 25.12 (2015), p. 1521-1536.

[ZG17] Xi Zhu et Diansheng Guo. « Urban event detection with big data of taxi OD trips : A time series decomposition approach ». In : *Transactions in GIS* 21.3 (2017), p. 560-574.