HEC MONTRÉAL

Étude comparative de modèles d'imputation de données manquantes basés sur les méthodes de factorisation matricielle dans un contexte spatiotemporel

par

Lamine Benhabiles

Aurélie Labbe
HEC Montréal
Directrice de recherche

Sciences de la gestion
(Spécialisation en intelligence d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

Décembre 2023

Résumé

Le présent mémoire vise à mener une étude comparative de différents modèles d'imputation de données manquantes basés sur les méthodes de factorisation matricielle dans un contexte spatiotemporel. L'étude explore divers modèles d'imputation, notamment l'imputation par la moyenne (MEAN), les modèles ARIMA (ARIMA), le Bayesian Probabilistic Matrix Factotization (BPMF), le Temporal Regularized Matrix Factorization (TRMF), le Bayesian Temporal Matrix Factorization (BTMF), le Bayesian Temporal Tensor Factorization (BTTF), le Kernelized Probabilistic Matrix Factorization (KPMF) ainsi que le Bayesian Gaussian Process Probabilistic Matrix Factorization (BGMF). Une attention particulière a été accordée aux modèles intégrant une modélisation fine de la matrice de variance-covariance des données dans le but de décrire adéquatement la structure spatiotemporelle inhérente aux données du transport routier. Pour ce faire, nous avons évalué ces modèles en termes de qualité de l'imputation et de vitesse de calcul. Une série de simulations a été menée pour quantifier la performance des différents modèles, en tenant en compte de l'effet du rang de la décomposition et de la proportion de données manquantes. De plus, l'étude examine l'impact de la structure des données manquantes sur l'efficacité des modèles en matière d'imputation.

Une introduction à l'analyse bayésienne est fournie, couvrant le théorème de Bayes et l'inférence bayésienne. Un exemple concret, le jeu de hasard pile ou face, illustrera ces concepts. Cet exemple sera suivi d'une initiation à la méthode de Monte-Carlo par chaines de Markov et à l'échantillonnage de Gibbs.

L'étude se base sur des données de trafic collectées à Seattle, capturant la vitesse des véhicules en mph (*miles per hour* [miles par heure]) sur plusieurs autoroutes majeures. Elle met en évidence la complexité spatiotemporelle dans les comportements du trafic. La présentation des données ainsi que leur analyse exploratoire sont détaillées : des diagrammes spatiaux et temporels sont utilisés pour visualiser les données et sont accompagnés d'une analyse approfondie de la dépendance spatiotemporelle et des composants des séries chronologiques.

Le plan de simulation et les résultats obtenus sont ensuite exposés. La conception des scénarios et la définition des mesures de performance sont abordées, suivies des résultats des simulations. Les diagnostiques graphiques et numériques sont analysés, incluant l'effet du rang de décomposition et des données manquantes sur les erreurs d'imputation, ainsi que les temps d'exécution des modèles étudiés. Enfin, le mémoire conclut par une discussion sur les résultats obtenus et leurs implications.

Dans ce contexte particulier, le choix entre les modèles étudiés dépend principalement du compromis entre la rapidité d'exécution, la précision de l'imputation recherchée, la tolérance aux valeurs extrêmes, le degré d'absence de données spécifique et la structure des données manquantes spécifiques. Les résultats montrent que l'imputation par la moyenne (MEAN) n'est pas fiable. Les modèles ARIMA (ARIMA) et le Bayesian Probabilistic Matrix Factotization (BPMF) semblent présenter des biais systématiques. Le Temporal Regularized Matrix Factorization (TRMF) affiche une performance relativement satisfaisante. Les modèles Bayesian Temporal Matrix Factorization (BTMF) et Kernelized Probabilistic Matrix Factorization (KPMF) offrent un bon équilibre entre précision et fiabilité dans un contexte spatiotemporel. Le modèle Bayesian Temporal Tensor Factorization (BTTF) n'a pas été évalué, puisque nos données sont de nature matricielle et que le BTTF requiert des données sous une forme tensorielle. Quant au Bayesian Gaussian Process Probabilistic Matrix Factorization (BGMF), son évaluation a été omise en raison de l'indisponibilité du code source de la part de l'auteur. Par ailleurs, le choix entre le Bayesian Temporal Matrix Factorization (BTMF) et le Kernelized Probabilistic Matrix Fac-

torization (KPMF) dépend d'une analyse minutieuse des priorités spécifiques et du pourcentage de données manquantes. On constate que le Kernelized Probabilistic Matrix Factorization (KPMF) excelle dans la réduction des erreurs extrêmes, ce qui se manifeste par une Root Mean Square Error (RMSE) globalement plus basse, pour les ensembles de données avec 30% et 60% de valeurs manquantes. Inversement, le modèle Bayesian Temporal Matrix Factorization (BTMF) présente une Root Mean Square Error (RMSE) légèrement plus élevée, une conséquence de sa plus grande sensibilité aux valeurs extrêmes. Cependant, dans des contextes où le taux de données manquantes atteint 90%, la performance du Kernelized Probabilistic Matrix Factorization (KPMF) se détériore significativement. Dans ces conditions, le modèle BTMF (Bayesian Temporal Matrix Factorization) affiche une RMSE considérablement plus basse que celle du KPMF. Par conséquent, le BTMF se révèle être le modèle le plus adapté et efficace pour l'imputation des valeurs manquantes dans des situations où le taux de données manquantes est extrêmement élevé. Néanmoins, il convient de noter que, malgré sa performance supérieure en termes de précision, le modèle Bayesian Temporal Matrix Factorization (BTMF) peut parfois rencontrer des problèmes de convergence et présenter une sensibilité accrue aux valeurs extrêmes. Ainsi, le choix entre le Kernelized Probabilistic Matrix Factorization (KPMF) et le Bayesian Temporal Matrix Factorization (BTMF) devrait être basé sur des critères spécifiques tels que la précision requise dans l'imputation des données, la tolérance aux erreurs de grande envergure, et le pourcentage de données manquantes présentes dans le jeu de données concerné.

Mots-clés

Analyse bayésienne, Méthodes de factorisation matricielle probabiliste, Méthode de Monte-Carlo par chaines de Markov (MCMC), Échantillonnage de Gibbs, Imputation de données manquantes, Modèles ARIMA, Modèles de séries temporelles, Modèles bayésiens.

Abstract

This thesis aims to conduct a comparative study of different missing data imputation models based on matrix factorization methods in a spatiotemporal context. This study explores various imputation models, including mean imputations (MEAN), ARIMA models (ARIMA), Bayesian Probabilistic Matrix Factorization (BPMF), Temporal Regularized Matrix Factorization (TRMF), Bayesian Temporal Matrix Factorization (BTMF), Bayesian Temporal Tensor Factorization (BTTF), Kernelized Probabilistic Matrix Factorization (KPMF), and Bayesian Gaussian Process Probabilistic Matrix Factorization (BGMF). Particular attention is given to models that incorporate a refined modeling of the variance-covariance matrix to adequately describe the spatiotemporal structure inherent in road transport data. To this end, we have evaluated these models based on the quality of imputation and computational speed. A series of simulations was conducted to quantify the performance of different models, taking into account the effect of the rank of the decomposition and the proportion of missing data. Furthermore, the study examines the impact of the structure of missing data on the models' imputation efficiency.

An introduction to Bayesian analysis is provided, covering Bayes' theorem and Bayesian inference. A concrete example, the coin toss game, will illustrate these concepts, followed by an introduction to Markov Chain Monte Carlo (MCMC) methods and Gibbs sampling.

This study is based on traffic data collected in Seattle, capturing vehicle speed in mph (miles per hour) on several major highways. It highlights the spatiotemporal

complexity in traffic behaviors. The presentation of the data and their exploratory analysis are detailed. Spatial and temporal diagrams are used to visualize the data and are accompanied by a thorough analysis of spatiotemporal dependence and time-series components.

The simulation plan and the results obtained are then presented. The design of the scenarios and the definition of performance measures are discussed, followed by the simulation results. Graphical and numerical diagnostics are analyzed, including the effect of decomposition rank and missing data on imputation errors, as well as the execution times of the studied models. Finally, the thesis concludes with a discussion on the results obtained and their implications.

In this specific context, the choice between the studied models primarily depends on the trade-off between execution speed, the desired accuracy of imputation, tolerance to extreme values, the specific degree of data absence, and the structure of the specific missing data. The results show that imputation by the mean (MEAN) is not reliable. The ARIMA (ARIMA) models and the Bayesian Probabilistic Matrix Factorization (BPMF) appear to exhibit systematic biases. The Temporal Regularized Matrix Factorization (TRMF) displays a relatively satisfactory performance. The Bayesian Temporal Matrix Factorization (BTMF) and Kernelized Probabilistic Matrix Factorization (KPMF) models offer a good balance between accuracy and reliability in a spatiotemporal context. The Bayesian Temporal Tensor Factorization (BTTF) model was not evaluated, as our data are of a matrix nature and BTTF requires data in a tensor form. As for the Bayesian Gaussian Process Probabilistic Matrix Factorization (BGMF), its evaluation was omitted due to the unavailability of the source code from the author. Furthermore, the choice between Bayesian Temporal Matrix Factorization (BTMF) and Kernelized Probabilistic Matrix Factorization (KPMF) depends on a careful analysis of specific priorities and the percentage of missing data. It is noted that Kernelized Probabilistic Matrix Factorization (KPMF) excels in reducing extreme errors, which is reflected by a globally lower Root Mean Square Error (RMSE) for datasets with 30% and 60% missing values.

Conversely, the Bayesian Temporal Matrix Factorization (BTMF) model presents a slightly higher RMSE, a consequence of its greater sensitivity to extreme values. However, in contexts where the missing data rate reaches 90%, the performance of the Kernelized Probabilistic Matrix Factorization (KPMF) deteriorates significantly. Under these conditions, the BTMF model (Bayesian Temporal Matrix Factorization) displays a considerably lower RMSE than that of the KPMF. Consequently, BTMF proves to be the most suitable and effective model for the imputation of missing values in situations where the rate of missing data is extremely high. Nevertheless, it should be noted that, despite its superior performance in terms of accuracy, the Bayesian Temporal Matrix Factorization (BTMF) model can sometimes encounter convergence issues and exhibit increased sensitivity to extreme values. Thus, the choice between Kernelized Probabilistic Matrix Factorization (KPMF) and Bayesian Temporal Matrix Factorization (BTMF) should be based on specific criteria such as the required accuracy in data imputation, tolerance to large-scale errors, and the percentage of missing data present in the dataset concerned.

Keywords

Bayesian analysis, Probabilistic matrix factorization, Markov chain Monte Carlo method (MCMC), Gibbs sampling, Data imputation, ARIMA models, Time series models, Bayesian models.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ACF AutoCorrelation Function

ADN Acide désoxyribonucléique

ARIMA AutoRegressive Integrated Moving Average

BPMF Bayesian Probabilistic Matrix Factorization

BTMF Bayesian Temporal Matrix Factorization

BTTF Bayesian Temporal Tensor Factorization

CO2 Dioxyde de carbone

CP Décomposition CANDECOMP/PARAFAC

DM Données manquantes

DMB Données manquantes par blocs

DMM Données manquantes mixtes

DMU Données manquantes unitaires

GPMF Bayesian Gaussian Process Probabilistic Matrix Factorization

KPMF Kernelized Probabilistic Matrix Factorization

MAP Maximum a posteriori

MCMC Chaines de Markov par Monte Carlo

MEAN Modèle d'imputation par la moyenne

MTQ Ministère des Transports du Québec

OCDE Organisation de coopération et de développement économiques

PACF Partial AutoCorrelation Function

PMF Probabilistic Matrix Factorization

RMSE Root Mean Square Error

STI Systèmes de transport intelligents

TRMF Temporal Regularized Matrix Factorization

Remerciements

C'est avec un grand plaisir et un profond sentiment de devoir que je remercie toutes les personnes ayant contribué, de près ou de loin, à l'aboutissement de ce mémoire. Je suis particulièrement reconnaissant envers ma directrice de recherche, la Professeure Aurélie Labbe, dont l'expertise a été une source d'apprentissage inestimable pour moi. Merci pour votre patience, votre disponibilité et votre dévouement !

Ma gratitude s'étend également à tous mes enseignants qui, tout au long de ma formation, m'ont guidé avec confiance et encouragement.

Je tiens à exprimer ma profonde gratitude envers ma famille, qui a joué un rôle essentiel dans mon parcours universitaire. En particulier, à ma chère mère qui n'a pas ménagé ses efforts pour me pousser à ne jamais me décourager et à poursuivre mon chemin, fixant ainsi notre destin. À toi, maman, je dédie ce travail !

À mes chères sœurs, pour votre amour indéfectible, et à mes amis, qui n'ont jamais hésité à m'apporter leur soutien et surtout leur compréhension.

Ce mémoire n'aurait pas été possible sans le concours précieux de toutes ces personnes, et je leur en suis profondément reconnaissant.

Chapitre 1

Introduction

Le transport routier est devenu un obstacle majeur au développement économique des régions et des pays (OCDE (2003)). En effet, la congestion du trafic est lourde de conséquences pour la qualité de vie des populations, pour l'environnement et pour l'économie. Les embouteillages infligent des coûts socio-économiques supplémentaires aux acteurs de la société, c'est-à-dire les individus, les entreprises et les instances gouvernementales (Didier (2020)). Ces coûts se manifestent dans l'augmentation du temps de transport, la consommation accrue de carburant, la pollution de l'environnement, la santé psychologique des usagers et l'augmentation des risques d'accidents (Moustakbal (2009)).

À l'heure actuelle, les zones métropolitaines du monde entier sont confrontées à des problèmes de congestion routière (OCDE (2006)). Une étude réalisée par le ministère des Transports du Québec, en collaboration avec la firme de consultants ADEC, a montré qu'à Montréal, la congestion a coûté 1,85 milliard de dollars canadiens en 2008. Ces coûts sont principalement dus aux retards des usagers qu'occasionnent les heures de pointe. De plus, des coûts supplémentaires sont associés aux émissions de polluants atmosphériques : ces coûts ont été estimés à 44,9 millions de dollars (MTQ (2014)). Ce problème n'est pas isolé : en 2011, aux États-Unis, les congestions routières ont coûté 121 milliards de dollars et ont été responsables de

l'émission supplémentaire de 56 milliards de livres de dans l'atmosphère (Schrank (2012)). À l'échelle mondiale, le Groupe d'experts intergouvernemental sur l'évolution du climat des Nations Unies (GIEC) a identifié le secteur du transport comme responsable de 23 % des émissions mondiales de gaz à effet de serre. Devant cette réalité, l'Organisation de coopération et de développement économiques (OCDE) insiste sur l'urgence d'investir dans les infrastructures urbaines pour gérer efficacement la congestion (PricewaterhouseCoopers (2013)).

De fait, la compréhension, de la part des gestionnaires d'infrastructures et des décideurs en matière de transport, du flux de circulation des véhicules sur le réseau routier s'avère nécessaire. Celle-ci permettrait d'équilibrer les impacts précédemment mentionnés et de soutenir une croissance économique durable. À cet effet, les systèmes de transport intelligents (STI) se présentent comme une solution prometteuse pour gérer efficacement les embouteillages (Kołodziej (2022)). Les STI permettent de créer une coopération entre les utilisateurs, les véhicules et les infrastructures routières. Ils sont utilisés pour exploiter les mesures de flux de trafic dans le but de surveiller le trafic, de prédire les tendances, ainsi que de planifier et de concevoir de nouvelles installations (Canada (2022)). Par exemple, une étude longitudinale portant sur les STI dans 99 zones urbaines aux États-Unis sur une période de 20 ans, de 1994 à 2014, a montré que le recours aux STI était associé à une réduction des coûts de plus de 4,7 milliards de dollars et à une réduction de la durée des trajets de plus de 175 millions d'heures par an. De plus, une meilleure gestion du trafic a non seulement réduit la consommation de carburant de 53 millions de gallons, mais a aussi réduit les émissions de CO2 de plus de 10 milliards de livres par an (Cheng (2020), Dirks (2010)).

Cependant, il existe un obstacle majeur à l'exploitation des STI. En effet, ces systèmes dépendent de l'utilisation de capteurs pour collecter une grande variété de données, telles que la vitesse et le flux des véhicules, la surveillance des conditions météorologiques routières et la détection des incidents de circulation. Or, ces capteurs peuvent être sujets à divers problèmes : défaillances techniques, problèmes de

connexion, intempéries ou même vandalisme.

Par conséquent, en raison des problèmes fréquents affectant les capteurs, les ensembles de données présentent souvent des lacunes, menant à la perte ou à la corruption des informations recueillies. Cela affecte négativement la précision des prévisions de flux de trafic et réduit l'efficacité des STI dans la gestion de la congestion.

En outre, le problème des données manquantes peut survenir pour de nombreuses raisons, quel que soit le format des données :

- Dans les enquêtes ou les sondages, il est fréquent que certains participants ne répondent pas à toutes les questions ou qu'il y ait des erreurs de saisie de données (Kovar (1995)) ;
- Dans les données géospatiales, en particulier dans le domaine immobilier, la présence de données manquantes est souvent le résultat d'actions indépendantes de plusieurs autorités publiques. Cette situation est exacerbée par la fusion non coordonnée de différentes sources de données (Khrulkov (2022)) ;
- Dans les jeux de données textuelles ou multimédias, les données manquantes peuvent se présenter en raison de problèmes de stockage ou de corruption de données (Pazhoohesh (2021)).

De ce fait, le problème des données manquantes est courant dans de nombreux domaines d'activité. Par exemple, le domaine des systèmes de recommandation et de filtrage collaboratif nécessite la résolution des problèmes d'imputation à grande échelle. Le concours pour le prix Netflix représente un bon exemple de cette situation. Lors du concours, Netflix a fourni des données issues de son système de recommandation de films. La proportion de valeurs manquantes variait grandement entre ces données, et certains films n'avaient reçu que trois notes. L'objectif de Netflix était alors d'imputer, ou de prédire, les valeurs manquantes afin d'améliorer la précision des recommandations et de fidéliser les utilisateurs pour les inciter à continuer de s'abonner au service (Bennett (2007)).

La bio-informatique est un autre domaine confronté au problème des données manquantes, particulièrement avec les données provenant des micropuces (*microarray*), qui présentent un défi majeur. Par exemple, un ensemble de données de méthylation de l'ADN se présente généralement sous la forme d'une matrice échantillon-par-gène, dans laquelle le nombre d'échantillons M est souvent bien inférieur au nombre de gènes N. Dans cette matrice, chaque élément (i, j) représente le niveau de méthylation pour la j^e région du génome dans l'échantillon i (Schein (2021)). Pour analyser ces données et appliquer une méthode de réduction de la dimensionnalité, il est nécessaire d'avoir un ensemble complet de données. Toutefois, dans la pratique, la matrice de données est incomplète, ce qui complique son analyse. Or, toutes les techniques de réduction de dimensions habituellement appliquées nécessitent l'imputation des données manquantes afin d'obtenir un ensemble complet de données ainsi que des résultats fiables et précis.

Le récent domaine du traitement d'images suscite aussi un intérêt croissant. Comme les autres domaines évoqués, il est touché par le problème des données manquantes. En science de l'imagerie, les images obtenues présentent souvent une dégradation indésirable causée par des facteurs externes, tels que l'environnement et l'équipement électronique, ou encore par des facteurs humains. La restauration d'image est donc un problème important touchant diverses applications, comme l'imagerie médicale, la télédétection et la surveillance vidéo. Plus précisément, une image est représentée sous forme matricielle : les numéros de lignes et de colonnes correspondent à l'emplacement vertical et horizontal et les valeurs représentent l'intensité des pixels sur l'image (Weigert (2018), Hippert (2020)). La restauration d'image vise donc à améliorer la qualité de l'image en imputant des valeurs aux pixels manquants.

En statistique, le problème des données manquantes se pose lorsque l'on dispose d'une matrice de données réelles et que quelques éléments (i, j) de la matrice sont absents. Cette situation est problématique, car de nombreux algorithmes ou modèles probabilistes sont conçus pour fonctionner avec des matrices complètes, ne serait-

ce que pour calculer des estimations de base telles que la moyenne, la médiane ou l'écart-type. En effet, l'absence, même d'un seul élément, peut empêcher l'utilisation de l'ensemble du jeu de données et forcer l'analyste à ignorer les lignes et/ou les colonnes incomplètes. En effet, la méthode consistant à ignorer les observations qui contiennent des données manquantes (méthode implantée par défaut dans de nombreux logiciels statistiques) peut entrainer une réduction significative de la taille des données à analyser. Par exemple, dans le cas d'une analyse de régression dans laquelle une observation (une ligne du jeu de données) présente une valeur manquante pour l'une des variables prédictives (ou covariables), toute l'observation est retirée avec la commande *lm()* dans R (R Core Team (2022)). Ce processus d'exclusion est souvent appelé « analyse des cas complets ». Or, cette approche conduit à une perte d'information ou à une fausse représentation des phénomènes étudiés. Elle risque également de fausser l'analyse et les décisions qui en découlent.

Enfin, dans tous les cas, il est souvent intéressant de remplacer les valeurs manquantes par des estimations appropriées. Cette manière de faire permet une analyse plus complète de l'ensemble de la matrice de données.

1.1 Mise en contexte

Dans le domaine du transport routier, plusieurs méthodes ont été développées pour remédier au problème des données manquantes (Faloutsos (2018), Shi (2018), Li (2018)). Certaines méthodes sont simples et rapides, tandis que d'autres sont plus complexes et lentes (Jabir (2018)). Dans cette étude, nous explorons plusieurs modèles de traitement des valeurs manquantes, allant des modèles simples, comme l'imputation par la moyenne, aux modèles plus élaborés, tels que les modèles de séries chronologiques (Cryer (1986)) et les modèles basés sur les méthodes de factorisation matricielle (Singh et Gordon (2008), Lin (2020)). Plus particulièrement, nous nous concentrons sur les modèles basés sur les méthodes de factorisation matricielle. Nous nous intéressons précisément aux modèles qui intègrent une modélisation élaborée de

la matrice de variance-covariance des données dans le but de décrire adéquatement la structure spatiotemporelle inhérente aux données issues du transport routier. Ces modèles sont particulièrement efficaces dans des contextes multivariés, où les variables affichent une forte corrélation (Joreskog (1973)). Ces derniers permettent non seulement de préserver la structure spatiotemporelle, mais aussi d'améliorer la qualité de l'imputation des données manquantes. Cette amélioration dans l'imputation des données manquantes permet une interprétation plus fiable des résultats et facilite la prise de décision (Xie (2023)).

L'objectif de la factorisation matricielle est de trouver les matrices $\boldsymbol{U} \in \mathbb{R}^{K \times M}$ et $\boldsymbol{V} \in \mathbb{R}^{K \times N}$ pour une matrice donnée, $\boldsymbol{R} \in \mathbb{R}^{M \times N}$, en considérant que : $\boldsymbol{R} \approx \boldsymbol{U}^\top \boldsymbol{V}$. Dans ces matrices K représente le rang de décomposition et K est beaucoup plus petit que M et N ($K \ll M, N$). Lorsque la matrice $\boldsymbol{R}$ est complète, la meilleure estimation est donnée par la décomposition de la valeur singulière. Cependant, en présence de valeurs manquantes, cette technique ne peut plus être utilisée pour fournir une solution exacte. Par conséquent, le problème des données manquantes n'offre pas de solution optimale et la recherche du meilleur algorithme d'imputation est toujours d'actualité. Plusieurs solutions alternatives, basées sur les méthodes de factorisation matricielle, ont été envisagées pour aborder ce problème (Lawrence (2009), (Agarwal (2010), Shan (2010)). Pensons, notamment, à la recherche de la meilleure approximation du rang K de la matrice de données partiellement observée sous une fonction de perte spécifique, laquelle se contruit ainsi :

$$\underset{\boldsymbol{U},\boldsymbol{V}}{min} \sum_{(i,j)\in\Omega} (R_{i,j} - \boldsymbol{U}_{:i}^\top \boldsymbol{V}_{:j})^2 + \lambda_{\boldsymbol{U}} \mathcal{R}_U(\boldsymbol{U}) + \lambda_V \mathcal{R}_{\boldsymbol{V}}(\boldsymbol{V}), \tag{1.1}$$

où Ω représente l'ensemble des indices des données observées et $\boldsymbol{U}$ et $\boldsymbol{V}$ sont les matrices que nous cherchons à estimer. $\boldsymbol{U}_{:i}$ est la i^e colonne de $\boldsymbol{U}$ et représente les caractéristiques latentes associées à la ligne i de $\boldsymbol{R}$. De même, $\boldsymbol{V}_{:j}$ est la j^e colonne de $\boldsymbol{V}$ et représente les caractéristiques latentes associées à la colonne j de $\boldsymbol{R}$. Les paramètres de régularisation $\lambda_{\boldsymbol{U}}$ et $\lambda_{\boldsymbol{V}}$ sont associés, respectivement, aux termes de

régularisation $\mathcal{R}_U(\boldsymbol{U})$ et $\mathcal{R}_V(\boldsymbol{V})$, qui permettent d'éviter le surapprentissage et/ou l'emploi de certaines structures spécifiques. En trouvant une estimation des deux matrices latentes, $\boldsymbol{U}$ et $\boldsymbol{V}$, nous pouvons obtenir une estimation de la matrice $\hat{\boldsymbol{R}}$ et, donc, des estimations pour chaque observation manquante $R_{i,j}$ dans $\boldsymbol{R}$.

Pour ce mémoire, nous disposons d'un jeu de données publiques recueilli par des détecteurs de trafic déployés sur quatre autoroutes de la région de Seattle. Ce jeu de données contient des mesures spatiotemporelles de la vitesse du réseau autoroutier pour une période d'un an à intervalle de temps constant de 5 minutes. Chaque observation de vitesse attribuée à une borne kilométrique représente une moyenne calculée à partir de plusieurs détecteurs à boucle situés sur les voies principales dans la même direction du trafic. Les données sont stockées sous la forme de matrice, nommée $\boldsymbol{R}$, de dimensions $M \times N$, où les M lignes représentent les 323 capteurs et où les N colonnes représentent le temps (365 jours x 24 heures x 12 intervalles de temps discret de 5 minutes). L'élément (i, j) de la matrice de données représente la vitesse observée sur le capteur i au temps j. Ces données ont des caractéristiques particulières, car elles combinent à la fois des attributs spatiaux, tels que la distance entre les capteurs, la direction vers laquelle le capteur est placé et la localisation du capteur, avec des attributs temporels, tels que la tendance de la vitesse moyenne au cours du temps et la périodicité de la vitesse observée à travers les jours de la semaine (Oberoi (2019)). Cette dépendance spatiotemporelle induit une forte corrélation entre les observations, ce qui nécessite l'introduction d'une modélisation fine de la matrice de variance-covariance des données pour ces observations. Grâce à une modélisation fine de la matrice de variance-covariance des données, il est possible d'obtenir des résultats plus robustes lors de l'imputation des données manquantes dans ce contexte spatiotemporel.

1.2 Objectif de la recherche

Notre recherche vise à mener une étude comparative des modèles d'imputation de données manquantes basés sur la méthode de factorisation matricielle décrite précédemment. Nous nous concentrons sur des modèles récents qui intègrent la structure complexe de la matrice de variance-covariance des données pour résoudre le problème de l'imputation des données manquantes dans un contexte spatiotemporel.

Nous souhaitons comparer les modèles en termes de qualité d'imputation et de vitesse de calcul. Nous analyserons l'impact du rang de décomposition et de l'absence de données sur les performances des modèles d'imputation choisis. De plus, nous évaluerons l'impact de la structure des données manquantes sur l'efficacité de ces modèles à imputer les valeurs manquantes. Pour ce faire, notre étude s'appuiera sur les données de trafic routier de la ville de Seattle mentionnées précédemment. Nous créerons différents scénarios artificiels de valeurs manquantes à partir des données originales pour effectuer nos analyses. Nous espérons que les résultats de notre recherche permettront de mieux comprendre les avantages et les limites de chaque modèle et d'identifier le modèle le plus approprié pour résoudre le problème de données manquantes dans le secteur du transport routier.

Le second chapitre de ce mémoire présente une introduction à l'analyse bayésienne. Cette approche repose sur la loi de Bayes, qui permet de mettre à jour l'information *a priori* à partir de nouvelles observations. Nous expliquerons cette approche et son utilisation dans le contexte de notre étude. Un exemple concret issu du jeu de hasard pile ou face illustrera ces concepts, suivi d'une introduction à la méthode de Monte-Carlo par chaines de Markov et à l'échantillonnage de Gibbs.

Le chapitre trois constitue notre revue de la littérature sur ce sujet. Il est consacré à la présentation des modèles retenus pour notre analyse. Chacun de ces modèles a fait l'objet d'un article scientifique récent. Nous détaillerons donc les hypothèses de chaque modèle pour mieux comprendre leur fonctionnement et leur utilisation dans le contexte de données spatiotemporelles.

Le chapitre quatre se concentre sur la présentation des données utilisées dans notre étude. Il permet de découvrir et de présenter les caractéristiques spatiotemporelles inhérentes aux données du transport routier : la distribution spatiale et temporelle, les structures complexes de la dépendance spatiotemporelle et les composantes des séries chronologiques.

Le chapitre cinq se concentre sur la description du plan de simulation pour des scénarios artificiels de valeurs manquantes et présente les résultats de notre étude suite à l'application de ces modèles. Nous décrirons en détail le paramétrage des modèles et les méthodes utilisées pour comparer leurs performances, ainsi que les résultats obtenus.

Le chapitre six conclura cette étude en présentant les avantages et les limites de chaque modèle. Nous partagerons également les enseignements tirés, et amorcerons une discussion sur nos résultats ainsi que sur les limites de notre recherche.

Enfin, nous espérons que notre travail aidera à mieux comprendre les défis associés à l'imputation des données manquantes dans un contexte spatiotemporel. Nous visons à proposer des orientations stratégiques claires concernant le choix du modèle approprié pour traiter les données manquantes. Cette démarche a pour but d'améliorer la prise de décision et d'ainsi renforcer la capacité des systèmes de transport intelligents à gérer efficacement la congestion routière.

Chapitre 2

Introduction à l'analyse bayésienne

Ce chapitre vise à introduire les concepts fondamentaux de l'analyse bayésienne, qui constitue le socle des modèles étudiés dans cette recherche. Les fondements théoriques sont principalement inspirés des travaux de Chen (2021), Stone (2013) et Martin (2018). Des méthodes d'inférence bayésienne spécifiques sont discutées en suivant les orientations de Parent (2007), Shalev-Shwartz (2014), Dax (2014), Mohamed (2011) et Pelgrin (2008). En outre, des exemples pratiques et des approches computationnelles sont adaptés de Murphy (2012), Boutahar (2015), Boutahar (2019) et des notes de cours de Nalborczyk (2020) et Dupuis (2007). Enfin, des cas d'étude et des exemples introductifs ont été adaptés de Donovan et Mickey (2019) et Jadeja (2021).

2.1 Théorème de Bayes

Le terme *bayésien* est dérivé du nom du mathématicien britannique Thomas Bayes (1702-1761). Dans son article *Philosophical Transactions of the Royal Society of London*, il se prête à l'examen des probabilités conditionnelles, c'est-à-dire, des probabilités que l'évènement A et l'évènement B se produisent de manière indépendante ou conditionnelle (O'Connor (2023)). La probabilité conditionnelle de

l'évènement A sachant B est notée $P(A|B)$ et définie comme suit :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \tag{2.1}$$

en supposant que B est un évènement de probabilité non nulle. De la même manière, si A est un évènement de probabilité non nulle, nous avons :

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{2.2}$$

De 2.1 et 2.2, nous déduisons le théorème de Bayes (1763) :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2.3}$$

où $P(A)$ et $P(B)$ représentent les probabilités marginales de A et B, respectivement.

2.2 Inférence bayésienne

Nous considérons ici une variable aléatoire, Y, dont nous observons la suite de réalisations $\boldsymbol{y} = (y_1, ..., y_n)$. Nous supposons que Y suit une certaine distribution dont la fonction de densité, notée $f(\boldsymbol{y}|\boldsymbol{\theta})$, dépend d'un vecteur de paramètres $\boldsymbol{\theta}$ que nous souhaitons estimer. Dans l'analyse bayésienne, le vecteur de paramètres $\boldsymbol{\theta}$ est considéré comme une variable aléatoire, dont la distribution *a priori* associée est notée $\pi(\boldsymbol{\theta})$. Le modèle probabiliste pour ces observations est donné comme suit :

$$\begin{aligned} Y &\sim f(\boldsymbol{y}|\boldsymbol{\theta}), \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2.4}$$

À l'aide du théorème de Bayes, nous pouvons combiner l'information associée à la distribution *a priori* et l'information apportée par les données observées afin d'obtenir une distribution *a posteriori* du vecteur de paramètres $\boldsymbol{\theta}$. Cette distribution est définie comme suit :

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})}, \tag{2.5}$$

$$\propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{2.6}$$

- $p(\boldsymbol{\theta}|\boldsymbol{y})$ est la distribution *a posteriori*, sachant $\boldsymbol{y}$. Elle résume l'incertitude portée aux valeurs du vecteur de paramètres $\boldsymbol{\theta}$.
- $f(\boldsymbol{y}|\boldsymbol{\theta})$ est la fonction de vraisemblance (*likelihood*) du vecteur de paramètres $\boldsymbol{\theta}$, une fois $\boldsymbol{y}$ observée, c'est-à-dire, la vraisemblance d'observer les données $\boldsymbol{y}$ en particulier compte tenu du vecteur de paramètres $\boldsymbol{\theta}$.
- $\pi(\boldsymbol{\theta})$ est la loi *a priori*. C'est la distribution qui décrit l'information *a priori*.
- $f(\boldsymbol{y})$ est la loi marginale de Y. C'est la probabilité d'observer les données $\boldsymbol{y}$, peu importe la valeur de $\boldsymbol{\theta}$: $f(\boldsymbol{y}) = \sum_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ lorsque $\boldsymbol{\theta}$ est discret et $f(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ lorsque $\boldsymbol{\theta}$ est continue. Le passage de 2.5 à 2.6 est justifié par le fait que $f(\boldsymbol{y})$ est une constante de normalisation permettant à $p(\boldsymbol{\theta}|\boldsymbol{y})$ d'avoir une intégration de 1.

L'inférence bayésienne utilise le théorème de Bayes pour estimer la distribution du vecteur de paramètres inconnu $\boldsymbol{\theta}$ en fonction de données observées $\boldsymbol{y}$. Cette méthode permet d'ajuster l'information *a priori*, représentée par la loi *a priori*, de manière cohérente à partir de la vraisemblance des données observées. La distribution *a priori* et la distribution *a posteriori* décrivent les incertitudes associés au vecteur de paramètres $\boldsymbol{\theta}$. Si la distribution *a priori* et la distribution *a posteriori* appartiennent à la même famille de distributions de probabilités, elles sont appelées « **distributions conjuguées** ». Dans ce cas, la distribution *a priori* est spécifiquement appelée « **distribution *a priori* conjugée** » pour la fonction de vraisemblance $f(\boldsymbol{y}|\boldsymbol{\theta})$.

Les modèles statistiques bayésiens peuvent aussi être **hiérarchiques**, c'est-à-dire que la distribution *a priori* $\pi(\boldsymbol{\theta})$ peut elle-même être décomposée en plusieurs lois conditionnelles et une loi marginale. Cette décomposition prend la forme suivante :

$\pi(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta}|\boldsymbol{\theta}_1)\pi_2(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)...\pi_n(\boldsymbol{\theta}_{n-1}|\boldsymbol{\theta}_n)\tau(\boldsymbol{\theta}_n)$, où $\boldsymbol{\theta}_i$ désigne un vecteur d'hyperparamètres (Boutahar (2015)). L'utilisation de ces modèles permet surtout de décrire l'information *a priori* sur $\boldsymbol{\theta}_i$ de manière plus subtile et plus complexe.

Pour illustrer le processus d'inférence bayésienne et ses concepts décrit précédemment, nous allons maintenant examiner un exemple pratique. Pour des raisons de simplicité, notre attention se porte sur l'inférence d'un unique paramètre inconnu, noté θ.

Comme présenté antérieurement, le théorème de Bayes permet de combiner l'information associée à la distribution *a priori* et l'information apportée par les données observées afin d'obtenir une distribution *a posteriori* du paramètre θ. Nous utiliserons, comme exemple, le jeu de hasard pile ou face, qui se joue avec une pièce de monnaie. Le principe du jeu est de lancer une pièce de monnaie plusieurs fois et d'enregistrer le nombre de fois où le côté « pile », ou côté « face », est obtenu.

Dans ce jeu, nous nous intéressons à la variable aléatoire Y_N, qui correspond au nombre de côtés « face » obtenus après N lancers indépendants successifs. L'univers Ω associé à cette expérience aléatoire se construit avec l'ensemble binaire $\{0, 1\}$, où $X = 1$ lorsque le côté « face » est obtenu et où $X = 0$ lorsque le côté « pile » est obtenu. À chaque essai, si on admet que $P(face) = \theta$, alors $P(pile) = 1 - \theta$. Sous les conditions que nous avons décrites, la variable aléatoire Y_N suit une loi binomiale paramétrée par N et θ. La taille de l'échantillon N est supposée connue et nous travaillons conditionnellement à sa valeur. Ainsi, la problématique qui nous intéresse consiste à déterminer si la pièce de monnaie est équilibrée (c'est-à-dire, si $\theta = 0.5$), ou si elle est biaisée en faveur de l'un des deux côtés (c'est-à-dire, si $\theta \neq 0.5$). Pour y répondre nous devons estimer au mieux la valeur inconnue du paramètre θ.

Pour estimer la valeur inconnue du paramètre θ, qui représente la probabilité d'obtenir le côté « face » lorsqu'une pièce de monnaie est lancée, nous supposons disposer de y, une réalisation de la variable aléatoire Y_N. Cette observation a été obtenue en simulant N variables aléatoires de Bernoulli $X_1, ..., X_N$ de paramètre

$\theta \in [0, 1]$, indépendantes et identiquement distribuées. Pour ce faire, nous avons utilisé la fonction *bernouilli.rvs()* du logiciel Python. Nous donnons en paramètre à la fonction une valeur biaisée en faveur du côté pile de $\theta = 0.65$ et une valeur de $N = 100$. La fonction retourne de façon aléatoire une séquence de 100 lancers et nous observons : $y = \sum_{i=1}^{N}(x_i) = 72$.

Dans le but d'illustrer la démarche bayésienne, nous allons tenter d'estimer la valeur de θ, qui a servi à générer ces données.

Selon le théorème de Bayes 2.5, nous aurons besoin de spécifier la fonction de vraisemblance et la loi *a priori* pour le paramètre θ.

La fonction de vraisemblance

Comme décrit précédemment, la fonction de vraisemblance désigne la distribution binomiale qui s'écrit sous cette forme :

$$p(y|\theta, N) = \frac{N!}{y!(N-y)!}\theta^y(1-\theta)^{N-y}. \tag{2.7}$$

Cette fonction de probabilité représente le nombre de succès dans une séquence de N expériences indépendantes, sachant la valeur du paramètre θ. Si nous connaissons la valeur de θ, la distribution binomiale nous donnera la distribution attendue du nombre de faces, mais comme nous ignorons la valeur du paramètre θ, nous lui attribuerons une loi *a priori*.

Loi *a priori*

La loi Beta est une distribution pour une variable continue, θ, prenant des valeurs dans l'intervalle [0,1]. Cette distribution est paramétrée par α et β, qui contrôlent la forme de la distribution. La fonction de densité de cette distribution , où $\Gamma(\cdot)$ est la fonction Gamma, est donnée comme suit :

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{2.8}$$

Il y a plusieurs raisons pour lesquelles la distribution Beta est souvent choisie comme distribution *a priori*. Tout d'abord, sa variable aléatoire est bornée entre 0 et 1, tout comme le paramètre recherché, ce qui en fait un choix naturel. En outre, la distribution Beta est très flexible et peut prendre de nombreuses formes différentes, y compris la distribution uniforme. Cela la rend très pratique pour exprimer une grande variété d'hypothèses *a priori*.

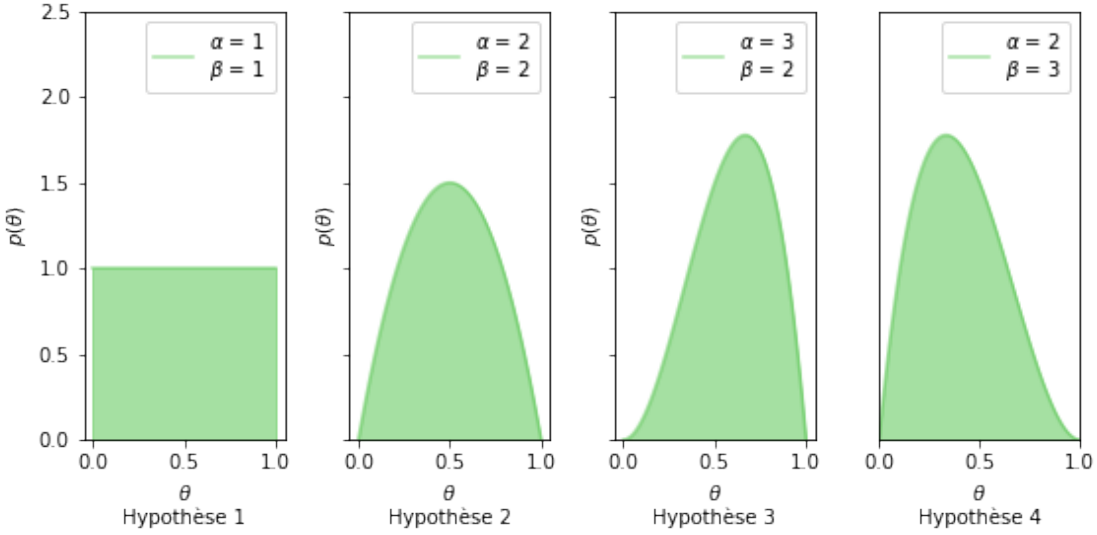

FIGURE 2.1 – Exemples d'utilisation de la distribution Beta pour exprimer des hypothèses *a priori*. Les hypothèses sont représentées par les valeurs des paramètres α et β : Hypothèse 1 : $\alpha = \beta = 1$, cela signifie qu'il n'y a aucune information *a priori* disponible. Hypothèse 2 : $\alpha = \beta = 2$, cela indique qu'il n'y a pas de biais particulier et que les deux résultats, « pile » et « face », ont une probabilité égale d'être observée. Hypothèse 3 : α est supérieur à β, cela implique un biais en faveur d'un résultat « face ». Hypothèse 4 : β est supérieur à α, cela indique un biais en faveur d'un résultat « pile ».

Dans notre contexte, la distribution Beta peut être utilisée pour exprimer des biais en faveur d'un résultat particulier en ajustant les valeurs de ses paramètres α et β. Par exemple, lorsque $\alpha = \beta = 1$, cela signifie qu'il n'y a aucune information *a*

priori disponible. Lorsque $\alpha = \beta = 2$, cela indique qu'il n'y a pas de biais particulier et que les deux résultats, « pile » et « face », ont une probabilité égale d'être observés. Si α est supérieur à β, cela implique un biais en faveur d'un résultat « face », tandis que si β est supérieur à α, cela indique un biais en faveur d'un résultat « pile ». La figure 2.1 illustre comment la distribution Beta peut prendre différentes formes en fonction des valeurs de α et de β, permettant ainsi d'exprimer une grande variété d'hypothèses *a priori*.

Loi *a posteriori*

Selon le théorème de Bayes, nous devons multiplier la fonction de vraisemblance par la distribution *a priori*. Nous devons donc multiplier la fonction de vraisemblance binomiale par la fonction Beta comme suit :

$$p(\theta|y) = \frac{N!}{y!(N-y)!}\theta^{y}(1-\theta)^{N-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{2.9}$$

Mathématiquement, nous pouvons simplifier cette équation en éliminant tous les termes qui ne dépendent pas de θ et les ordonner. Nous obtenons alors :

$$p(\theta|y) \ \alpha \quad \theta^{y+\alpha-1}(1-\theta)^{N-y+\beta-1}. \tag{2.10}$$

Cette expression a la même forme que l'expression de la distribution *a priori* 2.8, où $\alpha_{a\ posteriori} = \alpha_{a\ priori} + y$ et $\beta_{a\ posteriori} = \beta_{a\ priori} + N - y$. Ainsi, nous partons d'une distribution *a priori* Beta et nous aboutissons à une distribution *a posteriori* également Beta. Cette cohérence mathématique est nommée situation de conjugaison. En somme, la distribution *a posteriori* n'est autre que la distribution *a priori* Beta mise à jour. Elle fournit aussi une information fort utile sur l'usage des deux paramètres d'équilibrage y et $N-y$, qui ne sont rien d'autre que le nombre de « faces » et de « piles » observés, respectivement.

La figure 2.2 compare les résultats de quatre modèles bayésiens où chaque ligne représente le résultat d'un modèle avec une hypothèse *a priori* différente. Les hypothèses sont caractérisées par les valeurs des paramètres α et β telles qu'illustrées dans la figure 2.1. Chaque colonne représente trois ensembles de données correspondant à un nombre différent d'expériences : 10 lancers, 100 lancers et 200 lancers. La couleur verte représente la distribution *a priori*, la couleur rouge, la fonction de vraisemblance et la couleur bleue, la distribution *a posteriori*. Le nombre de lancers et le nombre de faces sont indiqués dans chaque graphique. La ligne noire verticale a une valeur de 0.65 indiquant la vraie valeur de θ, fixée ultérieurement. Le résultat de l'analyse bayésienne est une distribution *a posteriori* pour le paramètre θ, dont la valeur la plus probable est donnée par le mode de la distribution, qui est ici de 0.65, soit le résultat attendu pour tous les scénarios considérés. Nous remarquons également que la variance de la distribution *a posteriori* est proportionnelle à l'incertitude que le modèle a quant à la valeur du paramètre. En effet, la variance de la distribution *a posteriori* suit une tendance décroissante en fonction de l'augmentation du nombre de lancers : la variance de la distribution est plus grande lors de 10 lancers que lors de 100 lancers et celle observée après 100 lancers est supérieure à celle constatée après 200 lancers. Cette dynamique indique la diminution de l'incertitude concernant la valeur du paramètre θ à mesure que le nombre de lancers augmente. Plus précisément, une plus grande précision dans l'estimation de θ est obtenue lorsque la variance de la distribution *a posteriori* est réduite. Lorsque le nombre de lancers est suffisamment élevé, les quatre modèles, chacun doté de différentes hypothèses *a priori*, convergent vers une estimation commune de $\theta = 0.65$, la probabilité attendue d'obtenir le côté « face » lors du lancer d'une pièce de monnaie.

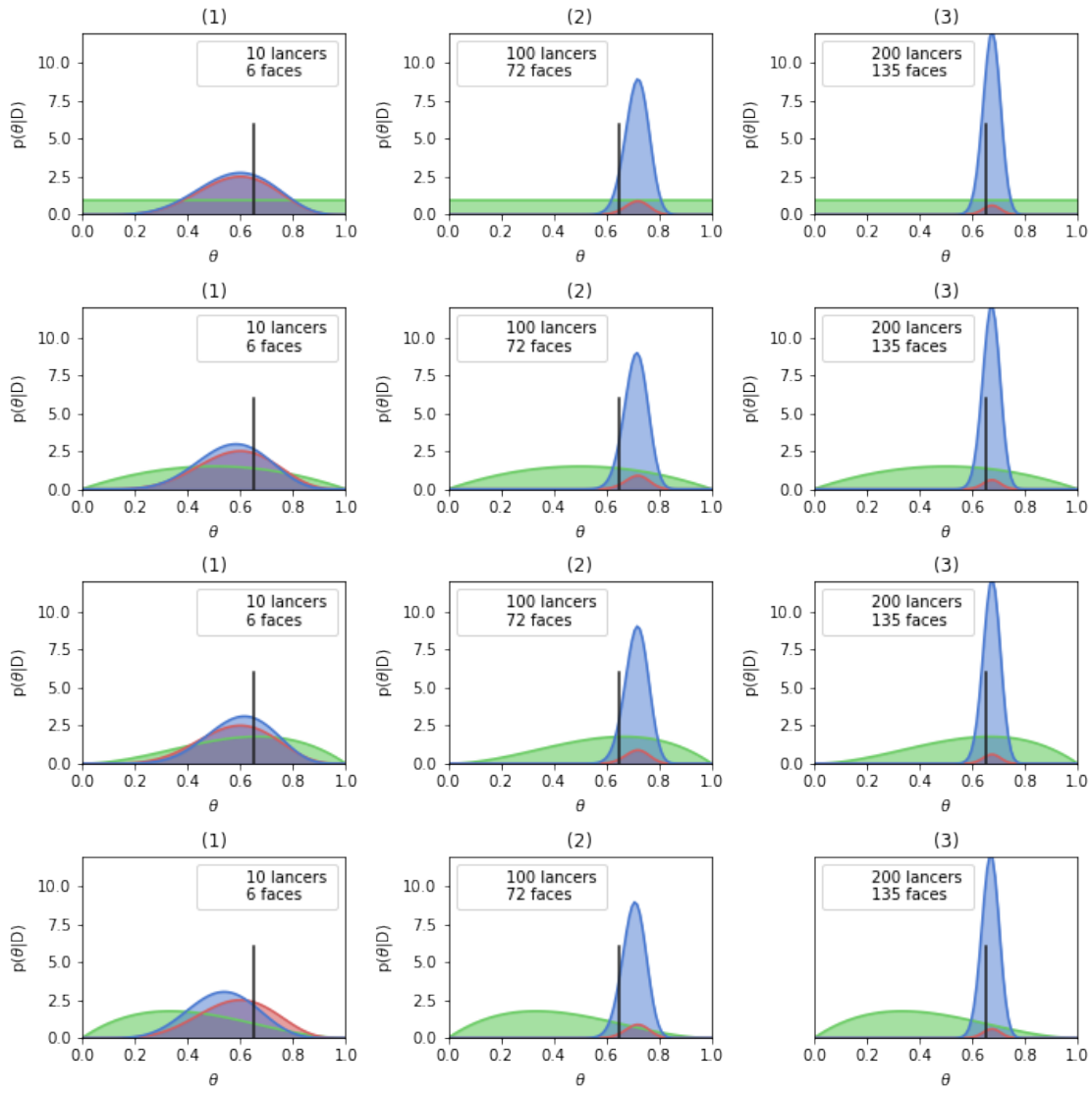

FIGURE 2.2 – Processus d'inférence bayésien selon différentes hypothèses *a priori*. Les courbes vertes représentent les distributions *a priori*, Les courbes rouges, les fonctions de vraisemblance et les courbes bleues, les distributions *a posteriori*. Les hypothèses *a priori* sont représentées par les valeurs des paramètres α et β : La ligne 1 correspond à l'hypothèse 1, où il n'y a aucune information *a priori* disponible. La ligne 2 correspond à l'hypothèse 2, où il n'y a pas de biais particulier et où les deux résultats, « pile » et « face », ont une probabilité égale d'être observée. La ligne 3 correspond à l'hypothèse 3, qui implique un biais en faveur d'un résultat de « face ». La ligne 4 correspond à l'hypothèse 4, qui indique un biais en faveur d'un résultat « pile ». Chaque colonne représente le résultat de trois expériences : 10 lancers, 100 lancers et 200 lancers.

2.3 Méthode de Monte-Carlo par chaines de Markov

Lorsque l'on souhaite estimer plusieurs paramètres, le calcul de la distribution *a posteriori* conjointe peut s'avérer très complexe, voire impossible, car la vraisemblance peut être difficile à calculer et la constante d'intégration peut nécessiter une intégration multidimensionnelle. Pour les modèles complexes de grande dimension, des algorithmes basés sur des simulations, comme les algorithmes de Monte-Carlo par chaines de Marcov (MCMC) (Neal (1993)), sont alors utilisés pour estimer ces paramètres de manière approximative, en surmontant les difficultés liées au calcul direct de cette loi.

Les algorithmes MCMC permettent de générer un échantillon de valeurs plausibles du vecteur de paramètres $\boldsymbol{\theta}$ selon sa distribution *a posteriori*. Chaque observation de cet échantillon est obtenue de manière itérative. Elle est le résultat d'un saut stochastique dans l'espace des paramètres en fonction de la valeur du paramètre obtenu à l'itération précédente. Il en résulte une dépendance markovienne entre les observations de l'échantillon, formant ainsi une chaine de Markov. Cette chaine est construite de manière à garantir la convergence vers une loi stationnaire, c'est-à-dire la loi *a posteriori* recherchée. Une fois la loi stationnaire atteinte, la chaine générée représente un échantillon de la loi *a posteriori* conjointe des paramètres et peut être utilisée pour estimer ces derniers (Lambert (2018)).

2.4 Échantillonnage de Gibbs

L'algorithme de Gibbs est un algorithme itératif de la famille des méthodes de Monte-Carlo par chaines de Markov (MCMC). L'échantillonnage de Gibbs est d'intérêt, surtout avec une distribution multivariée, car il est plus simple d'échantillonner à partir d'une distribution conditionnelle que de la marginaliser en intégrant une distribution conjointe. L'algorithme permet d'estimer de nombreux paramètres

considérant les autres paramètres fixes. Pour chaque itération MCMC, l'algorithme estime un seul paramètre à la fois, considérant les autres paramètres fixes. Pour ce faire, le théorème de Bayes est utilisé pour mettre à jour la distribution *a priori* vers une distribution *a posteriori*, puis pour échantillonner une proposition à partir de cette nouvelle distribution. Dans la même itération, l'algorithme passe au prochain paramètre, qui devient lui-même le paramètre d'intérêt. L'algorithme utilise le paramètre le plus récent pour créer la distribution conditionnelle de ce paramètre avant de procéder à son échantillonnage. Après un certain nombre d'itérations, les propositions sont utilisées pour estimer la distribution *a posteriori* conjointe, tout comme la distribution *a posteriori* marginale, pour chacun des paramètres.

Pour illustrer le fonctionnement de l'algorithme de Gibbs, supposons un modèle avec trois paramètres $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. Supposons également que nous observons plusieurs réalisations $\boldsymbol{y} = (y_1, ..., y_n)$ d'une variable aléatoire Y, dont la distribution est paramétrée par $\boldsymbol{\theta}$. Nous supposons qu'il est impossible de calculer une distribution *a posteriori* exacte, mais que nous sommes en mesure de calculer les trois distributions conditionnelles suivantes : $p(\theta_1|\theta_2, \theta_3, \boldsymbol{y}), p(\theta_2|\theta_1, \theta_3, \boldsymbol{y})$ et $p(\theta_3|\theta_1, \theta_2, \boldsymbol{y})$.

Nous supposons aussi que ces distributions sont assez simples pour pouvoir générer des échantillons pour chacune d'elles. Les instructions de l'algorithme de Gibbs s'écrivent comme suit :

Algorithm 1 Algorithme de Gibbs

1: Initialiser les paramètres $\theta_1^0, \theta_2^0, \theta_3^0$
2: **for** m=1 to n+N **do**
3: Simuler θ_1^m de $p(\theta_1^m|\theta_2^{(m-1)}, \theta_3^{(m-1)}, \boldsymbol{y})$
4: Simuler θ_2^m de $p(\theta_2^m|\theta_1^m, \theta_3^{(m-1)}, \boldsymbol{y})$
5: Simuler θ_3^m de $p(\theta_3^m|\theta_1^m, \theta_2^m, \boldsymbol{y})$
6: **end for**
7: Calculer $\hat{\theta}_{k\in\{1,2,3\}} = \frac{1}{N-n}\sum_{m=n+1}^{N}\theta_k^{(m)}$
8: Retourner $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$

L'algorithme commence par l'initialisation des paramètres inconnus $\theta_1, \theta_2, \theta_3$,

identifiés par $\theta_1^0, \theta_2^0, \theta_3^0$. Pour chaque itération (m), l'algorithme commence par simuler $\theta_1^{(m)}$ selon $p(\theta_1^m|\theta_2^{(m-1)}, \theta_3^{(m-1)}, \boldsymbol{y})$. Ensuite, il simule θ_2^m selon $p(\theta_2^m|\theta_1^m, \theta_3^{(m-1)}, \boldsymbol{y})$. Enfin, il simule θ_3^m selon $p(\theta_3^m|\theta_1^m, \theta_2^m, \boldsymbol{y})$. Ce processus itératif est répété (n+N) fois. Une fois les n premières itération terminées (atteinte de la loi stationnaire), les échantillons suivants sont utilisés pour calculer les estimations de Monte-Carlo $\hat{\theta_1}, \hat{\theta_2}, \hat{\theta_3}$ pour chacun des paramètres $\theta_1, \theta_2, \theta_3$, respectivement. Ainsi, la chaine de Marcov passe par deux phases : La phase de préchauffe (*burn-in*) qui correspond aux premières n itérations de l'algorithme, soit le temps nécessaire à la chaine de Marcov de converger vers sa loi stationnaire, et la phase d'échantillonnage de N itérations qui doit être assez longue pour une bonne estimation des paramètres (Donovan et Mickey (2019), Parent (2007)).

Chapitre 3

Description des modèles

Ce chapitre constitue notre revue de littérature qui couvre la description des modèles sélectionnés pour notre étude. Chaque modèle étudié y est présenté, en mettant l'accent sur les principes théoriques sous-jacents, ainsi que sur les hypothèses relatives à la modélisation des matrices de variance-covariance des données. Notre étude explore plusieurs modèles d'imputation, notamment l'imputation par la moyenne (MEAN), les modèles ARIMA (ARIMA), le Bayesian Probabilistic Matrix Factotization (BPMF), le Temporal Regularized Matrix Factorization (TRMF), le Bayesian Temporal Matrix Factorization (BTMF), le Bayesian Temporal Tensor Factorization (BTTF), le Kernelized Probabilistic Matrix Factorization (KPMF) ainsi que le Bayesian Gaussian Process Probabilistic Matrix Factorization (BGMF). Ces modèles seront appliqués à des données spatiotemporelles qui disposent de caractéristiques particulières, combinant à la fois des attributs spatiaux et temporels, telles que décrites précédemment. Par conséquent, nous accordons une attention particulière aux modèles qui intègrent une modélisation élaborée de la matrice de variance-covariance des données dans le but de décrire adéquatement la structure spatiotemporelle de ces données et, ainsi, obtenir des résultats plus robustes lors de l'imputation des données manquantes. Les modèles considérés comprennent des méthodes bien établies, innovantes et récentes.

3.1 Notation

Dans ce mémoire, nous utilisons les lettres minuscules pour désigner des scalaires (ex : x), les lettres minuscules en gras pour désigner des vecteurs (ex : $\boldsymbol{x} \in \mathbb{R}^M$) et les lettres majuscules en gras pour désigner des matrices (ex : $\boldsymbol{X} \in \mathbb{R}^{M \times N}$). Pour une matrice $\boldsymbol{X} \in \mathbb{R}^{M \times N}$, nous désignons la i^e ligne et la j^e colonne par $\boldsymbol{X}_{i:}$ et $\boldsymbol{X}_{:j}$, respectivement. Les éléments de la matrice sont notés par $X_{i,j}$. La norme Frobenius de $\boldsymbol{X}$ est désignée par $\|\boldsymbol{X}\|_F$. La transpose et la trace de $\boldsymbol{X}$ sont notées par $\boldsymbol{X}^\top$ et $tr(\boldsymbol{X})$, respectivement. Nous définissons la vectorisation de $\boldsymbol{X}$ comme sa concaténation de colonnes, c'est-à-dire en concaténant les vecteurs colonnes de $\boldsymbol{X}$ les uns en dessous des autres. Nous désignons cette vectorisation par $vec(\boldsymbol{X}) \in \mathbb{R}^{MN}$. Les opérations $\boldsymbol{A} \otimes \boldsymbol{B}$ et $\boldsymbol{A} \odot \boldsymbol{B}$ désignent le produit de Kronecker et la multiplication par éléments des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$, respectivement. Finalement, $\boldsymbol{I}_M$ désigne une matrice d'identité de dimensions $M \times M$ et $\mathbf{1}_M$ désigne un vecteur colonne de longueur M de 1.

Nous rappelons que nous disposons d'un jeu de données publiques recueilli par des détecteurs de trafic déployés sur quatre autoroutes de la région de Seattle. Ce jeu de données contient des mesures spatiotemporelles de la vitesse du réseau autoroutier pour une période d'un an à intervalle de temps constant de 5 minutes. Chaque observation de vitesse attribuée à une borne kilométrique représente une moyenne calculée à partir de plusieurs détecteurs à boucle situés sur les voies principales dans la même direction du trafic. Les données sont stockées sous forme de matrice, nommée $\boldsymbol{R}$, de dimensions $M \times N$, où les M lignes représentent les 323 capteurs, et les N colonnes représentent le temps (365 jours x 24 heures x 12 intervalles de temps discret de 5 minutes). L'élément (i, j) de la matrice de données représente la vitesse observée sur le capteur i au temps j.

3.2 Imputation par la moyenne (MEAN)

L'imputation par la moyenne (MEAN) est une méthode simple et couramment utilisée pour traiter les données manquantes dans des séries temporelles multivariées. Dans cette méthode, la matrice de données peut être considérée comme une collection de séries temporelles, c'est-à-dire que la série temporelle i correspond à la i^e ligne de la matrice de données $\boldsymbol{R}$, où chaque élément $R_{i,j}$ peut être observé ou manquant. Dans cette approche, si $R_{i,j}$ est observé, $\hat{R}_{i,j}$ prend la valeur de ce dernier. En revanche, si $R_{i,j}$ est manquant, $\hat{R}_{i,j}$ est égal à la moyenne des valeurs observées pour la série temporelle elle-même, i. La formule pour effectuer cette imputation est la suivante :

$$\begin{aligned} \hat{R}_{i,j} &= \begin{cases} R_{i,j} & \text{si } R_{i,j} \text{ est observé,} \\ \frac{1}{\eta_i} \sum_{j \in \Omega_i} R_{i,j} & \text{si } R_{i,j} \text{ est manquant,} \end{cases} \\ \eta_i &= |\Omega_i|, \end{aligned} \tag{3.1}$$

où Ω_i représente l'ensemble des indices j pour lesquels l'élément $R_{i,j}$ est observé pour la i^e série temporelle. η_i est le nombre d'éléments observés dans cette même série temporelle.

En somme, l'imputation par la moyenne assure que chaque élément manquant est remplacé par une constante qui représente la moyenne des valeurs observées appartenant à la même série temporelle que l'élément manquant.

3.3 Modèles de moyenne mobile intégrée autorégressive (ARIMA)

Les modèles de moyenne mobile intégrée autorégressive (ARIMA) ont été popularisés par Box (1976). Ces modèles ont pour but de prédire chaque valeur d'une série en fonction de ses valeurs antérieures. En d'autres termes, un modèle ARIMA

combine trois processus pour prédire les valeurs futures : un processus autorégressif, un processus de moyenne mobile et un processus d'intégration. Plus précisément, le processus autorégressif suppose que chaque point peut être prédit à partir d'un ensemble de points précédents pondérés par des coefficients, plus une erreur aléatoire. Le processus de moyenne mobile, quant à lui, suppose que chaque point est influencé par les erreurs des points précédents ainsi que par sa propre erreur. Enfin, le processus d'intégration suppose que chaque point présente une différence constante par rapport au point précédent. Les équations des processus s'écrivent :

- Autorégression AR(p) : $R_{i,j} = c + \phi_1 R_{i,j-1} + \phi_2 R_{i,j-2} + \cdots + \phi_p R_{i,j-p} + \epsilon_{i,j}$.
- Moyenne mobile MA(q) : $R_{i,j} = c + \theta_1 \epsilon_{i,j-1} + \theta_2 \epsilon_{i,j-2} + \cdots + \theta_q \epsilon_{i,j-q} + \epsilon_{i,j}$.
- Différenciation I(d) : $I = (1 - B)^d R_{i,j}$,

où $I(d)$ représente l'opérateur de différenciation d'ordre d, qui permet de rendre une série temporelle stationnaire en supprimant les tendances ou les motifs saisonniers. L'opérateur de retard, noté B, décale les observations d'un pas dans le temps.

L'équation du modèle ARIMA (p, d et q) combine les processus précédents pour prédire la valeur actuelle de la série en fonction de ses (p) valeurs passées, de ses (d) erreurs passées et de sa différence (q) avec les valeurs passées. Mathématiquement, ce modèle s'écrit comme suit :

$$\begin{aligned} R_{i,j} = c + \phi_1 R_{i,j-1} + \phi_2 R_{i,j-2} + \cdots + \phi_p R_{i,j-p} + \epsilon_{i,j} \\ -\theta_1 \epsilon_{i,j-1} - \theta_2 \epsilon_{i,j-2} - \cdots - \theta_q \epsilon_{i,j-q}, \end{aligned} \tag{3.2}$$

où l'équation ARIMA décrit $R_{i,j}$ en fonction de ses valeurs passées, des coefficients autorégressifs $\phi_1, \phi_2, \ldots, \phi_p$, d'une constante c et d'un bruit blanc $\epsilon_{i,j}$ avec une moyenne nulle et une variance constante. Pour une étude plus approfondie du modèle ARIMA et de ses applications, vous pouvez consulter Shumway (2017) et Courcot (2023).

3.4 Bayesian Probabilistic Matrix Factorization (BPMF)

Le Bayesian Probabilistic Matrix Factorization (BPMF) est un modèle de factorisation matricielle qui suggère d'adopter une approche bayésienne dans l'estimation des paramètres et hyperparamètres en intégrant des distributions *a priori* et des méthodes de Monte-Carlo par chaines de Markov (MCMC). Nous considérons que, pour une matrice de données $\boldsymbol{R} \in \mathbb{R}^{M \times N}$, nous utilisons la décomposition $\boldsymbol{R} \approx \boldsymbol{U}^\top \boldsymbol{V}$, où $\boldsymbol{U} \in \mathbb{R}^{K \times M}$ et $\boldsymbol{V} \in \mathbb{R}^{K \times N}$. Ici, K représente un rang de décomposition beaucoup plus petit que M et N ($K \ll M, N$). La structure distributionnelle de ce modèle se présente comme suit :

$$
\begin{aligned}
R_{i,j} &\sim \mathcal{N}(\boldsymbol{U}_{:i}^\top \boldsymbol{V}_{:j}, \alpha^{-1}), \\
\boldsymbol{U}_{:i} | \boldsymbol{\mu_U}, \boldsymbol{\Lambda_U} &\sim \mathcal{N}_M(\boldsymbol{\mu_U}, {\boldsymbol{\Lambda_U}}^{-1}), \\
\boldsymbol{V}_{:j} | \boldsymbol{\mu_V}, \boldsymbol{\Lambda_V} &\sim \mathcal{N}_N(\boldsymbol{\mu_V}, {\boldsymbol{\Lambda_V}}^{-1}), \\
\boldsymbol{\mu_U} | \boldsymbol{\Lambda_U} &\sim \mathcal{N}_M(\boldsymbol{\mu_0}, (\beta_0 \boldsymbol{\Lambda_U})^{-1}), \\
\boldsymbol{\mu_V} | \boldsymbol{\Lambda_V} &\sim \mathcal{N}_N(\boldsymbol{\mu_0}, (\beta_0 \boldsymbol{\Lambda_V})^{-1}), \\
\boldsymbol{\Lambda_U} &\sim \mathcal{W}(\boldsymbol{W_0}, v_0), \\
\boldsymbol{\Lambda_V} &\sim \mathcal{W}(\boldsymbol{W_0}, v_0),
\end{aligned}
\tag{3.3}
$$

où $\mathcal{N}(\mu, \sigma^2)$ est la distribution d'une loi normale avec une moyenne μ et une variance σ^2 et où $\mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ est la distribution d'une loi normale multivariée de dimension k avec un vecteur moyen $\boldsymbol{\mu}$ et une matrice de variance-covariance $\boldsymbol{\Sigma}$. La notation $\mathcal{W}(\boldsymbol{W_0}, v_0)$ représente la distribution d'une loi de Wishart avec v_0 degrés de liberté. $\boldsymbol{W_0}$ est une matrice d'échelle de dimension $K \times K$. En considérant les observations manquantes de $\boldsymbol{R}$, les fonctions de densité sont données comme suit :

$$
\begin{aligned}
p(\boldsymbol{R}|\boldsymbol{U},\boldsymbol{V},\alpha) &= \prod_{i=1}^{M}\prod_{j=1}^{N}\left[\mathcal{N}(R_{i,j}|\boldsymbol{U}_{:i}^{\top}\boldsymbol{V}_{:j},\alpha^{-1})\right]^{I_{i,j}},\\
p(\boldsymbol{U}|\boldsymbol{\mu_U},\boldsymbol{\Lambda_U}) &= \prod_{i=1}^{M}\mathcal{N}_M(\boldsymbol{U}_{:i}|\boldsymbol{\mu_U},{\boldsymbol{\Lambda_U}}^{-1}),\\
p(\boldsymbol{V}|\boldsymbol{\mu_V},\boldsymbol{\Lambda_V}) &= \prod_{j=1}^{N}\mathcal{N}_N(\boldsymbol{V}_{:j}|\boldsymbol{\mu_V},{\boldsymbol{\Lambda_V}}^{-1}),
\end{aligned}
\tag{3.4}
$$

où $I_{i,j}$ est la variable indicatrice qui prend la valeur 1 lorsque l'élément $R_{i,j}$ est observé et zéro dans le cas contraire.

Sachant les vecteurs latents $\boldsymbol{U}_{:i}^{\top}$ et $\boldsymbol{V}_{:j}$, le modèle BPMF suppose que chaque élément $R_{i,j}$ suit une loi normale univariée centrée sur le produit des deux vecteurs associés $\boldsymbol{U}_{:i}^{\top}$ et $\boldsymbol{V}_{:j}$ avec une certaine précision α. La fonction de vraisemblance est alors égale au produit de chacune des fonctions de densité individuelles des observations $R_{i,j}$.

Comme les matrices latentes $\boldsymbol{U}$ et $\boldsymbol{V}$ ne sont pas connues, le modèle BPMF leur attribue des distributions *a priori*. Plus précisément, le BPMF suppose que chaque colonne de la matrice $\boldsymbol{U}$ suit une loi normale multivariée centrée sur le vecteur $\boldsymbol{\mu_U}$ avec une certaine matrice de précision $\boldsymbol{\Lambda_U}$. De même, chaque colonne de la matrice $\boldsymbol{V}$ est supposée suivre une loi normale multivariée centrée sur le vecteur $\boldsymbol{\mu_V}$ avec une certaine matrice de précision $\boldsymbol{\Lambda_V}$.

Cependant, les hyperparamètres $\boldsymbol{\Theta_U} = \{\boldsymbol{\mu_U},\boldsymbol{\Lambda_U}\}$ et $\boldsymbol{\Theta_V} = \{\boldsymbol{\mu_V},\boldsymbol{\Lambda_V}\}$ sont également inconnus et doivent donc être modélisés. Pour ce faire, le modèle BPMF place une loi *a priori* normale-Wishart. Ainsi, les moyennes $\boldsymbol{\mu_U}$ et $\boldsymbol{\mu_V}$ sont supposées suivre une loi normale multivariée centrée sur $\boldsymbol{\mu_0}$ avec une variance définie par le produit d'un facteur de proportion β_0 multiplié par les matrices de précision $\boldsymbol{\Lambda_U}$ et $\boldsymbol{\Lambda_V}$, respectivement. Les matrices de précision $\boldsymbol{\Lambda_U}$ et $\boldsymbol{\Lambda_V}$ (l'inverse de la variance-covariance) sont elles-mêmes modélisées selon une distribution de Wishart avec les paramètres v_0 et $\boldsymbol{W_0}$. La distribution de Wishart est un *a priori* conjugué de la matrice de covariance inverse d'un vecteur aléatoire normal multivarié.

Le modèle BPMF spécifie des lois *a priori* Gaussienne-Wishart pour les hyperparamètres des facteurs spatiaux et pour les facteurs temporels, $\boldsymbol{\Theta_U}$ et $\boldsymbol{\Theta_V}$, respectivement. Le BPMF est dit Wishart-normal, car la distribution *a posteriori* sera une combinaison de la distribution normale et de la distribution de Wishart, comme présenté ci-dessous :

$$p(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\Theta_U}, \boldsymbol{\Theta_V} | \boldsymbol{R}) = \frac{p(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{V}, \alpha) p(\boldsymbol{U}|\boldsymbol{\Theta_U}) p(\boldsymbol{V}|\boldsymbol{\Theta_V}) p(\boldsymbol{\Theta_U}|\Theta_0) p(\boldsymbol{\Theta_V}|\Theta_0)}{\iint p(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{V}, \alpha) p(\boldsymbol{U}|\boldsymbol{\Theta_U}) p(\boldsymbol{V}|\boldsymbol{\Theta_V}) p(\boldsymbol{\Theta_U}|\Theta_0) p(\boldsymbol{\Theta_V}|\Theta_0)\, d\{\boldsymbol{U}, \boldsymbol{V}\}\, d\{\boldsymbol{\Theta_U}, \boldsymbol{\Theta_V}\}}, \quad (3.5)$$

où les hyperparamètres $\boldsymbol{\Theta_U} = \{\boldsymbol{\mu_U}, \boldsymbol{\Lambda_U}\}$ et $\boldsymbol{\Theta_V} = \{\boldsymbol{\mu_V}, \boldsymbol{\Lambda_V}\}$ représentent les paramètres des lois normales multivariées pour les lignes de $\boldsymbol{U}$ et les colonnes de $\boldsymbol{V}$, respectivement. $\Theta_0 = \{\boldsymbol{\mu_0}, v_0, \boldsymbol{W}_0\}$ représente les hyperparamètres connus du modèle BPMF. L'imputation d'une valeur prédite et/ou manquante $R^*_{i,j}$ est obtenue par la distribution marginale suivante :

$$p(R^*_{i,j}|\boldsymbol{R}, \Theta_0) = \iint p(R^*_{i,j}|\boldsymbol{R}, \boldsymbol{U}_{:i}, \boldsymbol{V}_{:j}) p(\boldsymbol{U}, \boldsymbol{V}|\boldsymbol{R}, \boldsymbol{\Theta_U}, \boldsymbol{\Theta_V}) p(\boldsymbol{\Theta_U}, \boldsymbol{\Theta_V}|\Theta_0)\, d\{\boldsymbol{U}, \boldsymbol{V}\}\, d\{\boldsymbol{\Theta_U}, \boldsymbol{\Theta_V}\}. \quad (3.6)$$

La difficulté du calcul de l'équation 3.6 est amplifiée, non seulement par le nombre de paramètres du modèle BPMF, mais aussi par l'implication la distribution *a posteriori* des paramètres $\boldsymbol{U}$ et $\boldsymbol{V}$. Cela nécessite de calculer le dénominateur du théorème de Bayes, ce qui rend le calcul analytique impossible. Ce calcul est donc estimé par la méthode MCMC, qui utilise l'approximation de Monte-Carlo, donnée comme suit :

$$p(R^*_{i,j}|\boldsymbol{R}, \Theta_0) \approx \frac{1}{Z} \sum_{z=1}^{Z} p(R^*_{i,j}|\boldsymbol{U}^z_{:i}, \boldsymbol{V}^z_{:j}), \quad (3.7)$$

où Z représente le nombre d'itérations.

Les échantillons $\{\boldsymbol{U}_{:i}^{z}, \boldsymbol{V}_{:j}^{z}\}$ sont générés par une chaine de Markov dont la distribution stationnaire est la distribution *a posteriori* à travers les paramètres et hyperparamètres du modèle BPMF, $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\Theta_U}$ et $\boldsymbol{\Theta_V}$.

Du à l'utilisation des distributions conjuguées pour les paramètres et hyperparamètres de ce modèle, il est facile d'échantillonner les distributions conditionnelles. L'algorithme de Gibbs est alors utilisé pour estimer les paramètres d'intérêts du modèle. Nous ne donnons pas les détails ici, mais vous pouvez consulter la section 3.3 de l'article original de Salakhutdinov (2008) pour une compréhension approfondie et des détails supplémentaires sur ce modèle.

3.5 Temporal Regularized Matrix Factorization (TRMF)

Le Temporal Regularized Matrix Factorization (TRMF) est un modèle qui permet une extension de la factorisation matricielle en intégrant les dépendances temporelles dans la matrice de facteurs latents $\boldsymbol{V}$. Nous rappelons que pour une matrice de données $\boldsymbol{R} \in \mathbb{R}^{M \times N}$, nous avons $\boldsymbol{R} \approx \boldsymbol{U}^{\top}\boldsymbol{V}$, où $\boldsymbol{U} \in \mathbb{R}^{K \times M}$ et $\boldsymbol{V} \in \mathbb{R}^{K \times N}$. K représente le rang de décomposition et K est beaucoup plus petit que M et N ($K \ll M, N$). En effet, pour résoudre le problème $\boldsymbol{R} \approx \boldsymbol{U}^{\top}\boldsymbol{V}$, nous pouvons résoudre l'équation mathématique suivante :

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{(i,j)\in\Omega} (R_{i,j} - \boldsymbol{U}_{:i}^{\top}\boldsymbol{V}_{:j})^2 + \lambda_{\boldsymbol{U}}\mathcal{R}_{\boldsymbol{U}}(\boldsymbol{U}) + \lambda_{\boldsymbol{V}}\mathcal{R}_{\boldsymbol{V}}(\boldsymbol{V}), \tag{3.8}$$

où Ω représente l'ensemble des indices des données observées. $\boldsymbol{U}_{:i}$ est la i^e colonne de $\boldsymbol{U}$ et représente les caractéristiques latentes associées à la ligne i de $\boldsymbol{R}$. De même, $\boldsymbol{V}_{:j}$ est la j^e colonne de $\boldsymbol{V}$ et représente les caractéristiques latentes associées à la colonne j de $\boldsymbol{R}$. Les paramètres de régularisation $\lambda_{\boldsymbol{U}}$ et $\lambda_{\boldsymbol{V}}$ sont associés aux termes de régularisation $\mathcal{R}_U(\boldsymbol{U})$ et $\mathcal{R}_V(\boldsymbol{V})$, respectivement. Ces termes permettent d'éviter

le surapprentissage et/ou d'encourager certaines structures spécifiques. Cependant, le choix courant du régularisateur $\|\boldsymbol{V}\|_F$ n'est pas adapté aux applications de séries temporelles, car il ne considère pas les relations temporelles entre les observations. Le TRMF propose d'utiliser un modèle de séries temporelles autorégressif pour modéliser explicitement les dépendances temporelles entre les colonnes de $\boldsymbol{V}$ sous forme d'une régularisation. L'idée derrière le choix de la régularisation de $\boldsymbol{V}$ est de considérer le modèle sous-jacent suivant :

$$\boldsymbol{V}_{:j} = \sum_{l \in \mathcal{L}} \boldsymbol{W}^{(l)} \boldsymbol{V}_{:j-l} + \boldsymbol{\epsilon}_{:j}, \tag{3.9}$$

où $\boldsymbol{W}^{(l)} \in \mathbb{R}^{K \times K}, l \in \mathcal{L}$ représente une matrice de poids associée à un indice de décalage l permettant de décrire la relation entre la colonne $\boldsymbol{V}_{:j}$ et la colonne précédente $\boldsymbol{V}_{:j-l}$. $\mathcal{L}$ est un ensemble, contenant les indices de décalage l, dénotant une dépendance entre le j^e et le l^e point temporel. $\boldsymbol{\epsilon}_{:j} \sim \mathcal{N}_N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_K)$ est un vecteur de bruit gaussien et $\boldsymbol{I}_K$ est une matrice d'identité de dimensions $K \times K$. En d'autres termes, chaque colonne $\boldsymbol{V}_{:j}$ est modélisée par la somme du produit entre l'ensemble de colonnes précédentes, identifiées par un ensemble de décalage $\mathcal{L}$ et par leurs poids respectifs. Le modèle TRMF est ainsi un modèle de séries temporelles qui est décrit comme suit :

$$\underset{\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Theta}}{min} \sum_{(i,j) \in \Omega} (R_{i,j} - \boldsymbol{U}_{:i}^\top \boldsymbol{V}_{:j})^2 + \lambda_{\boldsymbol{U}} \mathcal{R}_{\boldsymbol{U}}(\boldsymbol{U}) + \lambda_{\boldsymbol{V}} \mathcal{T}_{AR}(\boldsymbol{V}|\boldsymbol{\Theta}) + \lambda_{\boldsymbol{\Theta}} \mathcal{R}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}), \tag{3.10}$$

$$\begin{aligned} \mathcal{R}_{\boldsymbol{U}}(\boldsymbol{U}) &= \|\boldsymbol{U}\|_F^2, \\ \mathcal{T}_{AR}(\boldsymbol{V}|\mathcal{L}, \boldsymbol{W}, \eta) &= \frac{1}{2} \sum_{j=m}^{N} \left\| \boldsymbol{V}_{:j} - \sum_{l \in \mathcal{L}} \boldsymbol{W}^{(l)} V_{:j-l} \right\|^2 + \frac{\eta}{2} \sum_j \|\boldsymbol{V}_{:j}\|^2, \\ \mathcal{R}_{\Theta}(\boldsymbol{\Theta}) &= \|\boldsymbol{\Theta}\|_F^2, \end{aligned} \tag{3.11}$$

où m=1+max($\mathcal{L}$) et où $\eta > 0$ pour garantir une forte convexité de 3.11. $\boldsymbol{W} = \{\boldsymbol{W}^{(l)} \in \mathbb{R}^{K \times K} : l \in \mathcal{L}\}$ représente l'ensemble des matrices de poids en termes de

régularisation autorégressive. $\boldsymbol{\Theta}$ représente l'ensemble des paramètres à estimer dans le modèle TRMF.

La solution de l'équation 3.10 est obtenue par une procédure de minimisation alternée qui consiste à fixer tour à tour les variables $\boldsymbol{\Theta}$, $\boldsymbol{U}$ et $\boldsymbol{V}$ de manière à les optimiser séparément, sous le régularisateur temporel autorégressif $\mathcal{T}_{AR}(\boldsymbol{V}|\boldsymbol{\Theta})$. Ce régularisateur est paramétré par $\mathcal{L}$, $\boldsymbol{W}$ et η. Dans le but de réduire le nombre de paramètres à estimer dans chaque $\boldsymbol{W}^{(l)} \in \mathbb{R}^{K\times K}$, soit $|\mathcal{L}|K^2$ paramètres qui peuvent mener à un sur-apprentissage, le modèle TRMF propose une matrice diagonale $\boldsymbol{W}^{K\times\mathcal{L}}$, nécessitant d'estimer $|\mathcal{L}|K$ paramètres pour chaque matrice $\boldsymbol{W}^{(l)}$. Cette contrainte implique que les dynamiques temporelles des facteurs latents $\boldsymbol{V}_{:j}$ sont indépendantes les unes des autres. Autrement dit, cela signifie que chaque facteur latent influence uniquement son propre comportement dans le temps, sans interaction avec les autres facteurs latents. Plus précisément, chaque colonne de $\boldsymbol{V}$ à un instant j est prédite uniquement en fonction des valeurs passées de cette même colonne à des instants $j-l$, $l \in \mathcal{L}$, sans tenir compte des valeurs passées des autres colonnes. Ainsi, le modèle TRMF est simplifié et estimé en utilisant de moindres carrés alternés régularisés par graphe (GRALS). Cela permet de trouver les meilleures valeurs pour les différents paramètres pour ce modèle. Vous pouvez consulter l'article original de Yu (2016) pour une compréhension approfondie et pour plus de détails sur ce modèle.

3.6 Bayesian Temporal Matrix Factorisation (BTMF)

Le Bayesian Temporal Matrix Factorisation (BTMF) est un modèle qui s'inspire du Temporal Regularized Matrix Factorization (TRMF). Il tient compte des relations temporelles entre les différentes colonnes de la matrice de facteurs temporels. Nous rappelons également que $\boldsymbol{R}$, $\boldsymbol{U}$ et $\boldsymbol{V}$ sont définies de la même manière que

dans les modèles précédents (Bayesian Probabilistic Matrix Factorization [BPMF] et TRMF).

En vue d'adresser les contraintes du TRMF relatives à l'indépendance des facteurs temporels latents, le BTMF supprime la contrainte diagonale sur $\boldsymbol{W}^{(l)}, l \in \mathcal{L}$ et utilise un modèle autorégressif d'ordre d (Hyndman (2018)) avec un ensemble de décalage $\mathcal{L} = \{h_1, \ldots, h_d\}$ pour modéliser les colonnes de la matrice $\boldsymbol{V}$. De plus, le BTMF, tout comme le BPMF, est un modèle probabiliste bayésien dont la structure hiérarchique est présentée ci-dessous :

$$
\begin{aligned}
R_{i,j} &\sim \mathcal{N}(\boldsymbol{U}_{:i}^{\top}\boldsymbol{V}_{:j}, \tau_i^{-1}), \quad (i,j) \in \Omega, \\
\boldsymbol{U}_{:i} &\sim \mathcal{N}_M(\boldsymbol{\mu_U}, {\boldsymbol{\Lambda_U}}^{-1}), \\
\boldsymbol{\mu_U}|\boldsymbol{\Lambda_U} &\sim \mathcal{N}_M(\boldsymbol{\mu_0}, (\beta_0\boldsymbol{\Lambda_U})^{-1}), \\
\boldsymbol{\Lambda_U} &\sim \mathcal{W}(\boldsymbol{W_0}, v_0), \\
\boldsymbol{V}_{:j} &\sim \begin{cases} \mathcal{N}_N(\boldsymbol{0}, \boldsymbol{I}_K), & \text{si } j \in \{1,2,...,d\}, \\ \mathcal{N}_N(\sum_{k=1}^{d} \boldsymbol{A}_k \boldsymbol{V}_{:j-h_k}, \boldsymbol{\Sigma}), & \text{sinon.} \end{cases} \\
[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d]^{\top} &\sim \mathcal{MN}_{(Kd)\times K}(M_0, \boldsymbol{\Psi}_0, \boldsymbol{\Sigma}), \\
\boldsymbol{\Sigma} &\sim \mathcal{IW}(\boldsymbol{S}_0, v_0), \\
\tau_i &\sim Gamma(\alpha, \beta).
\end{aligned}
\tag{3.12}
$$

Le modèle BTMF considère que chaque observation est indépendante et identiquement distribuée selon une loi gaussienne, avec une précision τ_i spécifique pour chaque série temporelle $\boldsymbol{R}_{i:}$. La distribution *a priori* du vecteur $\boldsymbol{U}_{:i}$ est modélisée par une distribution gaussienne multivariée. Le modèle BTMF suppose une distribution normale des facteurs latents spatiaux. En effet, comme les paramètres $\boldsymbol{\mu_U}$ et ${\boldsymbol{\Lambda_U}}^{-1}$ sont inconnus, le modèle BTMF utilise une distribution *a priori* conjuguée Gaussienne-Wishart. En ce qui concerne la modélisation de la structure temporelle des facteurs latents $\boldsymbol{V}_{:j}$, ce modèle suppose une distribution normale multivariée avec un vecteur moyen autorégressif (VAR) de rang d, où $\boldsymbol{A}_k$ représente une matrice de coefficients $K \times K$ associée à un indice de décalage h_k. En pratique, ces matrices

sont supposées diagonales et le modèle BTMF utilise une distribution *a priori* conjuguée Matrice-Normale Inverse-Wishart pour ses coefficients. Ainsi, les dépendances temporelles sont modélisées à travers le vecteur autorégressif, où les coefficients des matrices $\boldsymbol{A}_k$ capturent la dynamique de la dépendance entre les différentes variables observées. Compte tenu de la complexité de la structure du BTMF, le modèle utilise la technique d'échantillonnage de Gibbs, qui fait appel à la méthode de Monte-Carlo par chaines de Markov (MCMC) pour estimer les paramètres. Grâce à l'utilisation de distributions *a priori* conjuguées, il est facile d'écrire toutes les distributions conditionnelles de tous les paramètres et hyperparamètres de manière analytique et de les estimer. La section 4.2 de l'article de Chen (2021) fournit plus de détails sur la manière dont l'inférence statistique est réalisée, y compris les équations et les méthodes utilisées pour estimer les paramètres.

3.7 Bayesian Temporal Tensor Factorization (BTTF)

Le modèle Bayesian Temporal Tensor Factorization (BTTF) est une extension du modèle Bayesian Temporal Matrix Factorisation (BTMF), conçu pour modéliser des séries temporelles tensorielles multidimensionnelles avec un ordre supérieur à 2 (ordre > 2).

Le BTTF utilise l'hypothèse selon laquelle les données d'intérêt peuvent être représentées sous forme de tenseur avec un ordre supérieur à 2. Un tenseur d'ordre supérieur à 2 a plus de deux dimensions, ce qui généralise le cas matriciel abordé précédemment. Dans un contexte pratique, nous considérons l'ensemble de données *TLC Trip Record Data*[1] qui enregistre des informations relatives à divers types de services de taxi. Cet ensemble de données compile des observations telles que les lieux de prise en charge et de dépose, ainsi que les heures de début de chaque trajet.

1. Source : `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`

Dans ce cas, ces données peuvent être représentées comme un tenseur d'ordre 3, que nous désignons par $\boldsymbol{R} \in \mathbb{R}^{M \times N \times T}$, où la première dimension, M, représente le lieu de prise en charge, la deuxième dimension, N, indique le lieu de dépose et la troisième dimension, T, correspond à l'heure de début du trajet. Concrètement, ce modèle utilise la décomposition CANDECOMP/PARAFAC, ou *décomposition CP*, qui permet de décomposer le tenseur $\boldsymbol{R}$ en une combinaison linéaire de matrices de facteurs (Pereira da Silva (2016)). Il est à noter que la factorisation matricielle peut être vue comme un cas particulier de la *décomposition CP* lorsque l'ordre du tenseur est égale à 2 (ordre = 2). Mathématiquement, la *décomposition CP* d'un tenseur $\boldsymbol{R}$ de rang de décomposition K est définie comme suit :

$$\boldsymbol{R} \approx \sum_{k=1}^{K} \boldsymbol{U}_{:k} \circ \boldsymbol{X}_{:k} \circ \boldsymbol{V}_{:k},$$

où $\circ$ dénote le produit extérieur entre les vecteurs, $\boldsymbol{U}_{:k} \in \mathbb{R}^{M}$, et où $\boldsymbol{X}_{:k} \in \mathbb{R}^{N}$ et $\boldsymbol{V}_{:k} \in \mathbb{R}^{T}$ représentent les colonnes k^e des matrices de facteurs $\boldsymbol{U} \in \mathbb{R}^{M \times K}$, $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ et $\boldsymbol{V} \in \mathbb{R}^{T \times K}$, respectivement. L'objectif de *la décomposition CP* est de trouver les matrices de facteurs qui minimisent l'erreur de reconstruction du tenseur original. *La décomposition CP* permet d'étendre le modèle BTMF vers des tenseurs

dont la structure hiérarchique est donnée comme suit :

$$
\begin{aligned}
R_{i,j,t} &\sim \mathcal{N}\left(\sum_{k=1}^{K} \boldsymbol{U}_{i,k}\boldsymbol{X}_{j,k}\boldsymbol{V}_{t,k}, \tau_{i,j}^{-1}\right), (i,j,t) \in \Omega \\
\boldsymbol{U}_{:i} &\sim \mathcal{N}_M(\boldsymbol{\mu_U}, {\boldsymbol{\Lambda_U}}^{-1}), \\
\boldsymbol{X}_{:j} &\sim \mathcal{N}_N(\boldsymbol{\mu_X}, {\boldsymbol{\Lambda_X}}^{-1}), \\
\boldsymbol{V}_{:j} &\sim \begin{cases} \mathcal{N}_T(\boldsymbol{0}, \boldsymbol{I}_K), & \text{si } t \in \{1,2,...,d\}, \\ \mathcal{N}_T(\sum_{k=1}^{d} \boldsymbol{A}_k \boldsymbol{V}_{:t-h_k}, \boldsymbol{\Sigma}), & \text{sinon.} \end{cases} \\
[\boldsymbol{A}_1, \ldots, \boldsymbol{A}_d]^\top &\sim \mathcal{MN}_{(Kd)\times K}(M_0, \boldsymbol{\Psi}_0, \boldsymbol{\Sigma}), \\
\boldsymbol{\Sigma} &\sim \mathcal{IW}(\boldsymbol{S}_0, v_0), \\
\tau_{ij} &\sim Gamma(\alpha, \beta).
\end{aligned}
\tag{3.13}
$$

Le modèle BTTF suppose que chaque élément suit une distribution gaussienne centrée sur le produit des trois vecteurs de composition avec un bruit pour chaque série temporelle. Cependant, lorsque les dimensions du tenseur sont grandes, il devient difficile de calculer le bruit pour chaque série temporelle. En pratique, ce modèle assume un bruit isotropique tel que $\tau_{i,j} = \tau$. Pour les facteurs spatiaux, le modèle BTTF utilise les mêmes distributions *a priori* que le BTMF, avec les paramètres et hyperparamètres modélisés par une distribution *a priori* conjuguée Gaussienne-Wishart. En ce qui concerne la modélisation de la structure temporelle des facteurs latents $\boldsymbol{V}_{:j}$, ce modèle suppose également la même modélisation que le modèle BTMF, soit une distribution normale multivariée avec un vecteur moyen autorégressif (VAR) de rang d, où $\boldsymbol{A}_k$ représente une matrice de coefficients $K \times K$ associée à un indice de décalage h_k. En pratique, ces matrices sont supposées diagonales et le modèle BTTF utilise une distribution *a priori* conjuguée matrice-normale inverse-Wishart pour ses coefficients. Ainsi, les dépendances temporelles sont modélisées à travers le vecteur autorégressif, où les coefficients des matrices $\boldsymbol{A}_k$ capturent la dynamique de la dépendance entre les différentes variables observées. Toutefois,

dans ce modèle, les matrices $\boldsymbol{U}$ et $\boldsymbol{X}$ sont considérées comme des matrices de facteurs spatiaux, tandis que la matrice $\boldsymbol{V}$ décrit la structure des dépendances temporelles. L'algorithme de Gibbs est utilisé pour adapter le modèle BTTF. Vous pouvez consulter l'article original de Chen (2021) pour une compréhension approfondie et des détails supplémentaires sur ce modèle.

3.8 Kernelized Probabilistic Matrix Factorization (KPMF)

Le Kernelized Probabilistic Matrix Factorization (KPMF) est un modèle qui intègre de l'information complémentaire dans le processus de factorisation matricielle. De fait, ce modèle cherche à capturer de manière explicite les structures de covariance non linéaires sous-jacentes à la matrice $\boldsymbol{R}$ en couvrant toutes les lignes et les colonnes de la matrice. Pour ce faire, ce modèle suppose que la forme fonctionnelle des matrices de variance-covariance, pour **les lignes** et **les colonnes** de $\boldsymbol{R}$, notées respectivement $\boldsymbol{S}^u \in \mathbb{R}^{M\times M}$ et $\boldsymbol{S}^v \in \mathbb{R}^{N\times N}$, est supposée connue grâce à la mise en place d'un noyau pour modéliser cette covariance. La structure hiérarchique du modèle est la suivante :

$$
\begin{aligned}
&R_{i,j} \sim \mathcal{N}(\boldsymbol{U}_{:i}^{\top} V_{:j}, \tau_i^{-1}), \\
&\boldsymbol{U}_{:,k} \sim \mathcal{GP}(\boldsymbol{0}, \boldsymbol{S}^u), \quad k \in \{1, ..., K\}, \\
&\boldsymbol{V}_{:,k} \sim \mathcal{GP}(\boldsymbol{0}, \boldsymbol{S}^v), \quad k \in \{1, ..., K\},
\end{aligned}
\tag{3.14}
$$

où $\mathcal{GP}$ désigne un processus gaussien déterminé par une fonction moyenne et une fonction de covariance. Les matrices de covariance $\boldsymbol{S}^u$ et $\boldsymbol{S}^v$ sont calculées en utilisant deux noyaux $s^u(\cdot, \cdot; \boldsymbol{\theta})$ et $s^v(\cdot, \cdot; \boldsymbol{\vartheta})$ avec des hyperparamètres $\boldsymbol{\theta}$ et $\boldsymbol{\vartheta}$, respectivement. Les noyaux s^u et s^v sont définis comme des fonctions de la distance entre les points associés aux facteurs latents.

Il existe plusieurs types de noyaux pour construire des matrices de variance-covariance adaptées. Dans notre contexte spécifique, nous disposons d'informations complémentaires qui se manifestent sous la forme d'une matrice d'adjacence présente dans notre ensemble de données. Nous présentons le noyau Regularized Laplacian Kernel, introduit par Smola (2003), où nous considérons les données de la matrice d'adjacence comme un graphe sans direction et sans poids, noté $\mathcal{G}$, avec des nœuds et des arêtes représentant les capteurs et leurs connexions. Les éléments de la matrice d'adjacence $\boldsymbol{A}^{M\times M}$ sont déterminés par $A_{i,j} = 1$, s'il existe une arête entre le i^e capteur et le j^e capteur, et $A_{i,j} = 0$ sinon. Une arête peut être considérée si les mesures de la vitesse des deux capteurs sont corrélées au-delà d'un certain seuil. Le noyau laplacien régularisé est exprimé sous la forme suivante :

$$s_{RL}(\boldsymbol{L};\theta) = (\boldsymbol{I} + \theta\boldsymbol{L})^{-1}, \tag{3.15}$$

où $\theta > 0$ est le paramètre de régularisation et où $\boldsymbol{I}$ est la matrice d'identité. La matrice Laplace $\boldsymbol{L}$ est définie comme étant la différence entre la matrice d'adjacence $\boldsymbol{A}$ et la matrice de degré $\boldsymbol{D}^{M\times M}$, où $\boldsymbol{D}$ est une matrice diagonale dont les entrées $d_i = \sum_{j=1}^{N} A_{ij}, i \in \{1, ..., N\}$ représentent les degrés des nœuds du graphe. Le noyau laplacien régularisé capture la régularité de la structure du graphe tout en tenant compte de la pénalisation de la variation entre les nœuds adjacents.

Ainsi, le modèle KPMF modélise les matrices $\boldsymbol{U}$ et $\boldsymbol{V}$ colonne par colonne. Pour chaque colonne, le modèle suppose une distribution multivariée *a priori*. Cette distribution est modélisée par un processus gaussien avec une fonction de moyenne égale au vecteur $\mathbf{0}$ et une fonction de noyau qui impose une structure aux matrices de variance-covariance $\boldsymbol{S}^u$ et $\boldsymbol{S}^v$, respectivement. En effet, celles-ci sont créées par un noyau qui prend en considération la distance entre le poids des points des lignes et/ou des colonnes. L'utilisation d'un noyau oblige les facteurs latents à capturer les covariances sous-jacentes entre les lignes et les colonnes simultanément. Ainsi, si deux lignes de la matrice $(\boldsymbol{R}_{i:}, \boldsymbol{R}_{i':} : i \neq i')$ sont interdépendantes, les facteurs latents

correspondants, partageront la même interdépendance. Il est à noter que, lorsque les matrices $\boldsymbol{S}^u$ et $\boldsymbol{S}^v$ sont diagonales, les lignes et/ou les colonnes dans les matrices $\boldsymbol{U}$ et $\boldsymbol{V}$, respectivement, sont indépendantes. Dans ce cas, le modèle KPMF se réduit au modèle Probabilistic Matrix Factorization (PMF) (Salakhutdinov (2010)). Cependant, si seulement une des matrices de variance-covariance est diagonale, la matrice latente correspondante peut être marginalisée, conduisant à un problème d'inférence Maximum a Posteriori (MAP) sur une matrice latente.

Le modèle KPMF utilise la descente de gradient pour minimiser la fonction de coût 3.16 et ajuster itérativement les valeurs des matrices $\boldsymbol{U}$ et $\boldsymbol{V}$ jusqu'à la convergence.

$$\begin{aligned} E = & \frac{1}{2\sigma^2} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(R_{i,j} - \boldsymbol{U}_{:i}^{\top} \boldsymbol{V}_{:j} \right)^2 \\ & + \frac{1}{2} \sum_{k=1}^{K} \boldsymbol{U}_{:k}^{\top} \boldsymbol{S}^{\boldsymbol{u}^{-1}} \boldsymbol{U}_{:k} \\ & + \frac{1}{2} \sum_{k=1}^{K} \boldsymbol{V}_{:k}^{\top} \boldsymbol{S}^{\boldsymbol{v}^{-1}} \boldsymbol{V}_{:k}. \end{aligned} \tag{3.16}$$

Cette descente de gradient permet au modèle KPMF de trouver les valeurs optimales des matrices $\boldsymbol{U}$ et $\boldsymbol{V}$, qui sont ensuite utilisés pour imputer les valeurs manquantes dans la matrice de données. Il est important de noter que $\boldsymbol{S}^u$ et $\boldsymbol{S}^v$ n'ont besoin d'être calculés qu'une seule fois, lors de l'initialisation. Vous pouvez consulter l'article original de Zhou (2012) pour une compréhension approfondie et des détails supplémentaires sur ce modèle.

3.9 Gaussian Process Probabilistic Matrix Factorization (GPMF)

Le Gaussian Process Probabilistic Matrix Factorization (GPMF) est un modèle inspiré par le Kernelized Probabilistic Matrix Factorization (KPMF). Contrairement

au KPMF, le GPMF peut apprendre automatiquement à la fois les hyperparamètres de la fonction noyau et les matrices de facteurs.

La structure hiérarchique de ce modèle est la suivante :

$$
\begin{aligned}
R_{i,j} &\sim \mathcal{N}(\boldsymbol{U}_{:i}^{\top} V_{:j}, \tau_i^{-1}), \\
\boldsymbol{U}_{:,k} &\sim \mathcal{GP}(\mathbf{0}, \boldsymbol{S}_k^u), \quad k \in \{1, ..., K\}, \\
\boldsymbol{V}_{:,k} &\sim \mathcal{GP}(\mathbf{0}, \boldsymbol{S}_k^v), \quad k \in \{1, ..., K\},
\end{aligned}
\tag{3.17}
$$

où $\mathcal{GP}$ désigne un processus gaussien déterminé par une fonction moyenne et une fonction de covariance. $\boldsymbol{S}_k^u$ et $\boldsymbol{S}_k^v$ sont calculées en utilisant deux noyaux, $s_k^u(\cdot, \cdot; \boldsymbol{\theta}_k)$ et $s_k^v(\cdot, \cdot; \boldsymbol{\vartheta}_k)$, avec des hyperparamètres $\boldsymbol{\theta}_k$ et $\boldsymbol{\vartheta}_k$, respectivement. Les noyaux s_k^u et s_k^v sont définis comme des fonctions de la distance entre les points associés aux facteurs latents. Le choix du noyau dépend de la nature des données et des relations temporelles que nous souhaitons modéliser. Chaque noyau est caractérisé par ses hyperparamètres. Pour la matrice de facteurs spatiaux, le modèle GPMF propose le noyau Regularized Laplacian Kernel : $s_{RL}(\boldsymbol{L}; \theta)$, où θ est l'hyperparamètre du noyau et où $\boldsymbol{L}$ est la matrice laplacienne normalisée. Il convient de noter que ce noyau a été initialement expliqué dans le contexte du modèle KPMF.

Concernant les facteurs temporels, le modèle GPMF propose deux choix pour le réglage des processus gaussiens ($\mathcal{GP}$) : exponentielle ou exponentielle au carré. Nous présentons les fonctions de noyau considérées dans ce modèle dans le tableau 3.1, où Δ représente la distance entre les points dans la dimension temporelle, où ϑ_k est l'hyperparamètre d'échelle qui contrôle la portée de la corrélation temporelle et où τ_k est l'hyperparamètre de variance qui détermine l'amplitude des variations temporelles.

Plus précisément, le modèle GPMF utilise le même noyau Regularized Laplacian Kernel avec l'hyperparamètre θ_k pour chaque colonne de $\boldsymbol{U}$. Pour les facteurs temporels de $\boldsymbol{V}$, ce modèle suppose utiliser K noyaux avec un hyperparamètre d'échelle ϑ_k et des hyperparamètres de variance respectifs $\{\tau_k : k = \{1, ..., K\}\}$. La structure

TABLE 3.1 – Fonctions de noyau *a priori* pour la dimension temporelle.

Fonction noyau	Paramètres	Formule
Exponentielle	$s_{Exp}(\Delta; \vartheta_k, \tau_k)$	$= \tau_k^2 \exp\left\{-\frac{\Delta}{\vartheta_k}\right\}$
Exponentielle au carré	$s_{SE}(\Delta; \vartheta_k, \tau_k)$	$= \tau_k^2 \exp\left\{-\frac{\Delta^2}{2\vartheta_k^2}\right\}$

descriptive de ces hyperparamètres est donnée comme suit :

$$\begin{aligned} \theta_k &\sim \mathcal{N}\left(\mu_\theta, \tau_\theta^{-1}\right), \quad k \in \{1, ..., K\}, \\ \vartheta_k &\sim \mathcal{N}\left(\mu_\vartheta, \tau_\vartheta^{-1}\right), \quad k \in \{1, ..., K\}, \\ \tau_i &\sim \text{Gamma}\left(a_0, b_0\right), \quad i \in \{1, ..., M\}. \end{aligned} \tag{3.18}$$

Les hyperparamètres du modèle $\boldsymbol{\Theta}_0 = \{\mu_\theta, \tau_\theta, \mu_\vartheta, \tau_\vartheta, a_0, b_0\}$ sont prédéfinis dans le processus d'échantillonnage et devraient incorporer les connaissances préalables sur une application spécifique. Néanmoins, la sélection de ces valeurs initiales a peu d'effet sur les résultats finaux lorsque le nombre d'itérations de mise à jour est suffisant, comme observé dans de nombreux schémas d'estimation bayésienne empirique.

Contrairement au KPMF, le BGMF intègre une distribution du bruit pour chaque ligne de la matrice de données $\boldsymbol{R}$, ce qui représente une modification importante par rapport au KPMF. Cette adaptation permet de mieux prendre en compte l'hétérogénéité des données réelles et d'obtenir des résultats plus précis. Le KPMF, quant à lui, utilise une approche de factorisation matricielle probabiliste qui impose une structure commune à toutes les colonnes de $\boldsymbol{U}$ et de $\boldsymbol{V}$, en utilisant une fonction de noyau unique avec deux paramètres distincts pour chaque matrice latente. En d'autres termes, le modèle suppose que toutes les colonnes de $\boldsymbol{U}$ et de $\boldsymbol{V}$ partagent la même structure de covariance. À l'inverse, le BGMF impose deux fonctions de noyau différentes pour modéliser les matrices de variance-covariance de $\boldsymbol{U}$ et de $\boldsymbol{V}$, où chaque colonne dispose de son propre paramètre. Cette approche

permet de capturer les relations spatiales et temporelles complexes qui peuvent exister entre les colonnes $\boldsymbol{U}$ et $\boldsymbol{V}$, en prenant en compte les différences de covariances entre les colonnes. Vous pouvez consulter l'article original de Lei (2022) pour une compréhension approfondie et des détails supplémentaires sur ce modèle.

3.10 Logiciels

Les codes sources des modèles sont mis à disposition du public et peuvent être consultés à l'adresse GitHub suivante : `https://github.com/xinychen/transdim`.

Chapitre 4

Présentation des données

Ce chapitre fournit une présentation détaillée des données utilisées et offre une première analyse visuelle à l'aide de diagrammes spatiaux et de séries temporelles. Nous explorons également les dépendances spatiales et temporelles, l'autocorrélation et l'analyse des composantes de la série chronologique. La méthodologie d'analyse des données spatiotemporelles utilisée dans ce chapitre s'inspire en partie du livre *Spatio-Temporal Statistics with R* de Wikle (2019).

4.1 Source de données

Les données disponibles ont été collectées par des détecteurs à boucle inductive déployés sur plusieurs autoroutes de la ville de Seattle, aux États-Unis, dont la I-5, la I-90, la I-405 et la SR-520. La boucle inductive est un dispositif utilisé pour détecter la présence de véhicules et collecter des données sur le trafic routier. Ce dispositif est composé d'un câble métallique en forme de boucle, enterré sous la chaussée. Lorsqu'un véhicule traverse cette boucle, ces roues perturbent le champ magnétique de la bobine. Cette perturbation est alors détectée par un équipement électronique. La figure 4.1 illustre le mécanisme des détecteurs à boucle inductive déployés sur une borne kilométrique. Les boucles inductives permettent de mesurer la vitesse des véhicules et de collecter des données en temps réel sur le trafic routier. Chaque

observation de vitesse attribuée à une borne kilométrique représente une moyenne calculée à partir de plusieurs détecteurs à boucle situés sur les voies principales dans la même direction du trafic. La mesure de vitesse est effectuée toutes les 5 minutes, ce qui correspond à 288 intervalles de temps par jour. Les données disponibles incluent 323 détecteurs à boucle, comme illustrées sur la figure 4.2. Cette dernière montre une carte routière de la région de Seattle, présentant quatre autoroutes principales et leurs itinéraires respectifs : I-5 (en rouge), I-405 (en jaune), SR-520 (en bleu) et I-90 (en violet). Les icônes bleues sur la carte indiquent les emplacements des capteurs. Les données couvrent une période d'une année complète, soit l'année 2015, pour un total de 33 953 760 observations. L'unité de mesure utilisée dans cette collecte de données est le mph (*miles per hour* [miles par heure]) et 1 mph équivaut approximativement à 1,60 km/h.

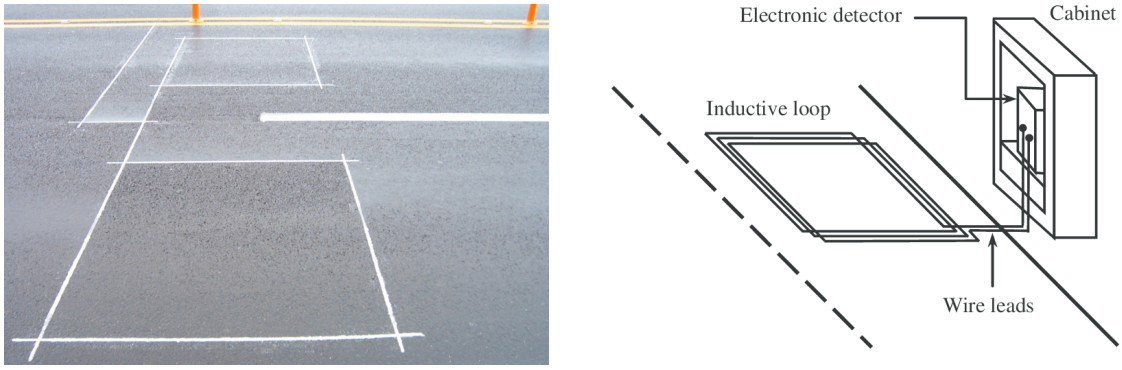

FIGURE 4.1 – Mécanisme des détecteurs à boucle inductive déployés sur une borne kilométrique [1] [2]

La figure 4.3 représente un aperçu du fichier de données. Les données de vitesse se présentent sous la forme d'une matrice. Chaque ligne correspond à un intervalle de temps et chaque colonne fait référence à l'identifiant d'un détecteur à boucle. Le nom de chaque entête de borne kilométrique contient 11 caractères : le premier caractère (« d » ou « i ») correspond à la direction (croissante ou décroissante), suivi

1. Source : `https://www.ecm-france.fr/`
2. Source : `https://www.researchgate.net/`

par 2 à 4 caractères représentant le nom de la route (par exemple, « 405 » identifie la route I-405). Les caractères « es » n'ont aucune signification. Enfin, les caractères 7 à 11 indiquent le poteau kilométrique (par exemple, « 15036 » indique le poteau kilométrique 150,36 du point d'origine).

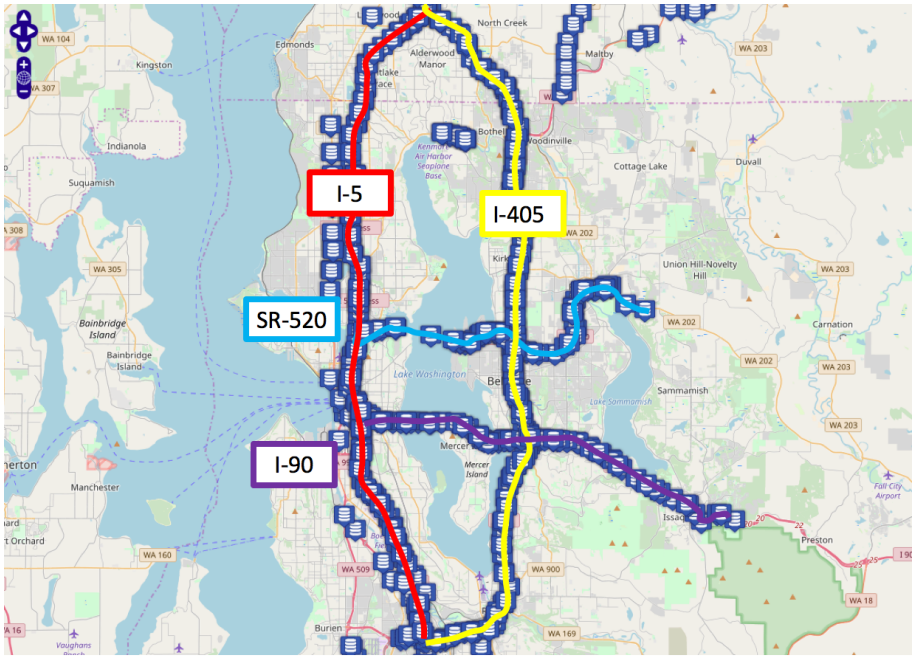

FIGURE 4.2 – Carte routière de la région de Seattle, présentant quatre autoroutes principales et leurs itinéraires respectifs : I-5 (en rouge), I-405 (en jaune), SR-520 (en bleu) et I-90 (en violet). Les icônes bleues sur la carte indiquent les emplacements des capteurs.

Parmi les ensembles de données disponibles, nous trouvons également :

1. **LoopSeattle2015A** : ce fichier contient la matrice d'adjacence numérique 323×323 qui décrit la structure du réseau des boucles.
2. **Cabinet Location Information** : ce fichier contient les données spatiales des détecteurs, notamment la latitude, la longitude et le poteau kilométrique.

Dans ce cas d'étude, nous nous intéressons aux données collectées pour les quatre premières semaines du mois de janvier 2015, c'est-à-dire du 1er au 28 janvier 2015 inclusivement. Ces données ont été agrégées sur 24 intervalles de temps, soit un enregistrement par heure, pour chaque jour. Pour agréger les données sur 24 intervalles

ID stamp	d005es15036	d005es15125	d005es15214	d005es15280	d005es15315	d005es15348	d005es15410
2015-01-01 00:00:00	61.939138	64.280883	62.077397	60.786423	63.120675	64.448315	63.411123
2015-01-01 00:05:00	59.232527	65.082450	64.808345	65.853953	59.206229	62.496716	65.992183
2015-01-01 00:10:00	61.991801	65.309123	64.803916	64.266082	62.239202	63.816610	60.196829
2015-01-01 00:15:00	62.480655	65.191651	67.206597	63.988427	65.808507	64.757556	62.011448

FIGURE 4.3 – Échantillon de données de vitesse de circulation. Chaque ligne représente une mesure associée à un horodatage précis. Chaque colonne correspond à un identifiant spécifique de capteur. Les valeurs reflètent les vitesses enregistrées en mph (*miles per hour* [miles par heure]) par chaque capteur à l'instant indiqué.

de temps, nous avons pris la moyenne des enregistrements de vitesse appartenant à la même heure.

Les données utilisées sont publiques et disponibles en libre accès sur le lien GitHub suivant : `https://github.com/zhiyongc/Seattle-Loop-Data`.

4.2 Exploration des données

Les données spatiotemporelles sont caractérisées par deux dimensions : l'espace et le temps. La dimension spatiale est généralement associée à une zone géographique définie, tandis que la dimension temporelle permet de saisir l'évolution des processus géographiques à des intervalles spécifiques $\mathcal{I}$, où $\mathcal{I}$ est un intervalle de temps qui peut être représenté sur une échelle continue ou discrète. Pour les besoins de cette étude, nous utilisons un ensemble discret, $\mathcal{I} = \{0, 1, ..., N\}$, où N est le nombre total d'observations temporelles, sur une période de 28 jours, avec des mesures prises toutes les heures, donnant lieu à un total de N=672 (28 jours $\times$ 24 heures) observations.

Pour explorer ces données, nous avons transformé leur format initial, dit « large », en format « long ». Chaque enregistrement comprend désormais l'identifiant du capteur, la date et l'heure de l'enregistrement de la vitesse, la valeur de la vitesse elle-même, le jour de la semaine, les coordonnées spatiales ainsi que la plage horaire.

4.2.1 Diagrammes spatiaux

La figure 4.4 montre une représentation graphique de la position des bornes fixes dans l'espace ainsi que l'état du trafic tout au long de la journée du 6 janvier 2015. Les données ont été recueillies à des intervalles d'une heure pour un total de 24 enregistrements dans la direction « d ».

Le schéma observé dans cette figure indique une stabilité de grande vitesse pendant les premières heures de la journée, de minuit à cinq heures du matin. À partir de six heures du matin, des points chauds apparaissent, principalement du nord vers le sud pour les autoroutes verticales (I-5 et I-405) et de l'est vers l'ouest pour les autoroutes horizontales (I-90 et SR-520).

Ainsi, au cours des quatre premières heures de pointe, de six heures à dix heures, la circulation devient plus dense en direction du centre-ville de Seattle. De onze heures à quatorze heures, la circulation reste perturbée, mais avec une intensité moindre, jusqu'à quinze heures. À partir de quinze heures, deux nouveaux points chauds apparaissent, le premier partant du centre-ville de Seattle et le deuxième au croisement des autoroutes I-405 et I-90. Le trafic commence à revenir à un flux régulier à partir de dix-neuf heures. La figure permet donc de visualiser clairement les changements dans l'état du trafic au fil de la journée, en particulier en ce qui concerne les heures de pointe et les zones les plus congestionnées.

La figure 4.5 illustre la position opposée à celle présentée dans la figure 4.4, en représentant la direction « i » dans le jeu de données. Tout comme la représentation précédente, cette figure montre une stabilité de la circulation à grande vitesse pendant les premières heures de la journée, de minuit à cinq heures du matin. Ce-

pendant, à partir de six heures du matin, des points de congestion apparaissent, principalement du sud vers le nord pour les autoroutes verticales (I-5 et I-405) et de l'ouest vers l'est pour les autoroutes horizontales (I-90 et SR-520). En outre, les données suggèrent que la I-90 et la SR-520 sont moins touchées par le trafic que la I-5 et la I-405.

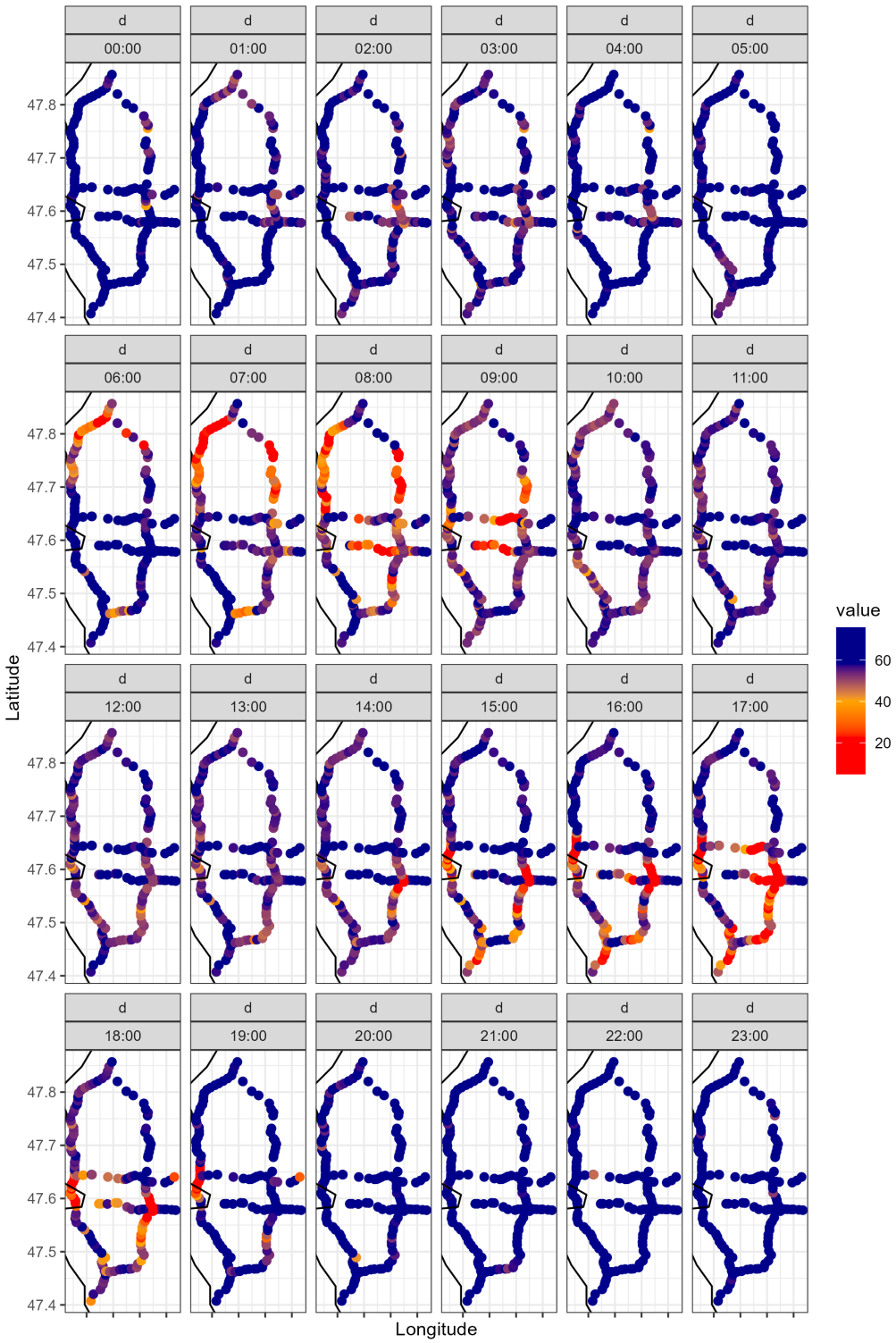

FIGURE 4.4 – Échantillon de données de vitesse de circulation enregistrées pour la direction « d », le 6 janvier 2015.

Les figures 4.6 et 4.7 présentent les points chauds observés sur l'ensemble de notre jeu de données de circulation automobile. Nous avons utilisé la vitesse minimale enregistrée pour chaque capteur au cours de la journée pour identifier les capteurs touchés par de grandes congestions. Chaque sous-figure représente donc la vitesse minimale enregistrée au cours d'une journée particulière.

L'ensemble des figures, qui couvre une période de 28 jours et les directions « d » et « i », révèle plusieurs observations. Tout d'abord, nous avons identifié un modèle récurrent où le trafic se déplace du nord vers le sud pour l'autoroute I-5 et du sud vers le nord pour l'autoroute I-405. Nous avons également observé que la congestion est plus importante dans les tronçons de croisement des autoroutes, tels que le croisement entre la I-5, la SR-520 et la I-90, ainsi que celui entre la I-405, la SR-50 et la I-90.

De plus, nous avons constaté que le trafic sur la I-5 touche principalement le tronçon nord, jusqu'au centre-ville : une fois le centre-ville dépassé, le trafic redevient plus ou moins régulier. En revanche, le trafic sur la I-405 est plus dense, et ce, sur toute sa longueur.

En outre, nous avons observé que les weekends sont caractérisés par un flux de circulation relativement stable et fluide, à l'exception des tronçons proches du centre-ville. En revanche, durant la semaine, le trafic congestionne les quatre autoroutes. Il est important de noter que le 1er et le 2 janvier sont considérés comme des weekends, car ce sont principalement des journées fériées.

En résumé, les figures 4.6 et 4.7 permettent d'identifier les points chauds et de comprendre les modèles de circulation automobile dans la région.

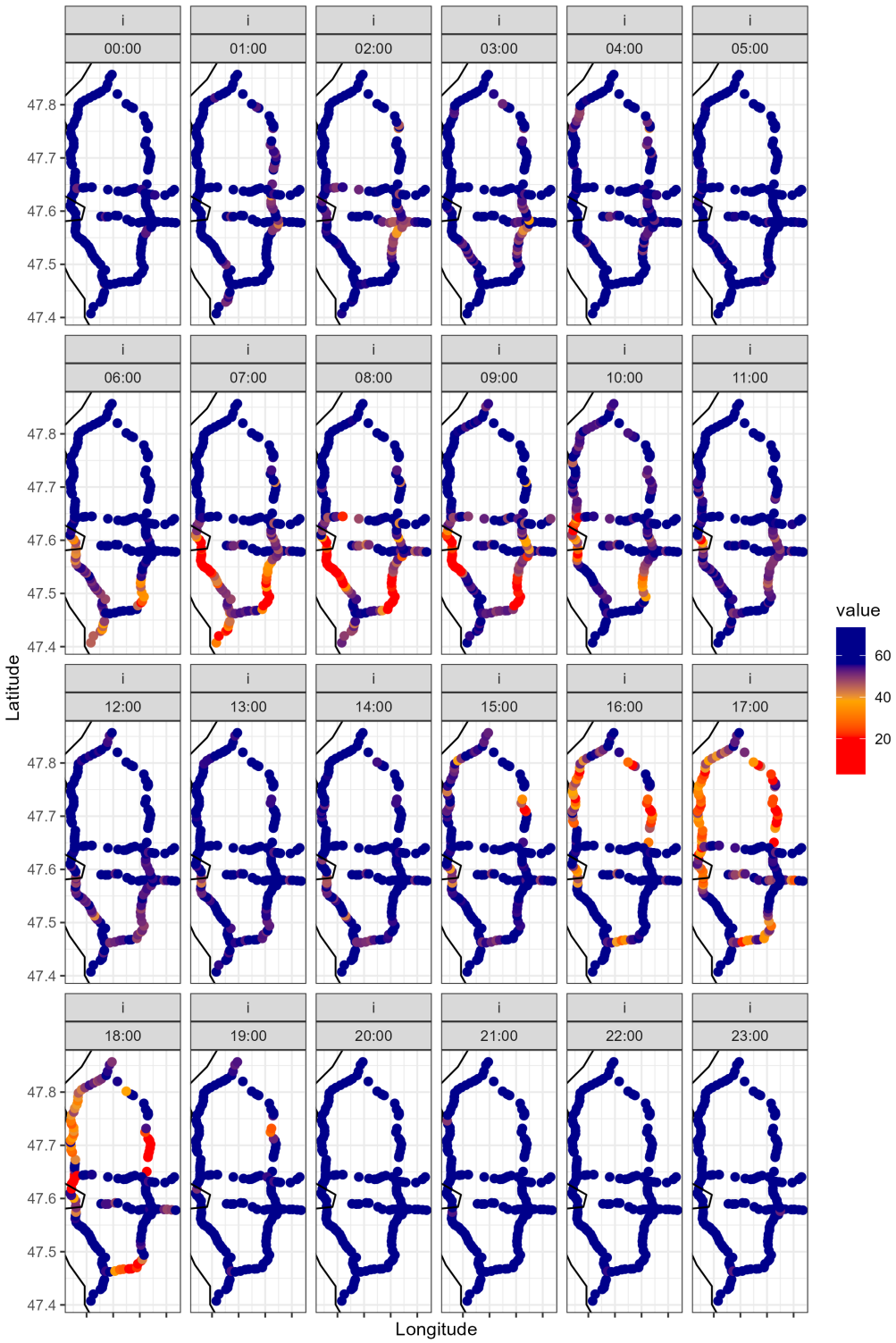

FIGURE 4.5 – Échantillon de données de vitesse de circulation enregistrées pour la direction « i », le 6 janvier 2015.

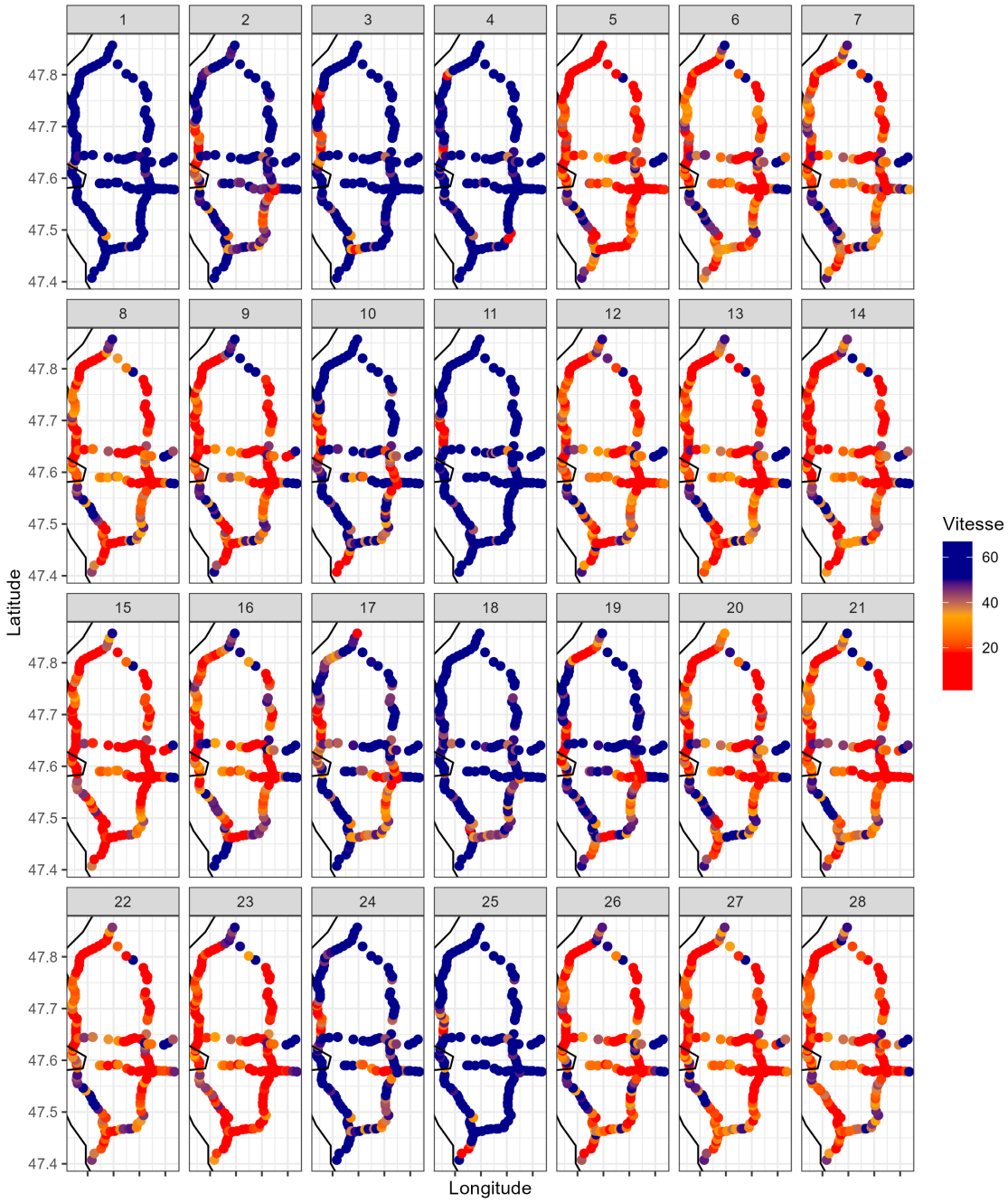

FIGURE 4.6 – Échantillon de données de vitesse de circulation minimale enregistrée par les capteurs pour la direction « d » sur les 28 jours du mois de janvier.

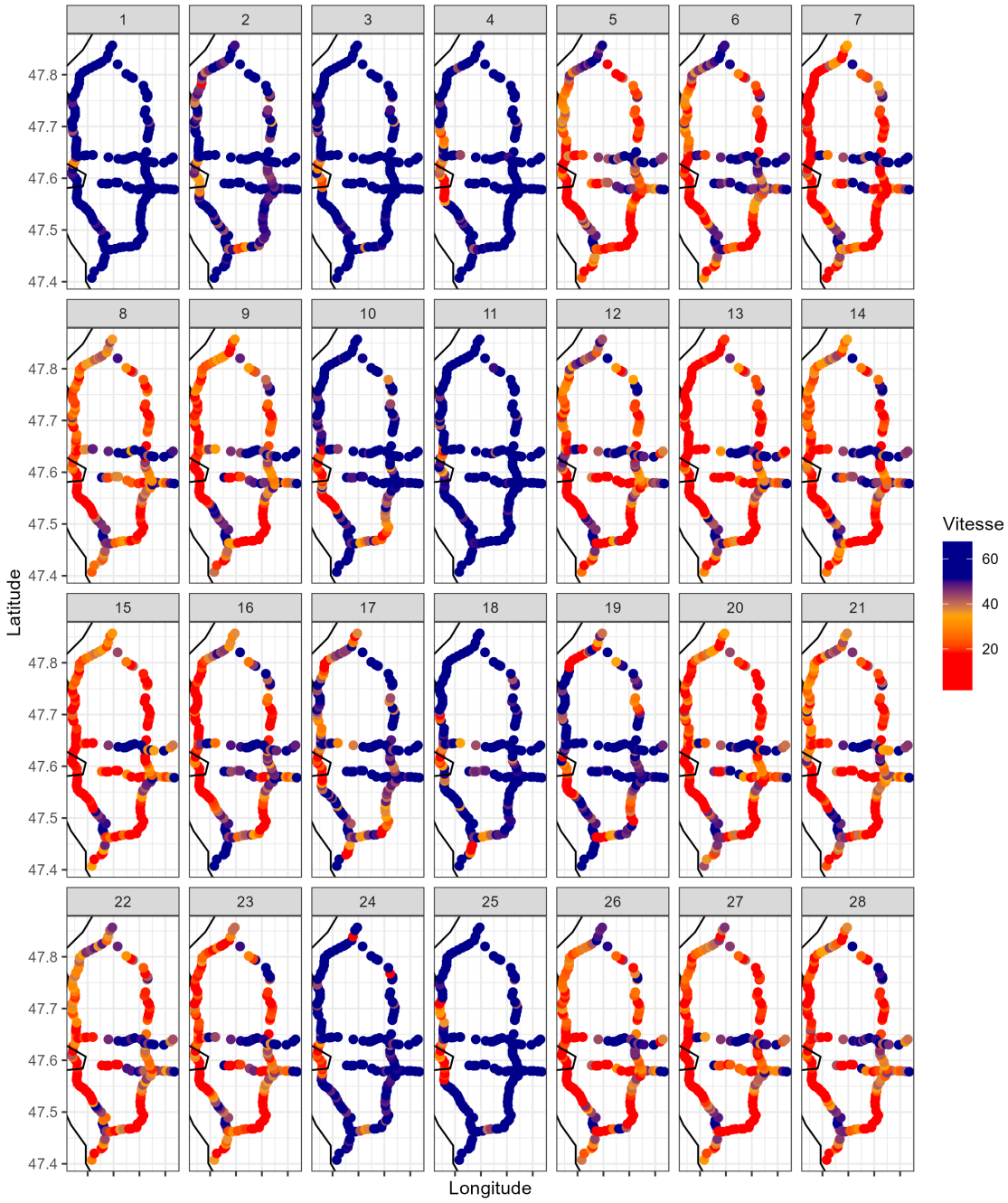

FIGURE 4.7 – Échantillon de données de vitesse de circulation minimale enregistrée par les capteurs pour la direction « i » sur les 28 jours du mois de janvier.

4.2.2 Diagrammes de séries temporelles

Le diagramme de séries temporelles nous permet d'étudier les changements de vitesse dans un espace fixe au fil du temps. Dans la figure 4.8, nous pouvons voir huit sous figures représentant chacune une série temporelle de la vitesse enregistrée à différents moments de différents capteurs voisins. De fait, pour chaque autoroute, nous avons choisi deux capteurs voisins de directions opposées, ce qui nous permet de comparer le comportement du trafic dans les deux sens pour une même autoroute.

Il convient de noter que les capteurs sont identifiés par la couleur de l'autoroute à laquelle ils appartiennent. Par exemple, les capteurs « d005es18449 » et « i005es18449 » font partie de la même autoroute, la I-005, mais dans des directions opposées (« d » pour descendant et « i » pour ascendant).

En observant les graphiques, nous pouvons voir que la vitesse de circulation varie considérablement, oscillant autour d'une moyenne de 60 mph, pour la I-5, la I-90 et la I-405, et autour d'une moyenne plus élevée, proche de 70 mph, pour la SR-520. Par ailleurs, la moyenne, pour les autoroutes I-90 et I-405, peut prendre une valeur aussi basse que 10 mph. Ainsi, la fluctuation de la vitesse de circulation n'est pas uniforme. Certains capteurs sont plus touchés par la congestion du trafic que d'autres : par exemple, le capteur « d405es02898 » est moins touché par la congestion du trafic que le capteur « d090es00430 », qui en souffre de manière quasi permanente.

En outre, les graphiques révèlent que le comportement dans les deux sens de la même autoroute peut être significativement différent. Le capteur « i520es00158 », bien qu'étant sur la même autoroute que le capteur « d520es00158 », ne souffre pas autant de la congestion du trafic que ce dernier. Ce phénomène est également observable entre les capteurs « d005es18449 » et « i005es18449 ». En somme, la figure 4.8 fournit des informations utiles concernant la circulation sur les autoroutes, en mettant en évidence les différences dans le comportement de la circulation dans chacun des sens de la même autoroute et en soulignant les capteurs les plus touchés par la congestion du trafic.

Les figures 4.9 et 4.10 montrent la position des bornes fixes pour chacun des capteurs. Chaque point rouge représente donc un capteur dont la série temporelle est présentée dans la figure précédente. Ces informations géographiques peuvent aider à comprendre la configuration de la route et à identifier les zones à problèmes en termes de congestion de trafic.

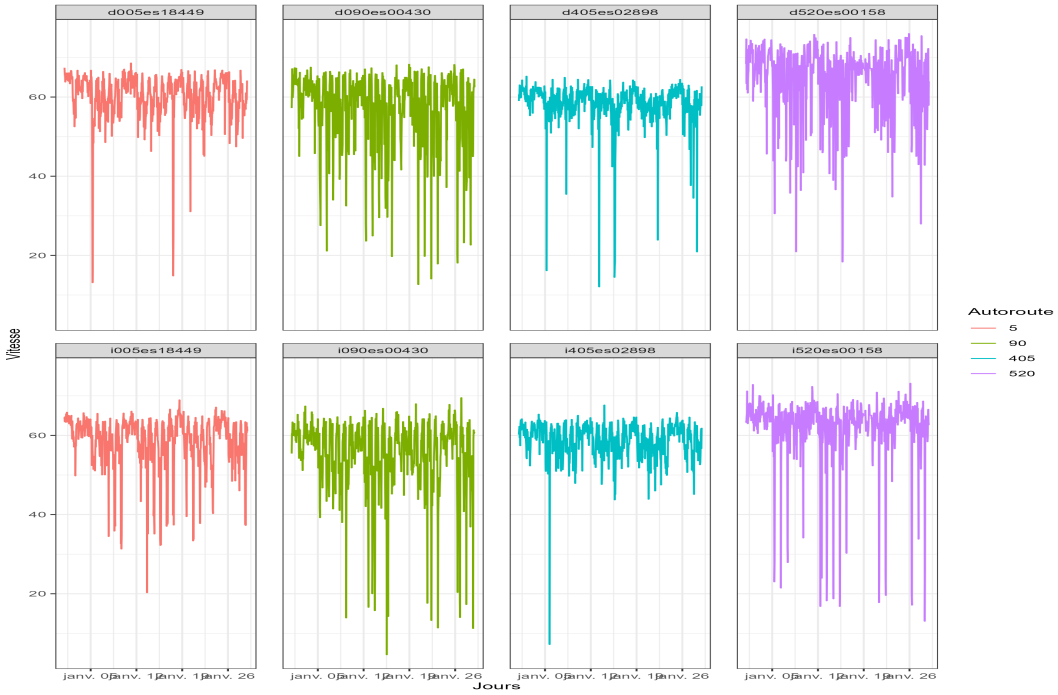

FIGURE 4.8 – Échantillon de données de vitesse de circulation. La figure présente huit sous-figures, où chaque sous-figure représente une série temporelle de la vitesse enregistrée le long d'une autoroute spécifique identifiée par une couleur. Les capteurs sont disposés en deux lignes, les capteurs du haut sont localisés dans le sens opposé des capteurs du bas et partagent la même localisation géographique. Les autoroutes sont identifiées par leur couleur respective.

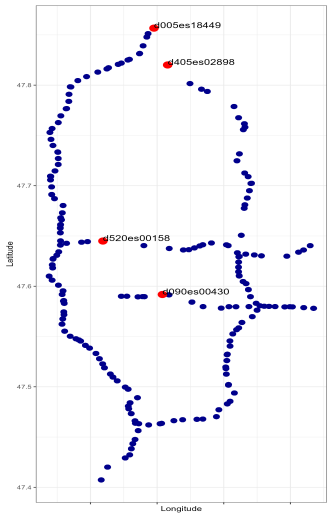

FIGURE 4.9 – Position des quatre bornes fixes dans la direction « d ».

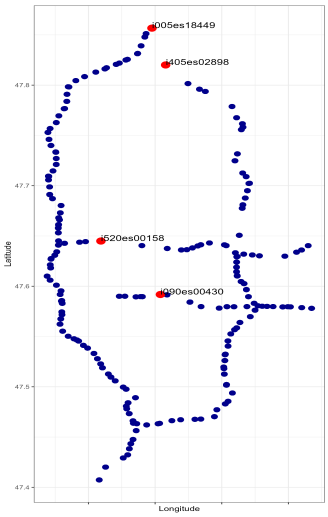

FIGURE 4.10 – Position des quatre bornes fixes dans la direction « i ».

4.3 Analyse exploratoire

4.3.1 Moyenne spatiale empirique

La vitesse est un paramètre essentiel de la circulation routière et peut être représentée par la vitesse moyenne temporelle ou spatiale. La moyenne spatiale empirique est obtenue en calculant la moyenne des points temporels pour une localisation donnée.

Dans notre cas, nous avons calculé la moyenne spatiale empirique en agrégeant la vitesse par capteur, sens de la circulation et plage horaire pour les 28 jours de données étudiées. Nous avons divisé la journée en trois plages horaires distinctes : de six heures à dix heures, de quinze heures à dix-neuf heures et les autres heures de la journée. La figure 4.11 illustre la variation de la vitesse moyenne en fonction du capteur et de la plage horaire.

En examinant la première plage horaire (six heures à dix heures) pour les deux sens de circulation, nous avons constaté que le trafic est congestionné au nord des autoroutes I-5 et I-405, tandis qu'il est fluide au sud. En revanche, pour la plage horaire suivante (quinze heures à dix-neuf heures), le trafic est plus fluide au nord

des autoroutes I-5 et I-405 qu'au sud. Nous avons donc observé une inversion du trafic, non seulement en fonction du sens de circulation, mais aussi en fonction de l'heure de la journée.

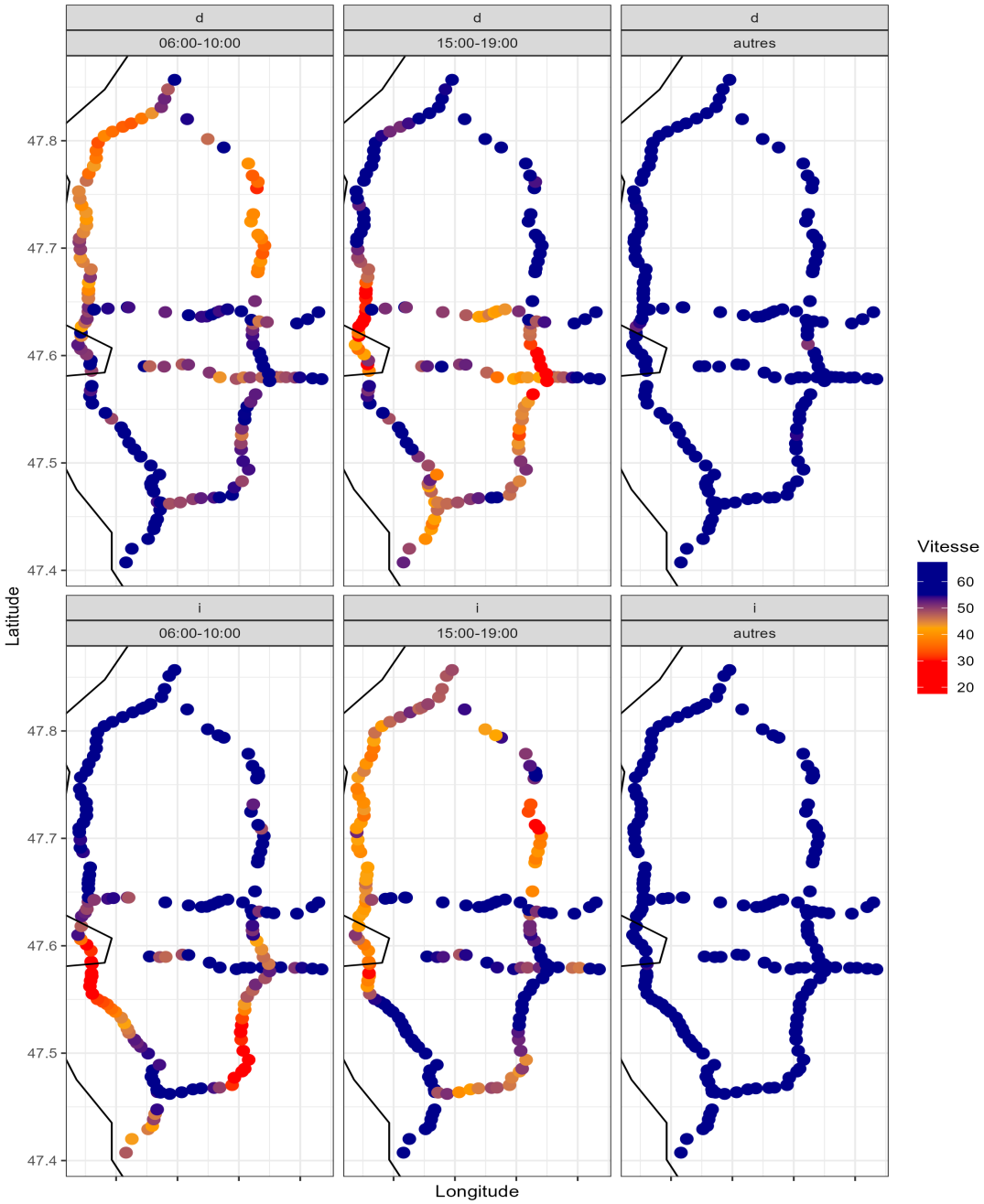

FIGURE 4.11 – Moyenne spatiale empirique par capteur en fonction de l'orientation et de la plage horaire.

4.3.2 Moyenne temporelle empirique

En plus de la moyenne spatiale temporelle, nous pouvons calculer la moyenne temporelle empirique qui consiste en l'agrégation de la moyenne de la vitesse enregistrée selon l'espace. L'équation est donnée comme suit :

$$\hat{\mu}(\boldsymbol{R}_{:,j}) = \frac{1}{M} \sum_{i=1}^{M} R_{i,j}. \tag{4.1}$$

La figure 4.12 illustre une représentation graphique détaillée de la série temporelle de la vitesse moyenne mesurée par les 323 capteurs disponibles, qui couvrent l'ensemble des emplacements spatiaux étudiés. En effet, cette figure permet d'observer la nature saisonnière de la vitesse moyenne pour les quatre autoroutes étudiées ainsi que les fluctuations temporelles en un point donné : ces fluctuations sont représentées par des couleurs différentes, qui peuvent être interprétées comme des variations spatiales.

Plus précisément, la série temporelle de la vitesse moyenne mesurée par les capteurs suit deux schémas saisonniers qui se répètent. Le premier schéma est hebdomadaire et se caractérise par un modèle redondant des semaines. En effet, on peut observer une augmentation de la vitesse durant les weekends et une fluctuation de la vitesse durant les jours de la semaine. Par exemple, au cours de la semaine du lundi 5 janvier au dimanche 11 janvier, nous observons une fluctuation de la vitesse pendant les jours ouvrables et une augmentation de la vitesse pendant le weekend. À l'intérieur de ce schéma saisonnier, on retrouve un second schéma, plutôt journalier, qui montre un sommet soutenu de la vitesse représentant les heures hors trafic et deux creux qui représentent les heures de pointe.

En outre, la figure 4.12 permet également d'observer les fluctuations temporelles en un point donné, qui mettent en évidence des variations de la vitesse moyenne en fonction de la localisation sur les autoroutes. Ainsi, nous pouvons observer des zones où la vitesse moyenne est plus élevée et d'autres où elle est plus faible.

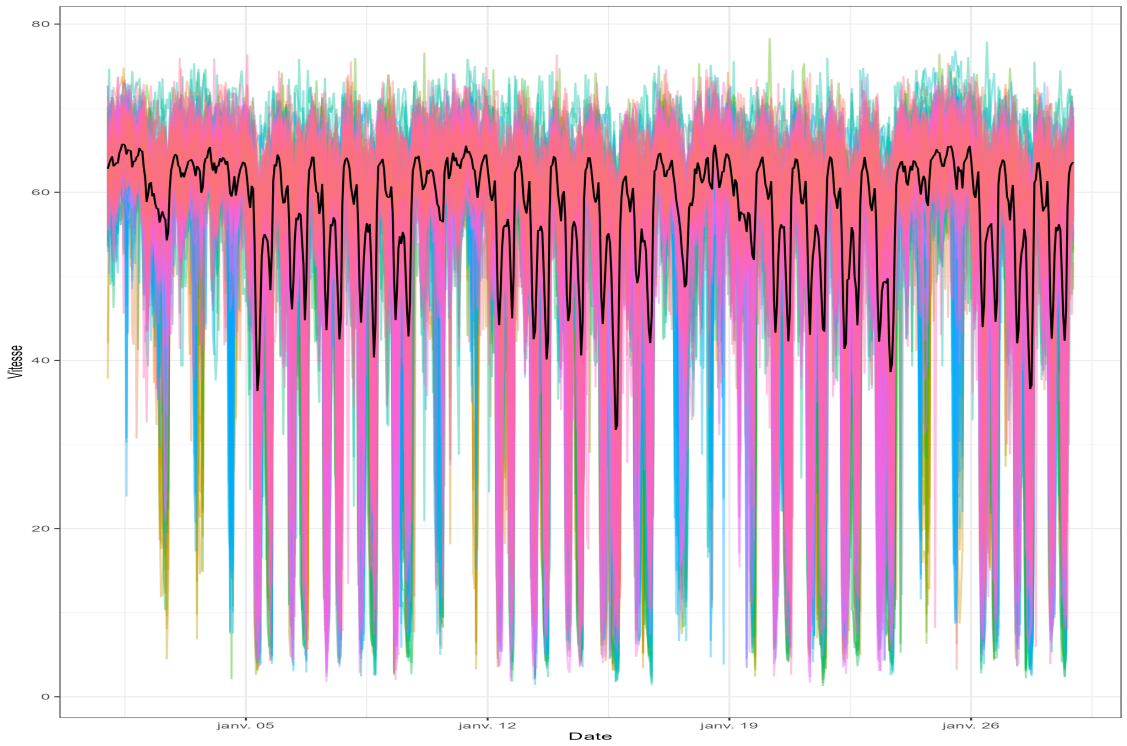

FIGURE 4.12 – Moyenne temporelle empirique. La figure illustre une représentation graphique détaillée des séries temporelles de la vitesse moyenne mesurée par les 323 capteurs disponibles, qui couvrent l'ensemble des emplacements spatiaux étudiés. Chaque série temporelle de vitesse moyenne est représentée par une courbe de couleur différente, permettant d'identifier les variations temporelles en un point donné. La moyenne temporelle empirique est représentée par une courbe en noir, qui correspond à la moyenne de la vitesse enregistrée agrégée selon l'espace.

4.3.3 Dépendance spatiale

La dépendance spatiale est un concept qui décrit la relation entre les valeurs de vitesse pour des paires d'emplacements distants. Selon la loi de Tobler (1970), « tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés ». Cette relation peut évoluer dans le temps en fonction de facteurs tels que le débit de circulation, les accidents de la route, etc. Comprendre comment un phénomène localisé se propage dans l'espace et dans le temps est essentiel pour évaluer comment les structures spatiales sous-jacentes se modifient à différentes périodes de temps.

Dans notre étude, nous avons utilisé un ensemble de données contenant les mesures de vitesse de 323 emplacements dans la région de Seattle. Les données comprennent le nom de la station, la latitude et la longitude de la station, ainsi que la moyenne des vitesses enregistrées agrégées par heure. Pour évaluer la dépendance spatiale des données, nous avons utilisé l'outil Spatial Autocorrelation (Global Moran's I), qui est une statistique inférentielle. Cela signifie que les résultats de l'analyse sont interprétés dans le contexte de l'hypothèse nulle, qui stipule que la vitesse moyenne est distribuée de manière aléatoire parmi les capteurs de la zone étudiée.

Pour calculer la dépendance spatiale, nous avons généré une matrice de pondérations de distance inverses. Pour ce faire, la distance euclidienne est utilisée pour calculer la distance entre les paires de coordonnées « Longitude » et « Latitude » en utilisant la fonction *dist()* de R. Nous avons ensuite utilisé la fonction *Moran.I()* du logiciel R pour calculer la *p-value*. Les résultats ont montré une *p-value* de 2.664535e-15, ce qui nous permet de rejeter l'hypothèse nulle selon laquelle il n'y a aucune autocorrélation spatiale dans la variable vitesse à un niveau de confiance alpha de 0,01. En conclusion, notre analyse a montré une dépendance spatiale significative dans les données de vitesse de la région de Seattle.

Les figures 4.13 et 4.14 représentent des cartes des corrélations spatiales entre les capteurs. La figure 4.13 est générée à partir de la matrice de corrélation entre les

capteurs calculée en utilisant la fonction *cor()* de R sur les 323 séries temporelles. Cependant, afin de mettre en évidence uniquement les corrélations significatives, nous avons filtré les valeurs pour ne conserver que celles supérieures ou égales à 0.5. Ainsi, la figure 4.13 représente visuellement les connexions spatiales significatives entre les capteurs. La figure 4.14 quant à elle, représente graphiquement la matrice d'adjacence fournie dans le jeu de données. Les connexions géographiques, dans les deux graphiques, sont représentées en fonction de l'information disponible, où chaque connexion, respectivement, représente une forte corrélation entre deux capteurs, illustrée par une arête. Nous remarquons que les deux cartes sont identiques, indiquant que la matrice d'adjacence est construite à partir de la matrice de corrélation. De plus, nous observons une similarité des valeurs en fonction de leur emplacement géographique. Cette similarité est caractérisée par des dépendances ou des interactions spatiales qui sont d'autant plus fortes que les localisations sont proches, ce qui confirme la loi de Toble. Cela signifie que lorsque l'on effectue des mesures de vitesse, ces mesures sont toutes liées les unes aux autres. L'intensité de leur lien fonctionne selon la distance qui les sépare, ce qui peut être particulièrement pertinent dans notre cas, puisque les quatre autoroutes ont été reconstruites à partir de ces matrices. En somme, ces cartes permettent de visualiser la corrélation spatiale entre les capteurs et de mieux comprendre les relations entre les différentes mesures de vitesse.

4.3.4 Dépendance temporelle

La dépendance temporelle réfère à la relation entre les observations enregistrées à différents moments dans le temps. Contrairement à la dépendance spatiale, qui peut être multidirectionnelle, la dépendance temporelle est unidirectionnelle. Cela signifie que les observations passées peuvent influencer les observations futures, mais pas l'inverse. Afin de poursuivre notre analyse de la dépendance temporelle, nous avons converti notre série spatiotemporelle multidimensionnelle en une série uni-

FIGURE 4.13 – Dépendance spatiale basée sur la matrice de corrélation. Chaque coefficient de corrélation supérieur à 0.5 établit une connexion, représentée par une arête.

FIGURE 4.14 – Dépendance spatiale basée sur la matrice d'adjacence. Une connexion est établie pour chaque valeur égale à 1 et représentée par une arête.

dimensionnelle en utilisant la moyenne temporelle empirique. En d'autres termes, pour chaque instant de temps, nous avons calculé la moyenne de toutes les valeurs enregistrées.

La figure 4.15 montre que les valeurs enregistrées pendant les jours de semaine (du lundi au vendredi) sont fortement corrélées entre elles, ce qui indique que ces jours présentent des comportements similaires. De même, les valeurs enregistrées pendant les jours du weekend (samedi, dimanche et fériés) sont également fortement corrélées entre elles, indiquant que ces jours présentent également des comportements similaires. Cependant, les valeurs enregistrées pendant les jours de semaine

ont tendance à être faiblement corrélées avec celles enregistrées durant les jours du weekend, indiquant des différences dans les habitudes et les comportements des gens pendant ces périodes. En outre, la forte corrélation observée entre les valeurs enregistrées lors des journées du weekend suggère que des facteurs spécifiques aux weekends peuvent avoir une influence significative sur ces valeurs.

La figure 4.16 montre un graphique représentant la corrélation entre les heures de la journée. Ce graphique révèle plusieurs plages horaires au cours desquelles la corrélation entre les heures est particulièrement forte. La première plage horaire qui attire notre attention est celle des heures de pointe du matin, de six heures à dix heures. Pendant cette période, la corrélation est la plus importante de toutes les plages horaires du jeu de données. Cela signifie que les évènements qui se produisent sur la route pendant ces heures ont tendance à se produire à des moments similaires tous les jours. La deuxième plage horaire, de quinze heures à dix-neuf heures, est également caractérisée par une forte corrélation entre ces heures. Cela suggère qu'il y a également une régularité dans le trafic pendant cette période de la journée. En outre, le graphique montre trois autres plages horaires où la corrélation entre les heures est notable : la plage horaire de minuit à quatre heures, de onze heures à quatorze heures et de vingt heures à vingt-trois heures. En somme, ces plages horaires révèlent une structure temporelle complexe dans le trafic routier.

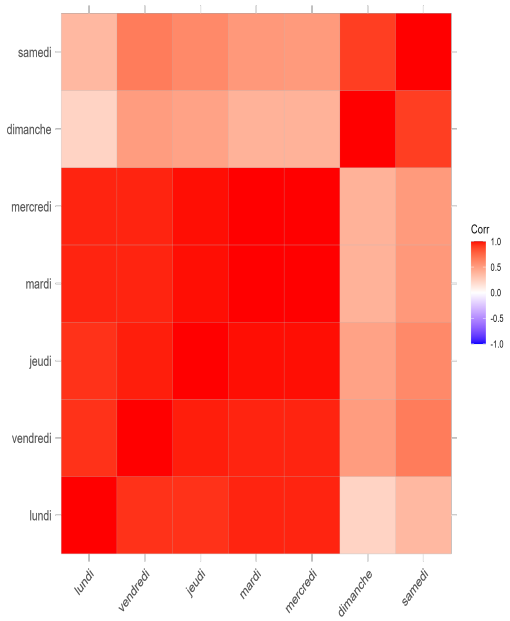

FIGURE 4.15 – Corrélation temporelle entre les jours de la semaine. Les jours de la semaine sont regroupés de manière à ce que ceux qui sont plus fortement corrélés soient proches les uns des autres sur le graphique. La figure représente une matrice carrée illustrant la corrélation temporelle entre les jours de la semaine. Chaque case de la matrice représente le degré de corrélation entre deux jours spécifiques. Les jours de la semaine sont étiquetés le long des axes horizontal et vertical de la matrice.

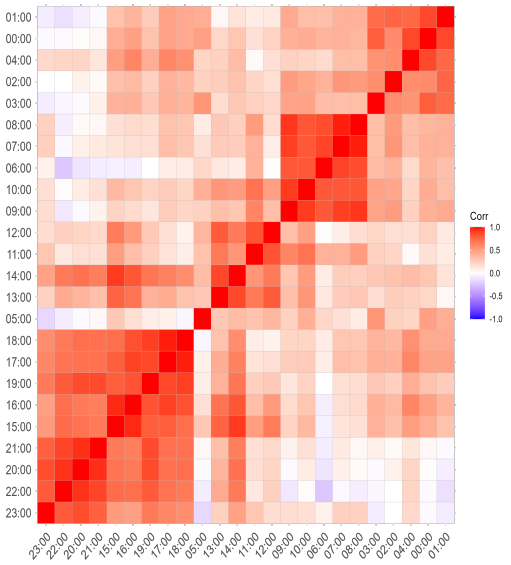

FIGURE 4.16 – Corrélation temporelle entre les heures de la journée avec regroupement des heures d'achalandage. Les heures de la journée sont regroupées de manière à ce que celles qui sont plus fortement corrélées soient proches les unes des autres sur le graphique. La figure représente une matrice carrée illustrant la corrélation temporelle entre les heures de la journée, avec un regroupement des heures d'achalandage. Chaque case de la matrice représente le degré de corrélation entre deux heures spécifiques. Les heures de la journée sont disposées le long des axes horizontal et vertical de la matrice.

4.3.5 Diagrammes d'autocorrélation

L'autocorrélation est la corrélation entre les valeurs d'une série temporelle sur des périodes successives. En d'autres termes, il s'agit de la mesure de la dépendance d'une valeur à une heure donnée avec les valeurs des heures précédentes. Cette dépendance est un élément clé dans de nombreux modèles de prévision.

Les graphiques Autocorrelation Function (ACF) et Partial Autocorrelation Function (PACF) sont deux façons de visualiser cette dépendance. L'ACF est une mesure de corrélation linéaire qui examine la relation entre une observation dans une série temporelle et les observations précédentes à différents décalages (lags). Le graphique ACF trace la corrélation en fonction du décalage sur l'axe horizontal et la valeur de corrélation sur l'axe vertical. Il permet ainsi d'identifier les tendances et les motifs périodiques dans la série temporelle. Les pics significatifs sur le graphique ACF indiquent les décalages où les observations sont fortement corrélées, suggérant une dépendance temporelle. D'autre part, le PACF mesure la corrélation directe entre une observation et une observation retardée, en éliminant l'influence des observations intermédiaires. Elle représente la corrélation partielle entre les observations à différents décalages. Le graphique PACF met en évidence les relations de dépendance directe entre les observations en montrant les pics significatifs qui ne peuvent pas être expliqués par les décalages précédents. Pour calculer la fonction ACF et la fonction PACF, nous avons utilisé les fonctions *acf()* et *pacf()* de R sur la moyenne temporelle empirique.

Ainsi, dans le graphique ACF de la figure 4.17, nous voyons les différents décalages temporels (t-1, t-2, t-3, etc.) sur l'axe horizontal, tandis que les corrélations entre t et les différents décalages sont représentés sur l'axe vertical. Notez que chaque graphique comporte deux lignes horizontales en pointillés qui représentent le seuil de signification. Les pics qui dépassent la ligne pointillée horizontale sont considérés donc comme significatifs. Par ailleurs, le PACF, présenté dans la figure 4.18, supprime toute corrélation indirecte pouvant être présente dans l'ACF. En somme,

le diagramme ACF indique que les pics significatifs apparaissent par vagues. Nous pouvons remarquer des pics significatifs positifs, suivis par des pics significatifs négatifs, et vice-versa. Par ailleurs, grâce au diagramme PACF, nous pouvons constater que le premier décalage (t-1) a la plus grande influence, mais que celle-ci diminue rapidement. Cela signifie que la vitesse enregistrée à une heure donnée est fortement corrélée à la vitesse enregistrée à l'heure précédente.

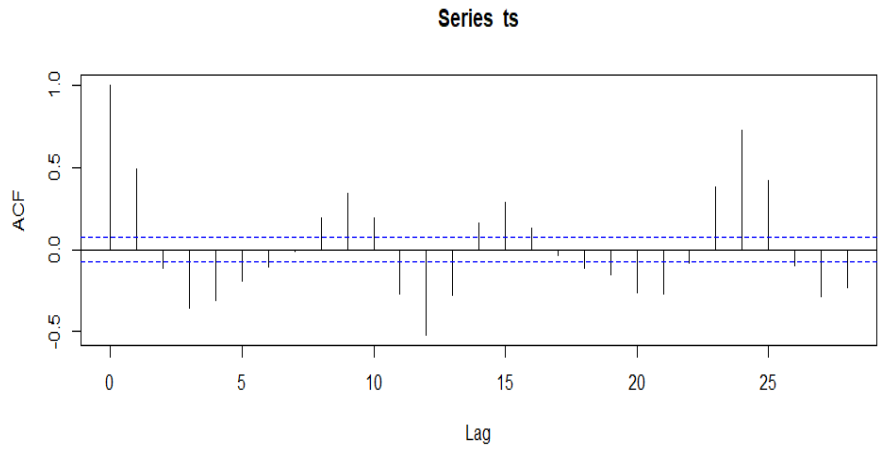

FIGURE 4.17 – Diagrammes d'autocorrélation ACF.

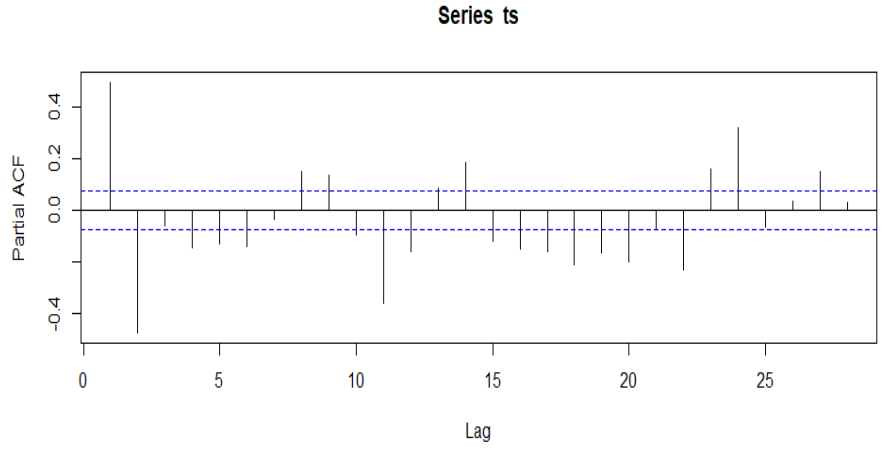

FIGURE 4.18 – Diagrammes d'autocorrélation PACF.

4.3.6 Composantes de la série chronologique

La décomposition saisonnière est le processus par lequel une série temporelle est décomposée en plusieurs composants, généralement, la tendance, la saisonnalité et les résidus. Ces composants constitutifs sont essentiels à l'analyse des séries chronologiques.

La figure 4.19 présente la décomposition de la série temporelle de la vitesse moyenne enregistrée par les 323 capteurs. Cette figure est divisée en quatre parties. La première partie montre les données observées telles qu'elles sont. La seconde partie affiche la courbe de tendance, suivie des composantes saisonnières et aléatoires sur une périodicité journalière. En analysant ces graphiques, on remarque que la tendance semble relativement stationnaire, mais qu'elle est légèrement bruitée. Pour améliorer la qualité de cette tendance, nous pouvons rendre la série stationnaire. Pour cela, nous pouvons calculer la différence entre chaque observation et son observation précédente avec un lag de 1.

La figure 4.20 présente la décomposition de la série temporelle stationnaire obtenue à partir de la série temporelle initiale. Cette figure montre que la saisonnalité est bien capturée avec des pics et des vallées horaires en forme de M. Cette décomposition nous permet de mieux comprendre la série temporelle et d'analyser les comportements saisonniers de la vitesse moyenne enregistrée par les capteurs.

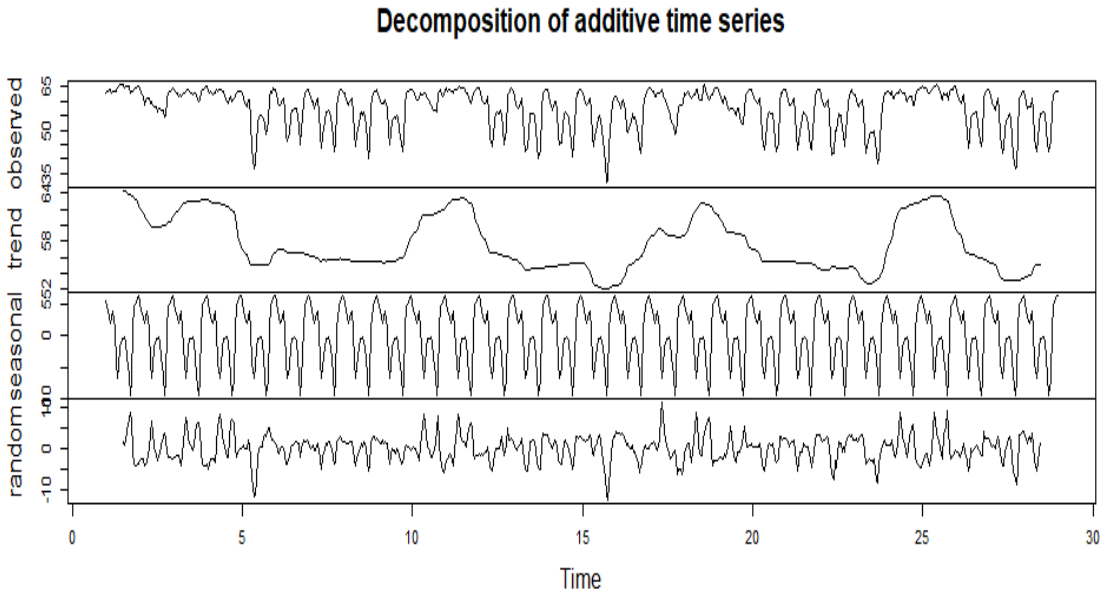

FIGURE 4.19 – Composantes de la série temporelle (avant la stationnarité).

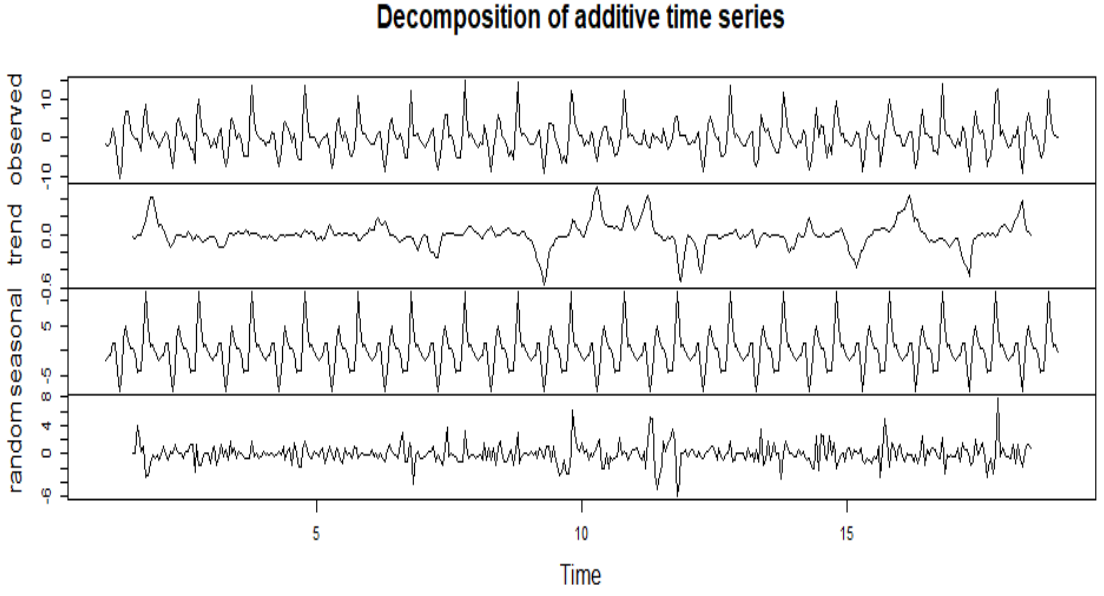

FIGURE 4.20 – Composantes de la série temporelle (après la stationnarité (Lag 1)).

Chapitre 5

Étude de simulation et résultats

Ce chapitre a pour objectif de comparer les méthodes d'imputation décrites au chapitre trois, intitulé « Description des modèles », sous différents scénarios de données manquantes. Nous souhaitons comparer les modèles en termes de qualité d'imputation et de vitesse de calcul. Nous rappelons que l'objectif de notre recherche est de mener une étude comparative des modèles d'imputation de données manquantes basés sur les méthodes de factorisation matricielle que nous avons décrites précédemment. Plus précisément, notre attention se porte sur les modèles Bayesian Probabilistic Matrix Factorization (BPMF), Temporal Regularized Matrix Factorization (TRMF), Bayesian Temporal Matrix Factorization (BTMF) et Kernelized Probabilistic Matrix Factorization (KPMF), qui intègrent la structure complexe de la matrice de variance-covariance des données pour résoudre le problème de l'imputation des données manquantes. Pour ces modèles, nous analysons l'impact du rang de décomposition et de l'absence de données sur les performances des modèles d'imputation choisis. De plus, nous évaluerons l'impact de la structure des données manquantes dans l'efficacité de ces modèles à imputer les valeurs manquantes.

5.1 Préparation des scénarios

Le jeu de données à notre disposition est exempt de données manquantes. Toutefois, pour réaliser notre étude, nous avons créé différents jeux de données artificiels selon trois scénarios distincts, chacun caractérisé par une structure différente de données manquantes : les données manquantes unitaires (DMU), les données manquantes par blocs (DMB) et les données manquantes mixtes (DMM).

— Les données manquantes unitaires (DMU) sont des données qui manquent de manière aléatoire dans le jeu de données. Cela signifie que les éléments manquants sont dispersés de manière aléatoire parmi les observations, sans suivre un schéma particulier ;

— Les données manquantes par blocs (DMB) impliquent que des blocs entiers de données soient manquants. Dans notre cas, des périodes de données complètes sur une période de 24 heures, équivalant à 24 éléments, sont supprimées du jeu de données. Ces périodes manquantes sont sélectionnées de manière aléatoire parmi les observations ;

— Les données manquantes mixtes (DMM) sont un mélange des deux scénarios précédents. Elles comprennent à la fois des données DMU et DMB.

Nous avons également varié le taux de valeurs manquantes p de 30 % à 90 %, en augmentant p progressivement par incréments de 10 % pour chacun des trois scénarios. Dans un scénario DMM, environ la moitié de ces données manquantes, soit $p/2$, sont de type DMU, tandis que l'autre moitié, également $p/2$, sont de type DMB. Cela nous permet d'obtenir un taux de valeurs manquantes de p. Nous avons créé 21 ensembles de données différents, soit sept ensembles de données pour chaque scénario, DMU, DMB et DMM, chacun ayant un pourcentage différent de données manquantes (30 %, 40 %, 50 %, 60 %, 70 %, 80 % et 90 %). Les données sont construites sous forme d'une matrice, $\boldsymbol{R}^{M\times N}$, où $M = 323$ lignes correspondent aux identifiants des capteurs et où $N = 672$ colonnes correspondent au temps (28 jours x 24 heures). L'élément (i, j) de cette matrice indique la vitesse moyenne observée

sur le capteur i à l'heure j. L'analyse des mesures de la moyenne et de l'écart-type de la vitesse des 21 ensembles de données artificielles est présentée dans le tableau 5.1. Ce tableau montre que les vitesses moyennes (μ) et les écarts-types (σ) de la vitesse ne sont pas affectés par la variation des proportions de données manquantes. En effet, les moyennes et les écarts-types restent relativement constants, quel que soit le pourcentage des données manquantes ou scénarios. La vitesse moyenne est estimée à 57.67 mph (*miles per hour* [miles par heure]) et l'écart-type est estimé à 11 mph.

TABLE 5.1 – Distribution des données selon les trois scénarios avec différentes proportions de données manquantes.

Scénarios	DM en %	Taille de DM	Statistiques de vitesse	
			Vitesse Moyenne (μ)	Écart-type de la vitesse (σ)
Données initiales	0 %	0	57.67	11.21
DMU	30 %	65 295	57.71	11.18
	40 %	86 969	57.72	11.26
	50 %	108 617	57.69	11.26
	60 %	130 236	57.83	11.16
	70 %	151 910	57.75	11.25
	80 %	173 796	57.92	11.10
	90 %	195 592	58.12	11.12
DMB	30 %	65 125	57.67	11.20
	40 %	86 836	57.67	11.21
	50 %	108 550	57.67	11.20
	60 %	130 246	57.70	11.19
	70 %	151 943	57.67	11.28
	80 %	173 649	57.87	11.14
	90 %	195 363	57.92	11.27
DMM	30 %	65 122	57.66	11.22
	40 %	86 828	57.69	11.19
	50 %	108 540	57.68	11.19
	60 %	130 237	57.67	11.20
	70 %	151 944	57.66	11.21
	80 %	173 656	57.70	11.16
	90 %	195 365	57.66	11.20

DMU : DM Unitaires ; DMB : DM par Blocs ; DMM : DM Mixtes.

5.2 Mesures de performance

Nous évaluons la performance des modèles en calculant l'écart entre les valeurs imputées et les données réelles originales pour chaque modèle considéré. Pour ce faire, nous définissons des données d'entrainement et des données de test pour chacun des 21 ensembles de données présentés précédemment. Pour chaque ensemble de données, les données de test représentent les données manquantes dans l'ensemble de données d'entrainement et la phase d'apprentissage se fait en l'absence de ces données.

Pour évaluer la qualité des valeurs imputées, nous commençons notre analyse en utilisant des méthodes de diagnostiques visuelles. Pour ce faire, nous utilisons des graphiques pour comparer visuellement la courbe des données observées avec les courbes des données imputées par les modèles étudiés afin de vérifier leur similitude. De plus, nous élaborons des graphiques de la distribution des erreurs, celle-ci consiste à noter la densité de l'écart entre les données imputées et les données observées afin de détecter les éventuelles différences dans la performance pour chaque modèle. Ces graphiques nous permettent de voir si les valeurs imputées sont biaisées ou si elles ont une distribution différente de celle des valeurs réelles.

Dans notre analyse, nous avons choisi d'utiliser des méthodes de diagnostique quantitatives en plus des diagnostiques visuels pour évaluer la performance de nos modèles d'imputation de données manquantes. En effet, nous avons choisi d'utiliser la mesure de la Root Mean Square Error (RMSE) pour évaluer la performance de nos modèles, car elle est largement considérée comme une mesure d'évaluation appropriée (Chai (2014)). Cette mesure prend en compte toutes les observations du jeu de données test et accorde plus de poids aux erreurs les plus importantes, permettant ainsi une évaluation plus précise de la qualité de la prédiction. En d'autres termes, elle évalue la différence quadratique moyenne entre les valeurs imputées et les valeurs réelles manquantes, ce qui permet de mieux évaluer la qualité de l'imputation. Une valeur plus faible de la RMSE est donc indicatrice d'une meilleure performance du modèle d'imputation.

Pour calculer la RMSE, nous avons considéré, pour chaque jeu de données, un échantillon ($\boldsymbol{y}_{\text{entrainement}}, \boldsymbol{y}_{\text{test}}$), chacun des jeux ayant un nombre spécifique d'observations noté respectivement par $N_{i,\text{entrainement}}$ et $N_{i,\text{test}}$. L'indice i représente le i^e jeu de données parmi les 21 jeux créés et présentés précédemment. Nous avons alors calculé la différence entre les valeurs imputées et les valeurs réelles pour chaque observation dans l'échantillon de test. Cette différence est élevée au carré et une moyenne est prise sur l'ensemble des observations dans le jeu de test. Enfin, nous prenons la racine carrée de cette moyenne pour obtenir la RMSE pour chaque modèle et chaque jeu de données. La RMSE est donnée par :

$$\text{RMSE}_i = \sqrt{\frac{1}{N_{i,\text{test}}} \sum_{j=1}^{N_{i,\text{test}}} (y_j^{\text{test}} - y_j^{\text{imputée}})^2}, \tag{5.1}$$

où y_j^{test} est la valeur réelle de l'observation j dans l'ensemble de test, $y_j^{\text{imputée}}$, qui est la valeur imputée de l'observation j. $N_{i,\text{test}}$ est le nombre total d'observations dans l'ensemble de test pour le i^e jeu de données.

En plus de la mesure de performance RMSE, nous nous intéressons au temps nécessaire pour générer les imputations comme un autre facteur important à prendre en compte lors de l'évaluation des modèles. En effet, le temps de calcul peut varier considérablement en fonction de la complexité du modèle et du rang de décomposition. En somme, notre intérêt est de trouver le modèle qui offre les meilleures performances en termes de qualité d'imputation et de temps de calcul.

5.3 Paramétrage des modèles

Pour le modèle Bayesian Probabilistic Matrix Factotization (BPMF), les paramètres $\boldsymbol{U}$ et $\boldsymbol{V}$ ont été initialisés de manière aléatoire pour l'échantillonnage de Gibbs. La précision des erreurs d'observation α^{-1} a été fixée à 0.1 après que nous ayons effectué une recherche par grille sur l'ensemble de valeurs [0.001, 0.01, 0.1, 0.4,

1, 2, 4]. Les paramètres *a priori*, pour les distributions normales de $\boldsymbol{\mu_U}$ et $\boldsymbol{\mu_V}$, ont été fixés à $\mathbf{0}$ pour la moyenne $\boldsymbol{\mu_0}$ et à 1 pour la précision β_0. La matrice d'échelle *a priori*, pour les distributions Wishart de $\boldsymbol{\Lambda_U}$ et $\boldsymbol{\Lambda_V}$, a été fixée à la matrice d'identité $\boldsymbol{W}_0$. Le nombre de degrés de liberté *a priori* pour les distributions Wishart de $\boldsymbol{\Lambda_U}$ et $\boldsymbol{\Lambda_V}$ a été fixé au rang de décomposition K.

Afin de déterminer les paramètres optimaux pour le modèle ARIMA, nous avons employé la fonction *auto_ arima()* de la bibliothèque *pmdarima* de Python. Cette fonction automatise la recherche des meilleurs paramètres en utilisant des critères d'information, tels que l'Akaike Information Criterion (AIC). Dans le processus de sélection des paramètres optimaux pour le modèle ARIMA, une grille de recherche a été mise en place pour explorer différentes combinaisons de p et q. Les ordres maximaux de ces paramètres ont été fixés à 5, ce qui signifie que la fonction *auto_ arima()* a évalué des modèles avec p et q variant de 0 à 5 : les paramètres retenus sont p = 2 , q = 2 et d = 1.

Pour le modèle Temporal Regularized Matrix Factorization (TRMF), une recherche par grille a été effectuée pour trouver les valeurs des hyperparamètres qui minimisent la RMSE. Nous avons exploré un ensemble discret de valeurs pour chaque hyperparamètre dans l'intervalle [0, 1]. Les valeurs optimales des hyperparamètres ont été fixées à : $\eta = 0.091$, $\lambda_w = 0.2728$, $\lambda_x = 0.4286$, et $\lambda_\theta = 0.4286$.

Pour le modèle Bayesian Temporal Matrix Factorization (BTMF), les paramètres incluent : la moyenne *a priori* μ_0 pour la distribution normale de $\boldsymbol{\mu_U}$ fixée à 0, la précision *a priori* β_0 pour la distribution normale de $\boldsymbol{\mu_U}$ fixée à 1, la matrice d'échelle *a priori* $\boldsymbol{\Lambda_U}$ fixée pour la distribution normale de $\boldsymbol{U}_{:i}$, la matrice d'échelle *a priori* $\boldsymbol{W_0}$ fixée pour la distribution Wishart de $\boldsymbol{\Lambda_U}$ à la matrice d'identité $\boldsymbol{I_R}$ et le nombre de degrés de liberté *a priori* v_0 pour la distribution Wishart de $\boldsymbol{\Lambda_U}$ fixé à K. Les paramètres *a priori* pour les distributions normales de $\boldsymbol{A}$ et de $\boldsymbol{\Sigma}$ ont été fixés à des matrices d'identité. Les hyperparamètres α et β, pour la distribution Gamma de τ_i, ont été fixés à 10^{-6}. L'ensemble de décalage $\mathcal{L}$, utilisé dans le modèle autorégressif d'odre d, est fixé à $\mathcal{L} = \{1, 2\}$.

Pour le modèle Kernelized Probabilistic Matrix Factorization (KPMF), une méthode de recherche par grille a été mise en place pour identifier les valeurs optimales des hyperparamètres. Nous avons spécifiquement exploré les paramètres β et γ pour les noyaux de diffusion et le Regularized Laplacian, en testant des valeurs dans l'intervalle $[0, 10^{-10}]$. Les valeurs optimales retenues pour ces paramètres sont $\beta = 0.01$ et $\gamma = 0.1$. De même, différentes combinaisons possibles d'hyperparamètres du modèle ont été testées à partir des ensembles de valeurs prédéfinis : α^{-1} = [0.01, 0.02, 0.03], η = [0.0005, 0.001, 0.0015] et θ = [0.05, 0.1, 0.15]. Les valeurs optimales des hyperparamètres obtenus, déterminés sur la base de la RMSE, sont : $\alpha^{-1} = 0.03$, η = 0.0015 et $\theta = 0.15$. Ces valeurs ont été choisies pour leur efficacité à capturer les structures de covariance non linéaires sous-jacentes à la matrice de données, tout en évitant un surajustement du modèle aux données d'apprentissage. Les paramètres et hyperparamètres spécifiques à chaque modèle sont résumés dans le tableau 5.2.

TABLE 5.2 – Paramètres et hyperparamètres des modèles.

Modèles	Paramètres et hyperparamètres
ARIMA	$p = 2$, $q = 2$, $d = 1$
BPMF	$\boldsymbol{U}$ et $\boldsymbol{V}$ aléatoires, $\alpha = 0.1$, $\mu_0 = 0$, $\beta_0 = 1$, $\boldsymbol{\Lambda_U} = \boldsymbol{\Lambda_V} = \boldsymbol{I}_K$, $v_0 = K$
TRMF	$\eta = 0.091$, $\lambda_w = 0.2728$, $\lambda_x = 0.4286$, $\lambda_\theta = 0.4286$
BTMF	$\mu_0 = 0$, $\beta_0 = 1$, $\boldsymbol{\Lambda_U}$ (Wishart, $\boldsymbol{I}_K$), $v_0 = K$, $\boldsymbol{A}$ et $\boldsymbol{\Sigma}$ ($\boldsymbol{I}_K$), α et β pour $\tau_i = 10^{-6}$, $\mathcal{L} = \{1, 2\}$
KPMF	$\beta = 0.01$, $\gamma = 0.1$, $\alpha^{-1} = 0.03$, $\eta = 0.0015$, $\theta = 0.15$

Enfin, nous avons fixé la phase de préchauffe pour les modèles bayésiens à 1000 itérations et la phase d'échantillonnage à 2000 itérations. Plus de détails seront fournis dans la section convergence des modèles bayésiens.

5.4 Résultats

Dans cette étude, nous avons examiné trois structures de données manquantes distinctes (données manquantes unitaires [DMU], données manquantes par blocs [DMB] et données manquantes mixtes [DMM]) et avons fait varier les pourcentages de données manquantes p de 30 % à 90 %, en augmentant p progressivement par incréments de 10 % pour chacun des trois scénarios. Nous avons testé six modèles d'imputation (imputation par la moyenne [MEAN], modèles ARIMA, Bayesian Probabilistic Matrix Factotization [BPMF], Temporal Regularized Matrix Factorization [TRMF], Bayesian Temporal Matrix Factorization [BTMF] et Kernelized Probabilistic Matrix Factorization [KPMF]) en utilisant différents rangs de décomposition K (5, 8, 10, 20, 30, 40 et 50). À titre de rappel, nous passons brièvement en revue les principales différences entre les modèles étudiés dans la manière de modéliser la structure de la matrice de variance-covariance des données. Le modèle MEAN est un modèle rudimentaire qui ignore toute structure de corrélation présente dans les données. Le modèle ARIMA, en revanche, considère la corrélation temporelle, mais uniquement au niveau des résidus du modèle de régression. Les modèles basés sur la factorisation matricielle, tels que le BPMF, le TRMF, le BTMF et le KPMF, intégrent explicitement une modélisation complexe des structures de variance-covariance. Le BPMF utilise un modèle bayésien hiérarchique en supposant une distribution Gaussienne-Wishart pour les facteurs spatiaux ainsi que pour les facteurs temporels. Le TRMF modélise les dépendances temporelles sous forme d'une régularisation autorégressive et le BTMF étend ce dernier en modélisant plus finement les dépendances spatiales et temporelles. Le BTMF suppose une distribution Gaussienne-Wishart des facteurs spatiaux et une distribution normale multivariée avec un vecteur moyen autorégressif (VAR) pour les facteurs latents temporels. Enfin, le KPMF modélise les matrices de variance-covariance des facteurs latents en utilisant des processus gaussiens. Ces derniers emploient une fonction de noyau qui considère les distances entre les points correspondant aux facteurs latents.

Il est à noter que le modèle Bayesian Temporal Tensor Factorization (BTTF) n'a pas été évalué, étant donné la nature matricielle de nos données. En effet, le BTTF requiert des données sous forme tensorielle. Quant au Bayesian Gaussian Process Probabilistic Matrix Factorization (GPMF), son évaluation a été omise en raison de l'indisponibilité du code source de la part de l'auteur. Nous rappelons également que tous ces modèles ont été initialement définis dans les sections 3.2 à 3.8 du chapitre trois intitulé « Description des modèles ». Ainsi, nous présentons ci-dessous les résultats obtenus pour chacune des sections.

5.4.1 Diagnostique graphique

Cette section est dédiée à l'évaluation graphique de la performance des modèles d'imputation des données manquantes dans le scénario DMM. Nous avons comparé les résultats des six modèles d'imputation à des niveaux de données manquantes de 30 %, 60 % et 90 %, nous avons aussi fixé le rang de décomposition à 30 pour les modèles BPMF, TRMF, BTMF et KPMF.

La figure 5.1 illustre les données manquantes provenant de deux capteurs, identifiés comme capteur 1 et capteur 2, présentant un taux de 30 % de données manquantes. Les performances des divers modèles sont représentées par des courbes de différentes couleurs. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF. Les données réelles sont représentées en noir. Dans cette figure, nous présentons sept segments de séries temporelles comportant des données manquantes, représentant chacun un bloc de 24 heures de données manquantes. Pour le capteur 1, ces segments sont numérotés de B1 à B3, tandis que, pour le capteur 2, ils sont numérotés de B4 à B7.

Dans l'ensemble, et pour les deux capteurs, nous constatons que quatre des modèles, BPMF, TRMF, BTMF et KPMF, suivent la distribution des données ob-

servées, contrairement aux modèles MEAN et ARIMA. Les performances du modèle MEAN sont représentées par une ligne horizontale, ce qui indique que ce modèle ne prend pas en compte la structure spatiotemporelle des données et se contente d'imputer la moyenne des données observées. Cela explique ses performances réduites en comparaison avec les autres modèles. En revanche, le modèle ARIMA semble suivre la distribution des données lorsque la structure des données manquantes est unitaire, à l'exception des blocs B1 à B7. Cette différence dans la performance peut être expliquée par le fait que les blocs B1 à B7 sont composés de données manquantes en série. Celles-ci semblent plus difficilement imputables comparé aux données manquantes unitaires, pour lesquelles le modèle a accès à plus d'information contextuelle. Il convient également de noter la présence de données extrêmes, avec des valeurs inférieures à 20 mph pour le capteur 2, ce qui peut potentiellement affecter la performance de tous les modèles d'imputation.

Ainsi, pour approfondir notre analyse, nous commençons par un zoom sur les trois blocs de données manquantes du capteur 1. La figure 5.2 illustre trois graphiques, chacun représentant un bloc de 24 heures de données manquantes, classés du haut vers le bas. Le premier graphique représente le bloc de données manquantes B1, tandis que le dernier graphique correspond au bloc de données manquantes B3.

Dans le premier bloc, nous observons que le modèle MEAN maintient une performance constante, tandis que d'autres modèles, comme le BPMF, le TRMF, le BTMF et le KPMF, présentent des performances comparables aux données observées. Cependant, les modèles BPMF et BTMF ont montré une légère surévaluation, avec des valeurs élevées, comme observées pour les points 22 et 25. L'analyse du deuxième bloc a révélé une variabilité accrue pour les modèles BPMF, TRMF et BTMF, comme observé aux points 64, 65, 73 et 78. En outre, le troisième bloc a mis en évidence que KPMF et le BTMF s'alignent étroitement avec les données observées, suggérant leur pertinence pour ces données.

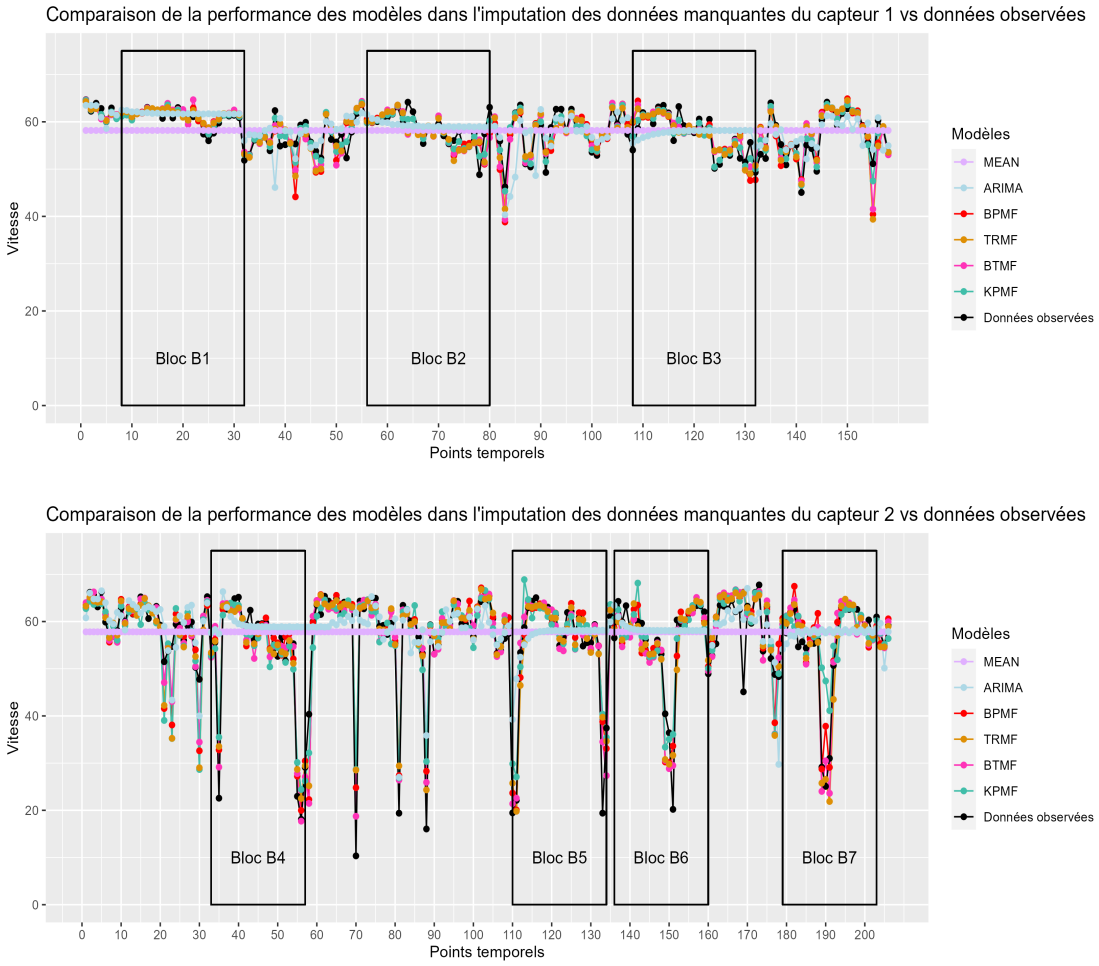

FIGURE 5.1 – Comparaison de la performance des modèles d'imputation sur les données manquantes pour les capteurs 1 et 2 dans le cas du scénario DMM. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF. Les données réelles sont représentées en noir.

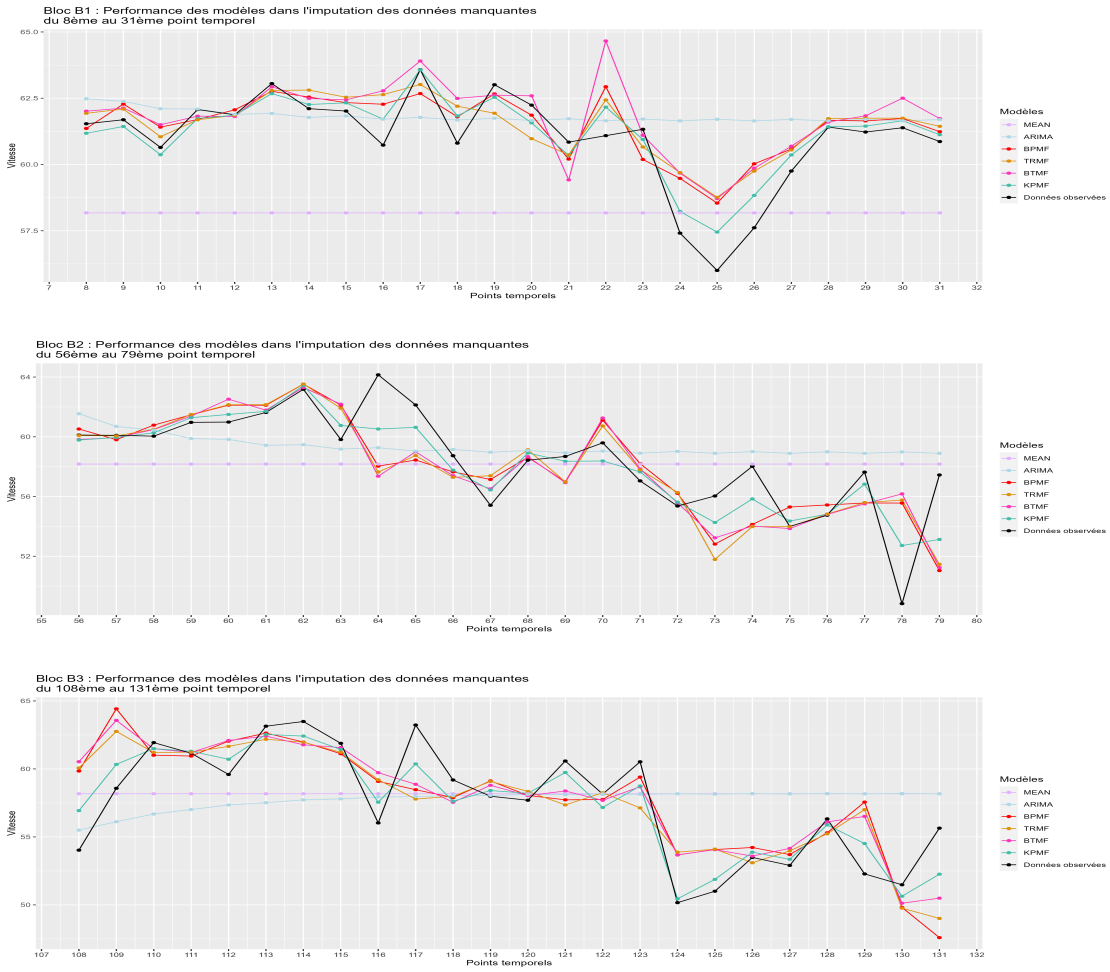

FIGURE 5.2 – Imputation des valeurs manquantes dans les blocs de 24 heures avec le scénario DMM et 30 % de valeurs manquantes : visualisation des blocs B1 à B3 classés par ordre croissant, un zoom sur les trois blocs de données manquantes du capteur 1 de la figure 5.1. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF. Les données réelles sont représentées en noir.

Nous poursuivons notre analyse en effectuant un zoom sur les quatre blocs de données manquantes du capteur 2 illustré dans la figure 5.3. Dans le quatrième bloc, il est observé que le modèle ARIMA présente une légère surévaluation de la moyenne par rapport au capteur 1. Cependant, le modèle BTMF se distingue en se rapprochant le plus des données observées, notamment aux points 35, 36, 37, 41, 46 et 48. Dans le cinquième bloc, tant le KPMF que le BTMF maintiennent une performance constante, à l'exception de quelques difficultés du KPMF à imputer des valeurs extrêmes, comme observée aux points 110 et 111. Le sixième bloc montre que les modèles BPMF et TRMF affichent des performances similaires, tandis que le KPMF continue à présenter une performance stable et des difficultés avec les valeurs extrêmes. Enfin, dans le septième bloc, les modèles BPMF, TRMF et BTMF maintiennent une précision soutenue dans leurs prédictions, observé aux points 183, 185, 198 et 201. Le KPMF présente encore quelques défis dans l'imputation des valeurs extrêmes, observé aux points 189, 190 et 191, comparativement au BTMF.

Nous avons noté une similitude dans les performances des modèles BTMF et KPMF lors de notre analyse. Cependant, le KPMF semble éprouver des difficultés plus marquées à imputer les valeurs manquantes du capteur 2 par rapport au capteur 1, surtout en présence de valeurs extrêmes.

Dans le but de mieux comprendre ces résultats, nous avons cherché à étudier les caractéristiques des données des deux capteurs. Pour ce faire, nous avons calculé l'écart-type des données observées pour chaque capteur et avons constaté une différence importante entre les deux capteurs. En effet, l'écart-type de la série temporelle est de 4 mph pour le capteur 1, tandis qu'il est de 10 mph pour le capteur 2. Ainsi, le capteur 2 a un écart-type 2.5 fois plus élevé que celui du capteur 1, ce qui se traduit par une présence accrue de valeurs extrêmes. Cette différence significative dans la variation des données entre les deux capteurs suggère que la performance du KPMF est fortement impactée par la variance des données interne pour chaque capteur pris individuellement.

En mettant de l'avant ces observations, le choix du modèle approprié semble

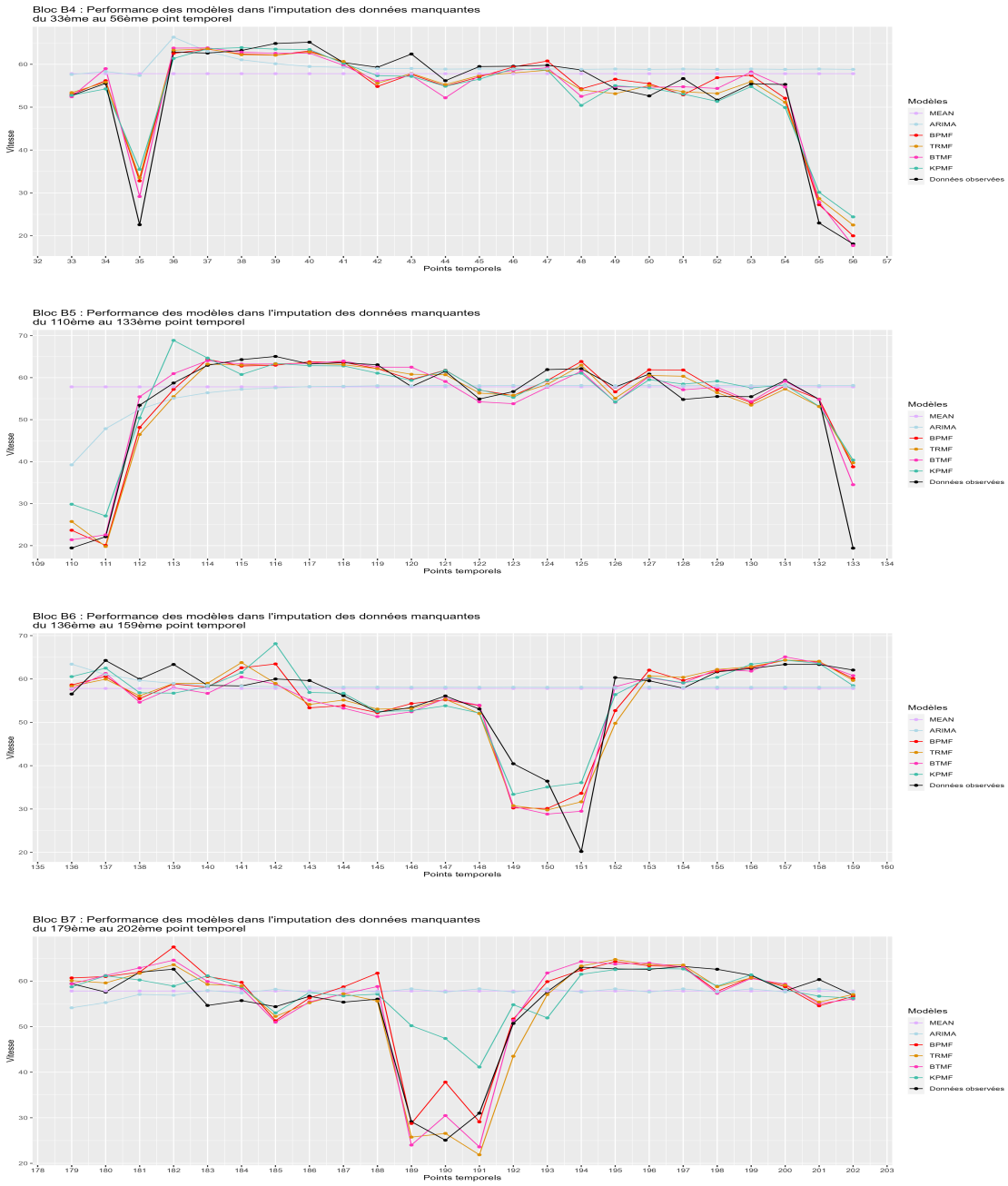

FIGURE 5.3 – Imputation des valeurs manquantes dans les blocs de 24 heures avec le scénario DMM et 30 % de valeurs manquantes : visualisation des blocs B4 à B7 classés par ordre croissant, , un zoom sur les trois blocs de données manquantes du capteur 2 de la figure 5.2. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF. Les données réelles sont représentées en noir.

dépendre, soit de la stabilité de l'imputation, soit de la qualité de l'imputation par rapport aux données observées. Le modèle KPMF pourrait potentiellement être fiable, en raison de sa stabilité à travers les blocs d'analyse, tandis que BTMF présente une bonne performance, près des données réelles. En revanche, les modèles BPMF et TRMF pourraient poser certains défis pour des tâches nécessitant une grande précision.

Nous voulons approfondir notre analyse quant à la qualité de l'imputation des données manquantes pour l'ensemble des modèles, spécifiquement dans le contexte du scénario DMM, et ce, pour l'intégralité des données imputées. Nous avons fixé le rang de décomposition à 30 et analysé les résultats d'imputation sur les jeux de données comportant un taux de données manquantes suivant : 30 %, 60 % et 90 %. Nous souhaitons comparer la performance des six modèles pour évaluer la densité de la différence entre les valeurs imputées et les valeurs réelles pour l'ensemble des données. Les graphiques de densité représentent l'écart entre les valeurs imputées et les valeurs réelles observées pour les différents modèles d'imputation de données. Cet écart est calculé comme suit : Valeur imputée - Valeur réelle. Nous utilisons *ggplot()* de R pour visualiser la distribution de ces écarts pour chaque modèle.

Nous nous attendons à ce qu'une distribution de la densité de la différence centrée sur zéro indique que le modèle a tendance à imputer des valeurs proches des données réelles. Une distribution de la différence plus pointue et plus élevée peut indiquer que le modèle impute des valeurs plus précises, avec des erreurs plus petites et moins fréquentes. Cependant, si l'objectif est de minimiser l'impact des valeurs extrêmes, alors la distribution qui a une hauteur plus basse, mais une queue plus large serait préférable, car elle indique que les erreurs extrêmes sont moins fréquentes.

La figure 5.4 illustre la densité des erreurs qui représente la différence entre les valeurs observées et les valeurs imputées pour les six modèles d'imputations. Ces résultats montrent que les distributions pour les modèles MEAN et ARIMA ne sont pas centrées sur zéro, ce qui indique que ces modèles ont tendance à produire des biais systématiques. Autrement dit, ces modèles ont tendance à imputer des valeurs loin

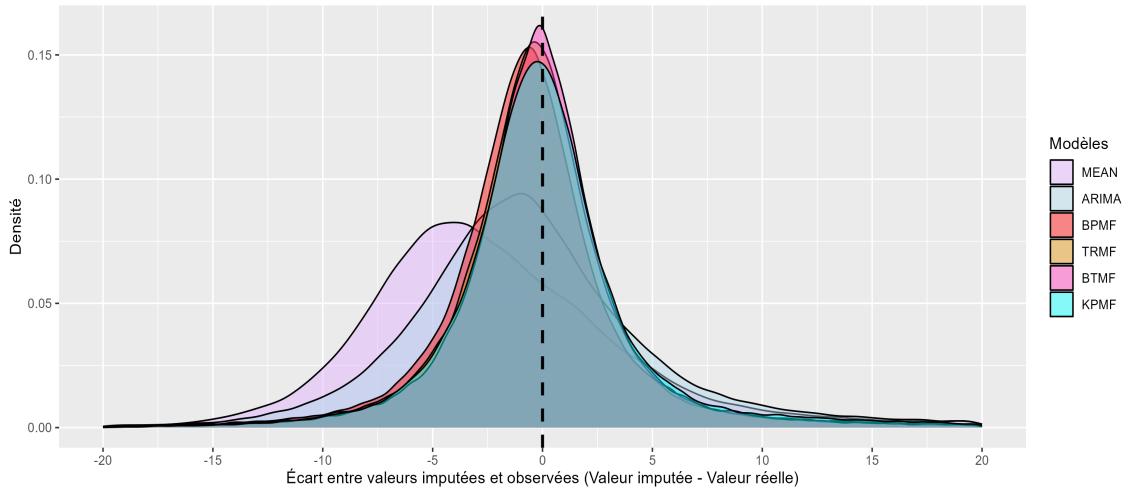

FIGURE 5.4 – Comparaison de la performance des modèles d'imputation sur les données manquantes dans le cas du scénario DMM avec 30 % de données manquantes. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

des données réelles. Toutefois, la distribution des erreurs du modèle ARIMA est plus proche de zéro que celle du modèle MEAN, ce qui implique que le modèle ARIMA est plus précis que le modèle MEAN. En revanche, le modèle BPMF présente une distribution symétrique, ce qui indique que les erreurs ont été réparties de manière uniforme autour de sa moyenne. Cependant, la moyenne de ses erreurs est un tout petit peu inférieure à zéro, ce qui implique que le modèle BPMF a tendance à sous-estimer les données manquantes. La distribution des erreurs du TRMF est presque similaire à celle du BPMF, mais elle est plus haute et plus proche du zéro, ce qui suggère que le TRMF a une meilleure précision que le BPMF. La distribution des erreurs pour le BTMF et le KPMF semblent très légèrement asymétriques vers la gauche, mais celle du BTMF est plus haute que celle du KPMF et plus centrée sur zéro. Cela indique que le BTMF est capable d'imputer des valeurs plus précises que le KPMF, mais il peut également produire des erreurs extrêmes plus fréquemment que ce dernier.

En augmentant le taux de données manquantes à 60 %, nous constatons, dans la figure 5.5, que la distribution des erreurs des modèles BPMF, TRMF, BTMF et KPMF sont toutes moins hautes que celles pour le taux de 30 % de données manquantes. Cela indique que les modèles ont plus de difficulté à imputer des données manquantes avec précision. Les résultats que montrent cette figure sont presque similaires à la figure précédente. Cependant, en comparant la distribution des erreurs entre le BTMF et le KPMF, on remarque que les distributions du KMPF et du BTMF sont légèrement moins centrées sur zéro. En outre, la distribution des erreurs du modèle BTMF est centrée autour d'une valeur légèrement négative proche de zéro et présente une hauteur plus élevée par rapport à celle du modèle KPMF. Cette caractéristique indique que le BTMF a potentiellement une meilleure aptitude à imputer avec précision les valeurs manquantes, comparé au KPMF. Cependant, la plus grande hauteur de la distribution des erreurs du BTMF suggère également une occurrence plus fréquente d'erreurs importantes, comparée au KPMF.

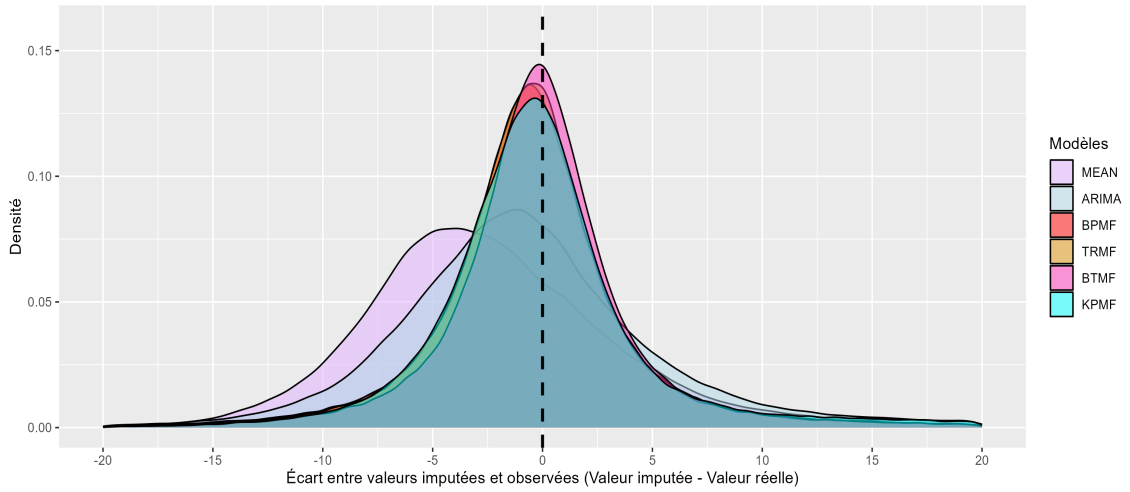

FIGURE 5.5 – Comparaison de la performance des modèles d'imputation sur les données manquantes dans le cas du scénario DMM avec 60 % de données manquantes. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

En augmentant le taux de données manquantes à 90 %, il est évident que l'imputation des valeurs manquantes devient plus difficile pour tous les modèles. Les résultats de la figure 5.6 montrent que la distribution des erreurs a encore baissé : plus que la distribution pour le taux de 60 % de données manquantes. De plus, on remarque que le KPMF est nettement éloigné de zéro : le graphique indique une performance beaucoup moins satisfaisante par rapport aux modèles BPMF, TRMF et BTMF. Le BTMF semble avoir une distribution des erreurs symétrique et plus haute que les deux autres modèles, ce qui suggère que l'imputation des données manquantes est plus précise. En revanche, la distribution des erreurs pour les modèles BPMF et TRMF sont un peu moins centrées sur zéro et asymétriques vers la gauche. Il est donc possible de conclure que le BTMF est le meilleur modèle pour l'imputation de données manquantes dans ce scénario.

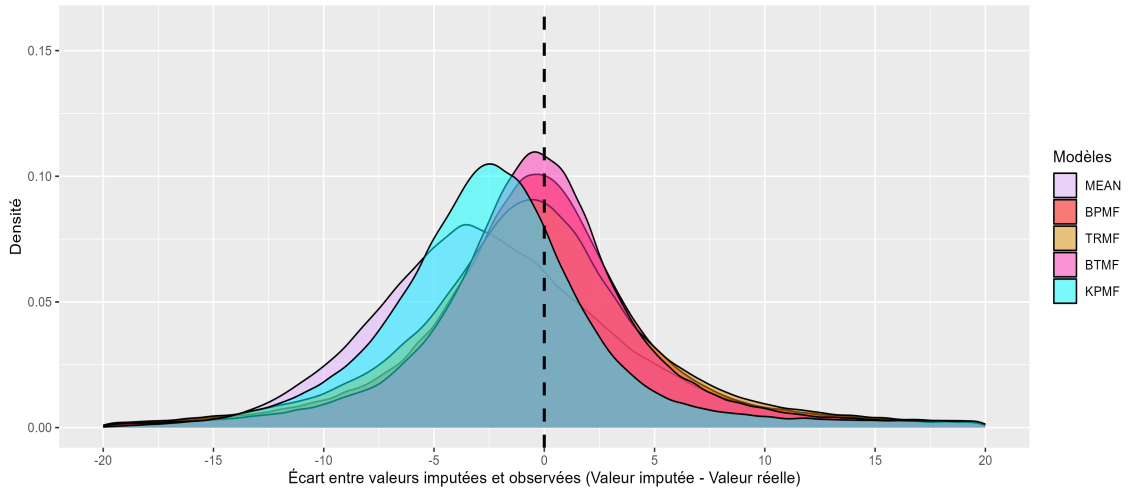

FIGURE 5.6 – Comparaison de la performance des modèles d'imputation sur les données manquantes dans le cas du scénario DMM avec 90 % de données manquantes. La courbe mauve clair correspond au modèle MEAN, la courbe bleu pâle représente le modèle ARIMA, la courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

En somme, le modèle BTMF se distingue par une distribution de la différence plus pointue que celle du KPMF, indépendamment du taux de données manquantes.

Cette caractéristique indique que le BTMF est généralement plus précis dans l'imputation des valeurs manquantes par rapport au KPMF. En revanche, le KPMF présente une distribution des erreurs plus basse avec une queue plus large, ce qui suggère une fréquence réduite d'erreurs extrêmes en comparaison avec le BTMF. De plus, en termes de performance, le BTMF impute des valeurs plus précises que le KPMF dans des situations extrêmes, telles que lorsque le taux de données manquantes atteint 90%.

À la lumière de ces observations, indiquant des niveaux de précision et de fréquence d'erreurs distinctes, nous souhaitons approfondir notre analyse en entreprenant un diagnostique numérique de ces distributions.

5.4.2 Diagnostique numérique

Dans cette section, nous nous intéressons à l'analyse des statistiques descriptives des performances des six modèles étudiés. Notre évaluation de la performance se base sur la mesure de l'écart entre les valeurs imputées par ces modèles et les valeurs réelles que nous avons présentées précédemment. Trois niveaux de pourcentage de données manquantes ont été étudiés : 30%, 60% et 90% avec un rang de décomposition fixé à 30 dans le cas du scénario DMM.

Dans le cadre de cette étude sur l'imputation de données, on s'attend à ce que l'augmentation du taux de données manquantes entraine une plus grande variabilité dans les valeurs imputées, mesurée par l'écart-type (σ). Cette hypothèse est basée sur l'idée que des niveaux élevés de données manquantes réduisent la quantité d'informations disponibles pour l'imputation, ce qui pourrait rendre le processus plus incertain et, donc, potentiellement plus variable. Par ailleurs, nous nous attendons à ce que l'amplitude de la distribution des erreurs soit plus étendue pour le modèle BTMF que pour le KPMF, une observation émise lors du diagnostique graphique de la section précédente.

Le tableau 5.3 présente des statistiques descriptives de l'écart entre les valeurs

imputées et les valeurs observées pour les modèles d'imputation étudiés dans le cas du scénario DMM avec le taux de données manquantes de 30 %, 60 % et 90 %. Chaque scénario indique le pourcentage de données manquantes et le modèle d'imputation utilisé pour imputer les données manquantes. Les statistiques descriptives de chaque modèle d'imputation sont fournies : la taille de l'échantillon des données imputées, l'écart-type (σ), le minimum, le premier quartile (Q1), la moyenne (μ), la médiane, le troisième quartile (Q3) et le maximum.

TABLE 5.3 – Tableau des statistiques descriptives de l'écart entre les valeurs imputées et les valeurs observées pour chaque modèle d'imputation dans le cas du scénario DMM. Les statistiques descriptives de chaque modèle d'imputation sont fournies : la taille de l'échantillon des données imputées, l'écart-type (σ), le minimum, le premier quartile (Q1), la moyenne (μ), la médiane, le troisième quartile (Q3) et le maximum.

DM	Modèles	Taille	σ	Minimum	Q1	μ	Médiane	Q3	Maximum
30 %	MEAN	65 122	10.75	-21.54	-5.72	0.14	-2.42	2.00	59.93
	ARIMA	65 122	10.34	-70.27	-3.85	0.79	-0.80	2.80	61.15
	BPMF	65 122	4.91	-46.60	-2.56	-0.49	-0.65	1.23	56.08
	TRMF	65 122	4.84	-44.51	-2.17	-0.09	-0.31	1.55	56.25
	BTMF	65 122	4.79	-41.88	-1.97	0.02	-0.16	1.64	57.32
	KPMF	65 122	5.21	-37.58	-2.02	0.03	-0.13	1.85	56.97
60 %	MEAN	130 237	10.79	-22.80	-5.86	0.13	-2.43	1.96	59.96
	ARIMA	130 237	11.10	-70.87	-4.22	0.79	-0.86	3.04	62.47
	BPMF	130 237	5.63	-69.11	-2.70	-0.38	-0.56	1.59	59.66
	TRMF	130 237	5.38	-46.01	-2.65	-0.30	-0.52	1.56	65.09
	BTMF	130 237	5.35	-52.94	-2.21	0.00	-0.19	1.82	59.21
	KPMF	130 237	6.24	-41.38	-2.58	-0.05	-0.27	1.89	58.39
90 %	MEAN	195 365	10.95	-24.87	-5.54	0.37	-2.05	2.43	62.27
	ARIMA	195 365	-	-	-	-	-	-	-
	BPMF	195 365	7.54	-62.21	-3.25	0.00	-0.30	2.61	69.18
	TRMF	195 365	7.80	-67.41	-3.85	-0.21	-0.57	2.69	90.53
	BTMF	195 365	7.07	-64.25	-2.96	0.04	-0.29	2.37	66.32
	KPMF	195 365	8.33	-33.09	-4.98	-0.87	-2.21	0.68	58.13

L'analyse des statistiques descriptives révèle que, dans le cas du scénario avec 30 % de données manquantes, le modèle MEAN semble moins fiable en raison de son écart-type élevé et de sa moyenne éloignée de zéro. Le modèle ARIMA, avec une moyenne de 0.79, manifeste un biais systématique en faveur de la surestimation des valeurs. À l'inverse, les modèles BPMF et TRMF, indiquent un biais en

direction de la sous-estimation, bien que celui du TRMF soit moins important. Par ailleurs, le modèle ARIMA présente la valeur minimale la plus basse (-70.27) et la valeur maximale la plus haute (61.15), ce qui pourrait suggérer la présence d'erreurs d'imputation extrêmes. Pour les modèles BTMF et KPMF, la proximité de leurs moyennes à zéro suggère une meilleure précision globale en termes d'imputation. De plus, ils se démarquent également en termes de fiabilité, comme en témoignent leurs valeurs de Q1 et Q3, qui sont plus proches de zéro par rapport aux autres modèles. Nous notons également que l'étendue des valeurs minimales et maximales est plus grande dans le modèle BTMF que dans le modèle KPMF, conformément à nos attentes.

En somme, le modèle MEAN ne semble pas fiable. Les modèles ARIMA et BPMF semblent avoir des biais systématiques, comme le montre leur moyenne éloignée de zéro. Le TRMF démontre une performance relativement satisfaisante. Les modèles BTMF et KPMF semblent offrir un bon équilibre entre précision, moyenne proche de zéro, fiabilité, écart-type plus faible et quartiles proches de zéro.

Dans le scénario à 60 %, le biais de surestimation reste presque inchangé entre les deux cas (0.79 contre 0.79) pour ARIMA, alors que le biais de sous-estimation augmente de manière significative dans le cas à 60 % (-0.09 contre -0.30) pour le TRMF. Les modèles BTMF et KPMF conservent leur proximité à zéro, indiquant une bonne précision. La distribution des erreurs semble globalement similaire entre les deux cas, bien que les quartiles indiquent généralement une légère augmentation de la dispersion des erreurs dans le cas à 60 %. Par ailleurs, la variabilité générale des erreurs d'imputation semble augmenter légèrement avec l'augmentation du pourcentage de données manquantes pour la plupart des modèles.

En somme, la présence accrue de données manquantes (60 % ou 30 %) n'affecte pas significativement le biais systématique des modèles, mais semble entrainer une légère augmentation de la variabilité des erreurs d'imputation, sauf pour le TRMF qui a connu une baisse significative de la moyenne, suggérant un biais vers la sous-estimation des valeurs. Les modèles BTMF et KPMF demeurent les plus stables en

termes de précision et de fiabilité à travers les deux scénarios. Le même constat est également fait en ce qui concerne l'amplitude des valeurs minimales et maximales pour les modèles BTMF et KPMF.

L'analyse des données avec 90 % de valeurs manquantes montre les quartiles Q1 et Q3 sont légèrement plus éloignés de zéro pour tous les modèles, indiquant une distribution plus large des erreurs. De fait, la variabilité a généralement augmenté pour tous les modèles, mais plus notablement pour le KPMF et le TRMF. Le KPMF montre une moyenne négative (-0.87), ce qui suggère un biais significatif vers la sous-estimation des valeurs. Ce biais est nettement plus grand par rapport au cas de 60 % où la moyenne était de -0.05. L'écart-type est de 8.33, reflétant une grande variabilité dans les erreurs d'imputation. Le BTMF, avec sa moyenne de 0.04, montre un biais beaucoup moins marqué, s'approchant presque de l'idéal de zéro. L'écart-type (7.07) est légèrement inférieur à celui du KPMF (8.33), ce qui pourrait indiquer une meilleure cohérence dans les erreurs d'imputation en faveur du BTMF. La médiane du BTMF est beaucoup plus proche de zéro (-0.29), ce qui peut indiquer une meilleure centralité des erreurs autour de zéro. Par ailleurs, le modèle ARIMA n'a pas pu produire de résultat dans le scénario avec 90 % de données manquantes, ce qui suggère que cette méthode pourrait ne pas être adaptée aux situations avec un pourcentage élevé de données manquantes.

Dans l'analyse des différents modèles d'imputation de données, et en fonction du taux de données manquantes, le MEAN présente une augmentation modérée de la moyenne et une légère augmentation de l'écart-type sur la même plage. Pour le modèle ARIMA, on constate une stabilité de la moyenne entre les scénarios à 30 % et à 60 % de données manquantes. Nous constatons aussi une légère augmentation de l'écart-type, sans variation importante dans les autres quantiles. Le modèle ARIMA n'a cependant pas pu produire de résultat dans le scénario avec 90 % de données manquantes. Quant au modèle TRMF, celui-ci montre une baisse de la moyenne et une augmentation significative de l'écart-type, avec des variations marquées dans les quantiles, surtout pour la valeur maximale à 90 % de données manquantes. En ce

qui concerne le modèle BPMF, nous constatons une augmentation significative de la moyenne et de l'écart-type entre 30 % et 90 % de données manquantes, avec des fluctuations dans les quantiles. Le modèle BTMF présente une moyenne proche de zéro à travers tous les scénarios et une augmentation de l'écart-type de 4.79 à 7.07. Le modèle KPMF montre une forte augmentation de la moyenne, passant de 0.03 à -0.87, et une augmentation significative de l'écart-type de 5.21 à 8.33, suggérant qu'il n'est pas adapté lors de scénarios extrêmes d'absence de données. Enfin, le BTMF semble offrir une meilleure performance en termes de biais et de centralité des erreurs d'imputation. Cependant, il présente des valeurs minimales et maximales plus extrêmes, ce qui pourrait indiquer des vulnérabilités aux valeurs aberrantes.

Dans notre analyse des statistiques descriptives de l'imputation de données manquantes, nous avions anticipé que l'augmentation du taux de données manquantes se traduirait par une plus grande variabilité dans les valeurs imputées et cette hypothèse a été confirmée par nos résultats. En effet, pour les modèles BPMF, TRMF, BTMF et KPMF, l'écart-type varie entre 4.79 et 5.21 pour le scénario à 30 % de données manquantes, entre 5.35 et 6.24 pour celui à 60 % de données manquantes et entre 7.07 et 8.33 pour le scénario à 90 % de données manquantes. Par ailleurs, le modèle BTMF se distingue par une distribution de la différence plus pointue que le KPMF, quelle que soit la proportion de données manquantes. Cela se traduit par une imputation de valeurs plus précises avec moins d'erreurs fréquentes. En revanche, le KPMF présente une distribution plus basse avec une queue plus large, indiquant une fréquence moindre d'erreurs extrêmes en comparaison avec le BTMF. Toutefois, le BTMF domine le KPMF en termes de performance, surtout dans le cas extrême où le taux de données manquantes atteint 90 %. Le KPMF ne semble donc pas être adapté à ce scénario extrême.

5.4.3 Effet du rang de décomposition sur la RMSE

Nous avons évalué les performances de quatre modèles (BPMF, TRMF, BTMF et KPMF) en utilisant sept rangs de décomposition différents (5, 8, 10, 20, 30, 40 et 50) dans trois scénarios (DMU, DMB et DMM), en présence de données manquantes avec des taux variés de données manquantes (30 %, 60 % et 90 %). Cette section met en évidence l'impact du choix du rang de décomposition sur la qualité de l'imputation, en termes d'erreur quadratique moyenne (RMSE).

Dans cette analyse de l'effet du rang de décomposition sur la RMSE, nous anticipons qu'à mesure que le rang de décomposition augmente, les modèles devraient être en mesure de capturer davantage d'informations contenues dans les données, ce qui, en principe, devrait conduire à une réduction de la RMSE. En d'autres termes, un rang de décomposition plus élevé devrait permettre aux modèles de mieux représenter la complexité des données, ce qui se traduirait par des imputations plus précises et, par conséquent, une RMSE plus faible. À l'inverse, une réduction du rang de décomposition devrait entrainer une perte d'information, ce qui pourrait se traduire par une augmentation de la RMSE. En effet, dans ce cas, les modèles auraient une capacité réduite à saisir la structure spatiotemporelle des données.

La figure 5.7 montre l'effet du rang de décomposition sur la RMSE pour un pourcentage fixe de données manquantes à 30 %, selon les différents scénarios, DMU, DMB et DMM, présentés de gauche à droite. Les résultats montrent une relation inverse entre le rang de décomposition et la RMSE pour les quatre modèles. En effet, plus le rang de décomposition augmente, plus la RMSE diminue. Toutefois, les modèles MEAN et ARIMA ne sont pas affectés par le choix du rang de décomposition et sont représentés par des lignes horizontales dans les graphiques ($\text{RMSE}_{\text{MEAN}}$ = (0.1071,0.1085,0.1074) et ($\text{RMSE}_{\text{ARIMA}}$ = (0.0912,0.1156,0.1037)), qui ont été omises pour des raisons de lisibilité.

Le modèle KPMF est le plus performant, quel que soit le rang de décomposition ou le scénario, tandis que le modèle BPMF est le moins performant dans ces mêmes

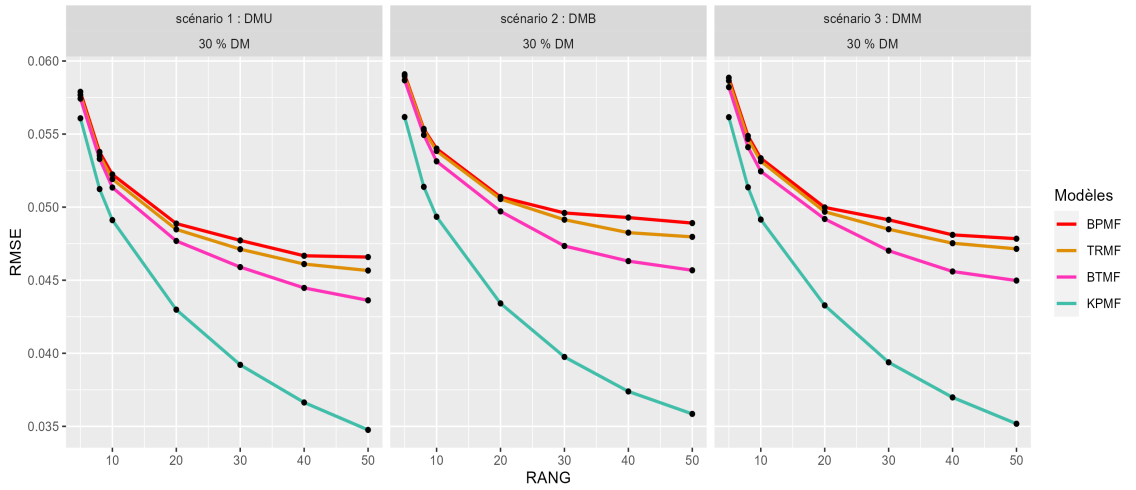

FIGURE 5.7 – Effet du rang de décomposition sur la RMSE pour les trois scénarios avec 30 % de données manquantes. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

deux cas. Pour des rangs de décomposition inférieurs à 10, les modèles BPMF, TRMF et BTMF ont des performances similaires, avec une légère supériorité pour le BTMF. Cependant, pour un rang de décomposition supérieur à 10, le TRMF devient plus performant que le BPMF, et le BTMF offre de meilleures performances que le TRMF. De plus, la RMSE est plus élevée pour le scénario DMB et plus faible pour le scénario DMU.

En augmentant le pourcentage de données manquantes à 60 % et en observant l'effet du rang de décomposition sur la RMSE dans la figure 5.8, nous constatons que le KPMF reste le modèle le plus performant, quel que soit le rang de décomposition ou le scénario, tandis que le BPMF demeure le moins performant. Les performances des modèles BPMF, TRMF et BTMF sont relativement proches lorsque le rang de décomposition est inférieur à 10. Cependant, pour un rang de décomposition supérieur à 10, les performances des modèles TRMF et BTMF sont nettement supérieures à celles du BPMF.

Nous remarquons également que, dans l'ensemble, la RMSE diminue à mesure que le rang de décomposition augmente. Cependant, cette constatation n'est pas

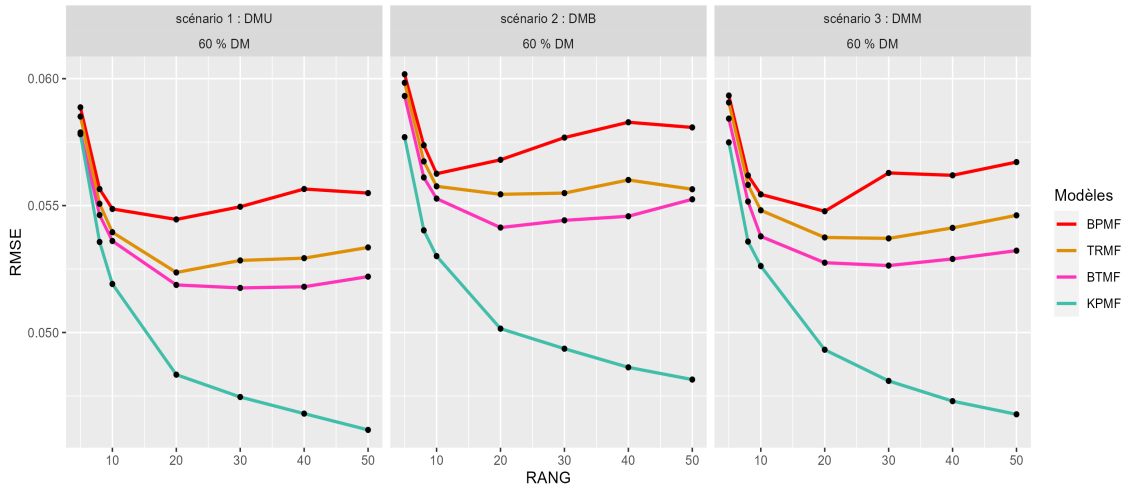

FIGURE 5.8 – Effet du rang de décomposition sur la RMSE pour les trois scénarios avec 60 % de données manquantes. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

vraie pour tous les modèles. En effet, nous remarquons qu'à partir d'une certaine valeur du rang de décomposition, dans ce cas le rang de décomposition égale à 20, la RMSE augmente lorsque le rang de décomposition augmente pour les modèles BPMF, TRMF et BTMF. Or, pour les modèles TRMF et BTMF, la RMSE augmente également avec le rang de décomposition, mais de manière moins prononcée. Pour le KPMF, cette relation inverse reste vraie, quel que soit le rang de décomposition. Cette observation met en évidence la sensibilité des performances des modèles, plus particulièrement des modèles KPMF et BTMF, aux variations du rang de décomposition et au pourcentage de données manquantes : le modèle KPMF se démarquant comme le plus performant sous les diverses valeurs du rang de décomposition. Ainsi, il est clair que l'augmentation du pourcentage de données manquantes a un impact sur l'effet du rang de décomposition sur la RMSE. Cependant, cet impact n'est pas égal sur tous les modèles. De fait, nous observons que la RMSE diminue à mesure que le rang de décomposition augmente. Par contre, les modèles BPMF, TRMF et BTMF présentent une augmentation de la RMSE à partir d'une certaine valeur du rang de décomposition, qui varie selon le modèle

et le scénario étudié. Pour les modèles BTMF et TRMF, le coude est atteint à un rang de décomposition de 20 dans tous les scénarios. Cela signifie que l'ajout de facteurs de décomposition au-delà de 20 n'améliore pas de manière significative les performances du modèle. En revanche, pour le modèle BPMF, le coude est atteint à un rang de décomposition de 20 dans les scénarios DMU et DMM, tandis qu'il est atteint à un rang de décomposition de 10 dans le scénario DMB. Ces résultats suggèrent aussi que les modèles BPMF et TRMF sont plus sensibles à ces conditions que les modèles BTMF et TRMF.

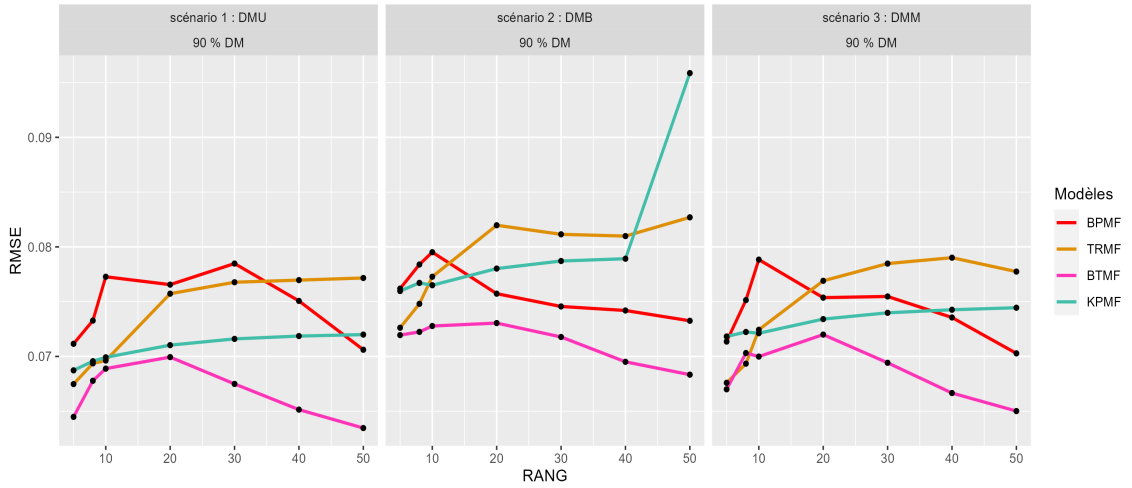

FIGURE 5.9 – Effet du rang de décomposition sur la RMSE pour les trois scénarios avec 90 % de données manquantes. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

En augmentant le pourcentage de données manquantes à 90 %, nous observons, dans la figure 5.9, des changements significatifs dans les performances des différents modèles. Le BTMF est le modèle le plus performant, quels que soient la valeur du rang de décomposition et le scénario, tandis que le modèle le moins performant varie en fonction du scénario et de la valeur du rang de décomposition. Les courbes de la RMSE pour les modèles BPMF et BTMF prennent une forme de U inversé, indiquant des fluctuations dans leurs performances. Les points d'inflexion observés dans les courbes de la RMSE, correspondant à des rangs de décomposition spécifiques, K=10

pour le BPMF et K=20 pour le BTMF, pourraient indiquer des valeurs de rang où les modèles réussissent à capturer des caractéristiques intrinsèques des données. En revanche, les courbes de la RMSE pour les modèles TRMF et KPMF suivent une forme logarithmique. La trajectoire croissante de la courbe logarithmique associée au modèle TRMF indique une sensibilité accrue du modèle au choix du rang de décomposition. En revanche, la pente plus modérée de la courbe associée au modèle KPMF suggère une robustesse et une plus grande latitude dans la sélection du rang optimal.

Dans le contexte de données manquantes à hauteur de 90 %, tous les modèles semblent éprouver des difficultés à imputer les données avec précision. Cela souligne l'impact significatif qu'un taux élevé de données manquantes peut avoir sur la capacité des modèles d'imputation à fournir des résultats fiables. Cependant, il est à noter que les valeurs les plus basses de RMSE sont obtenues lorsque la valeur du rang de décomposition est faible (K=5) pour les modèles TRMF et KPMF. Ainsi, il semble que, lorsque les données manquantes sont extrêmes, le rang de décomposition et la RMSE augmentent pour ces modèles. En revanche, pour les modèles BPMF et BTMF, la RMSE diminue au-delà d'un certain pic. Par ailleurs, à un taux de données manquantes extrêmes, la performance des modèles BPMF et BTMF s'améliore significativement avec un rang de décomposition élevé, soit un rang supérieur à 40 dans le cas des scénarios DMU, DMB et DMM.

Dans les trois scénarios de données manquantes (30 %, 60 % et 90 %), nous constatatons des différences notables dans la performance des modèles. En général, ils ont du mal avec le scénario DMB, où la RMSE est plus élevée. Cela indique des difficultés à combler les données manquantes qui se présentent sous la structure DMB. En comparaison, les modèles se débrouillent mieux avec le scénario DMU, où la RMSE est plus faible. Pour le scénario DMM, les résultats se situent entre les deux autres scénarios, avec des performances intermédiaires.

En somme, l'évaluation de quatre modèles d'imputation (BPMF, TRMF, BTMF et KPMF) a permis de constater que le choix du rang de la factorisation matricielle

est un paramètre décisif pour la qualité de l'imputation en termes de RMSE, surtout lorsqu'il s'agit de pourcentages élevés de données manquantes (90 %). Théoriquement, les modèles devraient être en mesure de capturer davantage d'information dans les données à mesure que le rang de décomposition augmente, ce qui devrait, en principe, conduire à une réduction de la RMSE. Inversement, une réduction du rang de décomposition entrainerait une perte d'information, et donc, une détérioration de la RMSE. Dans notre analyse, pour des taux faibles de données manquantes (30 %), la relation entre la RMSE et le rang de décomposition est inverse : une augmentation du rang de décomposition conduit à une diminution de la RMSE, quelle que soit la structure des données manquantes (scénarios). Cependant, nos observations révèlent une relation plus complexe lorsque le taux de données manquantes est élevé. Bien que cette relation soit effectivement linéaire pour des taux plus faibles de données manquantes, elle devient non linéaire lorsque le taux de données manquantes augmente. Nous avons observé que la RMSE affiche une courbe en forme de « U » dans le scénario où 60 % de données sont manquantes pour les modèles BPMF, TRMF et BTMF, et une forme de « U inversé » dans le scénario avec 90 % de données manquantes pour ces mêmes modèles. Ces observations sont en contradiction avec les attentes théoriques, qui postulent que la RMSE devrait diminuer de manière continue à mesure que le rang de décomposition augmente. Ce comportement inattendu en présence de taux élevés de données manquantes demeure non expliqué et suscite des discussions pour de futures investigations. Plusieurs hypothèses pourraient expliquer cette lacune : les modèles s'appuient de manière disproportionnée sur le faible ensemble de données disponibles pour ajuster leurs paramètres ou l'effet négatif des données manquantes surpasse l'effet bénéfique de l'augmentation du rang de décomposition, ce qui entraine des résultats imprévisibles. Par ailleurs, on constate que le KPMF excelle dans la réduction des erreurs extrêmes, ce qui se manifeste par une RMSE globalement plus basse, pour les ensembles de données avec 30% et 60% de valeurs manquantes. Inversement, le modèle Bayesian Temporal Matrix Factorization (BTMF) présente une RMSE légèrement plus élevée, conséquence de sa plus

grande sensibilité aux valeurs extrêmes. Cependant, dans des contextes où le taux de données manquantes s'élève à 90%, le BTMF montre une supériorité en termes de précision d'imputation. Ainsi, le choix entre le KPMF et le BTMF devrait être basé sur des critères spécifiques tels que la précision requise dans l'imputation des données, la tolérance aux erreurs de grande envergure, et le pourcentage de données manquantes présentes dans le jeu de données concerné.

5.4.4 Effet des données manquantes sur la RMSE

Pour tous les modèles, nous avons examiné trois scénarios de données manquantes (DMU, DMB et DMM) et nous avons fait varier les pourcentages de données manquantes p de 30 % à 90 % en augmentant p progressivement par incréments de 10 %. Nous avons également varié la valeur du rang de décomposition de 5, 8, 10, 20, 30, 40 et 50. Cette section permet d'évaluer la sensibilité des modèles à la proportion de données manquantes et leur performance en termes de RMSE. Il convient de rappeler que notre hypothèse initiale : alors que le taux de données manquantes augmente, nous devrions constater une augmentation de la RMSE, car il devient plus difficile pour le modèle de produire des estimations précises face à une importante quantité de données manquantes.

Les trois graphiques de la figure 5.10 représentent les performances des quatre modèles d'imputation (BPMF, TRMF, BTMF et KPMF) pour chaque scénario de données manquantes (DMU, DMB et DMM) avec un rang de décomposition fixé à 5. Les résultats montrent que le BPMF est le moins performant des quatre modèles, quel que soit le scénario. Le TRMF est plus performant que le BPMF et le BTMF est plus performant que le TRMF. Cependant, le KPMF est le modèle le plus performant jusqu'à un certain taux de données manquantes. En effet, les performances du KPMF varient selon les scénarios de données manquantes. Dans les scénarios DMU et DMB, le KPMF est le modèle le plus performant jusqu'à un taux de 70 % de données manquantes et devient le moins performant au-delà de ce

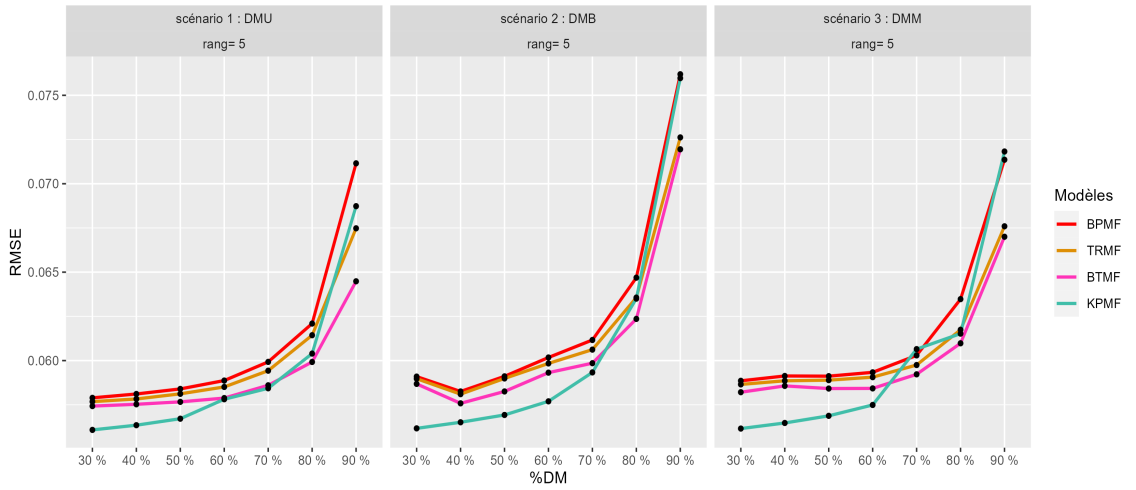

FIGURE 5.10 – Effet des données manquantes sur la RMSE pour les trois scénarios avec un rang de décomposition égale à 5. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

taux. Dans le scénario DMM, le KPMF reste le modèle le plus performant jusqu'à un taux de 60 % de données manquantes, avant de devenir le moins performant.

La figure 5.11 illustre une matrice présentant les performances des quatre modèles d'imputation (BPMF, TRMF, BTMF et KPMF) pour les trois scénarios (DMU, DMB et DMM), selon différentes valeurs du rang de décomposition (10, 30 et 50). Chaque rang de décomposition est représenté par une ligne, tandis que chaque scénario d'imputation est représenté par une colonne : DMU pour la première, DMB pour la deuxième et DMM pour la troisième.

En augmentant le rang de décomposition de 5 à 10, la première ligne de la figure 5.11, nous constatons que les performances des quatre modèles d'imputation n'ont pas significativement changé. Les performances du KPMF continuent de varier selon les scénarios de données manquantes. Dans les scénarios DMU et DMM, le KPMF reste le modèle le plus performant jusqu'à un taux de 70 % de données manquantes, mais devient le moins performant au-delà de ce taux. Il convient de noter que pour un rang de décomposition de 5, ce seuil était à 60 %. En revanche, dans le scénario DMB, le KPMF reste le modèle le plus performant jusqu'à un taux de 60 % de

données manquantes, avant de devenir le moins performant.

En augmentant le rang de décomposition de 10 à 30, comme illustré à la deuxième ligne de la figure 5.11, et de 30 à 50, à la troisième ligne de la figure 5.11, nous avons constaté que les performances du modèle KPMF n'ont pas significativement changé. En effet, à partir de la valeur 30 du rang de décomposition avec un taux de données manquantes supérieur à 80 %, la performance du KPMF diminue et le BTMF devient le modèle le plus performant en présence d'un manque extrême de données.

Nos résultats montrent que le BPMF est le moins performant des quatre modèles, quelle que soit la situation de données manquantes. Le TRMF se révèle plus performant que le BPMF, et le BTMF dépasse le TRMF en termes de performance. Cependant, le modèle KPMF se distingue en matière de performance jusqu'à un certain taux de données manquantes (< 80 %). Cette performance varie en fonction du scénario de données manquantes. Enfin, nos résultats confirment notre hypothèse de base selon laquelle l'augmentation du taux de données manquantes devrait entrainer une augmentation de la RMSE. Cette constatation est cohérente avec nos attentes initiales pour tous les modèles étudiés.

En examinant l'évolution des courbes RMSE en fonction du pourcentage de données manquantes de la figure 5.11, nous pouvons dégager des tendances dans le comportement des modèles, BPMF, TRMF, BTMF et KPMF, selon le rang de décomposition choisi.

Lorsque le rang de décomposition est inférieur à 20, toutes les courbes RMSE présentent une caractéristique commune, c'est-à-dire la forme exponentielle. Cela signifie que, plus le pourcentage de données manquantes augmente, plus la RMSE augmente. À un rang de décomposition égale à 5, nous pouvons repérer un coude dans les courbes RMSE entre 60 % et 80 % de données manquantes. Ce coude représente un seuil critique indiquant le point où la performance des modèles BPMF, TRMF et BTMF commence à se détériorer rapidement alors que le taux de données manquantes augmente. La première partie de ces courbes, où la RMSE reste relativement stable jusqu'au coude, suggère que ces modèles sont robustes face à une

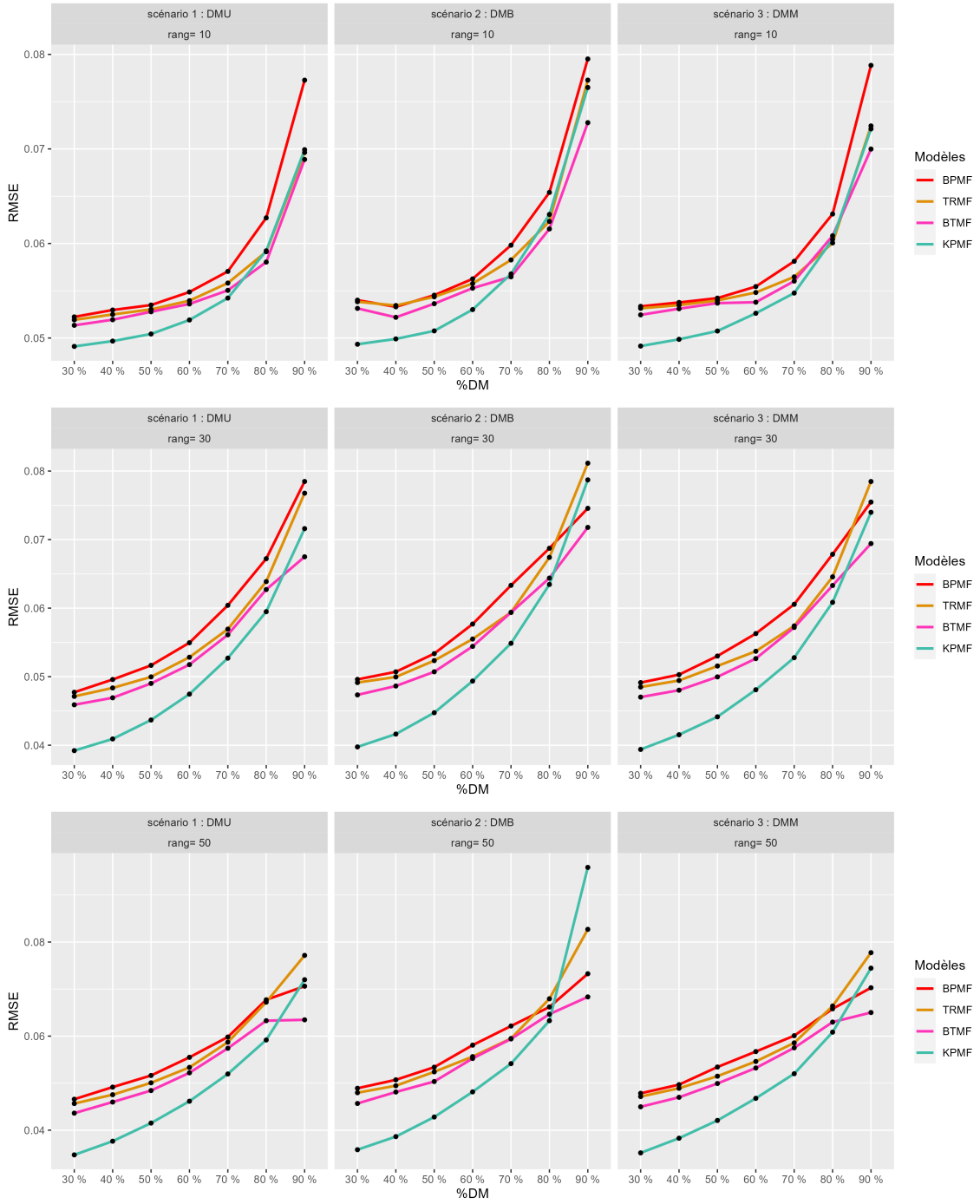

FIGURE 5.11 – Effet des données manquantes sur la RMSE pour les trois scénarios, selon différentes valeurs du rang de décomposition. Chaque rang de décomposition est représenté par une ligne, avec le premier rang de décomposition à 5, le deuxième à 10 et le troisième à 50. Chaque scénario d'imputation est associé à une colonne : DMU pour la première, DMB pour la deuxième, et DMM pour la troisième. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

certaine quantité de données manquantes. Cependant, une fois que le pourcentage de données manquantes dépasse 80 %, la RMSE augmente rapidement, signifiant que les modèles deviennent de plus en plus sensibles à l'absence de données.

Lorsque le rang de décomposition dépasse 30, l'ajout de données manquantes n'a plus le même impact exponentiel sur l'erreur. Les courbes exponentielles qui caractérisaient la croissance de la RMSE subissent un étirement horizontal vers la droite et s'aplatissent progressivement : le coude devient plus linéaire. Cela implique que, à un rang de décomposition très élevé, les modèles peuvent gérer une proportion plus importante de données manquantes avant que la RMSE ne commence à augmenter de manière significative. Par contre, cette gestion est moins efficace qu'à un rang de décomposition faible. Autrement dit, ils deviennent plus sensibles aux données manquantes.

Le changement de comportement des courbes RMSE des modèles BPMF, TRMF et BTMF, à mesure que le rang de décomposition augmente, révèle une relation entre l'augmentation du taux des données manquantes, le rang de décomposition et la rapidité d'augmentation de la RMSE. Ainsi, dans le but d'explorer cette relation, les graphiques de la figure 5.12 illustrent de manière visuelle les performances des différents modèles d'imputation de données pour différentes valeurs du rang de décomposition en fonction du taux de données manquantes. Chaque ligne de ces graphiques est spécifique à un modèle particulier. Chaque courbe représente un niveau de rang de décomposition spécifique, identifiée par une couleur mosaïque, et chaque scénario d'imputation est associé à une colonne : DMU pour la première, DMB pour la deuxième et DMM pour la troisième.

L'analyse des courbes de la RMSE révèle un phénomène commun à pratiquement tous les modèles. À un rang de décomposition de 5, nous constatons que la courbe de la RMSE démarre à un niveau relativement élevé. Cette courbe reste relativement stable jusqu'à ce qu'elle atteigne un point de pivot autour de 70 % à 80 % de données manquantes, après quoi elle présente une augmentation rapide. Pour un rang de décomposition de 10, la courbe de la RMSE commence à un niveau inférieur par

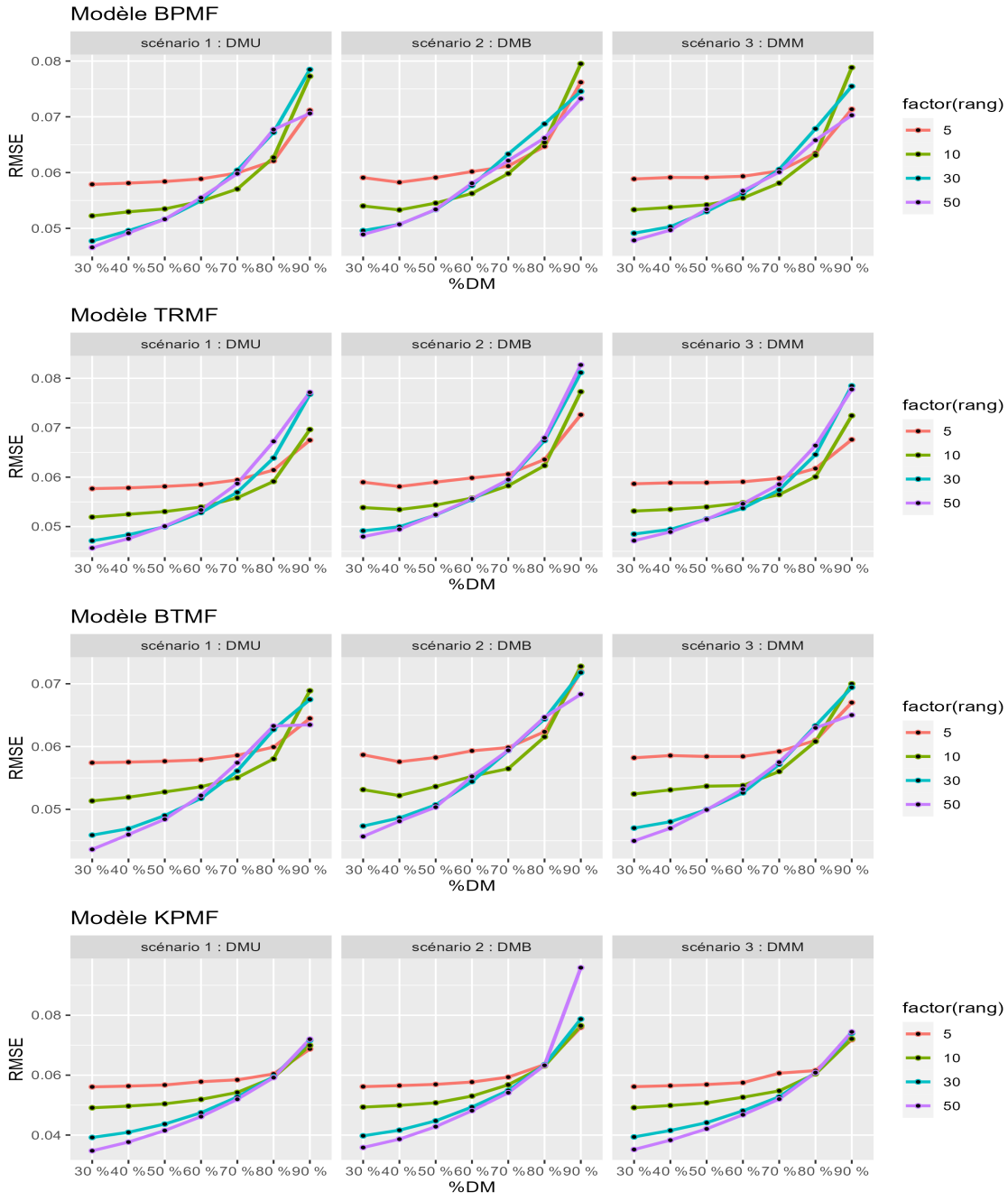

FIGURE 5.12 – Effet des données manquantes sur la RMSE pour chaque modèle et pour les trois scénarios selon différentes valeurs du rang de décomposition. Chaque modèle est représenté par une ligne, chaque scénario d'imputation est associé à une colonne : DMU pour la première colonne, DMB pour la deuxième et DMM pour la troisième. Chaque courbe représente un niveau de rang de décomposition spécifique. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

rapport à la courbe à rang de décomposition 5. Elle suit une trajectoire similaire jusqu'à ce qu'elle atteigne le même point de pivot, à environ 80 % de données manquantes. À partir de ce point, la courbe RMSE à rang de décomposition 10 devient plus abrupte. Notons également que les courbes de la RMSE pour les rangs de décomposition 30 et 50 présentent des comportements similaires. La courbe du modèle à rang de décomposition 50 démarre légèrement plus bas que celle à rang de décomposition 30 et elle devient plus raide lorsque le pourcentage de données manquantes atteint 80 %. Par ailleurs, les courbes de la RMSE semblent présenter le phénomène suivant : les courbes de rangs de décomposition plus élevés (30 et 50) ont tendance à croiser rapidement les courbes de rangs de décomposition inférieurs (5 et 10) à mesure que le pourcentage de données manquantes augmente. Cela indique que les courbes avec des rangs de décomposition plus élevés deviennent plus sensibles aux données manquantes à un stade précoce, ce qui se traduit par une augmentation plus rapide de la RMSE. En revanche, les courbes avec des rangs de décomposition plus faibles semblent présenter une certaine stabilité de la RMSE jusqu'à un certain point, ce qui signifie qu'ils sont moins sensibles aux données manquantes jusqu'à un seuil critique, au-delà duquel la RMSE commence à augmenter rapidement. Il semblerait que, à un certain taux de données manquantes, l'hypothèse liant l'augmentation du rang de décomposition et la baisse de la RMSE n'est pas vérifiée. En effet, nous observons que la réduction de la RMSE grâce à une valeur élevée du rang de décomposition, devient positive à un certain taux. À un taux élevé de données manquantes, un rang de décomposition élevé fait augmenter la RMSE de manière plus rapide.

Toutefois, nous remarquons que le modèle TRMF se comporte différemment. En effet, les courbes de la RMSE avec un rang de décomposition 50 sont restées plus basses que celles de rang de décomposition 30, ce qui signifie qu'à un rang de décomposition élevé, le modèle est plus stable à l'effet de l'augmentation des données manquantes, comparativement aux autres modèles.

En conclusion, notre analyse met en évidence deux tendances. L'augmentation

du rang de décomposition des modèles d'imputation est associée à une réduction de la RMSE, indiquant une amélioration de la précision de l'imputation. Cela confirme la notion selon laquelle un rang de décomposition plus élevé permet aux modèles de capturer davantage d'informations dans les données, conduisant ainsi à une meilleure performance. D'autre part, nous observons que l'augmentation du taux de données manquantes entraine une augmentation plus rapide de la RMSE, en particulier lorsque le modèle a un rang de décomposition élevé. Cette observation souligne la sensibilité accrue des modèles à rang de décomposition élevé à des niveaux élevés de données manquantes, ce qui peut rendre le processus d'imputation moins fiable à ces niveaux.

5.4.5 Temps de calcul

Comme mentionné dans la section précédente, nous nous attendons à une diminution de la RMSE à mesure que le rang de décomposition augmente, en raison d'une précision accrue dans la factorisation matricielle. À l'opposé, une réduction du rang de décomposition pourrait entrainer une perte d'information, susceptible de provoquer une augmentation de la RMSE. Toutefois, cette perte d'information se traduit par un gain sur le temps computationnel. Nous prévoyons donc que, lorsque le rang de décomposition des modèles diminue, le temps d'exécution devrait également diminuer de manière significative. Autrement dit, lorsque le rang de décomposition des modèles augmente, le temps d'exécution devrait également augmenter.

Dans la figure 5.13, nous explorons l'effet du rang de décomposition sur le temps d'exécution des quatre modèles d'imputation étudiés : BPMF, BTMF, TRMF et KPMF, pour le scénario DMM avec un rang fixé à 30. Chaque modèle est distingué par une couleur différente sur une courbe graphique.

Pour un rang de décomposition inférieur à 10, les temps d'exécution des modèles BTMF, TRMF et KPMF sont relativement similaires : la divergence dans leurs temps d'exécution commence à se manifester à partir d'un rang de décomposition

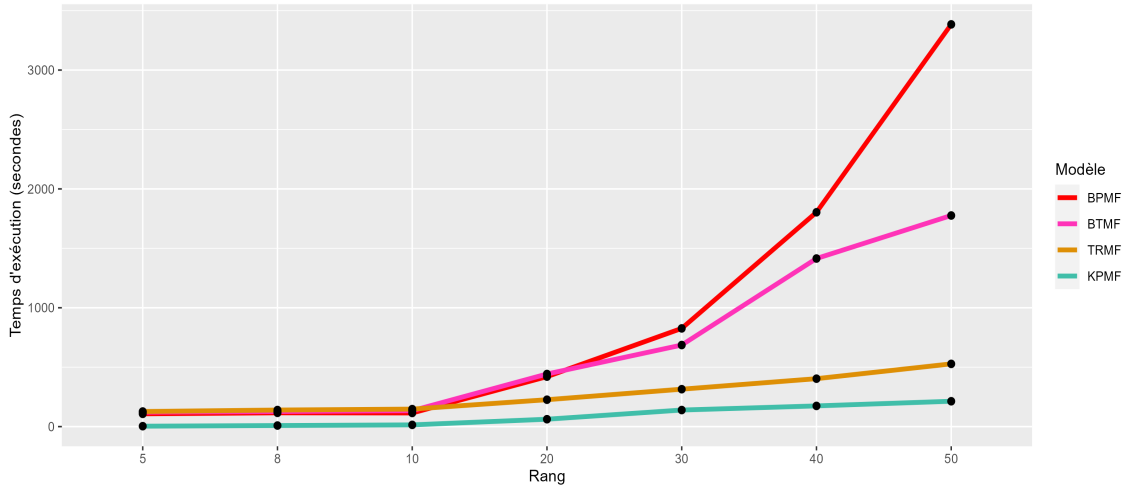

FIGURE 5.13 – Effet du rang de décomposition sur le temps d'exécution des modèles. La courbe rouge vif est associée au modèle BPMF, la courbe orange foncé représente le modèle TRMF, la courbe rose foncé est liée au modèle BTMF et la courbe verte correspond au modèle KPMF.

supérieur à 10. Pour un rang de décomposition supérieur à 10, nos observations empiriques confirment une relation positive entre le rang de décomposition et le temps d'exécution pour tous les modèles, ce qui se traduit par des courbes croissantes. Plus particulièrement, le BPMF présente une croissance exponentielle marquée en termes de temps d'exécution, tandis que les courbes associées au BTMF, TRMF et KPMF suivent une tendance de croissance logarithmique. De surcroit, la courbe logarithmique du BTMF indique une croissance encore plus rapide par rapport à celles du TRMF et du KPMF. En dernier lieu, il est à noter que, parmi les modèles étudiés, nous constatons que le KPMF affiche un temps d'exécution nettement plus court que le BTMF pour toutes les valeurs du rang de décomposition étudiées. Cela indique que le KPMF est plus efficace du point de vue computationnel que le BTMF. Le KPMF présente ainsi le temps d'exécution le plus court, ce qui peut être un critère déterminant dans le choix d'un modèle en fonction des exigences computationnelles.

En somme, ces observations confirment la relation linéaire : plus le rang de décomposition augmente, plus le temps d'exécution s'accroit. Par ailleurs, ces résultats complexifient la compréhension théorique traditionnelle, selon laquelle une augmen-

tation du rang de décomposition devrait invariablement entrainer une réduction de la RMSE. Ils mettent ainsi en lumière l'équilibre délicat qui doit être trouvé entre la perte d'information et l'efficacité computationnelle.

Les conclusions de cette analyse sur l'imputation de données manquantes dans le contexte de la factorisation matricielle révèlent plusieurs observations. Le modèle BTMF se distingue par sa capacité à produire des imputations plus précises avec une fréquence réduite d'erreurs, quel que soit le taux de données manquantes. Cependant, ce modèle révèle une variabilité accrue. En revanche, le modèle KPMF présente une stabilité caractérisée par une moindre occurrence d'erreurs extrêmes. Il montre des difficultés avec les valeurs extrêmes et une vulnérabilité lors de scénarios extrêmes d'absence de données. Par ailleurs, nous constatons que le temps d'exécution des modèles dépend du rang de décomposition choisi. Le modèle KPMF se distingue en présentant le temps d'exécution le plus court par rapport au BTMF. En augmentant le rang de décomposition, le KPMF suit une courbe logarithmique aplatie, tandis que le BTMF suit une courbe exponentielle. Le choix entre ces modèles dépend principalement du degré de précision souhaité, de la tolérance aux valeurs extrêmes, du degré de présence des données manquantes et du temps d'exécution. L'effet du rang de décomposition des modèles de factorisation matricielle sur la RMSE s'avère être plus complexe que prévu. Si une augmentation du rang de décomposition est théoriquement censée améliorer la précision des prédictions, cette étude a révélé des comportements non linéaires à des taux élevés de données manquantes. Les courbes en forme de « U » ou « U inversé » observées dans certains scénarios extrêmes remettent en question les attentes théoriques. D'autre part, nous observons que l'augmentation du taux de données manquantes entraine une augmentation plus rapide de la RMSE, en particulier lorsque le modèle a un rang de décomposition élevé. Enfin, cette analyse met en évidence la nécessité de concilier la précision des imputations et l'efficacité computationnelle lors du choix d'un modèle dans un contexte donné. Ce choix repose principalement sur le niveau de précision souhaité, la tolérance aux valeurs extrêmes, la fréquence des données manquantes et la contrainte

temporelle.

5.5 Convergence des modèles bayésiens

Les chaines de Markov ont des propriétés mathématiques qui assurent la convergence des algorithmes MCMC *sous certaines conditions* (Cowles (1996)). Cependant, il n'est pas possible de déterminer le nombre exact d'itérations nécessaires pour atteindre cette convergence. Bien qu'il n'y ait pas de garantie de convergence en un temps fini, plusieurs outils peuvent nous aider à identifier si une chaine de Markov n'a pas encore convergé vers sa loi stationnaire.

Pour évaluer la convergence de nos modèles bayésiens, nous avons utilisé plusieurs techniques, plus particulièrement la visualisation de la trace et la vérification de plusieurs chaines de Markov. La trace illustre les valeurs successives de la chaine de Markov et permet de détecter les signes de convergence. Nous avons également utilisé le diagnostique R-hat pour évaluer la convergence de nos modèles. Si R-hat est proche de 1, cela indique que les chaines ont convergé.

Dans le contexte de notre analyse sur la convergence des algorithmes bayésiens, basés sur la factorisation matricielle à faible rang de décomposition, nous avons observé des irrégularités concernant la stabilité des paramètres des matrices $\boldsymbol{U}$ et $\boldsymbol{V}$. Ces irrégularités sont devenues manifestes lorsque nous avons employé un nombre standard d'itérations, à savoir une phase de préchauffage (burn-in) fixée à 1 000 itérations et une phase d'estimation à 2 000 itérations. Ces défis de convergence ont été amplifiés en raison des configurations à complexité accrue, plus spécifiquement les scénarios DMB et DMM : ces scénarios sont caractérisés par une structure de données manquantes organisées en blocs. De plus, ces problèmes se sont accentués de manière significative lorsque le rang de décomposition du modèle était particulièrement faible ou lorsque la proportion de données manquantes était élevée et, plus particulièrement, en présence de la combinaison de ces deux facteurs.

Pour illustrer ce problème, nous avons choisi comme exemple une analyse dé-

taillée de la convergence du modèle BPMF avec un rang de décomposition fixé à 30 dans le contexte du scénario DMM avec une proportion de données manquante égale à 30 %. Pour ce faire, nous avons utilisé une période initiale de préchauffage de 2 millions d'itérations pour atteindre les valeurs les plus probables des paramètres. Cette extension du nombre d'itérations de préchauffage vise à accomplir deux objectifs : premièrement, atteindre les valeurs les plus probables des paramètres en accordant une marge de manœuvre statistiquement significative et, deuxièmement, approfondir notre compréhension des facteurs retardant le processus de convergence, c'est-à-dire la structure des données manquantes et le rang de décomposition du modèle. Nous avons ensuite enregistré les matrices de facteurs obtenues à partir de ces itérations et nous les avons utilisées pour initialiser 3 chaines de Markov. Pour ces trois chaines, nous avons fixé la phase de préchauffage à 5 000 itérations et la phase d'estimation à 15 000 itérations. La figure 5.14 illustre la trajectoire pour la phase de préchauffage de deux paramètres U[1,1] (en haut) et V[1,1] (en bas) et la figure 5.15 illustre la trajectoire des trois chaines de Markov pour la phase d'estimation. Nous remarquons que les graphiques montrent une fluctuation des valeurs à la hausse, puis à la baisse, ce qui indique que les chaines ne sont pas stationnaires.

Nous avons à vérifier la stabilité de la moyenne des deux paramètres choisis, à savoir U[1,1] et V[1,1]. Pour y parvenir, nous avons divisé chaque chaine en tranches de 1000 itérations et calculé la moyenne pour chaque tranche. Nous avons ensuite présenté les résultats sous la forme d'un diagramme de moustaches, comme indiqué dans la figure 5.16. Nous avons observé une différence notable de moyenne entre les trois chaines de Markov, ce qui indique une instabilité dans notre échantillonnage.

Le diagnostique de Gelman pour les chaines de Markov de notre modèle n'a pas pu être calculé, car la valeur R-hat est manquante (en utilisant la fonction *gelman.diag()* de R). Cela peut être dû à plusieurs causes potentielles. L'une des raisons courantes est la convergence insuffisante des chaines de Markov, ce qui peut être dû à une initialisation inappropriée des valeurs des paramètres ou une structure de modèle trop complexe. En présence d'une structure de données manquantes en

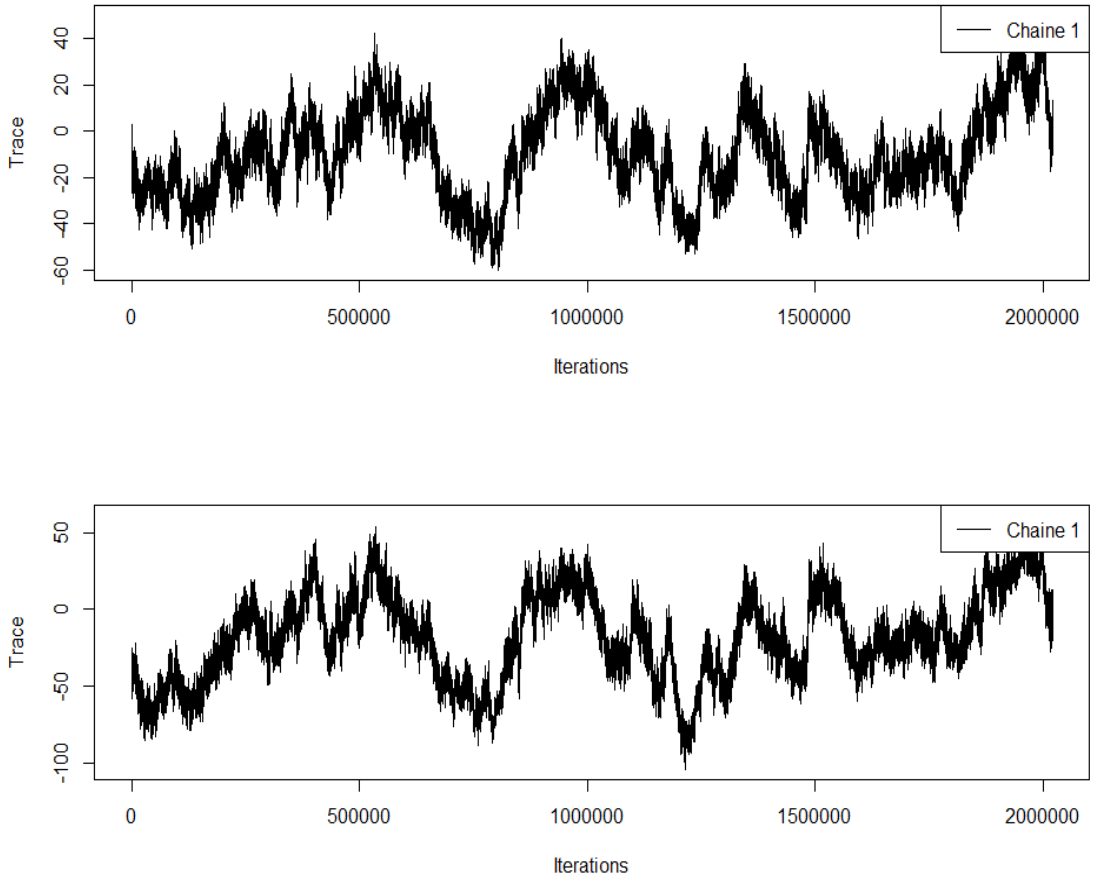

FIGURE 5.14 – Phase de préchauffage des paramètres U[1,1] (en haut) et V[1,1] (en bas), dans le cas du scénario DMM avec un rang de décomposition égal à 30 et 30 % de données manquantes.

blocs et d'un rang de décomposition faible, le processus de convergence peut être ralenti, rendant ainsi les chaines de Markov plus susceptibles de rester bloquées dans des régions de faible probabilité de l'espace des paramètres. Par ailleurs, il est à noter que le modèle BPMF se présente comme un modèle normale-Wishart, dans lequel les facteurs latents, tant spatiaux que temporels, sont modélisés de manière identique à travers un modèle bayésien hiérarchique. Par conséquent, la convergence insuffisante dans notre expérience pourrait suggérer que la convergence ne sera pas atteinte et révèle une inadéquation entre les hypothèses du modèle et la structure

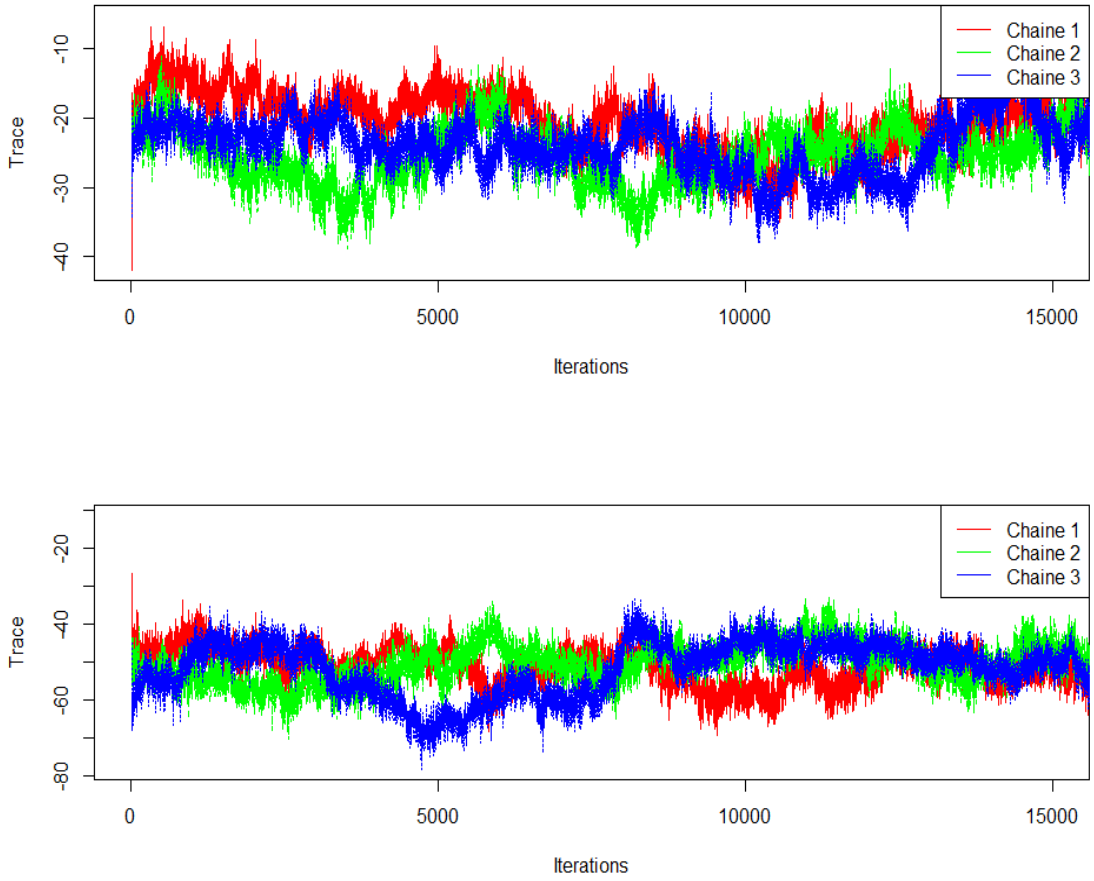

FIGURE 5.15 – Estimation des paramètres U[1,1] (en haut) et V[1,1] (en bas) selon trois chaines de Markov, dans le cas du scénario DMM avec un rang de décomposition égal à 30 et 30 % de données manquantes.

réelle des données spatiotemporelles.

Suite à de multiples tentatives d'optimisation de la convergence, incluant l'extension du nombre d'itérations et l'utilisation de matrices initialisées à partir des résultats d'autres modèles, aucun impact significatif n'a été observé sur la convergence. Ces constatations indiquent que des facteurs tels que la structure des données manquantes, le taux élevé de données manquantes et le choix du rang de décomposition du modèle peuvent constituer des obstacles substantiels à la convergence des modèles bayésiens. L'analyse de la convergence des modèles bayésiens semble égale-

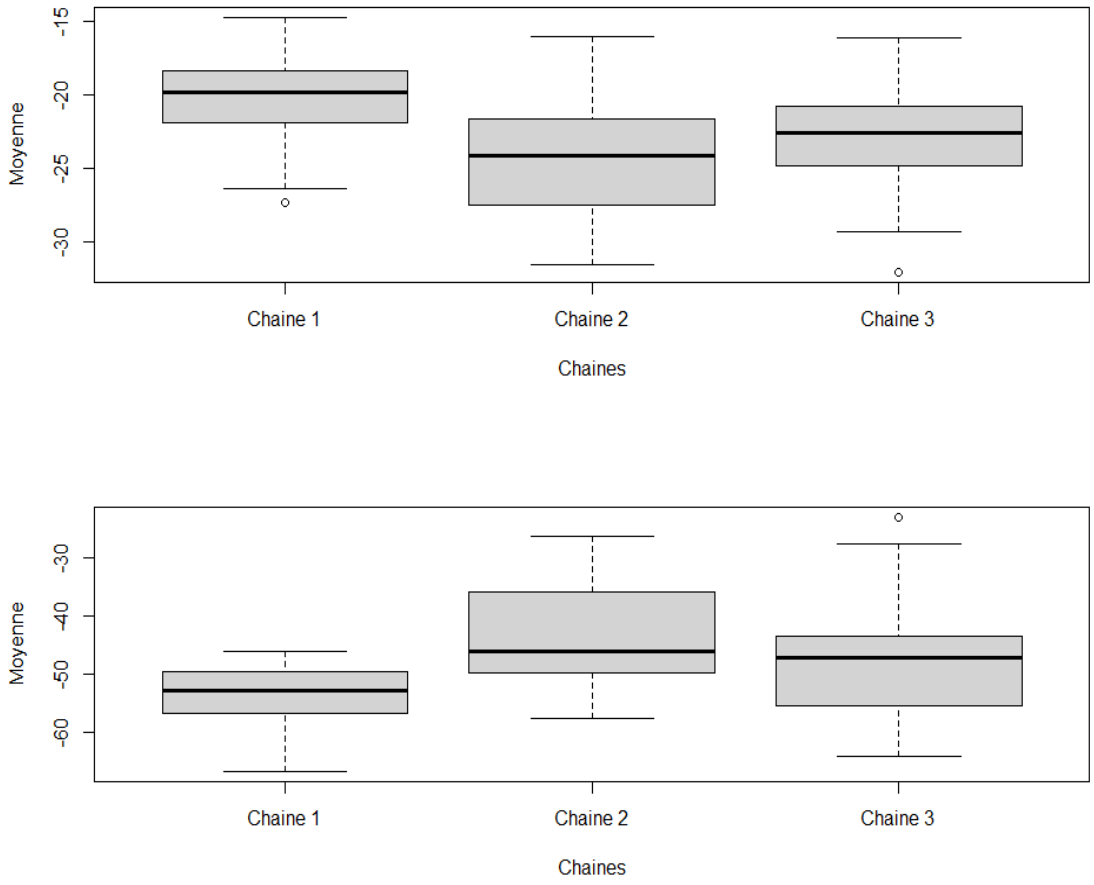

FIGURE 5.16 – Diagramme de moustaches des moyennes par tranche de 1000 pour U[1,1] (en haut) et V[1,1] (en bas) selon trois chaines de Markov pour le scénario DMM avec un rang de décomposition égal à 30 et 30 % de données manquantes.

ment révéler que la proportion de données manquantes, la structure de ces données manquantes et les hypothèses du modèle influencent conjointement l'estimation des paramètres *a posteriori*.

Il est également important de mettre en évidence l'influence du choix du rang de décomposition sur la convergence. Un rang de décomposition trop faible peut introduire une complexité excessive dans la structure des données et, de fait, compromettre la convergence. Par conséquent, le délai nécessaire pour la convergence constitue un défi significatif lors de l'implémentation des modèles bayésiens, parti-

culièrement lors des scénarios où les données manquantes sont prévalentes.

Pour conclure, nous avons établi la phase de préchauffage à 1 000 itérations et la phase d'estimation à 2 000 itérations, en conformité avec les normes standards d'évaluation de la convergence en modélisation bayésienne. Dans ce contexte, il convient de souligner que ces difficultés de convergence n'ont pas été observées dans le cas du modèle BTMF, comme le montre la figure 5.17. Ce modèle adopte une distribution Gaussienne-Wishart pour les facteurs spatiaux et une distribution normale multivariée avec un vecteur moyen autorégressif (VAR) pour les facteurs latents temporels. Cela permet une flexibilité dans la capture des relations non linéaires entre les éléments.

En somme, nos observations suggèrent que la proportion, la structure des données manquantes et le choix du rang de décomposition ont un impact significatif sur la convergence du modèle. De plus, une possible discordance entre les hypothèses sous-jacentes aux modèles et la structure intrinsèque des données pourrait exacerber ces défis de convergence.

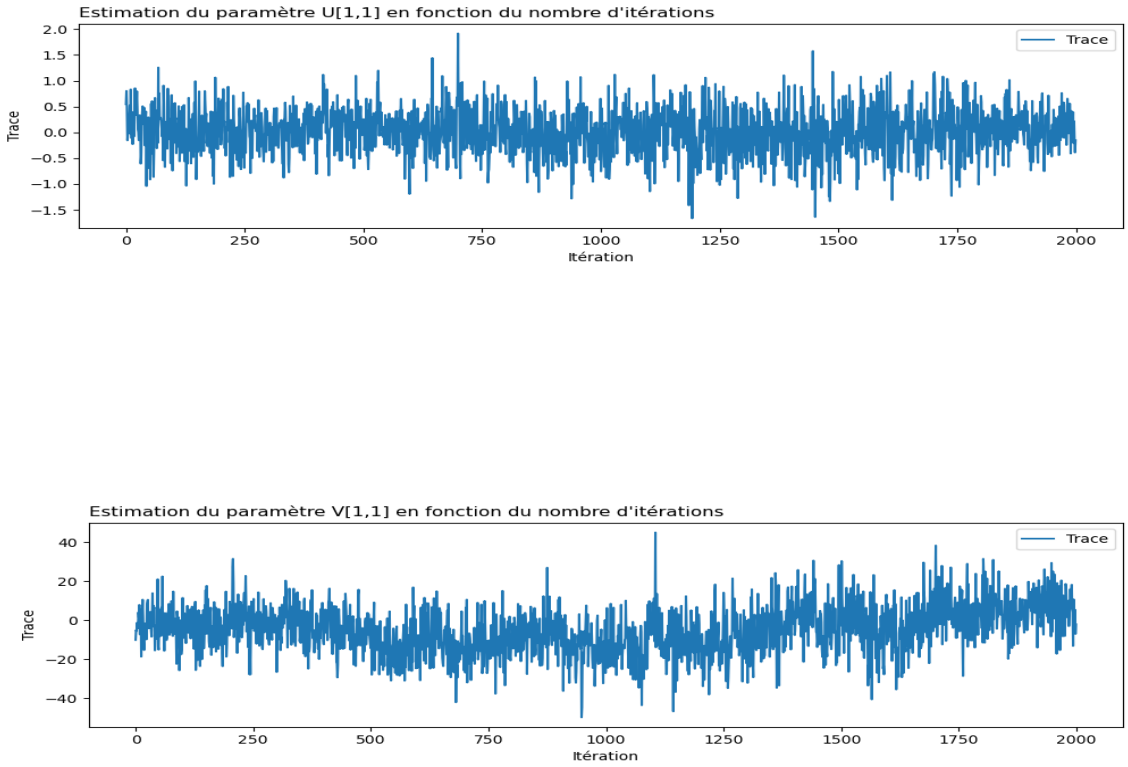

FIGURE 5.17 – Estimation des paramètres U[1,1] (en haut) et V[1,1] (en bas) pour le modèle BTMF dans le contexte du scénario DMM, avec un rang de décomposition fixé à 30 et 30 % de données manquantes.

Chapitre 6

Discussions et conclusion

Ce mémoire vise à effectuer une étude comparative des modèles d'imputation de données manquantes basés sur les méthodes de factorisation matricielle dans un contexte spatiotemporel. Cette étude explore différents modèles de traitement des données manquantes, du plus simple au plus complexe. Nous avons évalué ces modèles en observant la qualité de l'imputation et la vitesse de calcul. Une série de simulations a été menée pour quantifier leurs performances, en tenant compte de l'effet du rang de décomposition et de la proportion de données manquantes. De plus, notre étude examine l'impact de la structure des données manquantes sur l'efficacité des modèles en matière d'imputation.

Nous avons étudié les modèles suivants, qui varient tous légèrement quant à la façon dont ils modélisent la matrice de variance-covariance des données dans le but de traiter les données manquantes :

— **Imputation par la moyenne (MEAN)** : Ce modèle ne prend pas en compte la structure de corrélation présente dans les données. Il est plus rudimentaire et base son imputation sur des moyennes, sans considérer l'aspect spatial ou temporel des données ;

— **Modèles de moyenne mobile intégrée autorégressive (ARIMA)** : Bien que le modèle ARIMA soit principalement axé sur la modélisation des séries

temporelles, la structure de corrélation temporelle est considérée au niveau des résidus du modèle de régression ;

— **Bayesian Probabilistic Matrix Factorization (BPMF)** : Ce modèle correspond au premier modèle de factorisation matricielle probabiliste, dans lequel la structure de variance-covariance des données est modélisée à l'aide d'un modèle bayésien hiérarchique ;

— **Temporal Regularized Matrix Factorization (TRMF)** : Ce modèle se concentre principalement sur la décomposition de la matrice des données en matrices de facteurs latents, tout en modélisant explicitement les dépendances temporelles sous la forme d'une régularisation autorégressive ;

— **Bayesian Temporal Matrix Factorisation (BTMF)** : Ce modèle est une extension du modèle précédent, le TRMF, qui incorpore une modélisation plus riche des dépendances spatiales et temporelles. Le modèle suppose une distribution Gaussienne-Wishart des facteurs spatiaux et une distribution normale multivariée, avec un vecteur moyen autorégressif (VAR) des facteurs latents temporels. Le VAR lui-même est sujet à une distribution *a priori* conjuguée Matrice-Normale Inverse-Wishart ;

— **Kernelized Probabilistic Matrix Factorization (KPMF)** : Dans ce modèle, les matrices de variance-covariance associées aux facteurs latents jouent un rôle central dans la modélisation de la structure spatiotemporelle des données. Ces matrices de variance-covariance sont modélisées par des processus gaussiens utilisant une fonction de noyau. Cette fonction de noyau prend en compte les distances entre les points associés aux facteurs latents. Elle permet également une plus grande flexibilité dans la capture des relations non linéaires entre les éléments.

Ces modèles ont été testés avec des données de trafic routier collectées à l'aide de détecteurs à boucle inductive déployés sur plusieurs autoroutes de la ville de Seattle, aux États-Unis : la I-5, la I-90, la I-405 et la SR-520. Ces détecteurs enregistrent

principalement la vitesse des véhicules en mph (*miles per hour* [miles par heure]), où 1 mph équivaut à environ 1,60 km/h. Les données de trafic révèlent une structure spatiotemporelle complexe :

— De minuit à cinq heures du matin, les vitesses sont relativement élevées et stables ;

— La congestion apparait dès six heures du matin, avec des heures de pointe de six heures à dix heures en direction du centre-ville. De quinze heures à dix-neuf heures, deux nouveaux points de congestion se forment ;

— Les weekends présentent une circulation plus fluide, sauf aux abords du centre-ville. En semaine, la congestion est observée sur toutes les autoroutes, bien que les données de certains capteurs présentent moins de signes de congestion que d'autres ;

— Sur une base hebdomadaire, les vitesses augmentent les weekends et fluctuent en semaine. Sur une base quotidienne, deux baisses de vitesse indiquent les heures de pointe, tandis qu'un pic stable représente les heures creuses ;

— La moyenne spatiale empirique montre que les conditions de trafic varient en fonction du sens de la circulation et des plages horaires, avec des phénomènes d'inversion du trafic non seulement en fonction du sens de la circulation, mais aussi en fonction de l'heure de la journée. La variabilité spatiale est également prise en compte, puisque certaines zones affichent une vitesse moyenne élevée, et d'autres, une vitesse moyenne faible ;

— Une forte corrélation est observée à l'intérieur du groupe des jours de la semaine et de celui des jours du weekend, avec une corrélation intergroupe plus faible. Des plages horaires spécifiques, comme de six heures à dix heures et de quinze heures à dix-neuf heures, présentent une forte corrélation, révélant une régularité dans le trafic à ces moments. D'autres plages horaires significatives sont identifiées, notamment de minuit à quatre heures, de onze heures à quatorze heures et de vingt heures à vingt-trois heures. Les analyses de la fonc-

tion d'autocorrélation et de la fonction d'autocorrélation partielle révèlent une structure temporelle complexe, suggérant un effet d'inertie dans les données de la série temporelle.

Dans cette étude, nous avons évalué les six modèles d'imputation suivants : MEAN, ARIMA, BPMF, TRMF, BTMF et KPMF. Les évaluations ont été menées en considérant trois scénarios de données manquantes (DMU, DMB et DMM) et en variant le pourcentage de données manquantes p de 30 % à 90 %. Pour chaque scénario, nous avons augmenté progressivement p par incréments de 10 %. De plus, nous avons utilisé différentes valeurs de rang de décomposition matricielle K (5, 8, 10, 20, 30, 40 et 50), et ce, pour chaque modèle, afin d'obtenir une évaluation complète des performances.

En considérant les résultats de notre analyse sur l'imputation des données manquantes dans le contexte spécifique du transport routier, nous constatons que le choix du modèle approprié dépend de divers facteurs. Parmi ces facteurs figurent les objectifs spécifiques de l'analyse, la proportion et la structure des données manquantes, la disponibilité des ressources computationnelles, ainsi que les compromis à faire entre la précision des imputations et le temps d'exécution de chaque modèle. Cependant, à la lumière des conclusions tirées de notre étude, nous pouvons énoncer les principaux points suivants :

— **MEAN** : Ce modèle présente des performances inférieures en raison de son incapacité à prendre en compte la structure spatiotemporelle des données, ce qui le rend moins approprié ou pertinent pour l'imputation de données dans notre contexte ;

— **ARIMA** : Bien qu'il soit performant lorsque les données manquantes sont dispersées aléatoirement dans le temps et l'espace, le modèle ARIMA montre des limites lorsque les données manquantes sont regroupées en blocs. La forte variance de ces données semble donc affecter négativement la performance de cette méthode ;

— **BPMF** : Ce modèle donne des résultats acceptables dans certaines conditions, mais il montre ses limites lorsque la proportion de données manquantes devient plus importante. Nos résultats suggèrent que le BPMF ne se distingue pas particulièrement dans cette tâche, en particulier lorsque les taux de données manquantes sont élevés (entre 60 % et 90 %) ou lorsque celles-ci sont groupées en blocs ;

— **TRMF** : Le modèle TRMF offre un équilibre satisfaisant entre précision et efficacité computationnelle. Il a démontré de bonnes performances tout au long de notre étude et peut être considéré dans de nombreuses situations ;

— **BTMF** : Le BTMF s'est démarqué en termes de précision d'imputation des valeurs par rapport aux autres modèles. Il est particulièrement adapté aux données spatiotemporelles. Cette performance supérieure est d'autant plus remarquable lorsque le taux de données manquantes atteint le niveau extrême de 90 %. Néanmoins, le modèle présente des limites, comme la possibilité d'erreurs d'imputation significatives et la consommation computationnelle élevée ;

— **KPMF** : Le modèle KPMF maintient une performance stable, particulièrement lorsque les taux de données manquantes sont modérés. Plus précisément, il montre une occurrence moins fréquente d'erreurs extrêmes. Cependant, il présente des limites dans l'imputation de données manquantes lorsqu'il est confronté à un taux élevé de données manquantes, soit de 90 %, ce qui se traduit par une dégradation significative de sa performance. Le modèle KPMF s'est révélé être le modèle le plus rapide parmi ceux testés, ce qui en fait une option efficace lorsque le temps d'exécution est critique.

L'évaluation des modèles d'imputation basés sur les techniques de factorisation matricielle (BPMF, TRMF, BTMF et KPMF) a révélé l'impact de trois facteurs critiques sur la RMSE, à savoir le rang de décomposition, la proportion ainsi que la structure des données manquantes. Pour ouvrir la discussion sur nos résultats, nous pouvons résumer nos observations de la manière suivante :

— Les modèles basés sur les techniques de factorisation matricielle supposent généralement que le rang de la matrice est connu avant le calcul (Haeffele (2017)). Une telle hypothèse pourrait s'avérer fausse dans des applications réelles. Cependant, le choix du rang de décomposition influence directement la qualité de l'imputation, mettant en exergue l'importance de cette problématique. Dans notre étude, nous avons testé différentes valeurs du rang de décomposition afin d'évaluer leur impact sur la performance des modèles. Nous avons observé que le rang de décomposition offrant un équilibre optimal entre la qualité de l'imputation et la complexité du modèle varie selon les différentes structures de données manquantes étudiées. Cette variation souligne que le choix du rang de décomposition ne peut pas être standardisé et qu'il nécessite un examen attentif de l'ensemble du jeu de données en question. Par ailleurs, les modèles BPMF, TRMF et BTMF ont réagi plus sensiblement aux variations du rang, particulièrement lorsque le taux de données manquantes excédait 60 %. À l'opposé, le modèle KPMF a démontré une robustesse aux variations du rang lorsque le taux de données manquantes atteignait des niveaux élevés ;

— En présence d'un faible taux de données manquantes, l'augmentation du rang de décomposition des modèles d'imputation est associée à une réduction de la RMSE. Cela confirme l'hypothèse selon laquelle un rang de décomposition plus élevé permet aux modèles d'extraire davantage d'informations des données, conduisant ainsi à une meilleure performance ;

— Bien que la relation entre le rang de décomposition et la proportion de données manquante soit effectivement linéaire pour les taux plus faibles de données manquantes, nous observons une tendance non linéaire lorsque le taux de données manquantes est élevé. En effet, notre étude montre que l'augmentation du taux de données manquantes entraine une augmentation plus rapide de la RMSE, en particulier lorsque le modèle a un rang de décomposition élevé. Cette observation souligne que les modèles à rang de décomposition élevé sont plus

propices aux erreurs lorsque le taux de données manquantes est lui aussi élevé, ce qui peut rendre le processus d'imputation moins fiable. Par ailleurs, ces observations sont en contradiction avec les attentes théoriques, qui stipulent que la RMSE devrait diminuer de manière continue à mesure que le rang de décomposition augmente. Ce comportement inattendu de la RMSE, en présence des taux élevés de données manquantes, demeure inexpliqué et suscite des discussions pour de futures investigations. Plusieurs hypothèses pourraient expliquer cette lacune : les modèles s'appuient de manière disproportionnée sur le faible ensemble de données disponibles pour ajuster leurs paramètres ou l'effet négatif des données manquantes surpasse l'effet bénéfique de l'augmentation du rang de décomposition, ce qui entraine des résultats imprévisibles ;

— Nos résultats indiquent que le modèle KPMF (Kernelized Probabilistic Matrix Factorization) excelle particulièrement dans la minimisation des erreurs extrêmes, ce qui se traduit par une RMSE (Root Mean Square Error) globalement plus basse. Cette performance suggère une plus grande précision du KPMF dans l'imputation des valeurs manquantes par rapport aux autres modèles analysés, surtout dans des contextes où le pourcentage de données manquantes est faible ou modéré, comme dans les cas avec 30% et 60% de données manquantes. Cependant, il est important de souligner que lorsque le taux de données manquantes devient très élevé, comme dans le scénario à 90% de données manquantes, la performance du KPMF se dégrade significativement. Dans ces conditions, le modèle BTMF (Bayesian Temporal Matrix Factorization) affiche une RMSE considérablement plus basse que celle du KPMF. Par conséquent, le BTMF se révèle être le modèle le plus adapté et efficace pour l'imputation des valeurs manquantes dans des situations où le taux de données manquantes est extrêmement élevé.

— Notre analyse quantitative des données a révélé plusieurs dynamiques intéressantes en ce qui concerne la performance des modèles mesurée par la RMSE.

Premièrement, il a été observé que la RMSE tend à diminuer lorsque la valeur du rang de décomposition augmente, ce qui suggère une performance améliorée des modèles. Deuxièmement, une augmentation du taux de données manquantes entraine une hausse de la RMSE, indiquant une dégradation de la performance. Ceci suggère une interaction négative entre ces deux variables : le bénéfice obtenu en augmentant le rang semble être annulé par l'effet négatif d'un taux élevé de données manquantes. De plus, en présence de données manquantes, la RMSE semble augmenter plus rapidement lorsque la valeur du rang de décomposition est élevée. Cette interaction complexe entre le rang et le taux de données manquantes souligne l'importance de considérer ces deux facteurs en amont de l'évaluation de la robustesse et de l'efficacité des modèles ;

— Nos résultats montrent que la RMSE est généralement plus élevée dans les scénarios incluant des données manquantes par blocs (DMB et DMM) que dans le scénario de données manquantes unitaires (DMU). En effet, la disparition de blocs entiers de données crée une difficulté additionnelle pour les modèles d'imputation, car l'information environnante disponible est réduite. À l'inverse, dans le cas de la structure de données manquantes unitaires, où les données manquantes sont distribuées de manière aléatoire, les modèles ont accès à plus d'informations contextuelles. Cela permet une meilleure approximation des valeurs manquantes, ce qui se traduit par une RMSE plus basse. Un contexte riche en données adjacentes contribue à des imputations plus précises. Enfin, tous les modèles ont exhibé ces mêmes tendances, sauf le KPMF, qui a montré une sensibilité accrue dans le scénario DMB, en présence d'un taux élevé de données manquantes.

TABLE 6.1 – Comparaison de différents modèles en termes de qualité d'imputation (RMSE) et de temps d'exécution. Les indicateurs de performance sont notés de manière qualitative : un nombre croissant de signes plus (+) indique une meilleure performance, tandis qu'un nombre croissant de signes moins (-) indique une mauvaise performance.

Modèle	**Performance**	
	Qualité d'imputation (RMSE)	**Temps d'exécution**
MEAN	- - -	+ + +
ARIMA	- -	+ +
BPMF	-	- - -
TRMF	+	-
BTMF	+ +	- -
KPMF	+ + +	+

Dans l'ensemble, les modèles varient considérablement en termes de qualité d'imputation (RMSE) et de temps d'exécution. Comme en témoigne le tableau 6.1, bien que le KPMF et le BTMF se présentent comme les choix les plus pertinents, offrant un bon équilibre entre la qualité d'imputation et l'adaptabilité, seul le BTMF est spécifiquement conçu pour fonctionner efficacement en présence d'un taux élevé de données manquantes. Nous avons aussi constaté que le KPMF se distingue par sa capacité à minimiser les erreurs extrêmes, ce qui lui permet d'obtenir une RMSE globalement plus basse. En revanche, le BTMF présente une RMSE légèrement plus élevée en raison de sa sensibilité aux valeurs extrêmes. Ainsi, le choix entre le KPMF et le BTMF dépend des attentes vis-à-vis de la qualité de l'imputation, du temps d'exécution, du taux des données manquantes présent dans le jeu de données, tout comme de la tolérance du modèle envers les erreurs extrêmes.

Notre étude comparative des modèles d'imputation de données manquantes, comprenant MEAN, ARIMA, BPMF, TRMF, BTMF et KPMF, comporte tout de même plusieurs limitations qui méritent d'être soulignées :

— **La taille de l'échantillonnage** : Dans le cadre de notre étude, la taille de l'échantillon a été limitée à une période de quatre semaines. Cette contrainte

comporte certains inconvénients. D'un côté, cette durée restreinte peut empêcher la capture de la variation des données sur le long terme, surtout dans un contexte où les cycles saisonniers sont présents. Dans ce cas, un modèle complexe pourrait occasionner un sous-ajustement, tandis qu'un modèle plus simple risquerait le sur-ajustement (Brownlee (2020)). D'un autre côté, une taille d'échantillon réduite permet une analyse plus rapide, ce qui est particulièrement avantageux lorsque les ressources de calcul sont limitées.

— **La proportion et la structure des données manquantes** : Dans notre étude, nous avons analysé la complexité associée aux données manquantes sous forme de trois scénarios distincts : les données manquantes unitaires (DMU), les données manquantes par blocs (DMB) et les données manquantes mixtes (DMM). Premièrement, une limite importante de notre étude réside dans le fait que les blocs de données manquantes n'ont été examinés que sur une période de vingt-quatre heures, alors que, dans des conditions réelles, les blocs de données manquantes peuvent s'étendre sur des durées plus longues, allant jusqu'à des semaines, et peuvent potentiellement affecter la performance des modèles. Cependant, notre analyse monte l'impact des différentes structures de données manquantes, en mettant l'accent sur le scénario DMM, qui reflète de manière plus fidèle la variabilité des situations réelles. De plus, nous avons fait varier la proportion de valeurs manquantes p entre 30 % et 90 % par incréments de 10 % dans chacun des trois scénarios pour mieux comprendre la sensibilité des modèles. Enfin, pour garantir l'intégrité de notre analyse, nous avons modifié la graine du générateur de nombres aléatoires à chaque incrément de 10 % pour p, éliminant ainsi toute source potentielle de biais lors de la création des données artificielles ;

— **La nature des données spatiotemporelles** : Dans de nombreux cas pratiques, les scénarios spatiotemporels impliquent souvent des données non gaussiennes : le trafic internet, les décomptes, les réponses binaires ou encore le

traitement des valeurs extrêmes. Ces types de données présentent des défis en raison de leur distribution asymétrique et de leur hétérogénéité. Devant cette complexité, les modèles bayésiens hiérarchiques représentent une solution robuste. Ils sont particulièrement efficaces pour modéliser les matrices de variance-covariance des données, puisqu'ils en capturent la dépendance spatiale et temporelle (Beckers (2021)). Dans cette optique, nous avons étudié le modèle Kernelized Probabilistic Matrix Factorization (KPMF), qui explore les interactions non linéaires spatiotemporelles latentes dans les données, et avons ainsi enrichi notre analyse. Notre approche apporte donc à ce modèle une perspective complémentaire, qui renforce la manière dont celui-ci traite ces données complexes ;

— **La convergence des modèles bayésiens** : Dans notre étude, nous avons constaté que la convergence de l'algorithme de Gibbs peut s'avérer difficile lorsque la proportion de données manquantes est élevée, ce qui constitue une limitation importante dans l'interprétation des résultats à ce niveau. Cette difficulté a été constatée lors de notre analyse de la convergence du modèle BPMF. Celle-ci découle potentiellement du fait que les modèles bayésiens s'appuient sur des principes probabilistes pour estimer les paramètres : un faible volume de données ne fournit pas toujours suffisamment d'informations pour guider ces estimations vers une convergence fiable (Barhoumi (2006)). Elle met également en lumière l'importance d'une sélection judicieuse des distributions *a priori* dans la modélisation des structures temporelles et spatiales. Pour évaluer la convergence dans notre étude, nous avons employé plusieurs techniques, notamment l'analyse des traces de plusieurs chaines de Markov, le calcul de la moyenne par tranches de 1 000 estimations et le diagnostic de Gelman. Ces méthodes ont aidé à quantifier et à comprendre les défis liés à la convergence, comme détaillé dans la section « Convergence des modèles bayésiens » du chapitre « Études de simulation et résultats » ;

En guise de perspective, l'aspect insoluble du problème des données manquantes invite à une exploration continue des modèles actuels et à la recherche de solutions innovantes. Ceci dans le but de fournir des résultats toujours plus précis et pertinents relativement aux défis inhérents à l'imputation des données manquantes. De fait, nous avons noté que les modèles étudiés reposent sur l'échantillonnage de Gibbs pour inférer les paramètres des modèles. Toutefois, d'autres algorithmes d'inférence sont également pertinents, bien qu'ils n'aient pas été explorés ici. Par exemple, contrairement à l'algorithme de Gibbs, qui fournit une approximation numérique de la distribution *a posteriori*, les méthodes variationnelles ont l'avantage de fournir une approximation déterministe rapide de la distribution *a posteriori*. Ces méthodes peuvent offrir des solutions douées d'une précision comparable à l'échantillonnage de Gibbs, tout en étant plus rapides (Keribin (2010)). Par ailleurs, notre étude a mis en évidence que les modèles intégrant une modélisation autorégressive *a priori* de la structure temporelle, tels que le TRMF et le BTMF, ont tendance à présenter une meilleure qualité d'imputation des valeurs manquantes. Cette observation souligne l'importance du choix de la distribution *a priori* dans un contexte spatiotemporelle. De plus, notre étude a montré que le modèle KPMF performance mieux en termes de RMSE, soulignant ainsi l'efficacité des processus gaussiens dans la modélisation des structures de covariance des facteurs latents. Dans ce contexte, il est possible d'envisager une amélioration du modèle BTMF en ce qui concerne la modélisation des facteurs latents spatiaux. De fait, une modélisation sous-jacente des facteurs spatiaux par des processus gaussiens pourrait apporter des améliorations substantielles à la qualité d'imputation et au temps computationnel du modèle.

Enfin, nous espérons que les résultats de notre étude ont enrichi la compréhension des avantages et des limites que recèle chacun des modèles examinés. Nous espérons également avoir simplifié le processus de prise de décision en fournissant des orientations stratégiques claires concernant le choix du modèle approprié pour traiter les données manquantes, tout en mettant en évidence la complexité spatiale et temporelle des données de trafic. Cette démarche a pour but de renforcer la capacité

des systèmes de transport intelligents à gérer efficacement la congestion routière de même que d'améliorer la qualité de vie et la santé économique et environnementale d'une région ou d'un pays.

Bibliographie

Agarwal, D. e. B.-C. C. 2010, «Matrix factorization through latent dirichlet allocation», dans *WSDM*.

Barhoumi, M. A. 2006, *Traitement des données manquantes dans les données de panel : Cas des variables dépendantes dichotomiques*, Mémoire de maîtrise, Université Laval, Québec.

Bayes, T. 1763, «An essay towards solving a problem in the doctrine of chances», *Philosophical Transactions of the Royal Society of London*, vol. 53, p. 370–418. Origine du terme *bayésien*.

Beckers, T. 2021, «An introduction to gaussian process models», en ligne. `gpr.tbeckers.com`.

Bennett, J. e. S. L. 2007, «The netflix prize», dans *KDD Cup and Workshop at the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Jose, California, p. 35.

Boutahar, M. e. M. R.-C. 2015, *Initiation à la statistique bayésienne : Bases théoriques et applications en alimentation, environnement, épidémiologie et génétique*, ELLIPSES, ISBN 978-2340005013.

Boutahar, M. e. M. R.-C. 2019, *Méthodes en séries temporelles et applications avec R*, Sciences, Ellipse, Marseille, France.

Box, G. E. P. e. G. M. J. 1976, *Time Series Analysis : Forecasting and Control*, Holden-Day, 575 p..

Brownlee, J. 2020, «Impact of dataset size on deep learning model skill and performance estimates», https://machinelearningmastery.com/. Consulté le 2 octobre 2023.

Canada, T. 2022, «Its architectures», URL https://tc.canada.ca/en/road-transportation/innovative-technologies/its-architectures, consulté : 2023-10-24.

Chai, T. e. R. R. D. 2014, «Root mean square error (rmse) or mean absolute error (mae)», *Geoscientific model development discussions*, vol. 7, nº 1, p. 1525–1534.

Chen, X. e. L. S. 2021, «Bayesian temporal factorization for multidimensional time series prediction», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, nº 9, doi :10.1109/TPAMI.2021.3066551. Soumis le 14 Octobre 2019 (v1), dernière révision le 14 Mars 2021 (v2).

Cheng, M.-S. P. e. P. A. P., Zhi (Aaron). 2020, «Mitigating traffic congestion : The role of intelligent transportation systems», *Information Systems Research*, vol. 31, nº 3, doi :10.1287/isre.2019.0894, p. 837–855. URL https://doi.org/10.1287/isre.2019.0894, consulté le 20 octobre 2023.

Courcot, B. 2023, *Suivi et modélisation du potentiel hydrique du sol dans un contexte de stress climatiques : le cas d'une érablière à bouleau jaune à la marge nordique de sa distribution*, Mémoire de maîtrise en technologie de l'information, Téléuniversité, Québec, Canada.

Cowles, M. K. e. B. P. C. 1996, «Markov chain monte carlo convergence diagnostics : A comparative review», *Journal of the American Statistical Association*, vol. 91, nº 434, doi :10.1080/01621459.1996.10476956, p. 883–904. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476956.

Cryer, J. D. 1986, *Time Series Analysis*, Duxbury series in statistics and decision sciences, Duxbury Press, Boston, MA, ISBN 9780871509635, xi, 286 p..

Dax, A. 2014, «Imputing missing entries of a data matrix : A review», *Journal of Advanced Computing*, vol. 3, n° 3, doi :10.7726/jac.2014.1007, p. 98–222.

Didier, R. e. C. D. 2020, «Mobilités : Coûts moyens socio-économiques», `https://side.developpement-durable.gouv.fr/ACCRDD/doc/SYRACUSE/790514/mobilites-couts-moyens-socio-economiques`. Consulté le 2 octobre 2023.

Dirks, C. G. e. M. K., Susanne. 2010, «Smarter cities for smarter growth : How cities can optimize their systems for the talent-based economy», IBM Institute for Business Value.

Donovan, T. M. et R. M. Mickey. 2019, *Bayesian Statistics for Beginners : a step-by-step approach*, Oxford University Press, ISBN 9780198841296 (Print), 9780191876820 (Online), doi :10.1093/oso/9780198841296.001.0001.

Dupuis, J. 2007, «Statistique bayésienne et algorithmes mcmc», IMAT (Master 1), LSP-UPS.

Faloutsos, J. G. T. J. e. Y. W., Christos. 2018, «Forecasting big time series : Old and new», *Proceedings of the VLDB Endowment*, vol. 11, n° 12.

Haeffele, B. D. e. R. V. 2017, «Structured low-rank matrix factorization : Optimality, algorithm, and applications to image processing», .

Hippert, A. F. 2020, *Reconstruction de données manquantes dans des séries temporelles de mesures de déplacement par télédétection*, thèse de doctorat, Université Savoie Mont Blanc. URL `https://tel.archives-ouvertes.fr/tel-03507672`.

Hyndman, R. J. e. G. A. 2018, *Forecasting : principles and practice*, OTexts. [Online]. Available : `https://OTexts.com/fpp2`.

Jabir, M. 2018, *Comparaison de méthodes d'imputation des données manquantes appliquées à la base nationale sur les collisions*, Mémoire de maîtrise, HEC Montréal. Option Intelligence d'affaire, Maîtrise ès sciences en gestion (Msc).

Jadeja, D. 2021, *Communication of Uncertainty for Statistical Modelling of Election Outcomes*, Mémoire de fin d'études, HEC Montréal.

Joreskog, K. 1973, «Analysis of covariance structures», dans *Multivariate Analysis–III*, édité par P. R. Krishnaiah, Academic Press, ISBN 978-0-12-426653-7, p. 263–285, doi :https://doi.org/10.1016/B978-0-12-426653-7.50024-7. URL `https://www.sciencedirect.com/science/article/pii/B9780124266537500247`.

Keribin, C. 2010, «Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie», *Journal de la société française de statistique*, vol. 151, n° 2, p. 107–131. URL `http://www.numdam.org/item/JSFS_2010__151_2_107_0/`.

Khrulkov, M. E. M. e. S. A. M., Alexander A. 2022, «Approach to imputation multivariate missing data of urban buildings by chained equations based on geospatial information», dans *International Conference on Computational Science*, Springer, p. 234–247.

Kołodziej, C. e. C. G. e. G. D. e. W. A., Joanna et Hopmann. 2022, *Intelligent Transportation Systems - Models, Challenges, Security Aspects*, Springer International Publishing, ISBN 978-3-031-04036-8, p. 56–82, doi :10.1007/978-3-031-04036-8_3. URL `https://doi.org/10.1007/978-3-031-04036-8_3`.

Kovar, J. G. e. P. J. W. 1995, «Imputation of business survey data», dans *Business Survey Methods*, édité par D. B. A. C. M. C. e. P. Cox, B.G., John Wiley et Sons, Inc, New York, p. 403–423.

Lambert, B. 2018, *A Student's Guide to Bayesian Statistics*, SAGE Publications Ltd, London, United Kingdom, ISBN 978-1473916357. Ben Lambert - Imperial College London.

Lawrence, N. e. R. U. 2009, «Non-linear matrix factorization with gaussian processes», dans *ICML*.

Lei, A. L. Y. W. e. L. S., Mengying. 2022, «Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and kriging», *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, n° 10, p. 18 962–18 974.

Li, Y. e. C. S. 2018, «A brief overview of machine learning methods for short-term traffic forecasting and future directions», *SIGSPATIAL Special*, vol. 10, n° 1, p. 3–9.

Lin, X. e. P. C. B. 2020, «Optimization and expansion of non-negative matrix factorization», *BMC Bioinformatics*, vol. 21, n° 1, doi :10.1186/s12859-019-3312-5, p. 1–10. URL `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3312-5`.

Martin, O. 2018, *Bayesian Analysis with Python : Introduction to statistical modelling and probabilistic programming using PyMC3 and ArviZ, 2nd Edition*, 2e éd., Packt Publishing, 358 p..

Mohamed, S. 2011, *Generalised Bayesian Matrix Factorisation Models*, thèse de doctorat, University of Cambridge, St John's College.

Moustakbal, A. 2009, *L'Impact de la Congestion Routière sur l'Industrie du Camionnage dans la Région de Montréal*, Mémoire, Université du Québec à Montréal. Présenté comme exigence partielle de la Maîtrise en Administration des Affaires, Profil Recherche.

MTQ. 2014, «Évaluation des coûts de la congestion routière dans la région de montréal pour les conditions de référence de 2008», Rapport final, Ministère des Transports du Québec.

Murphy, K. P. 2012, *Machine Learning : A Probabilistic Perspective*, unknown éd., The MIT Press, ISBN 978-0262018029.

Nalborczyk, L. 2020, «Introduction à la modélisation statistique bayésienne, un cours en r avec le package brms», Distribué en date du 01-05-2020.

Neal, R. M. 1993, «Probabilistic inference using markov chain monte carlo methods», cahier de recherche CRG-TR-93-1, Department of Computer Science, University of Toronto.

Oberoi, K. S. 2019, *Modélisation spatio-temporelle du trafic routier en milieu urbain*, thèse de doctorat, Normandie.

OCDE. 2003, «Transport urbain de marchandises : les défis du xxi siècle», Technical report, OCDE. URL `https://www.itf-oecd.org/sites/default/files/docs/03deliveringf.pdf`.

OCDE. 2006, *OECD Territorial Reviews : Competitive Cities in the Global Economy*, OCDE, ISBN 9264027092, 446 p.. URL `https://www.oecd.org/`.

O'Connor, E. F., J. J. et Robertson. 2023, «Thomas bayes», `https://mathshistory.st-andrews.ac.uk/Biographies/Bayes/`. Consulté le 2 octobre 2023.

Parent, e. J. B. 2007, *Le Raisonnement Bayésien : Modélisation et Inférence*, Statistique et probabilités appliquées, Springer.

Pazhoohesh, A. A. R. D. e. S. W., Mehdi. 2021, «Investigating the impact of missing data imputation techniques on battery energy management system», *IET Smart Grid*, vol. 4, n° 2, p. 162–175.

Pelgrin, F. e. S. A. 2008, «Un regard bayésien sur les modèles dynamiques de la macroéconomie», *Économie & prévision*, vol. 183-184, p. 127–152.

PricewaterhouseCoopers. 2013, «Cities of opportunity : Building the future», `https://www.pwc.com/gx/en/capital-projects-infrastructure/publications/`

assets/pwc-cities-of-opportunity-building-the-future.pdf. Accessed : date-of-access.

R Core Team. 2022, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Salakhutdinov, R. e. A. M. 2008, «Bayesian probabilistic matrix factorization using markov chain monte carlo», dans *Proceedings of the 25th International Conference on Machine Learning*, ICML, Helsinki, Finland, p. 880–887.

Salakhutdinov, R. e. A. M. 2010, «Probabilistic matrix factorization», *Department of Computer Science, University of Toronto*.

Schein, A. N. H. W. e. P. F., Aaron. 2021, «Doubly non-central beta matrix factorization for dna methylation data», dans *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, vol. 161, PMLR, p. 1895–1904. URL https://proceedings.mlr.press/v161/schein21a.html.

Schrank, B. E. e. T. L., David. 2012, «TTI's 2012 Urban Mobility Report : Powered by INRIX Traffic Data», Texas A and M Transportation Institute. Available at http://mobility.tamu.edu.

Shalev-Shwartz, S. e. S. B.-D. 2014, *Understanding Machine Learning : From Theory to Algorithms*, Cambridge University Press, ISBN N978-1-107-05713-5.

Shan, H. e. A. B. 2010, «Generalized probabilistic matrix factorizations for collaborative filtering», dans *ICDM*.

Shi, X. e. D.-Y. Y. 2018, «Machine learning for spatiotemporal sequence forecasting : A survey», *arXiv preprint arXiv :1808.06865*.

Shumway, R. H. e. D. S. S. 2017, *ARIMA Models*, Springer International Publishing, Cham, ISBN 978-3-319-52452-8, p. 75–163, doi :10.1007/978-3-319-52452-8_3. URL https://doi.org/10.1007/978-3-319-52452-8_3.

Pereira da Silva, A. 2016, *Tensor techniques for signal processing : algorithms for Canonical Polyadic decomposition*, Theses, Université Grenoble Alpes ; Université Fédéral du Ceará. URL https://theses.hal.science/tel-01382042.

Singh, A. P. et G. J. Gordon. 2008, «A unified view of matrix factorization models», dans *Machine Learning and Knowledge Discovery in Databases*, édité par B. G. e. M. K. Daelemans, Walter, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-87481-2, p. 358–373.

Smola, A. J. e. R. K. 2003, «Kernels and regularization on graphs», dans *Learning Theory and Kernel Machines : 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, Springer, p. 144–158.

Stone, J. V. 2013, *Bayes' rule : a tutorial introduction to Bayesian analysis*, Sebtel Press, Lexington, KY, ISBN 9780956372840.

Tobler, W. R. 1970, «A computer movie simulating urban growth in the detroit region», *Economic Geography*, vol. 46, p. 234–240.

Weigert, U. S. T. B. A. M. A. D. A. J. e. a., Martin. 2018, «Content-aware image restoration : pushing the limits of fluorescence microscopy», *Nature methods*, vol. 15, n° 11, p. 1090–1097.

Wikle, A. Z.-M. e. N. C., Christopher K. 2019, *Spatio-Temporal Statistics with R*, Chapman & Hall/CRC, Boca Raton, FL. The book won the 2019 Taylor and Francis Award for Outstanding Reference/Monograph in the Science and Medicine category.

Xie, M. M. T. L. S. J. S. D. Z. Y. e. a., Peng. 2023, «Spatio-temporal dynamic graph relation learning for urban metro flow prediction», *IEEE Transactions on Knowledge and Data Engineering*.

Yu, N. R. e. I. S. D., Hsiang-Fu. 2016, «Temporal regularized matrix factorization for high-dimensional time series prediction», dans *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper_files/paper/2016/file/85422afb467e9456013a2a51d4dff702-Paper.pdf`.

Zhou, H. S. A. B. e. G. S., Tinghui. 2012, «Kernelized probabilistic matrix factorization : Exploiting graphs and side information», dans *Proceedings of the 12th SIAM International Conference on Data Mining*, SIAM, Anaheim, California, USA, p. 403–414, doi :10.1137/1.9781611972825.35.